Springer Series in the Data Sciences

Edited by:
Alicia L. Carriquiry
Judith M. Tanur
William F. Eddy
With Margaret L. Smykla

Statistics in the Public Interest

In Memory of Stephen E. Fienberg

Springer

Springer Series in the Data Sciences

Springer Series in the Data Sciences focuses primarily on monographs and graduate level textbooks. The target audience includes students and researchers working in and across the fields of mathematics, theoretical computer science, and statistics. Data Analysis and Interpretation is a broad field encompassing some of the fastest-growing subjects in interdisciplinary statistics, mathematics and computer science. It encompasses a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, including diverse techniques under a variety of names, in different business, science, and social science domains. Springer Series in the Data Sciences addresses the needs of a broad spectrum of scientists and students who are utilizing quantitative methods in their daily research. The series is broad but structured, including topics within all core areas of the data sciences. The breadth of the series reflects the variation of scholarly projects currently underway in the field of machine learning.

More information about this series at http://www.springer.com/series/13852

Alicia L. Carriquiry • Judith M. Tanur
William F. Eddy
Editors

With Margaret L. Smykla

Statistics in the Public Interest

In Memory of Stephen E. Fienberg

Springer

Editors
Alicia L. Carriquiry
Department of Statistics
Iowa State University
Ames, Iowa, USA

Judith M. Tanur
Department of Sociology
Stony Brook University
Stony Brook, New York, USA

William F. Eddy
Statistics Department
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

ISSN 2365-5674 ISSN 2365-5682 (electronic)
Springer Series in the Data Sciences
ISBN 978-3-030-75459-4 ISBN 978-3-030-75460-0 (eBook)
https://doi.org/10.1007/978-3-030-75460-0

Mathematics Subject Classification: 62-06

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Acknowledgments

Editing this book in memory of our friend, Steve Fienberg, was a bittersweet experience because even though we have already commemorated the fifth anniversary of his death, we still miss him dearly. This said, our work was also pleasant and rewarding thanks to the tremendous commitment and help we received from many individuals and institutions.

First, we wish to thank Donna Chernyk, our editor at Springer Nature, who made it all sound so straightforward and whose great attitude and positive responses to all our many queries enabled us to assemble a volume that we believe is worthy of Steve. Donna was always prompt, patient, and encouraging, and we are grateful to her for making our job so much easier.

Next, we gratefully acknowledge the Institute of Mathematical Statistics (IMS), the International Society for Bayesian Analysis (ISBA), the American Statistical Association (ASA), and Taylor & Francis, for their permission to reproduce material that had previously appeared in *Statistical Science,* the *ISBA Bulletin,* and *Chance.* Steve's younger son, Howard Fienberg, contributed many of the photos included in the last part in the book. Thank you, Howard! The *Tribune-Review* granted us permission to reprint two obituaries that appeared soon after Joyce Fienberg, Steve's wife, was murdered by a gunman who stormed the Tree of Life Synagogue in October of 2018. Heidi Sestrich, from the Department of Statistics and Data Science at Carnegie Mellon University, generously helped us to reformat some files to meet the publisher's requirements.

Several of the chapters in the book were sent out for anonymous review. Revisions suggested by these reviewers in all cases contributed to improving the readability and quality of the original manuscripts, so we wish to acknowledge their work even if they will remain unnamed.

Finally, there would not have been a book to edit without the excellent chapters contributed by a large number of authors. Just about everyone who was asked enthusiastically agreed to participate and the response was humbling. It is a testament to Steve's long and distinguished career that so many outstanding authors

from a variety of disciplinary areas were ready to honor him one more time. The collection of chapters is extraordinary, and reflect Steve's contributions, not only in terms of breadth but also in terms of scholarly quality. So thank you all so much.

Steve would have loved this book from cover to cover.

Prologue

Remembering Steve Fienberg

Obit for Stephen E. Fienberg written by Alicia Carriquiry, and reprinted with permission from the *ISBA Bulletin*, Vol. 24, No. 1, 2017.

Stephen E. Fienberg passed away on December 14, 2016, shortly after turning 74 years of age. He had been diagnosed with cancer about four years earlier, but kept such a demanding and productive schedule in spite of the disease that most of us were convinced that he would prevail in the end. Steve's death was a tremendous loss for statistics and for science in general, and he will be sorely missed.

Steve was born in Toronto, Canada, on November 27, 1942. In high school, it became obvious to him that he was good at, and greatly enjoyed, the sciences, in particular the mathematical sciences. Steve liked to tell that while his mother (who passed away in Toronto less than two years ago) thought that he was a genius, he was just a good student with an aptitude for mathematics and a passion for ice hockey. Steve went on to the University of Toronto, where he obtained a degree in mathematics in 1964. He applied to, and was admitted into, the doctoral program in statistics at Harvard University, and finished his PhD in 1968, under the supervision of Fred Mosteller.

Meeting Fred Mosteller and working closely with him in a variety of different projects was a life-changing experience for Steve. Mosteller at the time was a rare statistician in that he was genuinely driven by interesting applied projects. The fact that statistics could be brought to bear on so many other disciplines and to such good effect was a revelation, and these early experiences had a lasting impact on Steve's professional life. Steve had a profound respect and deep affection for Mosteller, and often spoke of how much he had learned from his years as a graduate student working with him.

After completing his PhD, Steve was recruited by William Kruskal, then Chair of the Department of Statistics at the University of Chicago, and began his career as an assistant professor. Kruskal, much like Mosteller, was also attracted to applications, and introduced Steve to many different faculty in a wide range of disciplines with

whom Steve began collaborating. In those days, political polling was becoming widespread, but polling methodology was not yet fully developed. Steve became intrigued by the political polling carried out by a local newspaper, and this interest led in part to many years of research in different aspects of survey sampling.

Even though Steve enjoyed his years in Chicago, he and his wife Joyce moved to Minnesota, largely for personal reasons. In Minnesota, Steve held his first administrative position as Chair of the Department of Applied Statistics at the Saint Paul campus of the University of Minnesota. From Minnesota, Steve and Joyce moved to Carnegie Mellon University, which Steve called his academic home and where he spent the rest of his professional life. Steve joined the Department of Statistics at CMU in 1980, and with the exception of a short stint as Provost of York University in Canada, he never left. A "Conversation with Steve" by two of Steve's dearest friends, Miron Straf and Judy Tanur, was published in *Statistical Science* in 2013, and is reprinted in this volume and includes many biographical details about Steve. It also paints a wonderfully warm picture of Steve as a person.

Steve's first research contributions were largely based on his dissertation research. Mosteller introduced Steve to a National Research Council study that was known as the "National Halothane Study," and which Steve described as a "giant contingency table." For his dissertation, Steve developed loglinear model theory and methods useful for the analyses of categorical data such as those collected in the study, and together with Yvonne Bishop and Paul Holland (also Mosteller students) published the well-known book *Discrete Multivariate Analysis* (1975), with the green covers. Throughout his career, Steve continued to advance the theory and implementation of loglinear models, but also built world-class research programs in privacy and confidentiality, machine learning, and algebraic statistics.

Steve was already interested in Bayesian theory by the time he arrived at CMU, but his career as a Bayesian statistician really took off then. Steve joined Jay Kadane and Morrie DeGroot when he came to CMU, and the three of them contributed to making the department a destination for Bayesians from all over the world. In the Mosteller and Kruskal tradition, Steve developed an interest in a wide variety of problems in other disciplines, and was instrumental in the creation and editing of journals with a focus on the principled application of statistics. These included the *Annals of Applied Statistics*, the *Journal of Privacy and Confidentiality*, and more recently, *The Annual Review of Statistics and Its Application*. Bayesians have much for which to be thankful to Steve. He was the second President of ISBA, and was largely responsible for attracting the funding for the ISBA 2000 World Meeting in Crete. He contributed the first article in the first issue of *Bayesian Analysis*, entitled "When did Bayesian inference become Bayesian?", a historical recount of the most important developments in Bayesian statistics between the time when Bayes' opus was published posthumously and the end of the last century. During what he called "the Bayesian Renaissance," Steve became a tireless and effective promoter of the Bayesian paradigm worldwide.

Possibly because of Mosteller's and Kruskal's influences, Steve's passion was to advance the principled and constructive use of statistics to solve real problems in other disciplines, preferably when those problems had a public policy implication.

Not long ago, Eric Lander, the renowned scientist and co-Chair of President Obama's Council of Advisors on Science and Technology (PCAST), referred to Steve as follows:

> Steve Fienberg is not just a statistician–he is a *public statistician*. He has brought his considerable statistical prowess to bear on problems of great public importance (emphasis added).

Steve's first forays into public policy began shortly after he arrived at CMU: he became involved with various government agencies on matters of data collection and data sharing, and joined the Committee on National Statistics (CNSTAT) soon after it was established. Through his work with CNSTAT (which continued throughout his career), Steve had an opportunity to positively impact the work at most (if not all) federal agencies in charge of collecting, synthesizing, and sharing official statistics.

After CMU, the institution in the US that most benefited from Steve's knowledge and dedication was the National Academies of Science, Engineering and Medicine (NASEM). Steve began participating in NASEM's activities in the mid-1980s, but became truly involved after his election to the National Academy of Sciences in 1999 (one of Steve's proudest professional accomplishments). Not only did Steve focus much of his efforts on the NASEM, he also motivated many of us to follow in his footsteps and view the NASEM as an effective vehicle to introduce positive change in society through science-based public policy and decision-making. Steve served the Academies in a variety of roles, but possibly the most consequential of those was his co-chairing of the Report Review Committee, which Steve viewed as an efficient means to ensure that every report published by the Academies was based on solid science and (as appropriate) on sound statistical reasoning.

I have had the privilege of calling Steve a friend for over 25 years, and his mentoring and efforts on my behalf changed the course of my professional life. A few years ago, Steve encouraged me (and Hal Stern and Karen Kafadar) to submit a proposal to establish a NIST Center of Excellence in Forensic Statistics, which would be located at our four institutions, with "headquarters" at Iowa State. Surprisingly to me (but not to Steve!) we were successful and obtained the funds to create the center in 2015. Steve was the intellectual leader, the one with the grand vision and the far-reaching ideas, and I have great hopes that the work on which we have embarked at the center will have a positive impact on society, because Steve was instrumental in setting us off on the right path. Hal, Karen, and I are tremendously thankful to Bill Eddy, who was Steve's close friend and colleague, for jumping in and picking up where Steve left off.

Steve was an affectionate and loyal friend, and he seemed to know everyone. But his world revolved around his wife Joyce and the rest of his family. Steve adored his grandchildren and loved spending time with them. He was particularly fond of having them all descend upon him and Joyce for extended summer visits. While not religious in the usual sense, Steve was proud of his Jewish heritage and culture and strongly believed in keeping the rituals and traditions, and in observing the holidays, as a means to nurture his sense of belonging and reinforce his ties to the Jewish community to which he felt so close.

Among his many other interests and activities, Steve always found time for his other "passions": ice hockey (which he continued practicing even into his 70s) and the *New York Times* crossword puzzle. He loved good food and fine wine (and single malt scotch) and was the instigator of the "Saturday Night Extravagant Dining" group (Jim Berger, Susie Bayarri, Merlise Clyde, Ed George, Dick De Veaux, Robert Wolpert (emeritus), Veronika Rockova, myself, and anyone else reckless enough to join us) during the Joint Statistical Meetings (JSM). But he was determined to encourage good dining habits among JSM goers long before then; remember *Belizaire*, anyone ?

Steve had a marshmallow core even though on occasion he could unsheathe the fangs. He was immensely patient with young faculty and students and with anyone who was really trying, but he did not suffer fools gladly. He loved a good competition but did his best to have the last word. He could be demanding, but he gave of himself generously and never ever expected anything in return. He was well respected by some, idolized by others, and ignored by no one, and sometimes he seemed invincible. His many friends will miss him dearly, for perhaps ever.

References

Fienberg, SE. 2006. When did Bayesian analysis become "Bayesian"? *Bayesian Analysis*, 1:1–40.

Straf, ML, Tanur, JM. 2013. A conversation with Stephen E. Fienberg. *Statistical Science*, 28:447–463.

In Memory of Joyce Fienberg

Much has happened since Steve died. Most horrifically, his wife of over 50 years, Joyce Libman Fienberg, was murdered in the worst anti-Semitic attack to be committed in the United States, when a gunman opened fire at the Tree of Life Synagogue in Pittsburgh on Saturday, October 27, 2018. Her funeral, on October 31, drew 1000 mourners.

Joyce and Steve met at the University of Toronto in 1963, where they were both enrolled. They married two years later and had two sons, Anthony and Howard. With boys in tow, Joyce and Steve traveled the world on behalf of Steve's career to various academic posts, conferences, research centers, and more.

Joyce earned a degree in social psychology at the University of Toronto, and in 1983 joined the Learning Research and Development Center (LRDC) at the University of Pittsburgh as a research specialist, analyzing learning in classrooms and museums, and studying the practices of highly effective teachers. Even though she held a full-time job until her retirement in 2008, she dedicated time to support Steve's career. While Steve worked tirelessly to bring out the best in his PhD

students, Joyce focused on making those students, and departmental post-docs and new faculty, feel more than welcome – special! – in what, for some, was a strange and unknown land: Pittsburgh, PA. Her home and her heart were always open to everyone, but in particular to anyone who was new in Pittsburgh and needed a friend. Students fondly remember the warmth of Joyce's gatherings in celebration of the Jewish holidays, where there was always a place at the Seder table regardless of religious affiliation.

Anyone who had the privilege of meeting Joyce was shocked and saddened by her death. Dr Gaea Leinhardt, who was Joyce's research partner for decades and her close friend, echoed the sentiments of many when she said: "Joyce was a magnificent, generous, caring and profoundly thoughtful human being."

On April 20, 2019, in honor of both Joyce and Steve, the Carnegie Mellon Department of Statistics and Data Science hosted the Stephen E. and Joyce Fienberg Memorial Lecture at CMU with guest lecturer Sir David Spiegelhalter. Later that year, the Stephen E. and Joyce Fienberg Professorship in Statistics and Data Science was conferred upon Rebecca Nugent, associate department head and co-director of undergraduate studies. It was a fitting tribute not only to Steve but also to Joyce, who was such a consequential member of the CMU community.

Joyce was Steve's rock, and the family they raised together was his greatest source of pride.

A Special Recognition for Margaret L. Smykla

While Carriquiry, Eddy, and Tanur are the editors of this volume, Margaret L. Smykla's contribution was critical to the success of the project and deserves special recognition. Margie met Steve earlier than several of us did and worked with him for decades. Her knowledge of the Department of Statistics and Data Science at Carnegie Mellon and of its inhabitants past and present is encyclopedic, and was a tremendous resource as we assembled the pieces for the book.

We are deeply grateful to Margie for her dedication and her efforts in support of this volume honoring Steve and Joyce.

The Genesis of This Book

Steve Fienberg produced a tremendous corpus of work in statistics and the social sciences writ large. His contribution is broad and deep, and laid the groundwork for the work of many others – students, colleagues, and total strangers alike. A quick search online results in over 72,000 (!) citations of his work and an h-index of 89. In the last year of his life, the number of citations his work received exceeded 20,000, confirming that Steve was at the top of his game until the very end.

We thought that editing a volume that included an eclectic but high-quality mix of contributions by a wide range of authors would honor Steve's legacy. In mid-2017, we contacted Donna Chernyk, an Editor with Springer Nature, who encouraged us to work on such a volume. While we had hopes of completing this project over a year ago, we are now delighted to see the light at the end of the tunnel and look forward to seeing the book in print before long.

We have organized the contributions into six sections:

- Theory and Methods for Categorical Data
- Bayesian Theory and Applications
- Statistics and the Law
- Causal Analyses
- Surveys and Censuses
- Official Statistics

A seventh section, entitled "Steve and Joyce As We Knew Them", includes short contributions by some of Steve and Joyce's closest friends, reminiscences contributed by former students, colleagues, and assorted others, and a collection of photographs to remember them by.

We are tremendously grateful to everyone who contributed a chapter for this volume. Steve would have loved to read every single one and would have found much to like. A huge proportion of those we approached with an invitation to write accepted and followed through. Because of page limits, there were countless other friends and colleagues of Steve's whom we did not contact. To everyone who would have liked to participate in this project, we apologize! If enough of you request a re-do, perhaps we can convince Springer to let us publish a Volume II with another 25 contributions!

While working to put this volume together, Judy, Bill, Margie, and Alicia spent a lot of time on zoom, discussing this project but also many other things. It is fair to say that we have all four enjoyed each other's company, have grown closer, and are grateful to have shared so much during the past three years. Steve would have been pleased.

The Ides of March, 2022

Alicia L. Carriquiry, William F. Eddy, Judith M. Tanur, and Margie Smykla

Contents

About the Editors

Alicia L. Carriquiry was born in Montevideo, Uruguay. Since 1990, she has been on the faculty in the Department of Statistics at Iowa State University. She is currently Distinguished Professor of Liberal Arts and Sciences, holds the President's Chair in Statistics, and is Director of the Center for Statistics and Applications in Forensic Evidence (CSAFE), a federally funded research center. Alicia is an elected member of the National Academy of Medicine, a fellow of the American Statistical Association, a fellow of the Institute of Mathematical Statistics, a fellow of the International Society for Bayesian Analysis, a fellow of the American Association for the Advancement of Science, and an elected member of the International Statistical Institute. She was named Technical Advisor for the Association of Firearm and Toolmark Examiners in 2018 and was elected to the American Academy of Forensic Sciences as an associate member in 2020. Alicia's research interests include measurement error modeling, survey sampling and Bayesian methods. In recent years, she has become interested in statistical learning algorithms and their application in various disciplines, in particular in forensic science and criminal justice. Alicia met Steve while she was still a doctoral student, and she could always count on him for friendship, good advice, and mentoring. He was a close and trusted friend to her for almost 30 years.

William F. Eddy was on the faculty of Carnegie Mellon University (CMU) from 1976 until his retirement as John C. Warner Professor of Statistics in 2020. He also held positions in machine learning, biological sciences, and at the Center for the Neural Basis of Cognition at CMU and in the Department of Neurobiology at the University of Pittsburgh. Bill was the founding editor of the *Journal of Computational and Graphics Statistics* (a joint publication of the American Statistical Association, the Institute of Mathematical Statistics, and the Interface Foundation of North America since 1991) and, together with Steve Fienberg, founded *CHANCE* magazine in 1988. He was very active in the National Academies and is the only person to have served as chairman of both its Committee on Applied and Theoretical Statistics and its Committee on National Statistics. He was named a

Lifetime National Associate of the Academies in 2007. Bill's late wife Connie and he were close friends of both Steve and Joyce Fienberg for more than 35 years.

Judith M. Tanur is Distinguished Teaching Professor Emerita, Department of Sociology, at Stony Brook University. She is a fellow of the American Statistical Association (ASA), the American Association for the Advancement of Science, and the Association for Psychological Science; an elected member of the International Statistical Institute; and a recipient of ASA's Founders Award. Judith was a member of the Committee on National Statistics of the National Research Council, chairing its Advanced Research Seminar on Cognitive Aspects of Survey Methodology. She was also an ASA/NSF/BLS Senior Research Fellow and served on the ASA Advisory Committee for Survey of Income and Program Participation (SIPP), as well as NSF (National Science Foundation) Panels and visiting committees. Judith has served on the Advisory Committee to the NSF Director for Social, Behavioral, and Economic Sciences and the Board of Directors of the Social Science Research Council (SSRC). In retirement she continues to serve on the Board of Trustees of NORC and has served on the Committee of International Advisors for the SSRC Vietnam Population Survey, as a member of the American Association for the Advancement of Science's On Call Scientists and their Committee on Scientific Freedom and Responsibility, and on the Advisory Committee for the American Association of Arts and Sciences project on Humanities Indicators. She worked for years with Steve Fienberg on an NSF-funded project on the parallels between randomized experiments and probability-based sample surveys, much of the work being done at her home in Montauk where Steve, Joyce, and their family visited regularly for many summers.

Margaret L. Smykla joined the Department of Statistics at Carnegie Mellon University in August, 1980 – at the same time Steve Fienberg came to CMU – and can remember boxes and boxes of his books being delivered to the Department's small central administrative office. She interacted and worked with Steve, his post-doctoral research scientists, and graduate students over the next several decades. In her early role as editorial assistant, she had the opportunity to work with Steve on *CHANCE* magazine, which he co-founded with Bill Eddy. Knowing she had writing aspirations, Steve insisted she write an article for *CHANCE*. His support and confidence in her abilities kicked off a freelance writing career which exists to this day (And thrills her mother that her writing degree from the University of Pittsburgh in 1978 is being put to good use!). During the many social events in which Margie participated, she had the chance to get to know Steve's wonderful wife, Joyce, and the couple's two fine sons, Anthony and Howard, and communicated with them even after Steve's death. The reminiscences and photos obtained on the occasions of Steve's 65th birthday and his Emeritus attainment were gathered and organized by Margie. She is a native of Pittsburgh and a grandmother, her favorite role of all. She also relishes the go-to role in the Department for anything alumni-related as, after 40 years, she knows just about everyone!

Part I
Introduction: Theory and Methods for Categorical Data

William F. Eddy

When Steve and his family moved to Pittsburgh in 1980, the first problem he talked to me about was inference in a 2×2 table (with ordered margins). I had always been interested in ordering problems, such as the distribution of order statistics and related matters. Superficially the parenthetical addition "with ordered margins" to the text description in the original problem Steve posed is only the slightest generalization of Steve's original research problem (1968, 1970 with John Gilbert) of inference in a 2×2 table. That original problem stemmed from US Census work by Deming and Stephan (1940) "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known."

In order to better understand the problem and Steve's focus on geometry when we met in 1980, I went home from that initial meeting, and in my basement I built a physical geometric model of the basic 2×2 tetrahedron (with barycentric coordinates) out of coat hangers using a tin-like solder to hold it together. Then I added the surfaces of independence using colored thread (red, white, blue, and black). If I remember correctly, those surfaces are hyperbolic paraboloids (or is it parabolic hyperboloids?). I was impressed by the geometry. That object (the coat hangers and thread) remained in my office until I retired in mid-2020. Early on it took regular trips to Steve's office as our discussions continued; later on it took longer forays down to Teddy Seidenfeld's office (I think Teddy used it as a teaching tool), but it always returned to its home.

Unfortunately, Steve and I never finished the project; I had drifted to some other things, and I didn't return with him to the original problem. I had gotten interested in the distribution of extremes of samples in two (and higher) dimensions (an order statistic sort of problem which led me to convex hulls and a whole different geometric world than the one that Steve was following). Steve maintained his interest in his original problem which fostered a whole generation of research and researchers. I do not know that Steve's ordering problem has ever been completely resolved.

Analysis of categorical data was Steve's starting point and had always been one of his primary interests since 1968. This section of the book contains three papers that revolve around that focus. The papers each stem directly from Steve's lifelong

interest. The three problems are wildly different, reflecting the diversity of Steve's approaches to the underlying common problem.

Larry Wasserman has chosen to branch out from the "usual" statistical problem in categorical data which consider inference, and Larry rather thinks about prediction. The prediction problem is different primarily because "point" prediction is simple in this discrete setting. This has led Larry to think about prediction of sets rather than points. He considers both asymptotically correct and exact methods and applies his ideas to several specific problems: including multinomials, regression models, and log-linear models.

Without covariates the problem is especially difficult; Larry points out that in the case of a uniform distribution, informative predictions are impossible. When the probability distribution is far from uniform and the sample size is large, then the confidence set concentrates on the "right" answer.

In Sect. 1.4 he turns to regression models. Because the object is to predict the set to which an observation belongs based on the observed values of the regressors, this becomes a classification problem. He has shown that many different estimators are successful; a parametric approach leads to logistic regression. An apparent problem with the approach occurs when the prediction set is empty. Larry considers estimators which produce non-empty sets, but he also argues that producing empty sets when the predictors attain rare values is a kind of warning sign concerning outliers.

In Sect. 1.5 he turns to Steve's favorite example: log-linear models. In this very interesting section, he argues that (a) sparse log-linear models lead to better interpretability and simultaneously conditional independence models (a goal of log-linear modeling) are not especially important to prediction. In Sect. 1.6 he focuses on random effects models and considers the problem of predicting a future set of observations, and in Sect. 1.7 he briefly considers a Bayesian method. Larry argues that his approach provides a useful fusion of frequentist and Bayes methods.

Capture-recapture (or "dual systems" estimation) has its roots in the classic problem of estimating how many fish are in a lake. It became of more general interest to statisticians because of the diversity of potential applications. The Lincoln-Petersen estimator, based on the hypergeometric distribution, is well described in Seber's book on animal abundance. It was used first by the Census in the evaluation of the 1950 US Census undercount and became a standard tool of Census evaluations. By the 1980s, many generalizations had been studied. So much so that the term "wily trout" became a not uncommon descriptor of members of a population with different probabilities of capture; I recall discussions with Steve and Bob Groves (later head of the Census Bureau) around that time or a little later that mentioned the wily trout. The NYT had used the term on August 12, 1964: "When an Angler Is After a Wily Trout, Patience Is Often His Best Weapon." And, I even just found, a previously unknown (to me) reference "Note. The Census Adjustment Case: Hunt for the Wily Trout" by James Pack, 1996, Jurimetrics, 31:1 35–51.

The paper by Manrique-Vallier, Ball, and Sadinle is a thorough review of the recent literature on this general problem and its application to estimation of war casualties. They begin by discussing the two basic assumptions of (1) independence

of samples and (2) homogeneity of rates. In Sect. 2.2, they explicitly discuss list dependence in the casualty estimation application, and also discuss the problem of homogeneity (the wily trout). In Sect. 2.3, they review some of the unresolved issues particularly in the current context of casualty estimation.

One final side note, I even found a reference to what I believe is a now defunct winery and restaurant in Canberra, Australia, called the Poachers Pantry that served a wine known as the Wily Trout. My geography is poor, but I believe the vineyard extended into the neighboring area of Murrumbateman. Steve would have loved the existence of the Poachers Pantry, and Belizaire (see the paper in this volume by Madansky) would have instigated a trip to sample the wine.

The paper by Gross, Karwa, and Petrovic provides the reader with an introduction to algebraic statistics, that part of statistics that uses the terminology and methods of modern algebra to gain insight into the structure and calculations of mathematical statistics. In the first section of this paper, they provide us with an introduction to Steve's view of the problems. As with any new topic, learning the language is the important first step, and Gross et al. provide a very nice summary of the basic ideas and terminology in Sect. 3.2. The novice reader will have to do some work to really "get" this unfamiliar territory.

Exponential random graph models (ERGMs) for the log-linear case are introduced in Sect. 3.3. These are, in general, network models and are the heart of the paper. A careful reading reveals that Fienberg worked on these models for most of his career. His early readers need to look more carefully here to see that he has continued to look at the same problems over and over. One can see that he understood more and more about these models as his vision clarified. Section 3.4 is the generalization beyond simple graphs, and yet here Fienberg puts his efforts into a problem that he was working on before he moved to Pittsburgh in 1980.

After defining log-liner ERGMs, the authors consider the special case of stochastic block models. For you, fans of the log-linear models, as popularized by Bishop, Fienberg, and Holland, these models have been around since the 1985 paper by Fienberg, Meyer, and Wasserman. The use of Markov bases made the connection between log-linear models and the non-negativity constraints mandatory until the 2010 paper with Petrovic, Rinaldo, and Fienberg obviated that constraint. Finally, near the end of this section, the authors show how Fienberg was able to introduce data into the algebraic framework. This has led to the development of Monte Carlo methods for fitting models.

Finally, in Sect. 3.4 the authors move beyond simple graphs. Through two examples they show that Fienberg's ideas applied to multiple authorship data lead to further insight. This allows the production of new algorithms for fitting these models to real data. The work started by Fienberg will clearly be continued by others.

Chapter 1
Prediction for Categorical Data

Larry Wasserman

1.1 Introduction

Steve Fienberg was well-known for his pioneering work on categorical data analysis. Based on his work, there was an explosion of papers on this topic. Most of this work focuses on questions related to inference. A problem that has received less attention is prediction. In this paper we review two methods for prediction of categorical data. However, our goal is not point prediction, but set-valued prediction as we explain in Sect. 1.2.

We consider several versions of this problem including: unstructured multinomials, multinomial regression, log-linear models, and random effects models. We mostly pursue frequentist approaches. But we also discuss an approach that fuses Bayesian and frequentist approaches in Sect. 1.7.

The Basic Idea Before plunging into the details, we first give a quick overview of the ideas. We will use an example where the data are continuous but the rest of the article is devoted to discrete data.

Suppose we observe n iid data points $Y_1, \ldots, Y_n$ where $Y_i \in \mathbb{R}$. Our goal is to construct a set C_n such that $P(Y_{n+1} \in C_n) \geq 1-\alpha$. Define residuals $R_i = |Y_i - \overline{Y}|$. One approach is to set $C_n = [\overline{Y}_n - c, \overline{Y}_n + c]$ where c is the $1-\alpha$ quantile of the residuals. Under some conditions on P, we have that $P(Y_{n+1} \in C_n) \to 1-\alpha$ as

L. Wasserman (✉)
Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, USA

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_1

$n \to \infty$. Note that we have made no parametric model assumptions. We can use any definition of "residual' that we like. The choice does not affect coverage but it does affect the length of the interval.

In fact, we could use a parametric model ($P_\theta \in \Theta$) to define residuals such as $R_i = 1/p_{\hat{\theta}}(Y_i)$. If the model is correct, we will get intervals that are nearly as short as possible. But the coverage is correct even if we have the model wrong.

However, if the model is complex (think high dimensional), then this approach fails. Sect. 1.2 considers an approach called conformal prediction that provides coverage under any conditions over all distributions. The idea is simple. We take a guess y at the value of Y_{n+1}. We form residuals on the augmented data $(Y_1, \ldots, Y_n, Y_{n+1})$ where $Y_{n+1} = y$. We can then use the residuals to test $H_0 : Y_{n+1} = y$. This gives a p-value $\pi(y)$. The process is repeated for every y. Now we invert the test to get $C_n = \{y : \pi(y) \geq 1 - \alpha\}$. Then $P(Y_{n+1} \in C) \geq 1 - \alpha$. This gives us the desired prediction set C_n without any assumptions on P.

These two approaches are explained in more detail in the rest of the paper.

Outline In Sect. 1.2 we review two methods for set-valued prediction. We apply the methods to unstructured multinomials in Sect. 1.2. We shall see that this is a nearly impossible problem unless the multinomial is mostly concentrated on a small set. In Sect. 1.2 we discuss categorical regression where the methods work very well. In Sect. 1.5 we briefly discuss log-linear models. Here we see that conditional independence—so vital for inference—does not help with prediction. We discuss random effects models in Sect. 1.6. In Sect. 1.7 we discuss prediction from a Bayesian point of view. Section 1.8 contains a discussion.

1.2 Two Prediction Methods

Here we review two methods for prediction. The common thread is that neither makes strong assumptions. Both methods start with a working model but the model is not assumed to be correct. Let

$$Y_1, \ldots, Y_n \sim P$$

be iid observations from a distribution P. Our goal is to construct set C_n which will be a function of $Y_1, \ldots, Y_n$. We would like C_n to have the property

$$P(Y_{n+1} \in C_n) \approx 1 - \alpha$$

for some pre-specified confidence level α. Here Y_{n+1} is a new observation.

The Asymptotic Method

The first method is an asymptotic method. Consider a parametric model $\mathcal{P} = (p_\theta;\ \theta \in \Theta)$. Let $\hat{\theta}$ be an estimate of θ such as the maximum likelihood estimator. We regard $\mathcal{P}$ as a working model. We do not require the model to be correct.

Given that we assume that the model may be wrong, how do we construct a prediction set? We can use a method proposed in Dunn and Wasserman (2018) which we now review. We make three assumptions:

(A1: Quasi-Consistency) There exists some $\theta_* \in \Theta$ such that $\hat{\theta} \xrightarrow{P} \theta_*$.
(A2: Smoothness) There exists $L > 0$ and $\epsilon_1 > 0$ such that, whenever $||\theta_1 - \theta_2|| \leq \epsilon < \epsilon_1$, $||p_{\theta_1} - p_{\theta_2}||_\infty \leq L\epsilon$.
(A3: Smoothness of Level Sets.) There exists $c > 0$ and $\epsilon_0 > 0$ such that, for every t and every θ, θ', for which $||\theta - \theta'|| \leq \epsilon < \epsilon_0$,

$$\mu(M(\theta, t)\Delta M(\theta', t)) \leq c\epsilon,$$

where μ is Lebesgue measure, Δ denotes the set difference, and $M(\theta, t) = \{y : p_\theta(y) > t\}$. The value θ_* in (A1) is usually the minimizer of the KL distance of $\int p(x) \log p(x)/p_\theta(x)dx$ but this is not required. Conditions (A2) and (A3) are smoothness assumptions that hold for typical parametric families.

Let $Z_j = p_{\hat{\theta}}(Y_j)$ and let $Z_{(1)} \leq \cdots \leq Z_{(n)}$ be the corresponding order statistics. Let $C_n = \{y :\ p_{\hat{\theta}}(y) \geq \hat{t}\}$ where $\hat{t} = Z_{(m)}$ and $m = \lfloor n\alpha \rfloor$. Then Dunn and Wasserman (2018) show that $\int_{C_n} dP = 1 - \alpha + o_P(1)$. Hence, $P(Y_{n+1} \in C_n) = 1 - \alpha + o_P(1)$. Note that C_n can be quite different from the set $D_n = \{y : p_{\hat{\theta}} > u\}$ where u is chosen so that $\sum_{y \in D_n} p_{\hat{\theta}} \approx 1 - \alpha$. The difference is that the validity of D_n depends on the model being correct. In contrast, C_n is valid even if the model is wrong. However, if the model is wrong, the set C_n may be large. In other words, validity of D_n depends on the correctness of the model. But model correctness only affects the size of C_n not its validity.

An Exact Method

An exact prediction set can be defined using conformal prediction (Lei and Wasserman 2014; Vovk et al. 2005). The idea is to start with a guess y for Y_{n+1}. We then test the hypothesis $H_0 : Y_{n+1} = y$ at level α. We do this for every $y \in \mathcal{Y}$ where $\mathcal{Y}$ is the sample space for Y. Then we invert the tests to get a confidence set for Y_{n+1}. Here are the steps in the procedure:

1. For each y do the following:
   (a) Let $\mathcal{A}(y) = \{Y_1, \ldots, Y_n, Y_{n+1}\}$ where $Y_{n+1} = y$. The set $\mathcal{A}(y)$ is called the augmented dataset.

(b) For each i, compute a (permutation invariant) score $R_i(y) = \phi(Y_i, \mathcal{A}(y))$.
(c) Compute the p-value

$$\pi(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1}).$$

2. Set $C_n = \{y : \pi(y) \geq \alpha\}$.

The set $\mathcal{A}(y)$ is called the augmented dataset. The set C_n has the remarkable property that $P(Y_{n+1} \in C_n) \geq 1 - \alpha$ for every distribution P. There is much flexibility in the choice of score $\phi(Y_i, \mathcal{A}(y))$. Again, we can use a working model $(p_\theta : \theta \in \Theta)$ and take $\phi(Y_i, \mathcal{A}(y)) = 1/p_{\hat{\theta}(y)}(Y_i)$ where $\hat{\theta}(y)$ is the MLE based on $Y_1, \ldots, Y_n, y$.

A question in all these methods is how to choose α. One approach is to treat the prediction problem like a typical estimation problem which suggests choosing $\alpha = 0.1$ or $\alpha = 0.05$. Another recommended choice is $\alpha = 0.5$ in which case, for obvious reasons, we call C_n the median prediction set.

1.3 Multinomials

Unstructured multinomials are one of the hardest cases for prediction sets. Imagine, for example, trying to predict Y_{n+1} when the distribution is uniform over $\{1, \ldots, k\}$. Constructing a reliable prediction set is virtually impossible unless we set $C_n = \{1, \ldots, k\}$. However, if the multinomial is sparse, there is some hope as we shall see. By a sparse multinomial we mean that the probability function is mostly concentrated on a small subset of the support.

Let $Y_1, \ldots Y_n$ where $Y_i \in \{1, \ldots, k\}$. The distribution is multinomial with probability function $p = (p_1, \ldots, p_k)$. Let $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_k)$ be the MLE. Hence, $\hat{p}_j = n_j/n$ where $n_j = \sum_i I(Y_i = j)$ for $j = 1, \ldots, k$.

If n is large compared to k and $\min_j p_j > 0$, then the asymptotic method is very simple. (Lebesgue measure has to be replaced by counting measure in (A3).) Define

$$\hat{t}_\alpha = \sup\Big\{t : \sum_j I(\hat{p}_j \geq t) \geq 1 - \alpha\Big\}.$$

Let $C_n = \{j : \hat{p}_j \geq \hat{t}_\alpha\}$. Then, from the result in Sect. 1.2, $P(Y_{n+1} \in C_n) = 1 - \alpha + o(1)$.

If p is far from uniform C_n will be non-trivial in the sense that C_n can be much smaller than $\{1, \ldots, k\}$ and if further the p_j's are distinct, then C_n asymptotically concentrates on the deterministic set $\Lambda = \{j : p_j \geq t_\alpha\}$ where $t_\alpha = \sup\{t : \sum_j I(\hat{p}_j \geq t) \geq 1 - \alpha\}$.

As mentioned above, if p is close to uniform, no useful predictions can be made. Indeed, if p is uniform, then, with very high probability, $C_n = \{1, \ldots, k\}$. There is a nonzero but vanishingly small probability that C_n can have fewer than k elements. But, as a random set, C_n is uniform over sets of size r, for every r. In any case, informative prediction is hopeless in this case.

1.4 Regression Models

So far we have only discussed unsupervised cases where there are no covariates. But often we have access to covariates in which case the data are of the form $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_i \in \mathbb{R}^d$ and $Y_i \in \{1, \ldots, k\}$. In this case the task is a bit different. We need to construct a function C_n mapping $\mathbb{R}^d$ to subsets of $\{1, \ldots, k\}$. We require C_n to satisfy $P(Y_{n+1} \in C_n(X_{n+1})) \geq 1 - \alpha$.

Unlike the unstructured multinomial case, these multinomial regression problems are an example where categorical prediction is very successful. Note that this is just classification. The set-valued prediction problem was studied in detail in Sadinle et al. (2017) (referred to as SLW in what follows). In this section we review the results in that paper.

Suppose that, in addition to the coverage requirement we want to minimize $\mathbb{E}|C_n(X_{n+1})|$ where $|A|$ denotes the number of points in the set. Then SLW showed that the optimal set is

$$C(x) = \{y : \ p(y|x) \geq t_\alpha\}, \tag{1.1}$$

where

$$t_\alpha = \sup\{t : \ P(p(y|x) \geq t) \geq 1 - \alpha\}.$$

This is an oracle prediction set since it requires knowing $p(x, y)$. In practice, we have to estimate p. The easiest approach is to form an estimate $\hat{p}(y|x)$ and plug this into (1.1). SLW consider a variety of estimators such as nearest neighbors and local polynomials. A parametric approach is to use the multinomial logistic model

$$p_\theta(y|x) = \frac{e^{x^T \theta_y}}{1 + e^{x^T \theta_y}}.$$

The extra error incurred by having to estimate $p(y|x)$ is $O(\epsilon_n + k\sqrt{\log n/n})$ where ϵ_n is the error in estimating $p(y|x)$. This reveals an interesting fact: the number of classes k can increase with n as long as $k_n = o(\sqrt{n/\log n})$.

The above methods provide asymptotically correct coverage. Also, the multinomial logistic method requires that the model is correctly specified. An alternative is

to use the exact method described earlier. It is straightforward to adapt this to the regression case. The steps are as follows:

1. For each (x, y) do the following:
   (a) Let $\mathcal{A}(x, y) = \{(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})\}$ where $X_{n+1} = x$ and $Y_{n+1} = y$.
   (b) For each i, compute a (permutation invariant) score $R_i(y) = \phi((X_i, Y_i), \mathcal{A}(x, y))$.
   (c) Compute the p-value

$$\pi(x, y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1}).$$

2. Set $C_n(x) = \{y : \pi(x, y) \geq \alpha\}$.

We then have that $P(Y_{n+1} \in C_n(X_{n+1})) \geq 1 - \alpha$ for all P. Hence, we have exact, finite sample coverage. An example of a score is $R_i = 1/p_{\hat{\theta}}(y|x)$. Another useful score is the residual $R_i = |1 - \hat{p}_{\hat{\theta}}(j|x)|$ for $Y_i = j$.

There is one drawback to all of these methods. For some values of x we might have $C_n(x) = \emptyset$. This is a consequence of trying to minimize $\mathbb{E}|C_n(X)|$. SLW provide two approaches to fix this problem. Our goal is to eliminate the set $\{x : C_n(x) = \emptyset\}$.

The first method proposed in SLW is to complete the set-valued classifier with a given baseline classifier, such as the Bayes classifier. Thus, let $c(\cdot)$ be a simple classifier such that $|c(x)| = 1$ for all x, and define

$$C^\dagger(x) = \begin{cases} C(x) & \text{if } C(x) \neq \emptyset \\ c(x) & \text{if } C(x) = \emptyset. \end{cases}$$

The second approach is called accretive completion. To explain this approach we need to use a modified version of validity, namely $P(Y \in C|Y = y) = \alpha_y$ where the α_y's are class specific confidence levels. Given $\boldsymbol{t} = (t_1, \ldots, t_K)$, denote $C_{\boldsymbol{t}} = \{(x, y) : p(y|x) \geq t_y\}$. For any $\{\alpha_y\}_{y=1}^k$, the optimal solution to the problem is $C_{\boldsymbol{t}}$ with $\boldsymbol{t}$ chosen such that $P\{Y \in C_{\boldsymbol{t}}(X)|Y = y\} = 1 - \alpha_y$. Under this solution, empty sets occur when $\sum_y t_y > 1$. Therefore, a sufficient condition for $C_{\boldsymbol{t}}$ to avoid empty sets is that $\sum_{y=1}^k t_y \leq 1$.

The method in SLW is to search for a set of thresholds $\{t_y\}_{y=1}^K$ that has the lowest ambiguity and guarantees no empty predictions as well as nominal coverage of $C_{\boldsymbol{t}}$:

$$\min_{\boldsymbol{t}} \mathbb{E}\{|C_{\boldsymbol{t}}(X)|\} \text{ subjectto } t_y \leq t_y^{(0)}, \forall y; \sum_y t_y \leq 1\,.$$

Solving this problem exactly is difficult. SLW provide a greedy method that approximates the solution. Basically each t_y is changed slightly, while increasing

the size of the prediction sets as little as possible. The process is repeated until there are no empty set predictions.

The bottom line is that there are effective methods to avoid null predictions.

On the other hand, this can be seen as a feature rather than a bug. For example, suppose we observe an X_{n+1} which is anomalous in the sense that we have not seen any X_i's close to X_{n+1} before. In this case, we might want to output the empty set since we should not extrapolate to outlying observations. In fact, Hechtlinger et al. (2018) show that using the score $p(x|y)$ instead of $p(y|x)$ actually encourages empty set predictions for outliers.

1.5 Log-Linear Models

In this section we consider prediction for log-linear models. Let $x = (x_1, \ldots, x_d)$. Recall that the model is

$$\log p_\theta(x_1, \ldots, x_d) = \sum_A \psi_{\theta_A}(x_A),$$

where A varies over all subsets of $\{1, \ldots, d\}$. The definitive reference on log-linear models is Bishop et al. (1975).

The saturated model corresponds to the unstructured multinomial as in Sect. 1.3 which, as we saw, is a difficult problem.

Usually, for log-linear models, the focus is on model selection corresponding to setting many of the higher-order terms (such as higher-order interactions) to 0. For inferential purposes this is crucial. Log-linear models can be represented as undirected graphs. Setting many terms to 0 leads to sparse and more interpretable graphs. In other words, sparsity leads to better interpretability.

Interestingly, sparseness does not necessarily help with prediction. To see this, suppose that we have two variables X_1 and X_2 where $X_1, X_2 \in \{1, \ldots, d\}$. The log-linear model is

$$\log p_\theta(x_1, \ldots, x_d) = \phi_\emptyset + \phi_1(x_1) + \phi_2(x_2) + \phi_{12}(x_1, x_2).$$

Now suppose we conduct model selection and conclude that the interaction $\phi_{12}(x_1, x_2) = 0$. This implies that X_1 and X_2 are independent which inferentially is an important finding. But if X_1 and X_2 both have a uniform distribution, then the joint distribution is uniform and informative prediction is impossible.

On the other hand, if the marginal distributions of X_1 and X_2 are far from uniform, then the joint distribution is highly concentrated and informative prediction sets are possible. Indeed, the oracle set $\{(x_1, x_2) : \ p(x_1, x_2) \geq t\}$ can be small. This simple example makes clear that it is the shape of the probability function that is crucial. Conditional independence—the main focus of log-linear models—is not central to prediction.

1.6 Random Effects

Random effects models provide interesting challenges for prediction. Suppose we have N data sets $\mathcal{X}_1, \ldots, \mathcal{X}_N$ where $\mathcal{X}_j = \{X_{j1}, \ldots, X_{nn_j}\}$. We assume that each variable is categorical taking values in $\{1, \ldots, k\}$. The distribution of the j^{th} dataset is the multinomial $p_j = (p_{j1}, \ldots, p_{jk})$. Denote the empirical distribution by $\hat{p}_j$. For simplicity we will assume that $n_1 = \cdots = n_N = n$.

There are two different prediction problems. The first is to predict a new observation on an existing distribution. This is where familiar tools like shrinkage come into play. The second, which we will focus on, is to predict a future set of observations $\hat{p}_{N+1}$.

Each $\hat{p}_j$ is a point on the simplex S_k. So the goal is to predict a new observation on S_k given N observations on S_k.

We can use the exact method in Sect. 1.2. A natural residual is $||\hat{p}_j - \overline{p}||_1$ where $\overline{p}$ is the average of the augmented data $\{\hat{p}_1, \ldots, \hat{p}_N, p\}$. Unfortunately, the method requires repeating the calculations for every p on the simplex which is computationally infeasible.

A fast approximation is available. Let $\overline{p} = N^{-1} \sum_{j=1}^{N} \hat{p}_j$. Let $R_i = ||\hat{p}_j - \overline{p}||_1$. Let t_α denote the α upper quantile of the R_i's. Finally, define $C_N = \{p : ||p - \overline{p}||_1 \leq t_\alpha\}$. If N and n are large, then $P(\hat{p}_{n+1} \in C_N) \approx 1 - \alpha$.

The set C_N is an L_1 neighborhood around $\overline{p}$ on the simplex. A challenge—which is faced whenever data are on the simplex—is to find a way to visualize the set C_N.

If a parametric model is available, then the problem gets much simpler. Suppose the model has the form $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$. First assume that the θ_j's are known. Then the problem is to predict θ_{N+1} from $\theta_1, \ldots, \theta_N$. We use the exact method to get a $1 - \beta$ prediction set A_N for θ_{N+1} using the augmented dataset $\theta_1, \ldots, \theta_N, \theta$. Now define $C_N = \bigcup_{\theta \in A_N} L_\theta$ where L_θ is a level set of the multinomial with probability content $1 - \gamma$. By the union bound, the coverage of C_N is at least $1 - (\gamma + \beta)$. Hence, choosing $\gamma + \beta \leq \alpha$ gives a $1 - \alpha$ prediction set. In practice, we use the $\hat{\theta}_j$'s rather than the θ_j's. This induces an error of $O_P(\min_j n_j^{-1})$.

As before, C_N is likely to be very large for the pure multinomial case. But the above method generalizes easily to the multinomial regression case where highly informative prediction sets are possible. See Dunn and Wasserman (2018) for examples in the case that Y is continuous.

1.7 Bayes

Steve Fienberg was, at heart, a Bayesian. But he was not dogmatic and was happy to use any tools that were useful for solving a problem. In that spirit, I will discuss a simple way to fuse Bayesian and frequentist inference.

Suppose we have a parametric model $\mathcal{P} = (p_\theta : \theta \in \Theta)$. Given a prior $\pi(\theta)$ we can, of course, compute the posterior $\pi(\theta|Y_1, \ldots, Y_n) = \mathcal{L}(\theta)\pi(\theta)/\int \mathcal{L}(s)\pi(s)ds$ where $\mathcal{L}(\theta) = \prod_i p_\theta(Y_i)$.

Now consider the predictive distribution

$$\pi(y|Y_1, \ldots, Y_n) = \int p_\theta(y)\pi(\theta|Y_1, \ldots, Y_n)d\theta.$$

Let $D_u = \{y : \pi(y|Y_1, \ldots, Y_n) \geq u\}$. We can choose u such that $P(Y_{n+1} \in D_u|Y_1, \ldots, Y_n) \geq 1 - \alpha$. Then D_u is a $1 - \alpha$ Bayesian predictive region.

But now consider the question: what is the frequentist probability that Y_{n+1} is in D_u? If the model is wrong or the prior conflicts with the likelihood, then the frequentist probability $P(Y_{n+1} \in D_u)$ can be much lower than $1 - \alpha$.

This is where the Bayesian-frequentist fusion comes in. Recall that the predictive sets C_n from Sect. 1.2 give regions with valid coverage for any choice of score function. So we can use $R_i = \pi(Y_i|Y_1, \ldots, Y_n)$ as a score and then apply either method from Sect. 1.2. If the prior and model are well chosen, then C_n will be similar to D_{u_α}. But if the model is wrong or the prior conflicts with the likelihood, the set C_n gets larger to compensate. This gives us the best of both worlds. Well specified models cause C_n to be close to the Bayesian solution. Otherwise, C_n still has coverage validity by essentially expanding the Bayes prediction set. Thus we have achieved a nice fusion of Bayesian and frequentist ideas for prediction.

1.8 Conclusion

Prediction used to be a small subset of inference. But these days, prediction has a much larger role in data analysis.

In this paper, we have reviewed techniques for prediction in models with categorical data. Our emphasis has been on set-valued prediction rather than point-valued prediction. This allows us to attach a confidence level to the predictions. We have even seen that it is possible to fuse frequentist ideas with Bayesian inference. I think Steve would have liked that.

Looking forward, there are many possible extensions of these ideas. For example, Hechtlinger et al. (2018) have applied these methods to deep neural nets. It would be nice to see more work in that setting. It would also be interesting to extend the methods in a way that is robust to outliers.

References

Bishop Y, Fienberg, S and Holland, P. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Dunn, Robin and Wasserman, Larry (2018). *Distribution-Free Prediction Sets with Random Effects*. arXiv preprint arXiv:1809.07441.

Hechtlinger, Y., Poczos, B. and Wasserman, L., (2018). *Cautious Deep Learning*. arXiv preprint arXiv:1805.09460.

Lei, Jing and Wasserman, Larry (2014). Distribution-free prediction bands for nonparametric regression. *Journal of the Royal Statistical Society.* Volume 76, pages 71-96.

Sadinle, Mauricio, Lei, Jing and Wasserman, Larry (2017). Sadinle, M., Lei, J. and Wasserman, L., (2018). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, pp.1–12.

Vovk, V., Gammerman, A. and Shafer, G., (2005). *Algorithmic learning in a random world*. Springer.

Chapter 2
Capture-Recapture for Casualty Estimation and Beyond: Recent Advances and Research Directions

Daniel Manrique-Vallier, Patrick Ball, and Mauricio Sadinle

2.1 Introduction

"How many people were killed?" This is perhaps the most basic quantitative question about the consequences of armed conflicts. While many groups attempt to create tallies of victims, these lists are usually subject to incomplete and non-representative registration. Difficulties faced by data-collection efforts include destroyed infrastructure, danger to field workers, suspicion of data collection by victim communities, among other factors. The result of such an approach can only be taken as an incomplete, non-representative sample with unknown biases, and it can only lead to a lower bound on the total number of casualties.

Capture-Recapture (CR), also known as Multiple Systems Estimation (MSE) in the context of human populations, is a family of statistical methods for estimating the size of closed populations, which can take advantage of the existence of multiple incomplete lists of casualties. CR methods vary in details and complexity, but they all rely on analyzing the patterns of inclusion-exclusion of individuals in the samples to estimate the probability of not being observed, and then the number of unobserved individuals. In the context of conflict casualty estimation, they were first used in Guatemala by Ball (2000), where researchers used three incomplete sources of information, which jointly documented more than 54,000 unique killings, and used CR to estimate the total to be more than 132,000. This analysis and several follow-

D. Manrique-Vallier (✉)
Department of Statistics, Indiana University, Bloomington, IN, USA
e-mail: dmanriqu@indiana.edu

P. Ball
Human Rights Data Analysis Group, San Francisco, CA, USA

M. Sadinle
Department of Biostatistics, University of Washington, Seattle, WA, USA

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_2

on projects helped support the case that the Guatemalan Army committed acts of genocide against the indigenous Mayan population (see Ball and Price (2018)). Other applications include deaths in the Kosovo war (Ball et al. 2002), in Peru (Ball et al. 2003), and Bosnia (Zwierzchowski and Tabeau 2010); see a review in Ball and Price (2019).

CR methods are best known for their application to animal abundance estimation, where they have developed considerably; see e.g. Otis et al. (1978) for a classic survey of methods in this domain. However, variants of these methods for estimating human populations have been in use for a long time—indeed the first documented application of a CR-like method, estimating the population of France, dates all the way back to the eighteenth century (Laplace 1786; Amorós 2014). In epidemiology CR has been applied to the estimation of prevalence of rare diseases and at-risk populations (e.g. International Working Group for Disease Monitoring and Forecasting 1995a;b; Okiria et al. 2019). CR is also used in census correction using post-enumeration sample surveys and dual-system estimators (Mulry and Spencer 1991), and administrative data and MSE (Zaslavsky and Wolfgang 1993). See Bird and King (2017) for a recent review.

As in many other areas of statistical methodology and practice, Steve Fienberg made crucial contributions to the development and adoption of CR. Prior to Steve's work, many researchers had expressed concerns about the ubiquitous assumption of independence between samples. However, it was not until the 1970s with the development of methods for discrete multivariate analysis (in which Steve also played a pivotal role) that Fienberg (1972) introduced the use of log-linear models as a feasible way to account for dependence among the samples. His log-linear CR approach has since become a standard part of the applied toolbox, especially in the estimation of human population sizes. Steve was also a champion of the use of CR as a tool for census correction, an area and debate in which he made important contributions (e.g. Darroch et al. 1993; Anderson and Fienberg 1999). In the specific application to casualty estimation in armed conflicts, Steve played a behind-the-scenes but direct role, as an advisor and mentor to many researchers, including two of the authors of this piece.

Several challenges arise when using CR methodologies for casualty estimation. Many of these problems are common to other applications, and there are readily available methodologies to address them. For example, dependence between lists can be addressed using a log-linear CR approach (Fienberg 1972; Bishop et al. 1975). Other problems in casualty estimation differ substantially from other contexts. For example, while several models for controlling individual heterogeneity in capture have been proposed in the ecology (e.g. Agresti 1994) and census correction (e.g. Darroch et al. 1993) literatures, most of them assume a symmetrical form of heterogeneity (for example, the Rasch model) which is not realistic in our context as lists are not exchangeable.

In this article we discuss the challenges of applying CR methods to the problem of estimating the total number of deaths in armed conflicts, and explore the advances of the last 15 years in the area. We also describe outstanding challenges and speculate possible solutions.

2.2 Capture-Recapture in Casualty Estimation: Challenges and Developments

The Capture-Recapture Approach

Consider a closed population of N individuals, and $J \geq 2$ incomplete lists taken from that population. In this context N will be the unknown number of victims of the armed conflict, and J the number of partial lists available. Let $x_{ij} = 1$ if individual i is recorded in list j and $x_{ij} = 0$ otherwise. We arrange all these indicators into individual-level vectors to form N individual *capture patterns* $\mathbf{x}_i = (x_{i1}, \ldots x_{iJ})$, one for each element of the population. For example, a pattern $\mathbf{x}_i = (0, 0, 1, 1)$ indicates that individual i was recorded by lists 3 and 4, but missed by the first two. We note that even though each individual has a capture pattern, any individual with pattern $\mathbf{0} = (0, \ldots, 0)$ is by definition unobserved. Our objective is to produce an estimate of how many individuals in the population belong to that class.

Capture-Recapture estimation of N is based on estimating the probability mass function $f(\mathbf{x}|\theta)$ for the capture patterns $\mathbf{x} \in \{0, 1\}^J$ from a sample truncated at $\mathbf{x} = \mathbf{0}$. We then use that model to predict $f(\mathbf{0}|\theta)$, and then N. In order for this to be possible it is necessary, at the very least, that whatever the model $f(\mathbf{x}|\theta)$ is, it can be estimated from the data, which, by definition, will never include the capture patterns $\mathbf{0}$. Conversely, even though we cannot observe $\mathbf{0}$, the model should make it possible to evaluate $f(\mathbf{0}|\theta)$. The model's other assumptions are mostly related to the specific form of the data generation process, and these will be encoded as specific parametric assumptions in $f(\cdot|\theta)$.

Two assumptions are commonly associated with CR estimation. The first one, *independence*, states that the probability of appearing in one list is not affected by having appeared in another list. The second one, *homogeneity*, requires that this probability distribution is the same for each individual in the population. These two can be expressed as the so-called *independence model*:

$$\mathbf{x}_i \overset{iid}{\sim} \prod_{j=1}^{J} p_j^{x_{ij}} (1 - p_j)^{1 - x_{ij}}, \quad i = 1, \ldots, N, \tag{2.1}$$

where p_j is the probability of appearing in list j. The independence model lies behind the earliest and most famous CR techniques, for example, the Petersen estimator,

$$\hat{N} = \frac{n_A \cdot n_B}{n_{AB}}, \tag{2.2}$$

where n_A and n_B are, respectively, the number of observed individuals in lists A and B, and n_{AB} is the number of individuals in common between the two.

Independence estimators like Petersen's are still occasionally useful—for instance, when both lists are independent simple random samples from the

population—but their assumptions are unrealistic in the casualty estimation setting. Specifically, probabilities of capture tend to vary, sometimes greatly, from individual to individual. From qualitative conversations with victim communities and grassroots human rights activists documenting abuses, we have learned that the two primary factors that affect the probability that an event will be observed are trust and logistics. Interviewers are asking survivors to relate events that are among the most traumatic situations that can happen to anyone. The survivors' willingness to report these events requires them to trust the interviewers. Conversely, if survivors perceive the interviewers as from rival political positions, they may choose not to disclose information to protect themselves. The second major influence on documentation dynamics is the logistical capability of each organization. Can the groups conducting documentation access the affected communities? Much mass violence occurs in remote areas. Groups that have interviewers willing to make arduous journeys may be better able to capture information in those locations. High-resource groups may be able to afford more and more adaptable vehicles, or in the case of the UN, helicopters.

List Dependence

Violations of the assumptions underlying (2.1) in the form of dependence between lists are common in casualty estimation. In the original Guatemala analysis (Ball 2000), researchers observed a form of negative dependence between two of the lists. The first of these dependent lists was the result of a qualitative investigation that took testimonies among Catholic religious communities conducted in the mid-1990s; the second was gathered by a coalition of NGOs, or nongovernmental organizations, mostly associated with the political left which took testimonies in the early 1990s among communities which had been part of the guerrillas' civilian base. Researchers noted that people in the religious communities that trusted the Catholic researchers would be less likely to report to the NGOs, and vice versa. In this scenario, individual witnesses prefer one documentation project to another, leading to negative list dependence.

Fienberg (1972) proposed to account for list dependencies through their explicit modeling as list-by-list interactions in log-linear models (see also Bishop et al. 1975). For example, using Bishop et al. (1975) notation, a model accounting for dependence between lists 1 and 2 when three lists are available would be

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)},$$

where $m_{ijk} = E[n_{ijk}]$ and n_{ijk} is the number of individuals with capture pattern $(i, j, k) \in \{0, 1\}^3$. A conditional maximum likelihood estimator for the undercount is given by the formula

$$\hat{n}_{000} = \frac{\hat{m}_{001}\hat{m}_{010}\hat{m}_{100}\hat{m}_{111}}{\hat{m}_{011}\hat{m}_{101}\hat{m}_{110}}. \tag{2.3}$$

This formula results from a no-second-order interaction assumption in the log-linear model, that is, $u_{123(ijk)} = 0$, necessary to ensure that the model is identifiable and that $f(\mathbf{0}|\theta)$ can be calculated.

Log-linear CR is a mature technology that has been used in several casualty estimation projects; e.g. Kosovo 2002 (Ball et al. 2002), Peru 1980–2000 (Ball et al. 2003), Guatemala 1982–1983 (Ball and Price 2018), and Bosnia in 1992–1995 (Zwierzchowski and Tabeau 2010). Nevertheless, this approach has several important limitations, and we will discuss some of them later in this article.

Heterogeneity

Differences in the probabilities of being listed due to individual traits are referred in the CR context as *heterogeneity of capture probabilities* (or "heterogeneity," for short). As discussed in the introduction, our experience has led us to believe that heterogeneity is the primary problem in CR applied to casualty estimation. Victims and witnesses of violence are subject to individual attributes that affect the listability of victims. We mentioned the problem of the degree of trust that witnesses put in different projects as one reason. Another important aspect is the social visibility of victims. Adults tend to be better known by their communities than children; authorities and famous people tend to be reported more than regular people; victims in remote rural locations tend to be less frequently reported than people in cities. All of these, and other more locally specific or less describable factors, contribute to the violation of the "equal distribution" assumption in model (2.1) and need special treatment. We now describe two approaches to deal with the effects of heterogeneity in CR.

The first approach to deal with heterogeneity is stratification (Sekar and Deming 1949). The idea is to use a discrete covariate that is known or suspected to be related to the source of heterogeneity to segment the population into homogeneous (or at least more homogeneous) sub-populations and estimate within them separately. A common choice in this context is place of death. For example, in Ball et al. (2003) researchers resorted to expert knowledge for dividing the Peruvian territory into 59 geographic strata roughly corresponding to known insurgent-counterinsurgent conflict dynamics, which were then treated separately. Other typical choices in casualty estimation are perpetrator agent (Ball et al. 2003) and period (Ball et al. 2002).

When properly executed, stratification can greatly help reduce the impact of heterogeneity; however, it also has important limitations. The most obvious one is the reduction of within-strata sample sizes. This reduces inferential power and can lead to identifiability problems. Another one is the need of data for the stratification

and of specialized knowledge about the relationship between heterogeneity and the available covariates.

A second, complementary approach to dealing with heterogeneity is through modeling. Sometimes heterogeneity manifests itself as list dependency and can be dealt with using log-linear and related methods. For example, even though we presented the case of Guatemala as an illustration of list dependency, a closer look reveals that the driver of said dependency were differences of listability due to individual traits, i.e. the level of trust each individual had on each documentation project.

A more direct modeling strategy is to directly represent the individual traits that lead to the differential capturability. This approach was first introduced by Sanathanan (1973) and was greatly developed in the context of animal abundance estimation as the model M_h and its variants (Otis et al. 1978). All these approaches have in common to introduce some form of individual-level random effect ω_i driving capturability:

$$\mathbf{x}_i|\theta, \omega_i \overset{ind}{\sim} f(\cdot|\theta, \omega_i), \qquad \omega_i \overset{iid}{\sim} H.$$

Most of the models developed in the animal estimation literature assume symmetric heterogeneity effects, that is, ω_i affects all lists in the same direction. An example of this structure is the Rasch model (Agresti 1994; Fienberg et al. 1999). This makes sense in ecology applications: if animals possess characteristics that make them difficult (or easy) to capture in general, they should be so for any trapping occasion.

Symmetric heterogeneity does not hold in casualty estimation. Different documentation projects often have different objectives, capabilities, and sympathies, resulting in different access to different types of victims. This means that the same individual traits ω_i may have different effects on different lists, sometimes in opposite directions. A dramatic case was observed in Peru (Ball et al. 2003). There victims of the Shining Path tended not to be captured by NGOs or the Ombudsman office, while victims of the armed forces tended to be favored by NGOs. In these cases symmetric effect models, like Rasch models, would be inadequate.

Models that allow for less constrained forms of heterogeneity have been proposed for casualty CR estimation. Manrique-Vallier (2016) proposed the use of Dirichlet process mixtures of independence models (NPLCM model). These models have been successfully used for re-analyzing heterogeneous data previously analyzed with log-linear models in Peru (Manrique-Vallier et al. 2019) and Kosovo (Manrique-Vallier 2016). It has also been used to estimate the total number of people who disappeared in the final 3 days of the Sri Lankan civil war (Ball and Harrison 2018); the number of women held as sexual slaves by Japanese authorities during World War 2 in Palembang, Indonesia (Ball et al. 2018); the number of people killed in drug-related violence in the National Capital Region of the Philippines (Ball et al. 2019a); and the number of social movement leaders killed in Colombia in 2016–2017 (Ball et al. 2019b). We discuss more about them in Sect. 2.3.

We believe that the use of flexible models that directly address heterogeneity, like NPLCM, is preferable to techniques that address the induced list dependency, like log-linear models. In Manrique-Vallier et al. (2019) we re-analyzed the Peruvian data from Ball et al. (2003) (plus an extra data source) using LCMCR and log-linear models. We noted that in cases in which results from the two approaches diverged, log-linear models were complex and required many interaction terms. We attribute this behavior to the fact that log-linear models can only address heterogeneity through its approximation using interaction terms and the no-highest-order interaction assumption. While in some cases a simple log-linear representation exists (as in the Guatemala and Kosovo cases), in others the necessary models will be highly complex and not identifiable. For a study on the relationship between LCMCR-type mixtures and log-linear models see Johndrow et al. (2017).

Model Selection

Even after selecting a family of models for CR estimation (e.g. log-linear), it is usually necessary to choose among many competing models. As in any other statistical problem, model selection can be performed using both knowledge about the problem and by formal model selection techniques.

An example of the use of substantive knowledge to guide model selection is presented in (Zwierzchowski and Tabeau 2010) for Bosnia and Herzegovina. There analysts for the International Criminal Tribunal for the Former Yugoslavia used twelve data sources (including eight enumerations of the names of people reported as dead) in a log-linear model to estimate the total of war casualties. They started by making dual-system estimates between pairs of systems. They noted which pairs seemed to produce plausible estimates, and which lead to substantially greater or lower than the plausible middle estimates. They decided that the pairs of lists that produced greater or lower estimates were those with substantial interactions. In the discussion, they describe how specific pairs of lists might be positively or negatively interacting. For example, they noted two projects that were both based in Sarajevo and both sampled deaths from Sarajevo with greater probability than deaths elsewhere. In the log-linear model, they included all the pairwise log-linear terms for the lists that they argued had substantial interactions. Naturally, such an approach is difficult to justify from a formal statistical point of view.

Formal model selection procedures have been prominently featured in casualty estimation studies using log-linear CR. The earliest of these (e.g. Ball et al. 2002; 2003) relied on exhaustive searches within the space of hierarchical log-linear models and were conducted based on the minimization of indexes that balanced parsimony with fit, like the BIC or the χ^2/df statistic. Although this is common practice in applied statistics and in CR in particular, the approach presents some important limitations. First, even though the model search is data-based (and therefore subject to sampling variability), estimation is performed conditioning on the selected model. This neglects the uncertainty associated with the model

selection process itself. Second, in many cases equally plausible models can produce substantially different results with no clear way of choosing one over the other. Finally, some families of models (like log-linear with a large number of lists) can be too large to fully evaluate.

Bayesian model averaging (BMA) avoids the model selection issue altogether. Instead of selecting one single "best" model, we average the posterior distributions of interest (in this case over the population size) over all models of a family, weighting by the posterior probability of the models themselves. Lum et al. (2010) used a BMA approach proposed by Madigan and York (1997) to estimate the number of fatal human rights violations in the department of Casanare, Colombia, in the period 1998–2007. The method of Madigan and York (1997) uses BMA in the space of decomposable graphical models, which is itself a sub-family of hierarchical log-linear models. Madigan and York (1997)'s method works in practice because discrete decomposable graphical models allow posterior estimation in closed form, and the number of models is not too large to evaluate provided the number of lists is small. However, BMA can become computationally challenging with large numbers of lists. Furthermore, as decomposable graphical models are a sub-family of hierarchical log-linear models, they also share some of their limitations; in particular, they might not be sufficiently flexible for modeling dependence induced by heterogeneity.

A different approach was taken by Manrique-Vallier (2016), who proposed the use of Dirichlet process mixtures of product-Bernoulli (independence) models. In this case the model is theoretically infinite-dimensional but has a structure that modulates the complexity of the mixture to what is needed to adjust well to the data. Similarly to BMA, this approach has the advantage of avoiding the model selection problem but avoids having to deal with several models to begin with. It also has the advantage of being computationally tractable, scaling easily to very large numbers of lists.

Recent advances notwithstanding, the problem of model selection in CR estimation presents a unique challenge. Any formal model selection procedure can only ensure that the models under consideration fit the observed data well enough to some criterion. However, since capture pattern $\mathbf{0}$ is unobservable by definition, there is no way of ensuring that a model that fits the observed part of the data well enough will lead to recovery of the true value of $f_T(\mathbf{0}|\theta_T)$ under the true model $f_T(\cdot|\theta_T)$. This is a well-known problem (see e.g. the discussion section in Bishop et al. (1975), Ch. 6) and the ultimate reason why the non-parametric CR problem is unidentifiable.

Practical Invisibility: α-Observability

An important assumption for CR estimation is that every individual in the population of interest must have a positive probability of being listed. With perhaps the exception of projects actively refusing to register particular victims or types of

victims, this condition is not difficult to meet in the context of casualty estimation: it is implausible that people can die or disappear without anybody at all noticing.

A related but less explored problem is when some individuals' probability of being listed is indeed positive, but very small. In these cases, even though the classical assumption of positive probability is satisfied, some individuals might be practically invisible to the sampling efforts. This phenomenon is especially problematic in heterogeneous populations, where it might be the case that we have several lists with plenty of data from individuals from easily observable subpopulations, but very few or none from less observable groups.

This problem was studied by Johndrow et al. (2019). They noted that this is an intrinsic problem in CR estimation under heterogeneity, and that an important consequence is that the estimation risk of the population size can, in many cases, be unbounded. As a compromise solution they proposed to abandon the objective of estimating the true population size and re-define the problem as estimating the number of individuals with a probability greater than an arbitrary threshold α of being observed ("α-observable").

2.3 Some Open Problems and Research Directions

Models and Extrapolation Assumptions

CR estimation is in its essence an extrapolation problem: use data from capture patterns in $\{0,1\}^J \setminus \{\mathbf{0}\}$ to estimate $f(\mathbf{0}|\theta)$. Since $\mathbf{0}$ is unobservable by definition, the way to project to this probability will be completely determined by the model $f(\cdot|\theta)$; this also means that truly non-parametric CR estimation is essentially impossible; see discussion in Manrique-Vallier (2016). Conversely, the way in which the probability $f(\mathbf{0}|\theta)$ relates to the rest of $f(\mathbf{x}|\theta)$, $\mathbf{x} \neq \mathbf{0}$, can neither be learned from data nor tested. The projection of the joint model to the unobservable part is related to the concept of *extrapolation distribution* in the missing data literature (Hogan and Daniels 2008).

Since statistical inference on the way in which observable patterns relate to the unobserved is impossible, selecting an appropriate model or family of models should be done in a way that best resembles the actual data generation process and with understanding of the implied extrapolation assumption. An important example is hierarchical log-linear models. As discussed in Sect. 2.2, log-linear models for J lists require an assumption of no $(J-1)$th-order interaction in order to be identifiable from data with capture patterns $\{0,1\}^J \setminus \{\mathbf{0}\}$. This condition itself defines the extrapolation assumption (from which the estimator in (2.3) is derived). The question then becomes: is this particular way of extrapolating reasonable for casualty estimation?

As explained in Sect. 2.2, we believe that in most casualty estimation problems, heterogeneity is the main driver of departures from the independence model and

so, with some exceptions, log-linear models are just an approximation to the true joint distribution of data. Therefore, even if the models fit the observed data well, the extrapolation assumption might not be appropriate for this problem. On the other hand, models that directly represent plausible heterogeneity structures, like NPLCM, might be more appropriate. Which models and extrapolation assumptions are better for different scenarios in casualty estimation is an open question that would benefit from additional research.

Data-Based Stratification

Stratification is often used as a first approach to tackle heterogeneity (see Sect. 2.2). The usual practice consists in using qualitative expert knowledge to find a partition of the population that could result in homogeneous sub-populations and estimate within each of them separately. Oftentimes, after trying a stratification scheme, some strata will still exhibit signs of residual heterogeneity. In these cases researchers sometimes revise the stratification scheme, adjust, and try again. For example, in Peru (Ball et al. 2003) researchers determined regional conflict dynamics and stratified accordingly. Then, after noting that model fitting in some of the regions was poor, they sub-divided them into smaller pieces forming a finer stratification scheme.

This iterative procedure seems natural and intuitive but is statistically problematic. Specifically, the process of looking at the results obtained under a stratification scheme to modify it is itself a data-based decision that is likely to alter the validity of inferences—similar to the so-called p-hacking problem (Gelman and Loken 2013). Manrique-Vallier et al. (2019) noted this problem in their re-analysis of the Peruvian data. They addressed it using a partial blinding procedure: two of the authors performed the calculations without sharing the results with the third, while the latter proposed sub-stratification schemes only based on external qualitative knowledge. This procedure partially addressed the risk of cherrypicking results based on what the researchers would want to see. However, the selection of which regions to sub-divide was still based on data-based evaluations of model fitting.

A possible alternative is to formally incorporate the stratification process into the modeling and estimation procedures. Let $\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}$ be the finest partition of the population we are willing to consider, determined from subject matter knowledge. Let us call these partitions *atomic strata*. Taking $\mathcal{Y}$ as the stratification scheme is equivalent to fitting M models $f(\cdot|\theta_{y_1}), \ldots, f(\cdot|\theta_{y_M})$ to each atomic stratum. On the other extreme, we can think of unstratified estimation as making the parameters of all M models equal, i.e. $\theta_{y_1} = \theta_{y_1} = \cdots = \theta$. In between, we can represent different stratification schemes as different agglomerations of atomic strata, where parameters are equal. For example, if we wanted to create a stratum that combines strata 1, 2, and 3, we would represent it by enforcing the restriction $\theta_{y_1} = \theta_{y_2} = \theta_{y_3}$. Using this idea we can think of performing simultaneous estimation of the stratification scheme and CR parameters (including

the population size) by specifying prior distributions that put positive mass into relevant groupings of atomic strata by enforcing equality on their parameters. This idea is similar to the method of Price et al. (2019) for the automatic combination of categories in logistic regression. This construction can also allow enforcing meaningful structures, like geographic or temporal contiguity by appropriately allowing equality among neighboring atomic strata.

Missing Data

As noted by Fienberg and Manrique-Vallier (2009), CR can be seen itself as a missing data problem. Indeed, many estimation methods are based on data- or sample-augmentation schemes that represent unobserved individuals as missing records—see e.g. Manrique-Vallier (2016). This makes it natural to combine CR with other forms of missing data problems and methods.

A frequent scenario in casualty estimation is when data for stratification is missing for some individuals. For example, in the study of the Peruvian conflict (Ball et al. 2003; Manrique-Vallier et al. 2019), about 10% of the records missed perpetrator attribution. As noted by Zwane and van der Heijden (2007)—who studied the problem for the special case of variables completely missing in some of the lists—in these cases the common practices of ignoring incomplete covariates, creating a special category out of them, or imputing "reasonable" values can be a source of either biases or too optimistic precision.

Manrique-Vallier et al. (2019) proposed a framework for Bayesian stratified CR estimation with incomplete stratification information in one covariate. They combined it with the model from Manrique-Vallier (2016) and used the resulting method to estimate deaths in the Peruvian conflict. The method is based on using a data-augmentation representation for both the unobserved individuals and the missing stratification which is then estimated using Markov Chain Monte Carlo simulation. At its core this method is based on a Missing at Random assumption (Little and Rubin 2002) whereby the information used to infer the missing stratification is obtained from records with similar capture patterns. A natural extension of this idea is to complement the information from the capture patterns with other variables. For example, in Manrique-Vallier et al. (2019) researchers had access to covariates that were not used in stratification (like age) which might be related to the missing stratification labels and could be used to better estimate them.

Data Copying Between Lists

An important exception to our belief of heterogeneity being the main driver of dependence between lists in casualty estimation is the case of sharing or copying of records between documentation projects. In these cases, in addition to gathering

first-hand information about casualties of a conflict, some projects directly incorporate data obtained by other projects into their listings. In our experience this is not a prevalent problem across the casualty estimation projects in which we have been involved. However, when it happens, its effect is noticeable. One example occurs in the conflict in Syria, where the Human Rights Data Analysis Group (HRDAG) has longitudinal access to lists put together by different projects. The databases are shared multiple times over time, as they are updated when new deaths are known and when additional information about previously reported deaths becomes available. In some cases, the overlap between two lists increases substantially between updates, where the newly overlapping records are found not to be present in one of the databases in the previous iteration, and the new records match exactly records in the other database. HRDAG, in conversation with one of the groups, learned that they copy published records from the others. This is a reasonable strategy for a group trying to maintain a comprehensive list, but it creates a strong positive dependence between the lists.

Copying of records between lists that also directly gather first-hand information is problematic in CR because it superimposes and confounds two data generation processes: the capture of individuals by documentation projects, and the relationship between those projects. From these, only the former process is useful for inferring the population size. Thus we need to somehow disentangle them. An ideal situation is that projects record the source(s) of each record so that we can identify which records have only been copied and remove them prior to statistical analysis. In the absence of such information we may try to model the copying process. This strategy will likely require external sources of information and/or strong and untestable assumptions to overcome non-identifiability. One of such possible additional sources of information can result from integrating the CR estimation and the record linkage process.

Internal Duplication

Typical multi-list CR methods (like all the ones that have been used for casualty estimation studies so far) only work with information about presence or absence in lists, in the form of vectors in $\{0, 1\}^J \setminus \{\mathbf{0}\}$. These vectors are usually the result of J-way record linkage among J lists, where individual lists are assumed to be free of duplicated records. In fact, in most projects an important part of the data preparation is making sure that the internal duplication within lists has been eliminated.

Internal duplication within lists carries plenty of useful information that can be lost during the "cleanup" process. In the same way in which the presence of an individual in more than one list is usually interpreted as an indication of a higher probability of being observed, repeated presence *in the same list* (or "duplication") can also contribute to the same conclusion. To take advantage of these data we need to create methods in which the multivariate capture patterns are not simply strings of zeros or ones, but of natural numbers, $\mathbf{x} \in \{0, 1, 2 \ldots\}^J$. A simple

version of such a model, assuming independence between two lists, has been proposed by Lerdsuwansri and Böhning (2018). However, the casualty estimation context is likely to require much more sophisticated multi-list models that represent plausible data generation scenarios and that can be integrated with other sources of information. An additional level of complication comes from the fact that, in practice, there will be uncertainty on which records are duplicates within a single dataset, that is, the counts $\mathbf{x} \in \{0, 1, 2 \ldots\}^J$ will be known with error (see e.g. Sadinle 2014; Steorts et al. 2016).

Record Linkage Errors

The capture patterns $\mathbf{x}_i \in \{0, 1\}^J \setminus \{\mathbf{0}\}$ are the essential input for all CR techniques. To obtain these we need to identify individuals that appear in multiple lists by linking their corresponding records. In the context of armed conflicts, witnesses, or victims of violence may report an event to different organizations at different points in time and with different degrees of detail. Unfortunately, reporting or collecting unique identifiers, such as national identification numbers, is rare in this context. This means that even the more basic question of how many unique casualties were reported to any one group cannot be easily answered, as it is often difficult to determine which records belong to the same individuals.

Probabilistic record linkage techniques (see e.g. Fellegi and Sunter 1969; Winkler 1988; Jaro 1989; Larsen and Rubin 2001; Sadinle and Fienberg 2013; Steorts et al. 2016; Sadinle 2017) take advantage of imperfect partial identifiers collected on the individuals, such as names and demographic information, dates and locations of the events. These pieces of information are usually subject to typographical and other types of errors, which lead to uncertainty in the correct way of linking the records. The result of the record linkage process will typically contain errors termed *false links* and *false non-links*, that is, records that were incorrectly linked and records that were incorrectly left unlinked, respectively. A false non-link can, for example, lead to a true capture pattern (0, 0, 1, 1) being incorrectly registered as two capture patterns (0, 0, 0, 1) and (0, 0, 1, 0); conversely, a false link can, for example, lead to two capture patterns (0, 0, 0, 1) and (0, 0, 1, 0) being incorrectly counted as (0, 0, 1, 1). Similar errors appear when record linkage techniques are used for duplicate detection within each list.

The effect of linkage errors is clearly seen in the Petersen estimator (2.2) in the case of two lists. Between the lists, false links will lead to higher n_{AB} and thereby lower population size estimates, whereas false non-links will lead to lower n_{AB} and higher population size estimates. Within each list, false links will lead to lower n_A and n_B and therefore lower population size estimates, whereas false non-links will lead to higher n_A and n_B and therefore higher population size estimates. For multiple lists, the specific impact of linkage errors will depend on the models being used.

Broadly speaking, the output of a linkage procedure can be seen as an estimator for the underlying correct way of linking the records. As every estimation procedure, the linkage is subject to sampling variability, and we are interested in "transferring" this "linkage uncertainty" into the population size estimation, with the goal of having final estimates that reflect the fact that the linkage is subject to error. Two strategies come to mind: a joint modeling strategy for both the linkage and the population size estimation, and a two-stage strategy where the output of probabilistic linkage is fed into the population size estimation. The first approach has been undertaken by Liseo and Tancredi (2011) and Tancredi and Liseo (2011), who created a joint Bayesian modeling strategy that combines a model for record linkage with a model for population size estimation; although their work focuses on the case of two lists, their strategy could in principle be extended to more general models. The second approach was undertaken by Sadinle (2018), who proposed a procedure called *linkage-averaging*, where the linkage and the population size estimation can be carried out in two separate stages, while still leading to proper Bayesian inferences under some conditions.

A characteristic of the joint modeling strategy is that the analyst will have to run the record linkage and the CR model jointly for each different CR model, which can be computationally intensive, whereas in the two-stage strategy the results from record linkage can be reused with different CR models. Another characteristic of both of these approaches is that their success is determined by the success of their record linkage and CR components. For example, if the record linkage model over-links or under-links, then the population size estimates will be lower or higher, respectively, with respect to what we would obtain under the correct linkage, regardless of whether one uses a joint model or a two-stage approach. Similarly, if the model for population size estimation is wrong, our estimates will be deficient regardless of whether one uses that model in a joint model or in a two-stage approach. Further research should be devoted to better understanding the properties of these strategies and to develop alternatives.

2.4 Final Comments

Our goal in this discussion was not to be exhaustive but rather to present some of the challenges, approaches, and directions we are most familiar with. CR for casualty estimation could benefit from developments in many other areas of statistics, such as model selection in regression problems, post-selection inference, small area estimation, and spatio-temporal modeling, just to name a few. Furthermore, CR techniques that are developed for estimating animal abundance in ecology, for corrections to census enumerations, or for disease prevalence estimation in epidemiology will also continue to be potentially useful for casualty estimation.

CR for casualty estimation is an area of research posing several technical challenges that have traditionally been bypassed in applications via ad-hoc solutions. More adequate solutions should account for the uncertainty in the correct ways of

modeling, extrapolating, stratifying, handling missing data, and deduplicating and linking records. Unfortunately, the flip side of better handling of these issues is that we will necessarily obtain casualty estimates with much broader uncertainty intervals. This can potentially mean that in certain situations the intervals will become too large to be practically useful. Nevertheless, it is desirable to have estimation methodologies that provide us with honest assessments of uncertainty and thereby avoid misleading and overconfident results.

References

Agresti, A. (1994), "Simple capture-recapture models permitting unequal catchability and variable sampling effort," *Biometrics*, 50, 494–500.

Amorós, J. (2014), "Recapturing Laplace," *Significance*, 11, 38–39.

Anderson, M. and Fienberg, S. E. (1999), "To sample or not to sample: The 2000 census controversy," *Journal of Interdisciplinary History*, 30, 1–36.

Ball, P. (2000), "The Guatemalan Commission for Historical Clarification: Intersample Analysis," in *Making the Case: Investigating Large Scale Human Rights Violations using Information Systems and Data Analysis*, eds. Ball, P., Spirer, H. F., and Spirer, L., American Association for the Advancement of Science, chap. 11.

Ball, P., Asher, J., Sulmont, D., and Manrique, D. (2003), "How many Peruvians have died? An estimate of the total number of victims killed or disappeared in the armed internal conflict between 1980 and 2000," AAAS. Report to the Peruvian Truth and Reconciliation Commission (CVR). Also published as Anexo 2 (*Anexo Estadístico*) of CVR Report.

Ball, P., Betts, W., Scheuren, F., Dudukovic, J., and Asher, J. (2002), "Killings and Refugee Flow in Kosovo, March–June, 1999," Report to ICTY.

Ball, P., Coronel, S., Padilla, M., and Mora, D. (2019a), "Drug-Related Killings in the Philippines," Tech. rep., Human Rights Data Analysis Group and the Stabile Center for Investigative Journalism.

Ball, P. and Harrison, F. (2018), "How many people disappeared on 17–19 May 2009 in Sri Lanka?" Tech. rep., Human Rights Data Analysis Group and the International Truth and Justice Project.

Ball, P. and Price, M. (2018), "The statistics of genocide," *CHANCE*, 31, 38–45.

Ball, P. and Price, M. (2019), "Using Statistics to Assess Lethal Violence in Civil and Inter-State War," *Annual Review of Statistics and Its Application*, 6:1, 63–84.

Ball, P., Rodríguez, C., and Rozo, V. (2019b), "Asesinatos de líderes sociales en Colombia en 2016–2017: una estimación del universo," Tech. rep., Human Rights Data Analysis Group and Dejusticia.

Ball, P., Shin, E. H.-S., and Yang, H. (2018), "There may have been 14 undocumented Korean "comfort women" in Palembang, Indonesia," Tech. rep., Human Rights Data Analysis Group and Transitional Justice Working Group.

Bird, S. M. and King, R. (2017), "Multiple Systems Estimation (or Capture-Recapture Estimation) to Inform Public Policy," *Annual Review of Statistics and Its Application*, 5.

Bishop, Y., Fienberg, S., and Holland, P. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press, reprinted in 2007 by Springer-Verlag, New York.

Darroch, J., Fienberg, S., Glonek, G., and Junker, B. (1993), "A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability," *Journal of the American Statistical Association*, 88, 1137–1148.

Fellegi, I. P. and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183–1210.

Fienberg, S. (1972), "The Multiple recapture census for closed populations and incomplete 2^k contingency tables," *Biometrika*, 59, 591–603.

Fienberg, S., Johnson, M., and Junker, B. (1999), "Classical multilevel and Bayesian approaches to population size estimation using multiple lists," *Journal of the Royal Statistical Society. Series A*, 162, 383–406.

Fienberg, S. E. and Manrique-Vallier, D. (2009), "Integrated methodology for multiple systems estimation and record linkage using a missing data formulation," *AStA-Advances in Statistical Analysis*, 93, 49–60.

Gelman, A. and Loken, E. (2013), "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time," Unpublished paper.

Hogan, J. W. and Daniels, M. J. (2008), *Missing Data in Longitudinal Studies*, Boca Raton: Chapman and Hall.

International Working Group for Disease Monitoring and Forecasting (1995a), "Capture-recapture and multiple-record systems estimation I: History and theoretical development," *American Journal of Epidemiology*, 142, 1047–1058.

—— (1995b), "Capture-recapture and multiple-record systems estimation II: Applications in human diseases," *American Journal of Epidemiology*, 142, 1059–1068.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414–420.

Johndrow, J. E., Bhattacharya, A., and Dunson, D. B. (2017), "Tensor decompositions and sparse log-linear models," *Annals of statistics*, 45, 1.

Johndrow, J. E., Lum, K., and Manrique-Vallier, D. (2019), "Low-risk population size estimates in the presence of capture heterogeneity," *Biometrika*, 106, 197–210.

Laplace, P. S. (1786), "Sur les naissances, les mariages et les morts," in *Histoire de L'Académie Royale des Sciences*.

Larsen, M. D. and Rubin, D. B. (2001), "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 96, 32–41.

Lerdsuwansri, R. and Böhning, D. (2018), "Extending the Lincoln-Petersen Estimator when Both Sources are Counts," in *Capture-Recapture Methods for the Social and Medical Sciences*, eds. Böhning, D., Van Der Heijden, P. G., and Bunge, J., Boca Raton, FL: Chapman & Hall/CRC, chap. 23, pp. 341–360.

Liseo, B. and Tancredi, A. (2011), "Bayesian Estimation of Population Size via Linkage of Multivariate Normal Data Sets," *Journal of Official Statistics*, 27, 491–505.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data: Second Edition*, New York: John Wiley & Sons.

Lum, K., Price, M., Guberek, T., and Ball, P. (2010), "Measuring Elusive Populations with Bayesian Model Averaging for Multiple Systems Estimation: A Case Study on Lethal Violations in Casanare, 1998–2007," *Statistics, Politics and Policy*, 1.

Madigan, D. and York, J. C. (1997), "Bayesian methods for estimation of the size of a closed population," *Biometrika*, 84, 19–31.

Manrique-Vallier, D. (2016), "Bayesian Population Size Estimation Using Dirichlet Process Mixtures," *Biometrics*, 72, 1246–1254.

Manrique-Vallier, D., Ball, P., and Sulmont, D. (2019), "Estimating the Number of Fatal Victims of the Peruvian Internal Armed Conflict, 1980-2000: an application of modern multi-list Capture-Recapture techniques," arXiv preprint, arXiv:1906.04763v2 [stat.AP, stat.ME].

Mulry, M. H. and Spencer, B. D. (1991), "Total Error in PES Estimates of Population," *Journal of the American Statistical Association*, 86, 839–855.

Okiria, A. G., Bolo, A., Achut, V., Arkangelo, G. C., Michael, A. T. I., Katoro, J. S., Wesson, J., Gutreuter, S., Hundley, L., and Hakim, A. (2019), "Novel Approaches for Estimating Female Sex Worker Population Size in Conflict-Affected South Sudan," *JMIR Public Health Surveill*, 5, e11576.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978), "Statistical inference from capture data on closed animal populations," *Wildlife monographs*, 3–135.

Price, B. S., Geyer, C. J., and Rothman, A. J. (2019), "Automatic Response Category Combination in Multinomial Logistic Regression," *Journal of Computational and Graphical Statistics*, 28, 758–766.

Sadinle, M. (2014), "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach," *Annals of Applied Statistics*, 8, 2404–2434.

—— (2017), "Bayesian Estimation of Bipartite Matchings for Record Linkage," *Journal of the American Statistical Association*, 112, 600–612.

—— (2018), "Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations," *Annals of Applied Statistics*, 12, 1013–1038.

Sadinle, M. and Fienberg, S. E. (2013), "A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems," *Journal of the American Statistical Association*, 108, 385–397.

Sanathanan, L. (1973), "A comparison of some models in visual scanning experiments," *Technometrics*, 15, 67–78.

Sekar, C. C. and Deming, W. E. (1949), "On a Method of Estimating Birth and Death Rates and the Extent of Registration," *Journal of the American Statistical Association*, 44, 101–115.

Steorts, R. C., Hall, R., and Fienberg, S. E. (2016), "A Bayesian Approach to Graphical Record Linkage and Deduplication," *Journal of the American Statistical Association*, 111, 1660–1672.

Tancredi, A. and Liseo, B. (2011), "A Hierarchical Bayesian Approach to Record Linkage and Size Population Problems," *Annals of Applied Statistics*, 5, 1553–1585.

Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 667–671.

Zaslavsky, A. and Wolfgang, G. (1993), "Triple-system modeling of census, post-enumeration survey, and administrative-list data," *Journal of Business & Economic Statistics*, 11, 279–288.

Zwane, E. and van der Heijden, P. (2007), "Analysing capture–recapture data when some variables of heterogeneous catchability are not collected or asked in all registrations," *Statistics in Medicine*, 26, 1069–89.

Zwierzchowski, J. and Tabeau, E. (2010), "The 1992–95 War in Bosnia and Herzegovina: Census-Based Multiple System Estimation of Casualties' Undercount," *Berlin: Households in Conflict Network and Institute for Economic Research*, 539.

Chapter 3
Algebraic Statistics, Tables, and Networks: The Fienberg Advantage

Elizabeth Gross, Vishesh Karwa, and Sonja Petrović

3.1 Introduction

Stephen Fienberg's early work on contingency tables [BFH74] relies on using intrinsic model geometry to understand the behavior of estimation algorithms, asymptotics, and model complexity. For example, in [Fie70], Fienberg gives a geometric proof of the convergence of the iterative proportional fitting algorithm for tables with positive entries. The result in [Fie70] relies on his study of the geometry of $r \times c$ tables in [Fie68] and his and Gilbert's geometric study of 2×2 tables [FG70]. This approach to understanding models would eventually fit within the field of algebraic statistics, a general research direction that would take hold in the 2000s, over 25 years after the publication of [Fie70] and the 1974 edition of Bishop, Fienberg, and Holland's book [BFH74], whose cover displayed the independence model for 2×2 tables as an algebraic surface.

The term "algebraic statistics" was coined in 2001 [PRW01] and generally refers to the use of broader algebraic—non-linear—and geometric—non-convex—tools in statistics. While the use of algebra and geometry had been long present in statistics, before the 2000s, linear algebra and convex geometry were the main tools used consistently. The field of algebraic statistics is now a branch of mathematical statistics that relies on insights from computational algebraic geometry, combinatorial

E. Gross
University of Hawai'i at Mānoa, Honolulu, HI, USA
e-mail: egross@hawaii.edu

V. Karwa
Temple University, Philadelphia, PA, USA
e-mail: vishesh@temple.edu

S. Petrović (✉)
Illinois Institute of Technology, Chicago, IL, USA
e-mail: sonja.petrovic@iit.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_3

geometry, and commutative algebra to improve statistical inference. As algebraic statistics matured and caught the attention of many researchers, Fienberg and his students and collaborators reformulated several fundamental statistical problems, e.g. existence of maximum likelihood estimators and ensuring data privacy, into the language of polyhedral and algebraic geometry. Today Fienberg's intuition and influence remain central to one of the principal applications in algebraic statistics: testing goodness of fit of log-linear models for discrete data. Within the last decade or so, much of his work in this area focused on log-linear *network* models. In this regard, Fienberg defined new models, explained how to represent relational data as contingency tables in order to apply tools from categorical data analysis, and addressed the problems of estimation, model fit, and model selection. This paper presents a brief overview of this line of work heavily influenced by Fienberg's vision, which continues to inspire us.

3.2 Geometry and Algebra of Log-Linear Models

Let us recall the basics and fix notation. Let $\mathcal{I} = [d_1] \times \cdots \times [d_k]$ be a finite set that indexes *cells* in a contingency table $u \in \mathbb{Z}_{\geq 0}^{d_1 \times \cdots \times d_k}$. The $(i_1, \ldots, i_k)$-cell counts the number of occurrences of the event $\{X_1 = i_1, \ldots, X_k = i_k\}$ for k categorical random variables with X_i taking values on a finite set $[d_i] := \{1, \ldots, d_i\}$. Log-linear models are probability distributions on the discrete set $\mathcal{I}$ whose sufficient statistics are given by marginals, i.e. subtables of the table u obtained by summing u across a subset the index set $\mathcal{I}$; since marginalization is a linear map, it can be presented as matrix multiplication. Specifically, a log-linear model for $\mathcal{I}$ is a linear exponential family defined by an $m \times |\mathcal{I}|$ matrix A, called the *design matrix*, taking the following form:

$$P_\theta(U = u) = \exp\{\langle Au, \theta \rangle - \psi(\theta)\}, \tag{3.1}$$

where $\theta \in \mathbb{R}^m$ is the vector of model parameters and $\psi(\theta)$ the normalizing constant. Note that specifying the matrix A completely specifies the contingency table model for $X_1, \ldots, X_k$, as it determines the vector of minimal sufficient statistics Au for the linear exponential family in (3.1). As is customary in algebraic statistics, we will denote the model (3.1) by $\mathcal{M}_A$.

Let us consider one of Fienberg's early favorite examples: the model of independence of two categorical random variables X_1 and X_2. Here, A is a $(d_1 + d_2) \times d_1 d_2$ matrix of the following form, where the first d_1 rows each have d_2 ones and the last d_2 rows contain d_1 copies of the $d_2 \times d_2$ identity matrix:

$$A = \left[\begin{array}{c|c|c|c} 1\ 1\ \cdots\ 1 & 0\ 0\ \cdots\ 0 & \cdots & 0\ 0\ \cdots\ 0 \\ 0\ 0\ \cdots\ 0 & 1\ 1\ \cdots\ 1 & \cdots & 0\ 0\ \cdots\ 0 \\ \vdots\ \vdots\ \ \vdots\ \ \vdots & \vdots\ \vdots\ \ \vdots\ \ \vdots & \vdots & \vdots\ \vdots\ \ \vdots\ \ \vdots \\ 0\ 0\ \cdots\ 0 & 0\ 0\ \cdots\ 0 & \cdots & 1\ 1\ \cdots\ 1 \\ \hline 1\ 0\ \cdots\ 0 & 1\ 0\ \cdots\ 0 & \cdots & 1\ 0\ \cdots\ 0 \\ 0\ 1\ \cdots\ 0 & 0\ 1\ \cdots\ 0 & \cdots & 0\ 1\ \cdots\ 0 \\ \vdots\ \vdots\ \ddots\ \vdots & \vdots\ \vdots\ \ddots\ \vdots & \vdots & \vdots\ \vdots\ \ddots\ \vdots \\ 0\ 0\ \cdots\ 1 & 0\ 0\ \cdots\ 1 & \cdots & 0\ 0\ \cdots\ 1 \end{array}\right].$$

The sufficient statistic for $\mathcal{M}_A$ is the vector of marginal counts (that is, table row and column sums). For a contingency table u, these counts are computed as:

$$A \begin{bmatrix} u_{11} \\ \vdots \\ u_{d_1 d_2} \end{bmatrix} = \begin{bmatrix} u_{1+} & \ldots & u_{+d_2} \end{bmatrix}. \tag{3.2}$$

In [Fie68], Fienberg describes the geometry of $\mathcal{M}_A$ in detail, describing the model of independence as the intersection of the *manifold of independence* with the probability simplex. In algebraic geometry, the manifold of independence is a Segre variety, a categorical product, which Fienberg describes explicitly by detailing the linear spaces corresponding to the product fibers, or in other words, every preimage of the map described by Eq. (3.2). In addition, the defining equations of the Segre variety corresponding to the independence model are stated in [Fie68] in statistical terms (see Section 4 of [Fie68]). These equations, which are polynomial equations in indeterminates that represent joint cell probabilities, are a key ingredient to assessing model fit.

Indeed, assessing model fit for log-linear models, and consequently, log-linear network models, is possible due to a fundamental result in algebraic statistics that establishes a connection between model-defining polynomials and sampling from the conditional distributions of log-linear models. The model-defining polynomials of interest are generating sets of polynomial ideals called *toric ideals* [Stu96], §4 and §5. The essential component, which binds together the statistical and algebraic, is the vector of (minimal) sufficient statistics for the log-linear exponential family, the vector Au in the definition above.

One way to perform goodness-of-fit testing for log-linear models, especially in sparse settings such as networks, is to perform Fisher's exact test (see e.g. Section 2.6 in [Agr92]). In many cases, however, it is infeasible to compute the exact conditional p-value, thus it is estimated using a Markov chain Monte Carlo (MCMC) method. The *exact conditional p-value* of a contingency table u is the probability that the Pearson's χ^2 statistic of a random data table is larger than

the Pearson's χ^2 statistic of the table u, conditional on the observed values of the sufficient statistics. The set of all tables with the same sufficient statistics as u is called *the fiber of u under the model* $\mathcal{M}_A$ and is defined as follows:

$$\mathcal{F}_A(u) := \{v \in \mathbb{Z}_{\geq 0}^{d_1 \times \ldots \times d_k} : Au = Av\}.$$

The naming of the reference set $\mathcal{F}_A(u)$ is derived from algebraic geometry: a fiber of a point of the linear map defined by A is its preimage under that map; in this case, we are considering the set of non-negative integer points in the preimage of the sufficient statistics Au. In order to perform the MCMC method to estimate the exact conditional p-value, a set of moves must be given, and this set of moves must connect all elements in the fiber $\mathcal{F}_A(u)$ so that the conditional distribution on the fiber can be sampled properly. Such a set of moves is called a Markov basis.

Definition 2.1 A *Markov basis* of the model $\mathcal{M}_A$ is a set of tables $\mathcal{B} := \{b_1, \ldots, b_n\} \subset \mathbb{Z}^{d_1 \times \ldots \times d_k}$ for which

$$Ab_i = 0$$

and such that for any contingency table $u \in \mathbb{Z}_{\geq 0}^{d_1 \times \ldots \times d_k}$ and for any $v \in \mathcal{F}_A(u)$, there exist $b_{i_1}, \ldots, b_{i_N} \in \mathcal{B}$ that can be used to reach v from u:

$$u + b_{i_1} + \ldots + b_{i_N} = v$$

while remaining in the fiber at each step:

$$u + \sum_{j=0}^{N} b_{i_j} \in \mathcal{F}_A(u), \text{ for } j = 1 \ldots N.$$

Note that the last requirement simply means that each move needs to preserve non-negativity of cells. As an example, let us consider the independence model with $N = 2$, $d_1 = 3$, and $d_2 = 3$. Then the fiber $F_A(u)$ for any u is a collection of 3×3 tables. Examples of three different Markov moves for the independence model in this setting are

1	−1	0
−1	1	0
0	0	0

,

−1	0	1
0	0	0
1	0	−1

,

0	0	0
0	1	−1
0	−1	1

.

It is hard to check *a priori* whether a given set of moves does in fact form a Markov basis for the model. However, the following foundational result from algebraic statistics allows one to compute a Markov basis by computing a generating set of a polynomial ideal.

Theorem 2.2 ([DS98]) *A set of vectors* $\mathcal{B} = \{b_1, \ldots, b_n\}$ *is a Markov basis of the log-linear model* $\mathcal{M}_A$ if and only if *the corresponding set of binomials* $\{x^{b_i^+} - x^{b_i^+}\}_{i=1,\ldots,n}$ *generates the toric ideal* $I_A := (x^u - x^v : u - v \in \ker_{\mathbb{Z}} A)$.

Considering again the independence model with $N = 2$, $d_1 = 3$, and $d_2 = 3$, the binomials associated with the three tables above are

$$x_{11}x_{22} - x_{12}x_{21}, \quad x_{13}x_{31} - x_{11}x_{33}, \quad x_{22}x_{33} - x_{23}x_{32}.$$

One can check that these three polynomials are not enough to generate the ideal I_A, and thus more moves are needed for a Markov basis.

Theorem 2.2 is a powerful result that connects categorical data analysis to algebra. By connecting network analysis to categorical data analysis, Fienberg was able to use the full force of this theorem for testing model fit of statistical network models.

3.3 Log-Linear ERGMs and Goodness-of-Fit Testing

As stated in the editorial piece [PSY19], Fienberg took joy in rediscovering old concepts from new points of view that gave them new interpretations and wider applicability; this was evident not only from his research articles and conference presentations, but various interviews, see, for example, [Vie15]. We follow his lead in the way we define *log-linear network models*.

Generally, a statistical network model is a collection of probability distributions over $\mathcal{G}_n$, the set of all (un)directed graphs on n vertices. The Fienberg approach to the analysis of statistical network models, dating back to the late '70s and early '80s, relies on explicitly making the connection to categorical data analysis by viewing graphs as contingency tables. For example, in [FW81a], Fienberg and Wasserman view a directed graph with n vertices as a $n \times n \times 2 \times 2$ table Y where $Y_{ij00} = 1$ if there is no edge between vertex i and j, $Y_{ij11} = 1$ if there is a reciprocated edge between i and j, $Y_{ij10} = 1$ if there is a non-reciprocated edge from i to j, and $Y_{ij01} = 1$ if there is a non-reciprocated edge from j to i, and all entries are 0 otherwise. Using this $n \times n \times 2 \times 2$ table, Fienberg and Wasserman then describe nine variants of a simple statistical network model, called the p_1 model [HL81], in terms of table marginals and show how these models can be fit using a version of iterative proportional scaling for multidimensional contingency tables. In addition, they also develop a variant of the p_1 model for K subgroups determined by nodal attributes, by collapsing the $n \times n \times 2 \times 2$ into a $K \times K \times 2 \times 2$ table; a precursor to the directed stochastic blockmodels.

The p_1 model and its variants described by Fienberg and Wasserman in [FW81b] are examples of log-linear ERGMs. Log-linear ERGMs are exponential family random graph models with a log-linear interpretation. Another example of log-linear models is stochastic blockmodels, which are given a contingency representation in

[FMW85]. Following the contingency table framework of the Fienberg approach, to define a log-linear ERGM, one chooses an embedding $\phi : \mathcal{G}_n \to \mathbb{R}^\ell$ such that for all $G = (V, E)$ we have $\phi(G) = \sum_{e \in E} \phi(e)$, and implicitly uses the embedding ϕ to represent G as a vector. For example, for directed graphs, a reasonable embedding would embed $\mathcal{G}_n$ into $\mathbb{R}^{n^2}$ and G would be represented by its vectorized adjacency matrix, while for undirected graphs $\mathbb{R}^{\binom{n}{2}}$ would work equally well. For directed graphs, a suitable embedding rooted in [FW81a] (see also [FW81b]) maps $\mathcal{G}_n$ into $\mathbb{R}^{n \times n \times 2 \times 2}$ by representing graphs by their vectorized $n \times n \times 2 \times 2$ Fienberg-Wasserman table as described above or a vectorized table of size $\binom{n}{2} \times 2 \times 2$ after removing redundant cells. These embeddings allow us to refer to graphs as vectors.

An exponential family random graph model, or an ERGM for short, is a collection of probability distributions on $\mathcal{G}_n$ that places the following probability on each graph $G \in \mathcal{G}_n$:

$$P_\theta(G) = Z(\theta) e^{\theta \cdot t(G)}, \tag{3.3}$$

where G is uniquely represented as a vector in $\mathbb{R}^\ell$, θ is a row vector of parameters of length q, the map $t : \mathbb{R}^\ell \to \mathbb{R}^q$ computes the sufficient statistics, and $Z(\theta)$ is a normalizing constant. The image of the sufficient statistic map t is a vector in which each entry is a network statistic used to specify the model, such as edge count, degree of a given vertex, number of edges in a given block of vertices, etc. When the sufficient statistic is a linear function on the entries of a natural contingency table representation of the graph, as in degree-based models or stochastic blockmodels, then the sufficient statistic map t can be described with a design matrix A and the model (3.3) takes the form of (3.1). When this happens, we call the model a log-linear ERGM.

Definition 3.1 We call an exponential family random graph model a *log-linear ERGM* if the sufficient statistic map t in the ERGM specification (3.3) is a *linear* map $t : \mathbb{R}^\ell \to \mathbb{R}^q$ from the space of graphs to the space of the minimal sufficient statistics of the model.

Log-linear ERGMs include degree-based models such as the β-model, models that include effects for reciprocity, such as p_1 models, and models for data with categorical nodal attributes, such as stochastic blockmodels. Since the sufficient statistic t is a linear map, dyadic independence is implied for a log-linear ERGM. Dyadic independence is another way to say that for each pair of vertices, i and j, the edge configuration (e.g. no edge between i and j, directed edge from i to j, directed edge from j to i, bidirected edge between i to j) is independent from the edge configuration between any other pair of vertices. Thus, we can fully specify a log-linear ERGM by specifying the distribution over each set of dyadic configurations.

Example 3.2 (Stochastic Blockmodels) Extremely popular in practice, this family of log-linear ERGMs models networks whose nodes are partitioned into groups–

blocks—according to some nodal attributes. For a directed network, each dyad can be in one of four states represented as follows: (0, 0) represents no edge, (1, 0) an edge from i to j, (0, 1) an edge from j to i, and (1, 1) a bidirected edge. Note that if the network is undirected, the model simply collapses to having only two dyadic states: (0,0) and (1,1). Denote by p_{ijkl} the probability of the dyad (i, j) to be in state (k, l).

Edge formation is governed by what Fienberg and Wasserman call choice parameters, denoted by δ^{rs}, and reciprocity effects ρ^{rs}. These parameters are defined on the level of blocks. In addition, Fienberg liked the use of an additional set of parameters λ_{ij} for normalization: ensuring that each dyad is observed in only one state at a time. Specifically, the model was defined in [FMW85] as follows:

$$\begin{aligned}
\log p_{ij00} &= \lambda_{ij} \\
\log p_{ij10} &= \lambda_{ij} + \delta^{b(i)b(j)} \\
\log p_{ij01} &= \lambda_{ij} + \delta^{b(j)b(i)} \\
\log p_{ij11} &= \lambda_{ij} + \delta^{b(i)b(j)} + \delta^{b(j)b(i)} + \rho^{b(i)b(j)},
\end{aligned} \tag{3.4}$$

where each node in the graph belongs to one of K blocks, $B_1, \ldots, B_K$, and $b(i)$ denotes the (known) block assignment of vertex i.

There are various special cases of stochastic blockmodels. For example, we can choose $\delta^{rs} = \delta + \alpha^r + \beta^s$ and $\rho^{rs} = \rho$, as in ([FMW85], Equation (2.10)). Then the model is the following special case:

$$\begin{aligned}
\log p_{ij00} &= \lambda_{ij} \\
\log p_{ij10} &= \lambda_{ij} + \delta + \alpha^{b(i)} + \beta^{b(j)} \\
\log p_{ij01} &= \lambda_{ij} + \delta + \alpha^{b(j)} + \beta^{b(i)} \\
\log p_{ij11} &= \lambda_{ij} + 2\delta + \alpha^{b(i)} + \alpha^{b(j)} + \beta^{b(j)} + \beta^{b(i)} + \rho.
\end{aligned} \tag{3.5}$$

In this setting, the sufficient statistics counted by the map t are the number of configurations for each dyad, the total number of edges, block in-degrees, block out-degrees, and the total number of reciprocated edges in the network. Here, the in-degree of block B_j (the number of edges that enter the block) is computed by adding in-degrees of all the nodes in the block, $d^{in}_{B_j} = \sum_{i \in B_j} d^{in}_i$. The out-degree is defined similarly.

Let us consider the space of directed graphs on $n = 3$ vertices $V = \{1, 2, 3\}$ with block structure $B_1 = \{1, 2\}$, $B_2 = \{3\}$, the design matrix A defining the linear map t would be as follows:

$$A = \begin{bmatrix} 1&1&1&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&1&1&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&1&1&1 \\ 0&1&1&2&0&1&1&2&0&1&1&2 \\ 0&1&1&2&0&1&0&1&0&1&0&1 \\ 0&0&0&0&0&0&1&1&0&0&1&1 \\ 0&1&1&2&0&0&1&1&0&0&1&1 \\ 0&0&0&0&0&1&0&1&0&1&0&1 \\ 0&0&0&1&0&0&0&1&0&0&0&1 \end{bmatrix}.$$

Let G be represented as a vector of length 12, where the first four entries correspond to the four possible dyadic configurations between vertices 1 and 2, the second four correspond to the four possible dyadic configurations between vertices 1 and 3, and the third four correspond to the four possible dyadic configurations between vertices 2 and 3. Then the first three rows of A count the number of configurations for each dyad (for simple graphs this count should always be one), the fourth row of A counts the total number of edges in G, the fifth and sixth rows count the block in-degrees, the seventh and eighth rows count the block out-degrees, and last row counts the total number of reciprocated edges in the network.

Example 3.3 (p_1 Models) The p_1-model for directed graphs was introduced by Holland and Leinhardt [HL81] and extended by Fienberg and Wasserman [FW81b]. It is a model that includes two nodal effects, one for popularity and another for expansiveness, and a reciprocation effect. Following Example 3.2, we denote p_{ijkl} the probability of the dyad (i, j) to be in state $(k, l) \in \{0, 1\}^2$. The dyadic probabilities for the p_1-model are specified as follows:

$$\begin{aligned} \log p_{ij00} &= \lambda_{ij}, \\ \log p_{ij10} &= \lambda_{ij} + \alpha_i + \beta_j + \delta, \\ \log p_{ij01} &= \lambda_{ij} + \alpha_j + \beta_i + \delta, \\ \log p_{ij11} &= \lambda_{ij} + \alpha_i + \alpha_j + \beta_j + \beta_j + 2\delta + \rho_{ij}. \end{aligned} \tag{3.6}$$

The parameters α_i and β_i record the rates at which the node i sends and receives links, while ρ_{ij} controls reciprocation. Note that the model specification includes additional parameters. Namely, there is δ, a density parameter and $\binom{n}{2}$ dyadic effects, λ_{ij}, which are normalizing constants as described in Example 3.2.

The p_1 model has three main variants that capture different reciprocation effects: zero reciprocation, constant reciprocation, and dyad-specific reciprocation, also referred to as differential reciprocity. For example, in the constant reciprocation case, $\rho_{ij} = \rho$ for all i, j. The sufficient statistics for the p_1-model with constant reciprocation consists of the number of edges, the in-degree sequence, the out-degree sequence, and the number of reciprocated edges.

The design matrix A for several small examples can be found in [PRF10].

While Fienberg's work allows for a transfer of technology from the contingency table literature to networks, the interpretability of models and model equivalence was not always immediate and required additional insight. As noted in [Hab81] and reiterated by Fienberg and co-authors in [FPR10], even simple ERGMs, such as the p_1 model, pose fundamental challenges to the practitioner even within the contingency table setting, especially when testing model goodness of fit. For example, as pointed out by Fienberg and co-authors in [PRF10], many network models such as the p_1 model are theoretically problematic, since, in these models, the number of parameters depends on the number of vertices. This means that as the population size grows, the model complexity also increases, unlike traditional statistical models, where the complexity is often fixed and independent of the sample size. Another challenge to using existing traditional methods from categorical data analysis in goodness-of-fit testing and model selection is that the data are naturally sparse, making standard asymptotic methods unreliable. Under such conditions, exact conditional tests are preferred for model selection and goodness-of-fit testing. However, as mentioned in the previous section, exact conditional tests pose their own difficult problems for networks, mainly since the exact distribution is over a space that is combinatorially large, and in most cases, innumerable. Finally, the contingency tables described by Fienberg and Wasserman are highly redundant and are subject not only to symmetric constraints but also product multinomial constraints, e.g. since each dyad can only be in one of the four possible configurations $Y_{ij00} + Y_{ij10} + Y_{ij01} + Y_{ij00} = 1$ for all $i \neq j$.

Fienberg was able to provide a work-around to the difficulties posed by exact conditional tests by using Markov bases and algebraic statistics. In 1998, Sturmfels and Diaconis published Theorem 2.2 [DS98]. Afterwards, the idea of using toric ideals for goodness-of-fit testing for various log-linear models gained traction, and about 10 years later, Fienberg, Petrović, and Rinaldo applied Theorem 2.2 to three of the main variants of the p_1 model in [PRF10], essentially introducing algebraic statistics to the field of network analysis. In particular, they describe Markov moves for each variant and its corresponding simplified model (the model obtained after forgetting the normalizing parameters). The work not only provided a breakthrough in goodness-of-fit testing for log-linear ERGMs but also had an impact in combinatorial commutative algebra. The toric ideals corresponding the p_1 model are connected to toric ideals of graphs, defined in [SVV94] (see also [Vil95] and [OH00]) and more generally, toric ideals of hypergraphs. Indeed, the results of [PRF10] provided an applied motivation for the systematic study of toric ideals of hypergraphs in the field of combinatorial commutative algebra (see e.g. [GP13, HT08, PS14, PTV19]).

Before [PRF10], Markov bases were always used in the setting where the only constraints on the contingency tables were that every entry needed to be non-negative. However, in the network setting, particularly in the case of a single sociometric relation, cells of the contingency tables are either 0 or 1 and there is only a single observation for each dyad. This was the first time in the Markov bases literature that sampling constraints of this form were directly incorporated in the study of Markov bases (note that related work [HT10], and relevant for the

problem here, on connecting tables with 0/1 entries appeared in the same volume). Fienberg and co-authors were able to effectively handle the network constraints by computing a minimal generating set of this ideal first and then by removing basis elements that violate the condition of one observation per dyad, which results in a product multinomial sampling scheme. Fienberg's idea of adding the normalizing parameters λ_{ij}s to the models directly enforced the $0/1$ constraint in sampling. In particular, if a move produced by a Markov basis computation is applicable to the observed network, in that it does not attempt to remove edges that are not present, then it will follow the sampling constraint in that it will not add an edge where there is one already. Examples of applicable and inapplicable moves for the p_1 model and the Sampson data depicted in Fig. 3.1 are shown in Figs. 3.2, 3.3, and 3.4.

It should be noted that Fienberg's idea to prune non-applicable moves was novel and paved the way for practical implementation of a goodness-of-fit test for log-linear ERGMs [GPS16]. Indeed, in [DFR+08], Fienberg and co-authors observed that Markov bases are *data independent*, meaning that they describe all the moves required to guarantee connectedness of any fiber; in other words, Markov bases do not depend on the observed network, only the model. This observation can help transform otherwise unwieldly sets of Markov moves into smaller and easier to manipulate sets of moves. For example, without pruning, the naive computation of a Markov basis for the p_1 model with constant reciprocation with 4 nodes has 80,610 moves, while the pruned Markov basis consisting of only elements applicable

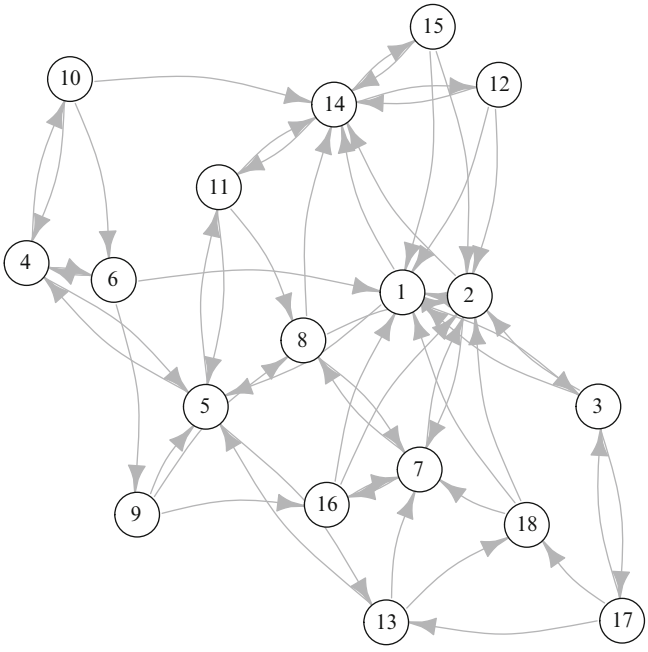


Fig. 3.1 The directed graph representation of Sampson's monastery dataset [Sam69]

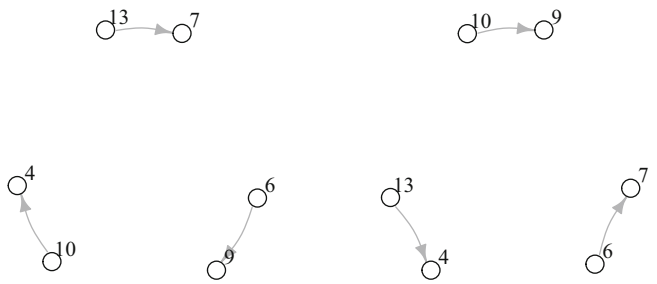


Fig. 3.2 A move from the Markov basis for the p_1 model with zero reciprocation. **Left:** Edges to remove. **Right:** Edges to add. This move can be applied to the network in Fig. 3.1 as it preserves node in-degrees and out-degrees. Note that edge $4 \leftarrow 10$ is reciprocated in the data, so after the move is applied, the total number of reciprocated edges is reduced by 1

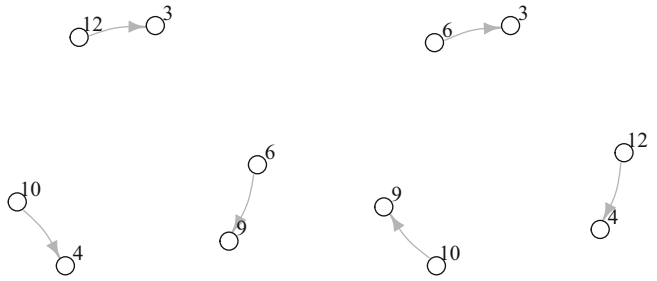


Fig. 3.3 A move from the Markov basis for the p_1 model with zero reciprocation. **Left:** Edges to remove. **Right:** Edges to add. However, this move cannot be applied to the network in Fig. 3.1 as the dyad $(3, 12)$ is observed in the state $(0, 0)$ rather than $(1, 0)$; that is, the edge $12 \rightarrow 3$ is not present, so it cannot be removed

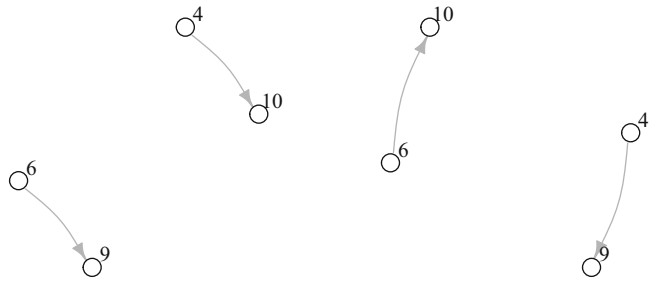


Fig. 3.4 A move from the Markov basis for the p_1 model with constant reciprocation. **Left:** Edges to remove. **Right:** Edges to add. This move can be applied to the network in Fig. 3.1. It preserves the number of reciprocated edges: the dyad $(4, 10)$ changes from reciprocated to directed edge, but the dyad $(6, 10)$ changes from directed to reciprocated

to simple networks and decomposed into essential building blocks, computed in [PRF10], has about 10 moves.

This idea was a starting point of departure from the algebraic status-quo approach, which is traditionally blind to data and as such leads to slow mixing times of the resulting Markov chains. After Fienberg's work in [PRF10], the main computational challenge remained open to make the theory useful for network data in practice. To this end, working within the *data dependent paradigm*, [GPS16] developed an algorithm to approximate the exact conditional p-value for log-linear ERGMs and implemented the algorithm for the p_1 model. The algorithm approximates the exact conditional p-value by using applicable Markov moves generated on an as-needed basis to move around the fiber. At each network in the chain, a goodness-of-fit statistic is computed and compared to the observed network. This adapted Metropolis-Hastings algorithm is described in detail in [GPS16].

For exposition and illustration of theoretical ideas, Fienberg saw great value in small data; for example, Sampson's monastery dataset [Sam69] (see Fig. 3.1) was the running example in [ABFX09] and also was an example dataset in Fienberg's survey of statistical network models with Goldenberg et al. [GZFA10]. Thus, the paper [GPS16] revisited the Sampson's monastery dataset and tested the fit of the p_1 model. The Sampson's monastery dataset, in Fienberg's words, was one of the reasons behind the construction of the Holland-Leinhardt p_1 model in the first place. However, this is not to say that Fienberg was not concerned with challenging big data problems, and the ideas described here do scale, e.g. [KP16] tests model fit for the β and p_1 models on co-authorship and citation networks of statisticians [JJ16] of about 3000 authors and 3000 papers. Finally, Fienberg was also an avid supporter of applications of statistics; it was he who suggested to the third author to study the Japanese corporate data set from *The New York Times* back in 2014 from the point of view of the p_1 model. As [Pet19] illustrates, the goodness-of-fit test confirms the Japanese Prime Minister's intuition.

3.4 Beyond Simple Graphs

The rapid increase of data-collecting mechanisms in recent decades has resulted in complex forms of network data, including multivariate and multi-agent networks. Still, in the growing field of network science, such data are still often represented in the form of a simple graph, mainly because simple random graph models are assumed to be easier to estimate and fit. However, such simplifications are not necessary with Fienberg's view of networks as contingency tables. This is because neither multiple observations on a single dyad, which increase cell counts in the table, nor multiway interactions, which increase table dimensions, present an additional layer of difficulty for estimation or testing model fit. On the contrary, the sampling algorithms based on Markov bases become easier, because the sampling constraint is relaxed.

One example of this simplification is when experiment data consisting of multiple observations is summarized as a simple graph by way of thresholding—preserving an edge between two nodes only if it was observed at least a fixed number of times. This happens very often in neuroscience and chemical reaction experiments. It is also often applied to social interactions data such as the co-authorship network in Fig. 3.5 below. In the co-authorship network, an edge (i, j) is present in the co-authorship graph if at least 4 joint papers were written by authors i and j. Why 4? This thresholding number of choices seems arbitrary at best (changing it may drastically change the structure of the graph), is done out of convenience, and in many applications results in significant information loss.

In [FMW80] and [FMW85], Fienberg, Meyer, and Wasserman set up the log-linear framework for multivariate directed graphs. We can think of a multivariate graph as a multi-layered network. For example, in the technical paper [FMW80], Fienberg, Meyer, and Wasserman consider a community of individuals and networks formed by three relations, information, money, and support; these relations are referred to as sociometric generators. In [FMW85], the authors develop extensions of [FMW80] to allow for covariates. Motivated by this, [RPF13] (see also [RPF10] for further details) study the *generalized* β-model for random graphs. They consider the log-linear model for undirected graphs whose sufficient statistics are node degrees, but they allow for the possibility that each dyad in the network be sampled a different number of times. Applying the geometric and combinatorial properties of log-linear models under product multinomial sampling schemes from [FR12], they derive necessary and sufficient conditions for MLE existence and discuss its asymptotics.

The second example of data simplification is also well illustrated using co-authorship data: it is common for multiway interactions to be collapsed to their

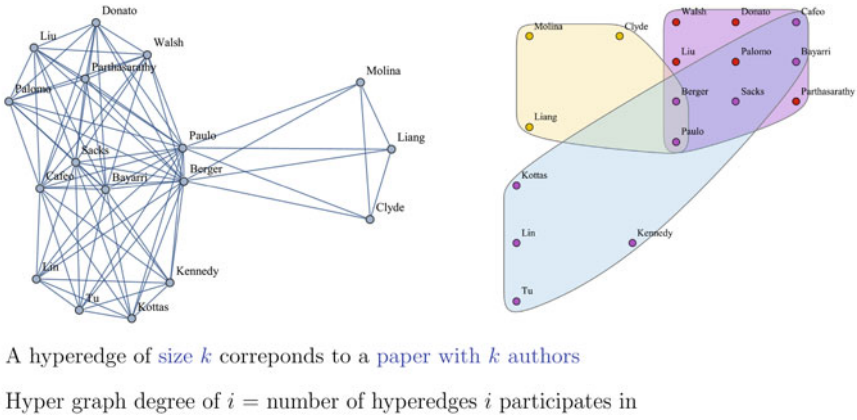


Fig. 3.5 The graph and the hypergraph representing *the same* co-authorship data. In the graph on the left, it is not clear at all that the data corresponds to exactly 3 published papers, for example, which is clear in the hypergraph on the right. Graphs in the figure adapted from [KP16]

induced pairwise interactions. However, most of the time, capturing the multiway interaction is more realistic and informative. Figure 3.5 shows how the information from data that naturally comes in form of a hypergraph is obscured when represented by the underlying graph. Indeed, once the first co-authorship network data for statisticians was collected and released in [JJ16], the last two authors set out to explore the effects of these data summaries. In [KP16], it is shown what information is lost by reducing the data to a simple graph by presenting multi-observation table data summaries, core-decomposition summaries, and hypergraph data summaries, all of which suggest possibly different conclusions than those from the derived simple graphs. For example, the authors considered the inner-most clique, that is, the largest completely connected subgraph, of the co-authorship graph where there is an edge between two authors if they coauthored at least 4 joint papers. While these authors have many neighbors, i.e. their nodes have a high degree, we argue that degree-based modeling on the simple graph does not capture everything behind the data. Specifically, Fig. 3.6 shows that the secret behind these cliques is a single many-author paper in both cases.

With the issues illustrated in [KP16] in mind, Fienberg and co-authors introduce the β model for random hypergraphs in [SSR+14], which builds upon and generalizes the well-studied β model for random graphs. Directly motivated by Fienberg's earlier foundational work, the authors provide two algorithms for fitting the model parameters, an iterative proportional scaling algorithm, and a fixed point algorithm. Furthermore, Fienberg and co-authors prove that both algorithms converge if the maximum likelihood estimator (MLE) exists, and they provide algorithmic and geometric ways of dealing the issue of MLE existence—one of Fienberg's favorite problems.

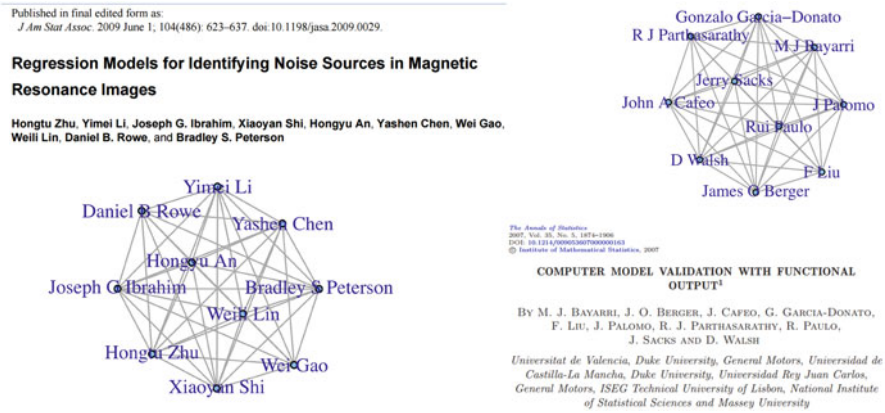


Fig. 3.6 The inner-most clique of each of the two co-authorship graphs studied in [JJ16]: each corresponds to a many-author paper. Graphs in the figure adapted from [KP16]

3.5 Closing Remarks

Fienberg always used to say how problems never go away, one just sees them under a new light. In this survey of Fienberg's work connecting categorical data analysis and algebraic statistics to network science, we hope we illustrated, in essence, this sentiment of continual discovery and rediscovery.

Acknowledgments Author Elizabeth Gross is supported by NSF DMS-1620109, and Sonja Petrović was partially supported by NSF DMS-1522662. Karwa was partially supported by NSF TRIPODS+X grant number 1947919.

References

ABFX09 Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. In *Advances in Neural Information Processing Systems*, pages 33–40, 2009.

Agr92 Alan Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153, 1992.

BFH74 Yvonne M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. Springer, 1974.

DFR+08 Adrian Dobra, Stephen E. Fienberg, Alessandro Rinaldo, Aleksandra Slavković, and Yi Zhou. Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation and disclosure limitation. In *In IMA Volumes in Mathematics and its Applications: Emerging Applications of Algebraic Geometry*, pages 63–88. Springer Science+Business Media, Inc, 2008.

DS98 Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, 26(1):363–397, 1998.

FG70 Stephen E. Fienberg and John P Gilbert. The geometry of a two by two contingency table. *Journal of the American Statistical Association*, 65(330):694–701, 1970.

Fie68 Stephen E. Fienberg. The geometry of an $r \times c$ contingency table. *The Annals of Mathematical Statistics*, 39(4):1186–1190, 1968.

Fie70 Stephen E. Fienberg. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 41(3):907–917, 1970.

FMW80 Stephen E Feinberg, Michael M Meyer, and Stanley Wasserman. Analyzing data from multivariate directed graphs: An application to social networks. Technical report, University of Minnesota, 1980.

FMW85 Stephen E. Fienberg, Michael M. Meyer, and Stanley S. Wasserman. Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80(389):51–67, 1985.

FPR10 Stephen E. Fienberg, Sonja Petrović, and Alessandro Rinaldo. *Algebraic statistics for p_1 random graph models: Markov bases and their uses*, volume Papers in Honor of Paul W. Holland, ETS. Springer, 2010.

FR12 Stephen E. Fienberg and Alessandro Rinaldo. Maximum likelihood estimation in log-linear models. *Annals of Statistics*, 40(2):996–1023, 2012.

FW81a Stephen E. Fienberg and Stanley S. Wasserman. Categorical data analysis of single sociometric relations. *Sociological methodology*, 12:156–192, 1981.

FW81b Stephen E. Fienberg and Stanley S. Wasserman. Discussion of Holland, P. W. and Leinhardt, S. "An exponential family of probability distributions for directed graphs". *Journal of the American Statistical Association*, 76:54–57, 1981.

GP13 Elizabeth Gross and Sonja Petrović. Combinatorial degree bound for toric ideals of hypergraphs. *International Journal of Algebra and Computation*, 23(6):1503–1520, 2013.

GPS16 Elizabeth Gross, Sonja Petrović, and Despina Stasi. Goodness-of-fit for log-linear network models: Dynamic Markov bases using hypergraphs. *Annals of the Institute of Statistical Mathematics*, 2016. https://doi.org/10.1007/s10463-016-0560-2.

GZFA10 Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

Hab81 Shelby J. Haberman. An exponential family of probability distributions for directed graphs: Comment. *Journal of the American Statistical Association*, 76(373):60–61, March 1981.

HL81 Paul W. Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, 76(373):33–65, 1981.

HT08 Huy Tài Hà and Adam Van Tuyl. Monomial ideals, edge ideals of hypergraphs, and their graded Betti numbers. *Journal of Algebraic Combinatorics*, 27(215–245), 2008.

HT10 Hisayuki Hara and Akimichi Takemura. Connecting tables with zero-one entries by a subset of a Markov basis. In Marlos Viana and Henry Wynn, editors, *Algebraic Methods in Statistics and Probability II*, volume 516 of *Contemporary Mathematics*. American Mathematical Society, 2010.

JJ16 Pengsheng Ji and Jiashun Jin. Coauthorship and citation networks for statisticians. *Annals of Applied Statistics*, 10(4):1779–1812, 2016.

KP16 Vishesh Karwa and Sonja Petrović. Coauthorship and citation networks for statisticians: Comment. invited comment on the paper by Jin and Ji. *Annals of Applied Statistics*, 10(4):1827–1834, 2016.

OH00 Hidefumi Ohsugi and Takayuki Hibi. Compressed polytopes, initial ideals and complete multipartite graphs. *Illinois J of Mathematics*, 44(2):391–406, 2000.

Pet19 Sonja Petrović. What is ... a Markov basis? *Notices of the American Mathematical Society*, 66(7):1088—1092, 2019.

PRF10 Sonja Petrović, Alessandro Rinaldo, and Stephen E. Fienberg. Algebraic statistics for a directed random graph model with reciprocation. In M. Viana and H. Wynn, editors, *Algebraic Methods in Statistics and Probability II*, volume 516 of *Contemporary Mathematics*, pages 261–283. American Mathematical Society, Providence RI, 2010.

PRW01 Giovanni Pistone, Eva Riccomagno, and Henry Wynn. Computational commutative algebra in discrete statistics. *Contemporary Mathematics*, 287:267–282, 2001.

PS14 Sonja Petrović and Despina Stasi. Toric algebra of hypergraphs. *Journal of Algebraic Combinatorics*, 39(1):187–208, 2014.

PSY19 Sonja Petrović, Aleksandra Slavkovic, and Ruriko Yoshida. " Old wine in new bottles," and some more new wine – Stephen Fienberg's influence on algebraic statistics. *Journal of Algebraic Statistics*, 10(1, Issue in honor of Stephen E. Fienberg):i–vi, 2019.

PTV19 Sonja Petrović, Apostolos Thoma, and Marius Vladoiu. Hypergraph encodings of arbitrary toric ideals. *Journal of Combinatorial Theory, Series A*, 166(11–41), 2019.

RPF10 Alessandro Rinaldo, Sonja Petrović, and Stephen E. Fienberg. On the existence of the MLE for a directed random graph network model with reciprocation. Technical report, 2010. http://arxiv.org/abs/1010.0745.

RPF13 Alessandro Rinaldo, Sonja Petrović, and Stephen E. Fienberg. Maximum likelihood estimation in the Beta model. *Annals of Statistics*, 41(3):1085–1110, 2013.

Sam69 S. Sampson. *Crisis in a cloister*. Unpublished doctoral dissertation, Cornell University, 1969.

SSR+14 Despina Stasi, Kayvan Sadeghi, Alessandro Rinaldo, Sonja Petrović, and Stephen E. Fienberg. Beta models for random hypergraphs with a given degree sequence. In *Proceedings of 21st International Conference on Computational Statistics*, 2014.

Stu96 Bernd Sturmfels. *Grobner bases and convex polytopes*, volume 8. American Mathematical Soc., 1996.

SVV94 Aaron Simis, Wolmer V. Vasconselos, and Rafael H. Villarreal. On the ideal theory of graphs. *Journal of Algebra*, 167(20):3890416, 1994.

Vie15 Statistics Views. "In most places, not only are statisticians not in control of Big Data efforts and data science, but sometimes they are totally excluded or at best, marginalised." An interview with Stephen Fienberg. http://www.statisticsviews.com/details/feature/8555981/In-most-places-not-only-are-statisticians-not-in-control-of-Big-Data-efforts-and.html, 10 Nov 2015.

Vil95 Rafael H. Villarreal. Rees algebras of edge ideals. *Communications in Algebra*, 23(9):3513–3524, 1995.

Part II
Introduction: Bayesian Theory and Applications

Alicia L. Carriquiry

Steve was an extraordinary statistician. What made him really stand out even among the upper echelon of his peers were his command of both statistical theory and methods and his profound commitment to expanding the principled use of statistics in other disciplines. His interest in the application of statistics in other fields predated – by decades – what the profession has come to accept: that statistics is the common language of science and as such plays an important role in the scientific and the policy realms.

During his long career, Steve worked on a broad set of problems, which sometimes were of mainly scientific interest, but other times arose from practical questions in public policy. Indeed, he often found a role for statistics and statisticians in areas where some of us found none. I clearly remember, when in the early 2000s, Steve and his then student Adrian Dobra set out to estimate the size of the World Wide Web! The common thread in much of Steve's work begins with a good understanding of the subject matter area, and ends with relevant findings that address the questions posed by the investigator. Along the way, Steve and collaborators often had to develop "fit-for-purpose" statistical tools when none existed. Steve had an almost unlimited capacity to learn about new areas and quickly grasp where statistics would be helpful; therefore, he really excelled when playing in a wide variety of sandboxes. Steve was a committed Bayesian, but his main goal was always to answer the question of interest to the best of his ability. If that involved using frequentist methods when those were more appropriate (or convenient), he was happy to do so.

The four manuscripts in this section of the book would have been of great interest to Steve. Like Steve's work, they represent a broad array of topics and approaches, but what they have in common is the levelheaded, forward-looking use of statistics to answer a challenging question. Steve was good friends with most of the contributors in this section. His relationship with Tom Louis, Julia Mortera, and Philip Dawid went back 40 years, and while the friendship with Ed George, Xiao-Li Meng, and Veronika Rockova was "only" about 20–25 years old, he had enormous affection and respect for all of them equally. He would have been delighted to know that his friends had contributed great papers to the volume in his memory.

Baines et al. work on an astrophysics problem, where the data consist of noisy observations arising from a complex process described by a deterministic system. In their specific application, the question is how to model photometric measurements of stars as a function of properties including mass, age, and metallicity of the stars. The state of the art requires using a series of look-up tables in the calculations, which decrease efficiency and contributes to underestimation of uncertainties. The authors present a flexible Bayesian hierarchical model to analyze stellar populations that avoids the tables, but that requires the development of a computational framework that is up to the task.

Louis' wonderful contribution makes us think once again about the advantages of "going the ways of Bayes." Why isn't everyone a Bayesian? This was the question posed by Efron in 1986, to which Louis provides a clear, nuanced, multipart response that is fun to read and touches upon many important topics including p-values, multiple testing, and model diagnostics. The final sections of the paper describe a major project in which both Louis and Fienberg participated and that has major policy implications. The Center for Medicare and Medicaid Services (CMS) provides risk-adjusted, hospital-specific measures to compare the performance of a hospital to a national standard. This is clearly a difficult endeavor and a prime target for criticisms by stakeholders ranging from patient advocates to the corporations with an interest in hospital care. To shore up its methodology and avoid some of the criticism, CMS partnered with the Committee of Presidents of Statistical Societies (COPSS) to address the statistical issues and make recommendations. Louis includes a step-by-step description of the process followed by the committee that brings the committee's work to life and provides a wonderful example of the type of principled, well-reasoned, science-based approach that characterized all of Steve Fienberg's contributions.

Mortera and Dawid revisit a problem in which Steve worked many years ago and that involves the evaluation of the uncertainty associated to forecasts. They consider both the case of a single forecaster and that of opinion pools, where the forecasts of multiple experts are combined in some way, and that typically leads to better prediction. In the last part of the manuscript, Mortera and Dawid discuss prediction markets, where individuals trade predictions on future events with unknown outcome, much like other financial instruments are traded in their respective markets. They consider different scenarios regarding the order in which the various traders reveal their probability of a future event, and where the information used by each of the traders may or may not be available to everyone else.

Finally, Rockova and George revisit the problem of model selection using a novel approach that involves determinantal point processes to construct a discrete prior distribution for the regression coefficients in the model. The prior they propose is derived from the spike and slab prior, obtained by rescaling the spike and slab prior by a factor proportional to the determinant of a kernel matrix that decreases the probability of subsets of similar variables. The resulting prior favors small models, but in addition penalizes sets of explanatory variables that exhibit high collinearity. As is the case with the other contributions in this section, the Rockova and George paper is novel, is readable, and introduces methodology that will have wide application.

Chapter 4
Colorful Stars and Black Boxes: Bayesian Analysis of Stellar Populations

Paul D. Baines, Xiao-Li Meng, Andreas Zezas, and Vinay Kashyap

4.1 Stellar Archeology

This article details a statistical analysis of a complex problem in astronomy and astrophysics, with a broader aim to suggest strategies and methodologies for similar "black-box" problems in physical sciences and beyond. For reasons that will become clear, the specific problem we address is sometimes known as the "stellar archeology" problem. The archeological analogy provides a nice overview of the scientific objective: we are interested in estimating the age of objects (stars) from measurements of their attributes (brightness).

P. D. Baines
Department of Statistics, University of California Davis, Davis, CA, USA
e-mail: pdbainesredsox@ucdavis.edu

X.-L. Meng (✉)
Harvard University, Department of Statistics, Cambridge, MA, USA
e-mail: meng@stat.harvard.edu

A. Zezas
University of Crete, Department of Physics, Institute for Astrophysics, Foundation for Research and Technology-Hellas, Heraklion, Greece

High Energy Astrophysics Division, Center for Astrophysics, Harvard & Smithsonian, Cambridge, MA, USA
e-mail: azezas@physics.uoc.gr

V. Kashyap
High Energy Astrophysics Division, Center for Astrophysics, Harvard & Smithsonian, Cambridge, MA, USA
e-mail: vkashyap@cfa.harvard.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_4

Astronomy has a long history of using and developing statistical methodology to analyze experimentally collected data (e.g., see [8]). Despite the inability to directly manipulate the processes being studied, astronomers and astrophysicists have amassed a large body of knowledge by both indirect observation of the underlying processes and the construction of physics-based models. As the understanding of underlying physical systems develops, observed data typically can be characterized as noisy observations of a complex physical process involving the parameters of inferential and scientific importance. The link between parameters of interest and observed data is provided by problem-specific knowledge, often in the form of a system of partial differential equations (PDEs). This characterization is common in many problems in astronomy, as well as other scientific fields such as the environmental sciences. When the driving systems cannot be solved analytically, or are particularly computationally expensive, the relevant community often relies on lookup tables, describing the expected observation (i.e., mean) for a variety of input parameters. Given the huge amount of expertise devoted to developing these models, the analysis of observed data often lags behind. Statistics can play an important role in such settings, although the resulting computation can be challenging. There has been increased interest in this type of problem, where one or more components in the model are a "black-box," lookup table, or computer-model output [3, 4, 17, 20].

In this article, we present an example of such a problem: a hierarchical Bayesian analysis of photometric data. The objective is to infer stellar properties such as the mass, age, and metallicity of individual and collections of stars. The mapping between the scientifically interesting properties (mass, age, and metallicity) and the observed data (photometric measurements) is governed by a series of isochrone tables: lookup tables derived under an assumed physics model. Isochrone tables are traditionally named after the location of the research groups that computed them: commonly used versions include the Geneva [13], Padova [14], and Dartmouth [6] isochrones. The highly structured mapping poses challenges for traditional computational methods, as discussed in Sect. 4.2. In Sect. 4.3 we present a generalizable and robust algorithm for posterior sampling that does not rely on any specific properties of the isochrone mapping. By avoiding the use of isochrone-specific fixes, we seek an algorithm that can successfully adapt to new lookup tables and could be applied to a wider class of problems. Combining the ideas of different augmentation schemes in [22] with an energy-based partition proposal distribution in the spirit of the Equi-Energy sampler [11], our "Equi-Expectation" MCMC scheme is both efficient and scalable to large datasets. The performance of competing sampling schemes is detailed in Sect. 4.4, together with an application to the 47 Tucanae dataset. In practice there are often uncertainties in the choice of deterministic physical model, and hence we investigate the issue of selecting between competing sets of isochrones in Sect. 4.5. The rest of Sect. 4.5 discusses some future directions and concludes.

4.2 Color-Magnitude Diagrams and Isochrone Tables

Photometric measurements are obtained by a detector, pointed at a particular region of the sky. Sources such as stars emit photons, which, together with additional background photons are counted whenever they pass through the detector. In crude terms, by counting the number and energy of photon arrivals at a particular detector location in a particular time interval, the photon counts can be calibrated to obtain the spectrum of a given source. The spectrum of a source represents the intensity across a continuous range of wavelengths and, as such, these observations can be expensive to obtain. An alternative is to use optical filters that allow only photons within a specified wavelength band/range to pass through. The measurements can then be thought of as estimating the integral of the spectrum over a small wavelength range. Depending on the number of bands, this approach yields a small number of measurements representing the brightness of the source that are both cheaper to obtain and simpler to analyze than their spectral counterparts. The brightness of a source in a photometric band such as B (blue) is also known as its B-band magnitude. Colors can be obtained as differences between magnitudes: for example, the color $B - V$ represents the difference in B- and V-band (visual) magnitudes. In light of this property we are able to freely switch between colors and magnitudes, and the analysis of Sect. 4.3 can be conducted across different combinations of colors and magnitudes.

To relate observed photometric data to the relevant physical quantities such as the age, mass, and metallicity of the stars, we use a theoretical collection of *isochrones*. The term "isochrone" is typically used to refer to the curve defined by tracing out the expected color and magnitude for stars of a fixed age and metallicity, for different initial masses. More generally, an isochrone can be viewed as a function that, given the physical properties of the star (mass, age, metallicity), returns the brightness of the star in a variety of different photometric bands. The metallicity of a star describes the relative abundance of elements such as oxygen and iron with respect to hydrogen.

The top panel of Fig. 4.1 displays all of the combinations of initial mass and age that appear in the (Padova) isochrone tables. The bottom panel of Fig. 4.1 displays the expected V-band magnitude and $B - V$ color of stars with a metallicity $Z = 0.004$, at each of the tabulated points of the isochrones. These plots correspond to the input and output spaces, with the isochrone mapping (i.e., the "black box") between them. The color of each point in the plot indicates the age of the star, with younger stars typically being hotter and brighter than their older counterparts. The plot of color against magnitude is known as the Color-Magnitude Diagram (CMD), and forms the basis of the use of photometric data to infer stellar properties. Here, however, we use CMD to refer to the more general setting including higher-dimensional photometry and arbitrary (non-degenerate) color-magnitude combinations.

The initial mass of a star is a crucial factor in determining the evolution of the star. As stars age they burn off their component elements in order from the lightest

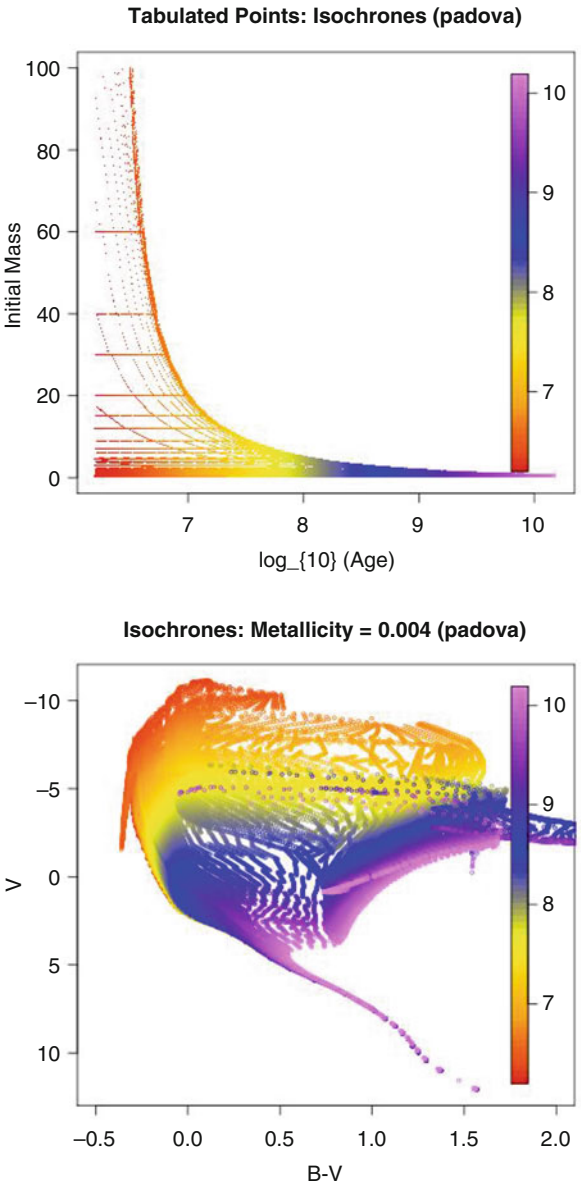


Fig. 4.1 Isochrone plot for (Top) the input/parameter space: initial mass and age, and, (Bottom) the output/observation-space: V-magnitude and $B - V$ photometry, for stars of metallicity $Z = 0.004$ from the Padova isochrones. The color of each point represents the age in $\log_{10}$-years (i.e., from $10^{6.0}$ to $10^{10.2}$ years), with the color-scale given on the right-hand side of each plot

to the heaviest, beginning with hydrogen and helium fusion. Since the chemical composition of the star is one of the determining factors in its photometry, and we have a physics model for the stellar evolution process, we can attempt to infer the age, initial mass, and metallicity of the star from observed data.

The bottom-left portion of the CMD (Fig. 4.1, bottom) is known as the *main sequence*. This is where stars spend most of their lives, usually before evolving into either a brown or white dwarf. On the main sequence, there are many different combinations of mass, age, and metallicity that produce the same expected photometry, leading to a degeneracy in measurements. Therefore, taken in isolation, the mass and age of a given star may or may not be identifiable. The applications we consider are those where we are interested in estimating the properties of a "cluster" of stars. Typically, by a cluster of stars we mean a collection of stars located in a similar physical location, and at a similar distance from the detector. Despite the individual-level potential non-identifiability, by combining observations, we can proceed to draw inference about both cluster-level and individual stellar properties. In addition to these challenges with identifiability, small changes in mass and age can potentially produce large changes in expected photometry, depending on the region of the CMD in which the star falls. These problems all add to the complexity of both the physical modeling and statistical analysis, but they are by no means unique to stellar archeology. We therefore believe that the strategies and methodologies in this article have general implications.

4.3 Hierarchical Modeling and Computation

Model Specifications

A photometric observation of source i, typically a star, is a vector of observed values in a combination of colors and magnitude bands, denoted by $Y_i \in \mathbb{R}^p$, where p is the number of bands. The (Gaussian) measurement errors from the detector are typically well understood, in the sense that variances are traditionally taken to be known for each band. Without loss of generality, we can assume unit variance (i.e., working with standardized Y_i). Here we allow for the measurement errors to be correlated across bands or colors: the correlation structure is assumed to be constant among all stars and is modeled with a weakly informative prior. Given the intrinsic properties of the stars, the measurement errors are assumed independent across different stars. The lower-level data generating process is thus given by

$$Y_i \big| M_i, A_i, Z \overset{ind}{\sim} N\left(f(M_i, A_i, Z), \mathbf{R}\right) \qquad i = 1, \ldots, n, \tag{4.1}$$

where M_i and A_i are the (initial) mass and age of the star, Z is the metallicity of the cluster and $f(M_i, A_i, Z)$ is a vector of the expected photometry of the star, found from the isochrone tables, and is standardized the same way as Y_i is. For

all applications here we consider Z to be known from external knowledge, as is standard in the astrophysics literature, although extending the model to include unknown metallicity is conceptually straightforward. The correlation matrix $\mathbf{R}$ is assumed to be the same across all observations, following a common strategy for balancing between model adequacy and model complexity (e.g., [2, 12]).

The literature on CMDs has assumed almost exclusively that stars in the same subpopulation (e.g., cluster) have the same age, and sought the best-fit isochrone based on this single age (e.g., [21]). This comes despite knowledge in many contexts that the spread in stellar ages is sizable. Our approach remedies this problem but retains model simplicity by placing a common structure on the ensemble of star ages. Allowing flexibility of individual parameters yet utilizing the common structure motivates the following model. We assume the "population" of $\log_{10}$ ages for a given cluster to be Gaussian (equivalently, age is $\log_{10}$-normal, not standard log-normal):

$$A_i|\mu_A, \sigma_A^2 \overset{iid}{\sim} N(\mu_A, \sigma_A^2), \tag{4.2}$$

where 10^{A_i} is the age of the star in years. The traditional single-age approach amounts to imposing $\sigma_A^2 = 0$ and finding a "best" choice of μ_A, the parameter of primary inferential importance. Here μ_A characterizes the theoretical mean age (on the $\log_{10}$-scale) of the collection of stars, while σ_A^2 specifies the intra-cluster variability of the individual ages. By modeling the distribution of individual star ages, we can potentially detect outlying stars or multi-cluster populations corresponding to multi-burst star formation processes. Although such discoveries are feasible when we move beyond the single-age paradigm, estimation in multi-clusters contexts should be redone with explicit multi-cluster models, as we discuss in Sect. 4.3.

To complete the model specification, we use the conjugate hyperprior:

$$\mu_A|\sigma_A^2 \sim N\left(\mu_0, \sigma_A^2/\kappa_0\right), \qquad \sigma_A^2 \sim \text{Inverse}-\chi^2\left(\nu_0, \sigma_0^2\right). \tag{4.3}$$

Typically we have prior knowledge that the stars in a given dataset are all of a similar, though not necessarily identical, age. The prior mean and variance for σ_A^2 are $m_0 \equiv \nu_0\sigma_0^2/(\nu_0 - 2)$ and $\tau_0^2 \equiv 2m_0^2/(\nu_0 - 4)$, respectively. Therefore, in this setting σ_A^2 is given a prior where ν_0 is large, and σ_0^2 is set to the expected within-cluster variance of the individual stellar ages. The isochrone mapping is both highly nonlinear and degenerate in that many different parameter values lead to the same expected photometry. As a result, there is typically insufficient information in the data alone to give meaningful answers. The inclusion of external knowledge from previous literature or standard astrophysics theory is an important tool in breaking these degeneracies. Indeed, the entire statistical model represents a translation of scientific understanding into a collection of modeling assumptions, and the Bayesian framework makes this task relatively straightforward. But we are

mindful of the need to check prior sensitivity and more appropriately to quantify inferential uncertainty.

The initial mass of a star, together with its metallicity, is one of the primary factors that determine how that star will evolve. The initial masses of stars are known to have a distribution, or *initial mass function* (IMF) that, for stars above a threshold M_{brk}, typically around one solar mass, is described by a power-law with parameter $\alpha = 2.5$ [18]. For stars below the threshold, the distribution of masses is considered to be uniform. However, we are interested in placing a prior on a star in our dataset, not the population of all stars. For a star of a given age we know *a priori* that for a star of that age to potentially be observed, it must have a mass within a certain range of values. As can be seen in the top panel of Fig. 4.1, stars with a large initial mass have a shorter lifespan than those with smaller initial mass. This leads to constraints on the support of the joint distribution of mass and age, with the support defined by the tabulation in Fig. 4.1. In light of this, and to ensure our prior includes only feasible (M_i, A_i) pairs we assume a distribution of the form:

$$p(M_i|A_i) = \begin{cases} 0 & M_i < M_{\text{min}} \\ k & M_{\text{min}} < M_i \leq M_{\text{brk}} \\ \frac{\alpha-1}{M_{min}} \left(\frac{M_i}{M_{min}}\right)^{-\alpha} & M_{\text{brk}} < M_i \leq M_{\text{max}}(A_i) \end{cases}, \tag{4.4}$$

where $M_{max}(A_i)$ is the maximum possible mass for an "observable" star of age A_i, as determined by the theoretical isochrones. M_{min} is selected to be the minimum mass that is scientifically reasonable for the dataset, or from the set of theoretical isochrones (usually 0.8 solar masses) and does not vary with age. The prior distribution on $\mathbf{R}$ is taken to be uniform across all positive definite correlation matrices. Note that this is not uniform on the correlation parameters, but will typically be close to uniform since the number of observed bands, p, is relatively small (e.g., single digit); see [2].

Posterior Inference and Sampling

The model specified by (4.1)–(4.4) yields a joint posterior distribution of dimension $d = 2(n+1) + 0.5p(p-1)$. In practice, however, we are typically most interested in the marginal posterior distribution of μ_A and σ_A^2. In real applications, n is usually on the order of tens of thousands, although the size of dataset can be anywhere from hundreds to millions of stars. The large amount of structure among the posterior distributions of the parameters poses a challenge to standard methods of approximating posterior quantities of interest. We now describe a Markov chain Monte Carlo (MCMC) scheme to sample from the posterior distribution. We utilize a Metropolis-within-Gibbs scheme, which sequentially draws from the $d - n$ full conditional distributions of each component of $\mathbf{R}$, of $\{(M_i, A_i), i = 1, \ldots, n\}$, and of (μ_A, σ_A^2); here we have only $d - n$ full conditional distributions because we draw

(M_i, A_i) jointly. Performing the first and last of these updates is straightforward. Hence, the greater interest is in sampling the mass and age of individual stars, given the observed photometry and stellar cluster parameters.

First we describe updates for the hyperparameters (μ_A, σ_A^2) and the correlation matrix $\mathbf{R}$. By conjugacy, the full conditional posterior distributions of the cluster-level parameters reduce to

$$\sigma_A^2|\mu_A, \mathbf{A} \sim \text{Inverse}-\chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + S_A^2(\mu_A) + \kappa_0(\mu_A - \mu_0)^2}{\nu_0 + n}\right)$$

$$\mu_A|\sigma_A^2, \mathbf{A} \sim N\left(\frac{\kappa_0\mu_0 + n\bar{A}}{\kappa_0 + n}, \frac{\sigma_A^2}{\kappa_0 + n}\right),$$

where $\mathbf{A} = \{A_i, i = 1, \ldots, n\}$, $S_A^2(\mu_A) = \sum_{i=1}^n (A_i - \mu_A)^2$, and $\bar{A} = \frac{1}{n}\sum_{i=1}^n A_i$.

To draw $\mathbf{R}$, we use component-wise Metropolis–Hastings updates, with a uniform proposal over the range of values that result in a valid (positive definite) correlation matrix. It is shown in [2] that when we change one correlation at a time, the positive definiteness constraints reduce to solving a quadratic equation to find the conditional support of the correlation. While this method can be inefficient for large correlation matrices, typically the number of observed bands is small in our application, and hence the proposal is rapid to compute and performs well in most settings.

To sample from the conditional posterior distributions $p(M_i, A_i|\mu_A, \sigma_A^2, \mathbf{Y})$ for $i = 1, \ldots, n$, we need to construct a proposal that is robust to both multi-modality and many different types of nonlinear dependencies that can be induced by different regions of the CMD. Figure 4.2 displays the "likelihood" of an old star as a function of initial mass and age, i.e., the contribution to the posterior from Eq. (4.1).

Ideally, to achieve this, we would utilize an energy-based sampler in the spirit of the Equi-Energy sampler of [11]. In its full incarnation the Equi-Energy sampler proceeds by constructing "energy bands" that attempt to empirically partition the full parameter space into posterior contours. Given the dimensionality, constructing such energy rings for the full $2(n + 1) + 0.5p(p - 1)$ dimensional posterior is infeasible in practice, as is constructing full energy bands for subsets of conditional distributions. Since the contours of the conditional posterior distributions depend on the conditioned values, it would be necessary to re-compute the partition for every star across every iteration. Nevertheless, we now describe how we can explicitly utilize the tabulated component of the posterior distribution to pre-compute a single partition that can be used across all conditional distributions (M_i, A_i), independently of the conditioning variables. By constructing the partition in this way, we retain the fundamental location-independent nature of the Equi-Energy sampler.

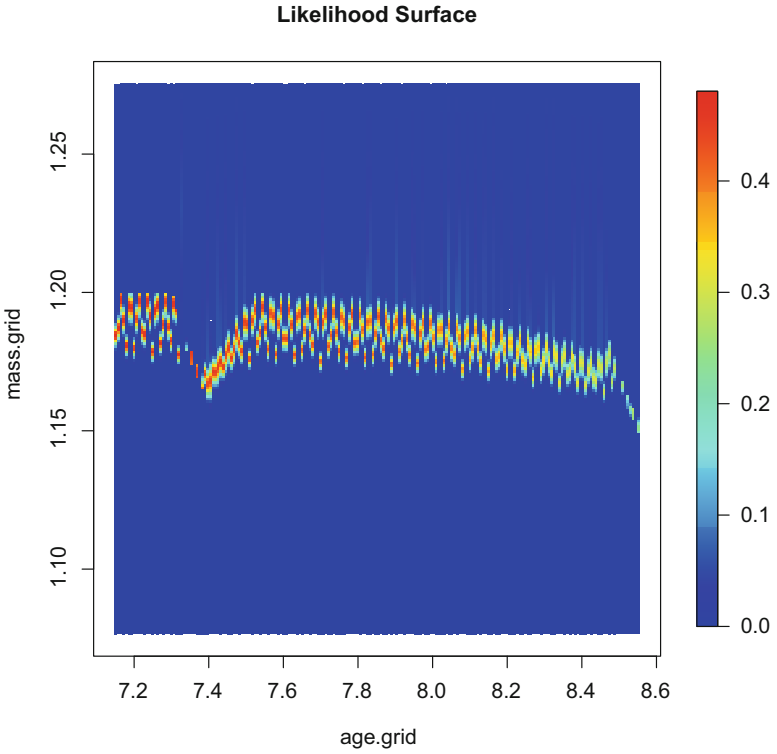


Fig. 4.2 The "likelihood" surface as a function of initial mass and age for a typical observation of an old star (10^{10} years)

A Partition Strategy Inspired by Equi-Energy Sampler

To avoid the additional complications of the CMD application, we first describe the construction of the proposal distribution for a simplified example. Consider two input (physical) parameters x and y that are related to two output quantities u and v on which measurements (with error) can be made. In the context of the CMD example x and y might correspond to the mass and age of an individual star, and u and v might correspond to two photometric bands. Suppose that the expected output for each of 2601 different combinations of input parameters (a regular grid of 51 unique values for each parameter) are given in a lookup table. The grid of input values is shown in the left-hand panel of Fig. 4.3. Our proposal distribution will be constructed from a partition of the parameter space: typically formed by polygons with corners at tabulated input points. The right-hand panel of Fig. 4.3 shows a possible partition of the input grid, obtained by Delaunay triangulation [16]. For each distinct polygon (triangle) we take the centroid as a "representative" of that region. Next, we compute the output value corresponding to the centroid. Since each

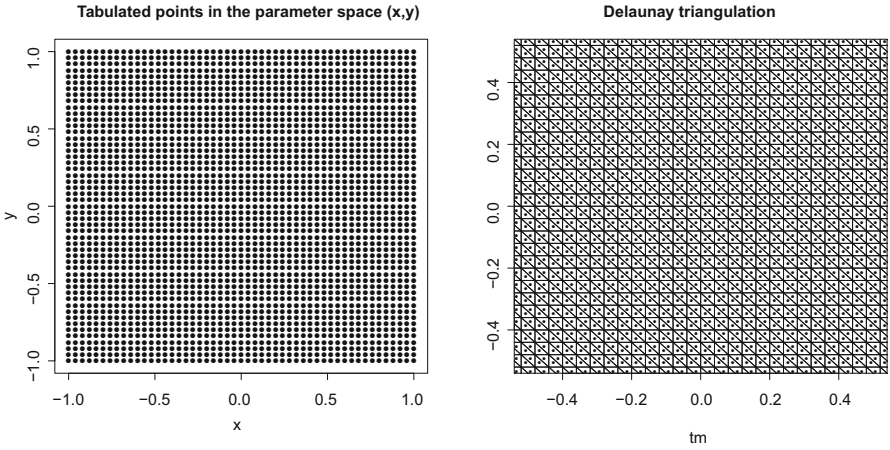


Fig. 4.3 (L) Tabulated combinations of the input parameters x and y. (R) Partition of the parameter space (x, y) into non-overlapping triangles, and centroids of those triangles. The right plot depicts only a subset of the parameter space, as the triangulation is regular. Vertices of the partitioning triangles are tabulated (x, y) points

vertex is a tabulated input point, the interpolated output value corresponding to the centroid is a distance-weighted average of the output values at the vertices.

For this example, we consider the following (isochrone) mapping from parameters to data space:

$$u = a_u(y - c_u)^2 + \sin(y) - |y + x| \tag{4.5}$$

$$v = a_v(x - c_v)^2 + \sin(x) + |y - x|, \tag{4.6}$$

where we select $a_u = 0.8$, $a_v = 1.2$, $c_u = -0.55$, and $c_v = 0.05$. Here we take the dimension of the data space to match the dimension of the parameter space, although this is not required. The mapping for an arbitrary point in the input, (x, y)-space, to the output, (u, v)-space, is done by interpolating the points in (u, v)-space corresponding to nearby points in (x, y)-space. For the interpolation to make sense in practice we require local continuity of the mapping between neighboring points, i.e., the tabulation must be sufficiently high resolution to enable safe interpolation of nearby values. This is not a restrictive requirement; all methods of analysis for CMDs rely upon sufficiently high-resolution tables. For the toy example, we now proceed as if the functional form of the mapping were not known: only tabulated values and interpolation are used.

Figure 4.4 illustrates some of the properties of the functional mapping from the parameters to the data. Firstly, as with the isochrone tables, the mapping is non-invertible: multiple (x, y)-values can lead to the same (u, v)-value, as manifested by intersecting cross-sections in the bottom panel of the figure. As with the isochrone tables, portions of the mapping are potentially invertible, but we want to derive a general method that does not rely on this fact. Secondly, as a result of this non-

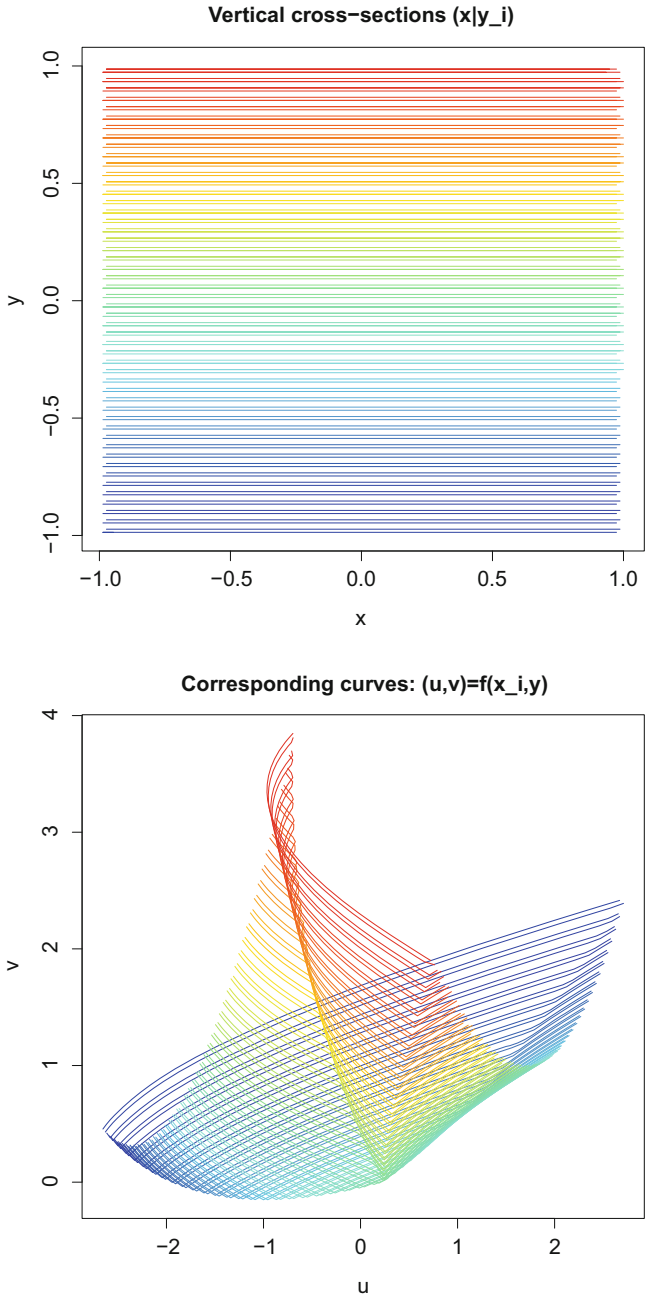


Fig. 4.4 (Top) Cross-sections of the parameter space, colored according to the fixed value of x. (Bottom) Each of the cross-sections maps to a curve in $(u, v) = f(x, y)$ as defined in Eqs. 4.5 and 4.6, where each curve is plotted in the same color as its corresponding cross-section

invertibility, similar observed data can arise from disjoint regions of the parameter space, hence the need to construct an efficient proposal distribution. Also, the mapping is not differentiable along the planes $x - y = 0$ and $x + y = 0$, also mimicking similar behavior to the isochrone tables.

The centroids of the input partition have their corresponding counterparts in the output space, shown in Fig. 4.5. The important observation is that, in a likelihood setting, similar expected values in the output space correspond to similar values of the target distribution. Hence, regions in the input space that correspond to nearby centroids in the output space will have similar likelihood values. In mapping back to the input space, we have essentially constructed a crude approximation to the inverse of the (many-to-one) mapping from input to output. The primary advantage of these "Equi-Expectation" contours is that they are expressed in a functional form. That is, given an arbitrary input point (x_0, y_0), we have instant access to a set of points with "similar" expected values, without knowledge of the observed data or conditioning parameters.

Exact contours of the likelihood surface depend on the observed data, and hence require fresh computation for each observation. However, we can form a random-walk style proposal in the output space that produces approximately location-independent moves in the input space. For sufficiently high-resolution tables, the regions of the input space that are nearby in terms of their expected output value will have similar values of the likelihood. Larger distances between points in the output space correspond to larger differences in likelihood, with the Euclidean distance providing a natural metric when observations are made with Gaussian measurement errors. In practice, computing distances between all of the (u, v)-centroids is computationally expensive if the tables are very high resolution, as the isochrones tables are. So, to reduce the computational burden, we define "similar" in this context by running a clustering algorithm on the centroids in (u, v)-space, and tracking the accompanying (x, y)-clusters. In the CMD example, these clusters correspond to values of mass and age that have similar expected photometry: essentially the banded inversion of the isochrone mapping, f, as in (4.1). In general, the dimensions of the input and the output spaces do not need to match, and we can have input parameters defined on $\mathbb{R}^k$ mapping to an output space on $\mathbb{R}^p$ or subsets thereof.

A Proposal Distribution for (M_i, A_i) via Ancillary Augmentation

Now we return to the CMD example and address additional implementation challenges. Unlike the toy example, the isochrones are given on an irregular (m, a) grid, so the choice of partition is not immediate. The partition can either be constructed manually or using a standard technique such as the Delaunay triangulation [16] of the input vertices; we use the latter method for all applications presented here. Figure 4.6 shows an example of the partition formed using Delaunay

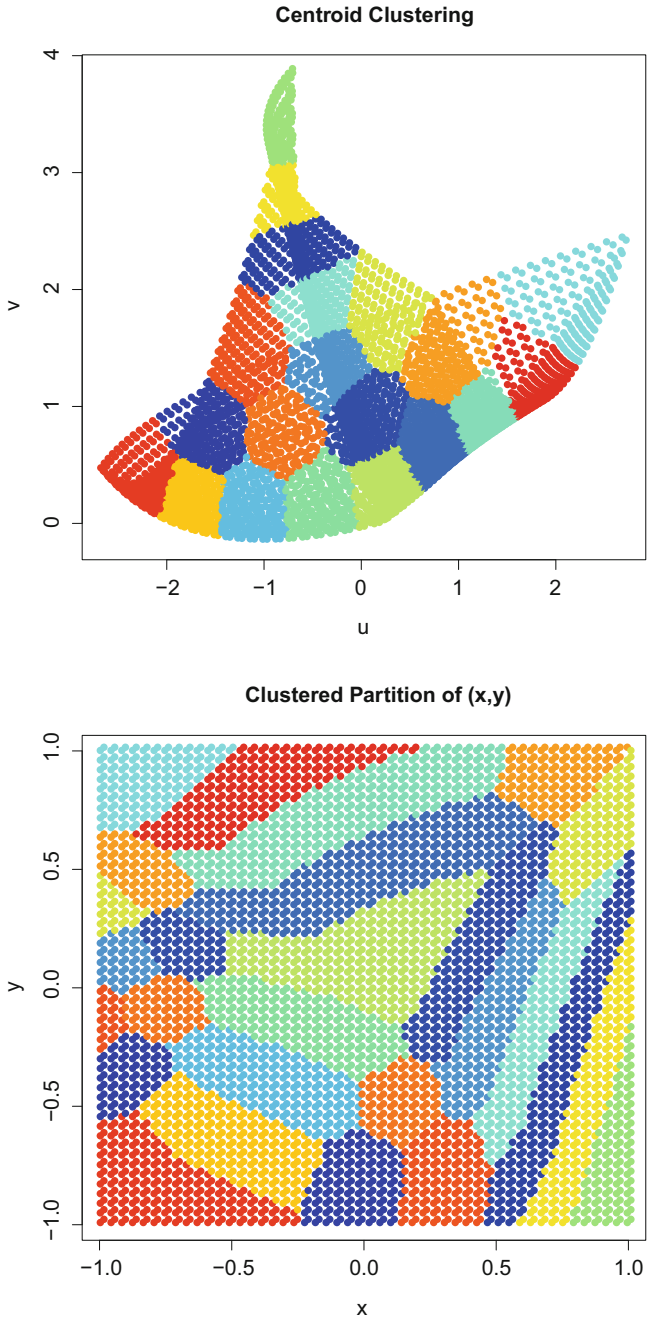


Fig. 4.5 (Top) Interpolated values $(u, v) = f(x, y)$ for each of the centroids defined by the (x, y)-partition in Fig. 4.3. The values are clustered into 50 groups containing "similar" output values, each represented with a different color. (Bottom) The corresponding clustered partition of the input. The color of each triangle reflects the cluster to which its centroid belongs

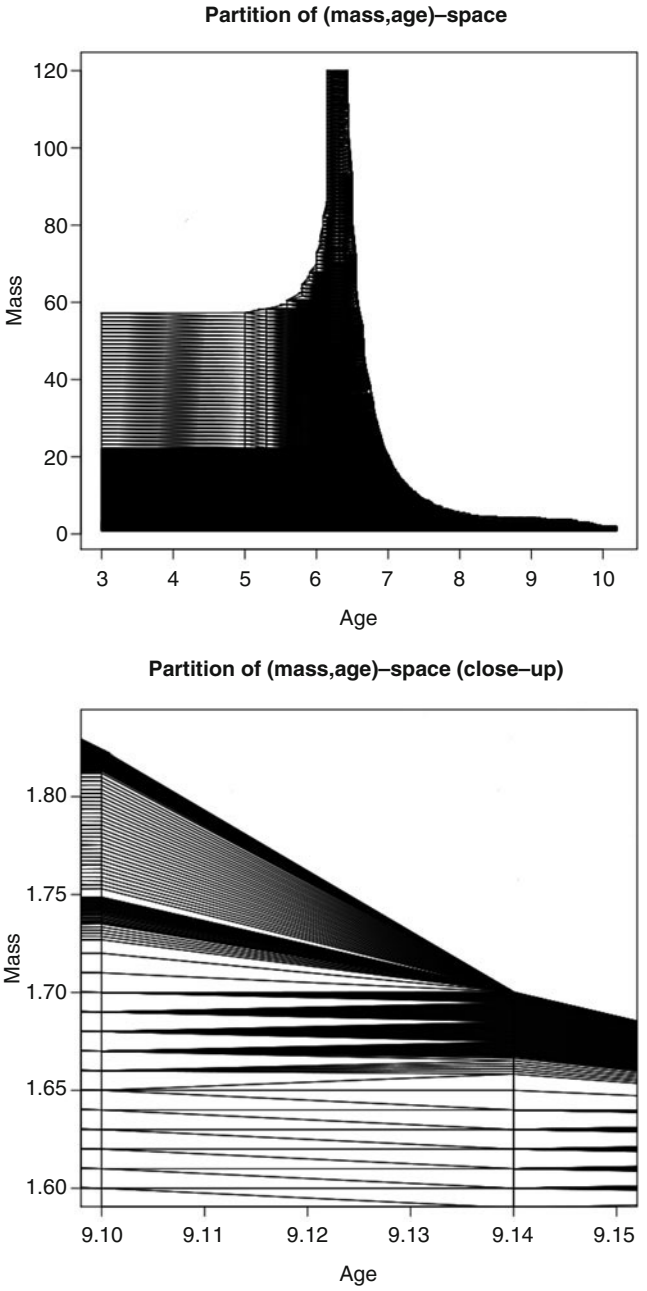


Fig. 4.6 (Top) Partition of the parameter space using the Delaunay triangulation, and (Bottom) a close-up of the partition. Note that blank regions in the upper right of each plot correspond to infeasible mass-age combinations

triangulation for the Geneva isochrones. Note that the tabulation is very irregular, with much higher resolution tabulation for masses close to the maximum allowable mass for each given age. Each corner of the polygon corresponds to a tabulated value that has a corresponding vector of expected photometric values. As discussed, assuming the isochrone tabulation is of sufficiently high resolution, the implied isochrone values within a given polygon can be approximated by interpolation of the (vector) values at the corner points. Next, taking the centroid of each polygon as a representative of that particular region of the parameter space, we proceed to construct approximate "contours" of the conditional posterior distributions that correspond to these centroid values. Each centroid is comprised of two components: (i) a pair of mass and age values (m_j, a_j), and (ii) an (interpolated) isochrone value describing the expected photometry with the given mass and age $f(m_j, a_j) \in \mathbb{R}^p$.

After running the clustering algorithm on the photometry vectors at the centroids, each cluster is simply a list of polygons defining a collection of possibly disconnected regions of the parameter space. For computational simplicity, we use k-means clustering to form C clusters. Given a set of C clusters of polygons, we can quantify approximate measures of the "distance" between points in each pair of distinct clusters. Finally, after reparameterization, we ensure that nearby clusters, as quantified by their distance in the observed photometric bands, will provide similar values of the conditional posterior—yielding a proposal that enables both location-independent movement throughout the mass-age parameterization, and approximate contour-based sampling for *all* of the n independent conditional distributions $p(M_i, A_i|\mu_A, \sigma_A^2, \mathbf{Y})$. As presented however, the motivation for the partitions was that they allow location-independent exploration of multiple modes and diverse regions of the parameter space. However, in our hierarchical model we must deal with the additional contributions from the informative prior distributions in Eqs. (4.2) and (4.4). To do this, we perform the proposal using the *ancillary parameterization* [22]. For applications where the lowest mass stars are above the IMF break-point $M_{brk} = M_{min}$, this becomes

$$\tilde{A}_i = \Phi\left(\frac{A_i - \mu_A}{\sigma_A}\right), \quad \tilde{M}_i = \frac{M_{min}^{-(\alpha-1)} - M_i^{-(\alpha-1)}}{M_{min}^{-(\alpha-1)} - M_{max}(A_i)^{-(\alpha-1)}}, \tag{4.7}$$

where $\Phi(x)$ is the CDF for the standard normal variable. Under this augmentation scheme the model becomes

$$Y_i|\tilde{\mathbf{M}}, \tilde{\mathbf{A}}, \mathbf{R}, \mu_A, \sigma_A^2 \sim N\left(f(\tilde{M}_i, \tilde{A}_i, \mu_A, \sigma_A), \mathbf{R}\right), \tag{4.8}$$

$$\tilde{A}_i|\mu_A, \sigma_A^2 \sim \text{Unif}[0, 1], \qquad \tilde{M}_i|\tilde{A}_i, \mu_A, \sigma_A^2 \sim \text{Unif}[0, 1], \tag{4.9}$$

$$\mu_A|\sigma_A^2 \sim N\left(\mu_0, \ \sigma_A^2/\kappa_0\right), \qquad \sigma_A^2 \sim \text{Inverse}-\chi^2\left(\nu_0, \sigma_0^2\right). \tag{4.10}$$

Hence, the conditional distribution of any given individual mass-age pair reduces to

$$p(\tilde{m}_i, \tilde{a}_i | \mu_A, \sigma_A^2, \mathbf{Y}) \propto \exp\left\{-\frac{1}{2}\text{tr}\left(\mathbf{R}^{-1}\tilde{\mathbf{F}}_i\right)\right\}, \quad (\tilde{M}_i, \tilde{A}_i) \in [0,1]^2, \tag{4.11}$$

where

$$\tilde{\mathbf{F}}_i = \left(\mathbf{y}_i - f(\tilde{m}_i, \tilde{a}_i, \mu_A, \sigma_A^2)\right)\left(\mathbf{y}_i - f(\tilde{m}_i, \tilde{a}_i, \mu_A, \sigma_A^2)\right)^\top. \tag{4.12}$$

By essentially placing all of the additional non-likelihood terms inside the mapping between sufficient and ancillary augmentation, we can help facilitate the improved performance of our likelihood-based proposal distribution. The impact of the transformation can be seen by the relative differences in areas between corresponding regions of the parameter space, i.e., the Jacobian. If the current state of the MCMC chain for star i is (m_i, a_i), which is contained in polygon k, in cluster l, then we implement the partition-based proposal as follows:

Algorithm 1 [For Computing the Proposal Distribution]

1. Select a cluster l^* with probability $p^C_{ll^*}$.
2. Select a polygon k^* from within cluster l^* with probability $p^W_{l^*k^*}$.
3. Propose a point (m_i^*, a_i^*) uniformly within polygon k^*, and map to $(\tilde{m}_i^*, \tilde{a}_i^*)$.

□

By encouraging moves between nearby clusters we can effectively explore different regions of the parameter space with similar photometry, and hence, similar likelihood. Note that although the partition is constructed in terms of the stellar mass and age, the transformation defined by Algorithm 1 is one-to-one and monotonic, and hence it forms a valid partition in the ancillary parameterization for any values of μ_A and σ_A^2. However, the transformation is not affine and the partition no longer consists of polygons. The transition probability corresponding to Algorithm 1 is given by $q\left((\tilde{m}_i, \tilde{a}_i), (\tilde{m}_i^*, \tilde{a}_i^*)\right) = p^C_{ll^*} p^W_{l^*k^*} |J(\tilde{m}_i, \tilde{a}_i)| / |\mathcal{U}_{k^*}|$, where $J(\tilde{m}_i, \tilde{a}_i)$ is the Jacobian of the transformation from the sufficient to ancillary augmentation evaluated at the proposed state, and $|\mathcal{U}_{k^*}|$ is the area (in the sufficient augmentation) of the k^*-th unique polygon within cluster l^*. There is some freedom in choosing both the cluster-to-cluster and within-cluster proposal probabilities. For the cluster-to-cluster probabilities we compute the centroid of all centroids of polygons within the cluster, providing an approximate "center" of the cluster (in $\mathbb{R}^p$) and then compute Euclidean distances between all cluster centers. The cluster-to-cluster proposal probabilities are then selected to be $p_{ll^*} = \exp\left\{-d^2(x_l, x_{l^*})/\beta\right\}$, where $d(\cdot,\cdot)$ is the Euclidean distance, x_l and x_{l^*} are the cluster centroids in photometric-space, and β is a tuning parameter controlling how freely we propose to move to nearby clusters.

All distances here are computed with respect to the Euclidean norm in photometric-space independently of the mass-age location, thus imitating the

posterior distribution and allowing free movement across modes in terms of the stellar mass and age. For example, two distant regions of the parameter space that produce the same expected photometry would be placed in the same cluster, and the proposal distribution as constructed provides a high probability of proposing to move between these disconnected regions. We note that the acceptance probability is strongly influenced by the area of the polygon in the ancillary scheme; an artifact of the ancillary transformation (4.7). This choice of cluster-to-cluster probabilities mimics a random-walk Metropolis proposal, "centering" the proposal around the current cluster, and proposing to move to regions of the parameter space with a probability that reflects the similarity of the photometry to the photometry at the current state.

As discussed, the size of the polygon in the ancillary parameterization is a function of the hierarchical structure and could also be accounted for in selecting the within-cluster probabilities if desired. For example, if the within-cluster proposal probability is chosen to be proportional to the area of the region in the ancillary parameterization, then it yields a uniform proposal over the area defined by the cluster. In practice this is implemented by computing the polygon areas in the sufficient parametrization and incorporating the Jacobian term. If this approach is taken then, since the mapping between parametrizations depends on the hyperparameters μ_A and σ_A^2, the Jacobian terms within the cluster would need to be recomputed at each iteration. Uniform proposals across the cluster do not require this extra computation but can be less efficient. In practice we combine this proposal distribution with a random-walk proposal of the form: $(m_i^*, a_i^*)^\top \sim N((m_i, a_i)^\top, \text{Diag}(\lambda_1, \lambda_2))$, where λ_j are proposal variances that can be tuned to achieve desired acceptance rates. This combination of proposal helps to facilitate both rapid local and global exploration of the posterior distribution. Since the correlation between mass and age depends on the region of the CMD, we do not attempt to approximate the correlation between the variables. In our experience, there is little performance change when using either the cluster-based or random-walk proposal distributions between 20–80% of the time.

Checking the Effectiveness of Our Proposal

To understand the impact of the transformation, and the resulting proposal distribution, we begin by examining the components of the posterior distribution within the original (m, a)-parametrization. Figure 4.7 shows the prior (Top) and posterior (Bottom) surfaces for an individual star. The posterior surface is obtained by combining the likelihood surface in Fig. 4.2 with the prior as shown. As we can see from the bottom panel of Fig. 4.7, the posterior surface is challenging to sample efficiently from, particularly given that the presence and scale of any large-scale ellipsoidal trends can vary dramatically across stars. In light of this, to retain computational robustness to the form of isochrone being used, and

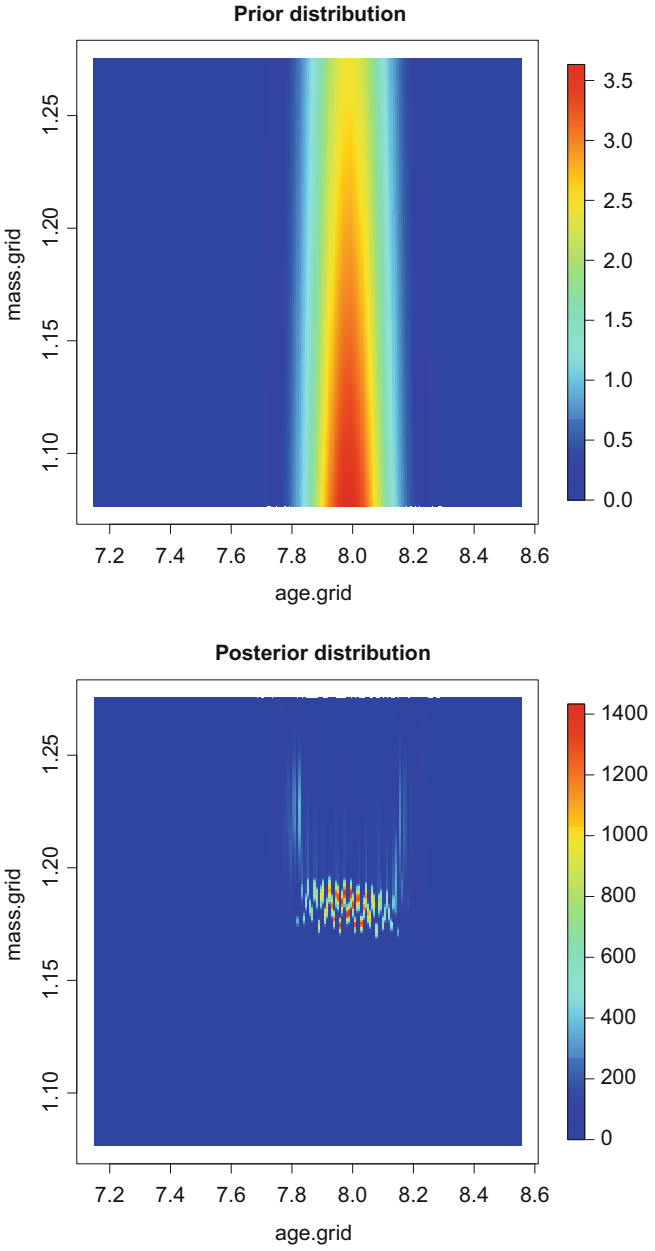


Fig. 4.7 (Top) Conditional prior surface for $(M_0, A_0)|\mathbf{Y}, \mu_a, \sigma_a^2$, and, (Bottom) conditional posterior surface for $(M_0, A_0)|\mathbf{Y}, \mu_a, \sigma_a^2$. The corresponding likelihood surface is given in Fig. 4.2

to maintain generality for non-isochrone settings, we avoid making observation-specific approximations to these conditional posterior distributions.

Our alternative approach, using the ancillary transformation, is depicted in Fig. 4.8. The top plot displays the transformation of the posterior distribution in Fig. 4.7 to the ancillary parametrization. Since the posterior distribution in the ancillary parametrization is simply a rescaling of the likelihood surface, we can observe the similarity in structure to Fig. 4.2. The bottom plot of Fig. 4.8 displays a proposal distribution obtained using our algorithm. The current state of the chain is highlighted by the black dot, and the proposal the proposal distribution mimics the contours of the ancillary posterior, albeit wrongly centered around the current state of the MCMC algorithm due to the random-walk style as implemented here. However, for this particular example the variance of our proposal is considerably greater than is desirable. This is the result of the observation falling in a region of insufficiently high resolution relative to the observational errors. This lack of resolution also illustrates the limitation of the "equi-expectation" approximation for low-resolution tables. Given higher resolution tables (or observations with higher measurement variance), and thus a higher resolution polygon-cluster proposal distribution, we will steadily obtain more appropriate contours and variance in the proposal.

As an analogous alternative to the random-walk style proposal, an independence style proposal could also be used where the cluster weights are proposed based on the distance between the observed photometry and the cluster centroids. This strategy would likely be more effective than the cluster-based alternative, but the large amount of computation required for each star and at each iteration renders it considerably more computationally expensive. In seeking the optimal trade-off between improved mixing and implementation speed, we elect not to pursue this further.

Although the clustering of polygons is not, in principle, necessary, the large number of polygons ($>200{,}000$) makes the construction of a polygon-to-polygon proposal more challenging, more memory-intensive, and more time-consuming. By adding the clustering of polygons, we need only store the much smaller $B \times B$ cluster-to-cluster proposal probability matrix, and possibly the within-cluster proposal probabilities (although this is not required for uniform proposals within the cluster).

Addressing Block Correlations

The proposal distribution for the individual stellar masses and ages is useful only in helping to sample efficiently from the series of conditional posteriors. As should be anticipated from the hierarchical specification of the model, there remains a large posterior correlation between (μ_A, σ_A^2) and $(A_1, \ldots, A_n)$.

To help address these problems, we embed our sampler within a parallel tempering (PT) framework [9] to facilitate easier movement around the posterior space.

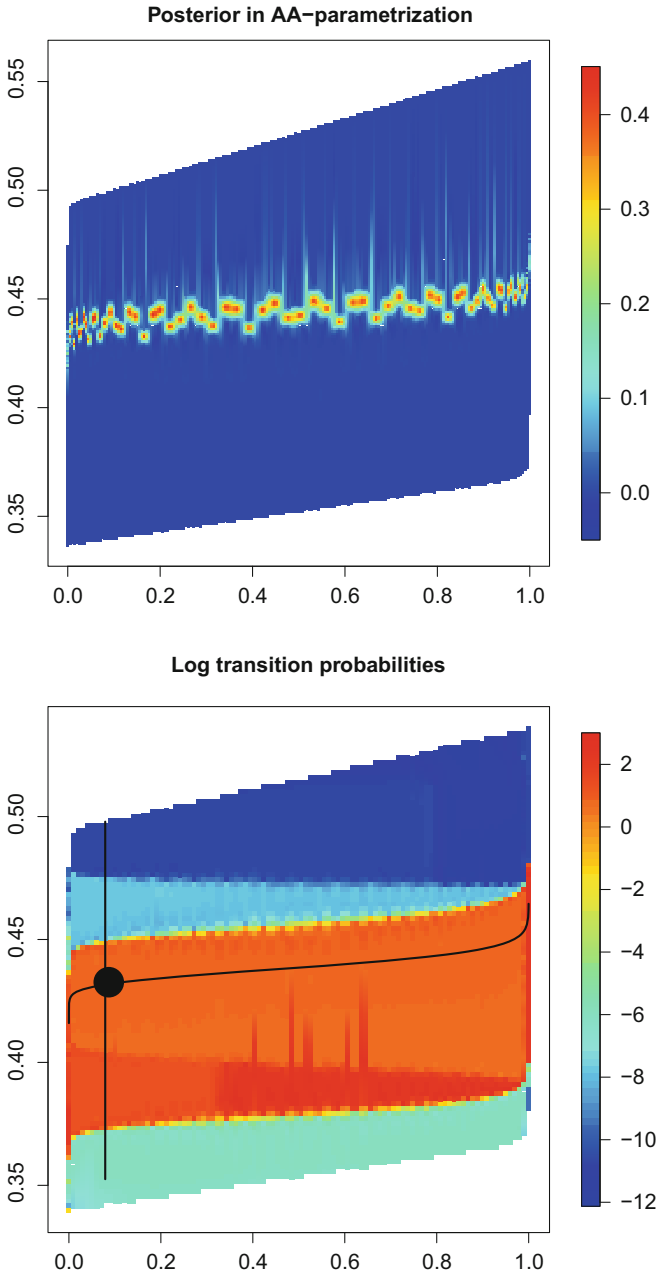


Fig. 4.8 (Top) Conditional posterior surface for $(\tilde{M}_0, \tilde{A}_0)|\mathbf{Y}, \mu_A, \sigma_A^2$, and, (Bottom) log-transition probabilities for (M_0, A_0) obtained by applying algorithm 1. The current state is shown by the black dot. The proposal mimics a random-walk across contours of the posterior surface

To sample $p(\theta)$ with energy $H(\theta) = -\log p(\theta)$, PT proceeds by constructing a sequence of tempered distributions, $\{p_1(\theta), \ldots, p_N(\theta)\}$ of the form $p_j(\theta) \propto \exp\left\{-H(\theta)/T_j\right\}$, with $T_N > \ldots > T_1 = 1$. By applying a series of monotonic transformations to the full posterior density, the full conditional densities in the Gibbs sampler are also transformed in an identical manner. An attractive feature of parallel tempering is that the modified conditional distributions require only trivial modifications. The tempered conditional posterior of (μ_A, σ_A^2) is given by

$$p_j(\mu_A, \sigma_A^2|\mathbf{M}, \mathbf{A}, \mathbf{R}, \mathbf{Y}) \propto \exp\left\{-\frac{1}{2T_j\sigma_A^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu_A - \mu_0)^2 + \sum_{i=1}^{n}(A_i - \mu_A)^2\right]\right\} \cdot (\sigma_A^2)^{-\left(1+\frac{(\nu_0+n+3-2T_j)/T_j}{2}\right)}$$

The conjugate marginal/conditional formulation can be shown to yield:

$$\sigma_A^2|\mathbf{M}, \mathbf{A}, \mathbf{R}, \mathbf{Y} \sim \text{Inverse}-\chi^2\left(\frac{\nu_n}{T_j}, \frac{\nu_0\sigma_0^2 + (n-1)s_A^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{A} - \mu_0)^2}{\nu_n}\right), \tag{4.13}$$

$$\mu_A|\sigma_A^2, \mathbf{M}, \mathbf{A}, \mathbf{R}, \mathbf{Y} \sim N\left(\frac{\kappa_0\mu_0 + n\bar{A}}{\kappa_0 + n}, \frac{T_j\sigma_A^2}{\kappa_0 + n}\right), \tag{4.14}$$

where $\nu_n = \nu_0 + n + 3(1 - T_j)$ and $s_A^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(A_i - \bar{A}\right)^2$. Since ν_n cannot be negative, we must impose $\max_j T_j < (\nu_0 + n + 3)/3$. Typically either n, ν_0 or both are large, and hence this condition is not generally restrictive.

4.4 Empirical Investigations

Simulation Studies

Given the complex properties of the isochrone tables, it is important to validate that the sampling algorithm can reliably converge to the correct posterior distribution. See [21] for an illustration of typical complications when using MCMC with isochrone tables. We approach this with an aggregate check of coverage properties. That is, we simulate many datasets from the model, and we then fit the model to obtain posterior intervals and check nominal and actual coverage are consistent. This is a special case of the more general framework in [5]. For this aggregate check we simulate 1000 datasets from the model for each of the parameter configurations detailed in Table 4.1.

Table 4.1 Details of coverage simulations used to validate the algorithm described in Sect. 4.3

Sim.	n	p	Bands	μ_0	σ_0^2	ν_0	κ_0	Number of datasets
1	100	3	(M(U),B,V)	8.1	0.01	1,000,000.0	1.0	1000
2	1000	3	(M(U),B,V)	9.2	0.05	1000.0	100.0	1000
3	100,000	3	(M(U),B,V)	6.6	0.16	1000.0	10.0	1000

Table 4.2 Coverage properties of the different sampling algorithms for simulation configuration 1 ($\mu_0 = 8.1$)

	1%	2.5%	5%	25%	50%	75%	95%	97.5%	99%
M_i									
MH	0.8	2.1	4.3	24.4	49.0	73.9	94.0	96.9	99.0
MH + PT	0.7	2.0	4.1	23.4	49.0	74.2	94.9	97.6	99.1
PC	0.7	2.1	4.3	24.3	48.8	74.2	94.4	97.1	99.0
A_i									
MH	1.7	3.4	5.5	25.2	48.0	72.3	92.5	95.2	96.9
MH + PT	1.2	2.9	5.0	26.2	49.4	75.2	94.7	97.6	98.9
PC	1.2	2.8	4.9	25.4	49.3	75.4	95.0	97.6	99.0
μ_a									
MH	1.7	2.9	5.0	24.8	46.9	72.0	91.6	94.7	96.3
MH + PT	0.7	2.2	4.3	25.8	49.3	76.5	95.3	97.7	99.2
PC	0.7	1.9	4.2	25.1	48.1	75.0	94.6	97.3	99.0
σ_a^2									
MH	0.7	1.8	3.4	24.2	48.0	75.1	95.2	97.1	99.1
MH + PT	0.7	1.5	2.8	21.5	45.9	72.4	95.0	97.0	99.5
PC	0.7	1.7	3.4	23.8	48.2	75.2	95.1	97.3	99.2

For each of the three settings, there are four different MCMC schemes:

1. **(MH):** Vanilla scheme using only random-walk proposals for the individual masses and ages, without tempering,
2. **(MH + PT):** As in (MH), with additional parallel tempering,
3. **(PC):** The Polygon-Cluster scheme of Algorithm 1, without tempering,
4. **(PC + PT):** As in (PC), with additional parallel tempering.

Due to limited computational resources, we did not implement the 4th scheme. In all cases we combine the results from four chains, and each algorithm is run for approximately the same total CPU time across the four chains. Remaining tuning parameters such as the variance of the random-walk proposal, the number of clusters and the cluster-to-cluster "variance" parameter β were chosen after pilot runs on a subset of the datasets. Table 4.2 shows the coverage properties for a subset of the parameters for simulation number 1.

The first, simplest method struggles to effectively sample the tails of the posterior distributions, particularly for the main parameter of interest, and most computationally challenging parameter, μ_a. Adding in tempering we do better

in many cases, but some potentially worrisome discrepancies between actual and nominal coverage still remain, even with tempering. The third method, using the cluster-based partition proposal without tempering appears to do slightly better than the standard approach both with and without parallel tempering. Although we do not directly employ the combination of the cluster-based proposal with parallel tempering across all 1000 datasets, we do recommend such an approach for the analysis of a single dataset. While on aggregate the differences between the approaches do not appear to be drastic, the results for any given dataset can differ by a non-negligible amount. Brute force numerical integration for a subset of the datasets suggests that the cluster-based proposal and the cluster-based proposal with tempering better capture the tails of the distribution, although we defer a fuller analysis for future investigation.

Results from configurations 2 and 3 are very similar to those presented above and omitted for brevity. One important difference that we note here though is the size of the dataset; configuration 3 analyzes 1000 datasets of 100,000 observations each: an important test of the scalability of our approach. We run the analysis for each dataset for a maximum of 24 h: a reasonable computational cost for such large-scale analysis.

NGC 104: 47 Tucanae

"47 Tuc" is a globular cluster estimated to be 13,000–17,000 light years away, originally discovered by Abbe Nicolas Louis de Lacaille in 1751 [7]. Being the second largest and second brightest globular cluster, it has been extensively studied in recent years. Examples include [10] and [19]. Here we reanalyze photometric data to investigate possible age differences within the cluster and to assess the sensitivity of estimates to the choice of hyperprior. The 47 Tuc data consists of 1,697 observed stars in V and B bands ($p = 2$), with no missing data. For the analysis here we consider the distance modulus to be fixed at 13.33, although the extension to estimating the distance modulus is, in principle, straightforward. Figure 4.9 shows the data that we analyze: each dot corresponds to a star, with accompanying measurement error. The colored dots in the figure represent the theoretical isochrones: our model essentially seeks a distribution over these curves that best represents the 47 Tucanae cluster. CMD-based estimates of the (single $\log_{10}$) age of 47 Tuc typically range from 9.95 to 10.10 ($9.0 - 12.5$ billion years). In light of this, we select the hyperparameters for the analysis to reflect the estimates and uncertainty ranges in the literature:

$$\mu_0 = 10.025, \quad \kappa_0 = \frac{9}{64}, \quad \nu_0 = 1000 \text{ and } \sigma_0^2 = 0.03^2. \tag{4.15}$$

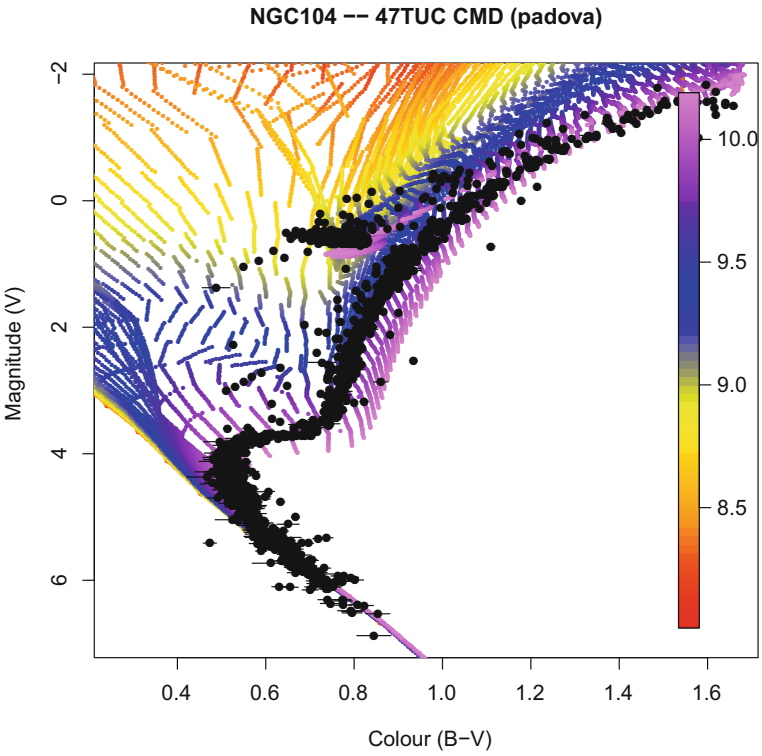


Fig. 4.9 Color-Magnitude Plot of the 47 Tucanae dataset. Each black dot represents a star in the dataset, colored points represent tabulated theoretical isochrone values. The color of each point represents its corresponding age, the mass, and metallicity of each point is not shown

These correspond to approximately

$$\mu_A \sim N(10.025, 0.08^2), \qquad \sigma_A^2 \sim N(0.03, 8.1 \times 10^{-10}).$$

The analysis was performed using the Polygon-Cluster proposal distribution for the individual mass-age distributions, and a ladder of 8 logarithmically spaced tempering distributions. We run a total of 10 chains, each for approximately 24 h, and combine the results for estimation. Relevant convergence diagnostics were checked, but we omit the details for brevity. Figure 4.10 shows the posterior median and 95% intervals for each of the stars in the dataset, sorted by increasing posterior median. We clearly see a heavy left-tail: a collection of approximately 100 stars that appear to have a lower age than the rest of the cluster. An alternative cruder but simpler visualization is given by simply plotting a histogram of the posterior medians of the individual stars, as shown in Fig. 4.11.

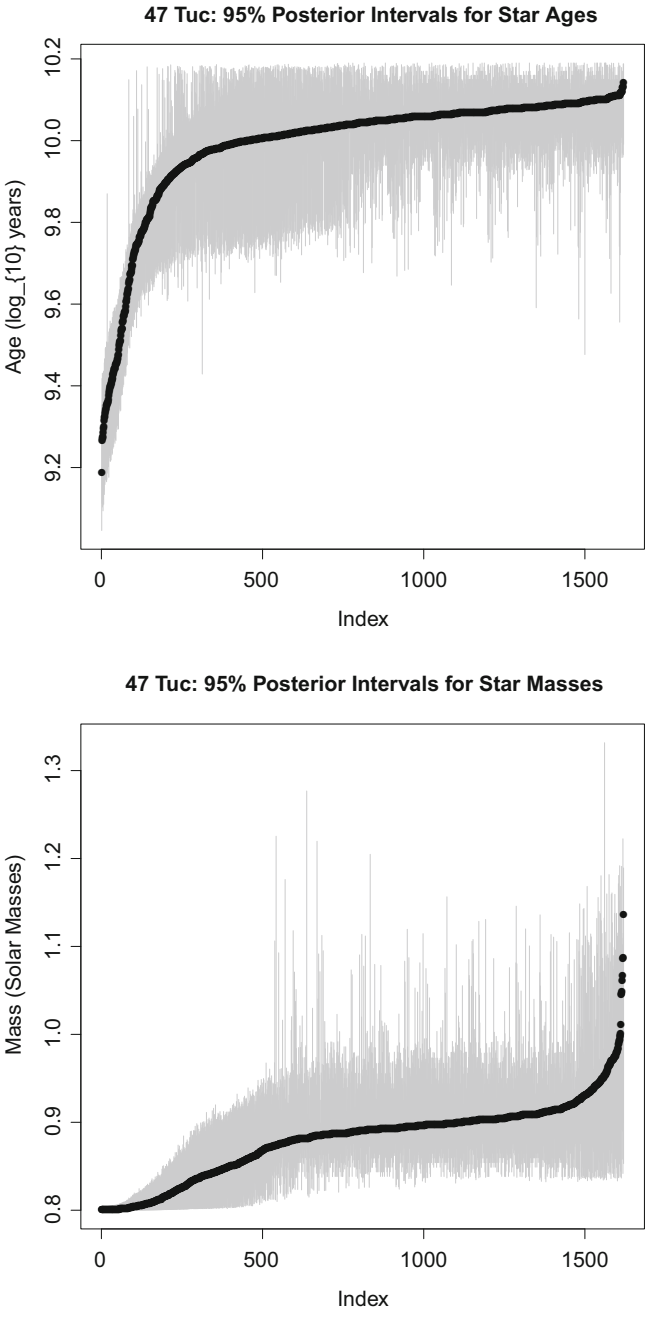


Fig. 4.10 Posterior intervals for the individual stellar ages a_i (Top) and masses m_i (Bottom). The stars are sorted in order of the posterior median, shown as a black dot: the accompanying 95% intervals are shown as gray bars

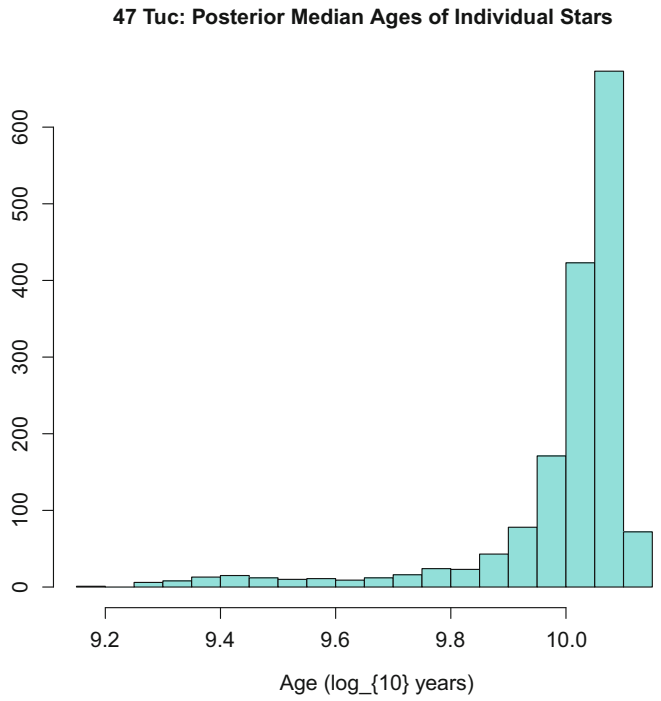


Fig. 4.11 Histogram of the posterior medians of all stars in the dataset. While providing a simpler visual, the distribution of posterior medians is a less complete representation than the posterior intervals of Fig. 4.10

The long-held belief has been that globular clusters are formed in a single burst from a single cloud of material. Based purely on the data and analysis here, however, there is a suggestion that 47 Tuc may contain multiple star formation bursts. Alternative explanations for the phenomena in Fig. 4.10 include contamination by foreground stars, misspecification or uncertainty in the distance modulus, or bias induced by extinction: see Sect. 4.5 for more details. Examining Fig. 4.10, there appear to be two bursts of star formation approximately 3 Gyr (billion) years apart at 7.9 Gyr and 11.5 Gyr ago, respectively. Recent independent work [1] using different techniques also suggests multi-burst SF in 47 Tuc, although analysis with higher quality multi-band photometric data would be required before drawing such scientific conclusions. Importantly, however, the flexibility in our model provides sufficient richness to be able to investigate previously untestable assumptions. Indeed, it is this additional flexibility and the appropriate modeling of uncertainty that is the primary contribution of statistical research in astrophysics.

4.5 Extensions and Future Work

Uncertainty in the "Black Box"

All of the previous analysis was predicated on the assumption that the "black box" describing the relationship between the physical parameters and the observed data (i.e., the isochrone mapping) is correct. In practice there is also uncertainty in these mappings, and we now consider some approaches to investigate this. Ideally, uncertainty in the mappings would be proliferated down through the mapping in the form of uncertainties in previously fixed quantities, i.e., essentially creating an expanded black-box/lookup table incorporating both different inputs and different physical assumptions. In practice, however, this is rarely feasible without access to the models that generate the lookup tables. In light of this, we consider a simpler problem: comparing competing sets of isochrones. For simplicity we consider comparison of two competing black boxes, although the extension to the comparison of more than two is straightforward.

Given two competing models, $\mathcal{M}_1$ and $\mathcal{M}_2$, differing only by the specific choice of isochrone table, i.e., f in Eq. (4.1), we specify prior probabilities for each model: $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$. In all cases here we begin with a neutral prior, selecting $p(\mathcal{M}_1) = p(\mathcal{M}_2) = 0.5$. The posterior model probabilities are then given by

$$p(\mathcal{M}_1|\mathbf{Y}) = \frac{1}{1 + \frac{p(\mathcal{M}_2)}{p(\mathcal{M}_1)} \frac{\int p(\theta)p(\boldsymbol{\mathcal{Y}}|\theta,\mathcal{M}_2)d\theta}{\int p(\theta)p(\boldsymbol{\mathcal{Y}}|\theta,\mathcal{M}_1)d\theta}}, \tag{4.16}$$

thus requiring only additional computation of the ratio of normalizing constants for the two competing posterior distributions. More general model comparisons allow for different priors $p(\theta|\mathcal{M}_j)$, although in our application the prior $p(\theta)$ is the same for both models. Meng and Wong [15] show how one can estimate ratios of normalizing constants using the bridge sampling identity:

$$\frac{c_2}{c_1} = \frac{\mathbb{E}_1\left[q_2(\theta)\alpha(\theta)\right]}{\mathbb{E}_2\left[q_1(\theta)\alpha(\theta)\right]}, \tag{4.17}$$

where α is an arbitrary function providing a "bridge" between the two densities. They also provide a fast-converging iterative scheme to approximate the estimator under the optimal α. Note that the expectations of each unnormalized posterior are taken with respect to the other model. Therefore, if posterior samples are available for the two competing models, then implementing this model comparison boils down to the evaluation of the unnormalized posterior density for each draw from its rival model.

Multi-Cluster Models

As discussed in the context of the 47 Tucanae analysis, there are potentially applications where we want to allow for the possibility of multiple stellar clusters. The model defined by (4.1)–(4.4) is explicitly designed for single-cluster populations, although one possible generalization is conceptually straightforward. We could consider replacing (4.2) by an alternative mixture distribution:

$$A_i | H_i, \mu_{A,H_i}, \sigma^2_{A,H_i} \overset{ind}{\sim} N\left(\mu_{A,H_i}, \sigma^2_{A,H_i}\right), \quad H_i \in \{1, 2, \ldots, K\}, \tag{4.18}$$

where H_i is the cluster membership of star i. In most applications H_i would be given a uniform prior. When combined with identical priors on the cluster-specific hyperparameters, the posterior is defined only up to label switching. While, in principle, the number of clusters K could also be estimated, this would likely be fixed as part of the analysis. The additional computational burden induced by (4.18) rests primarily in the additional block correlations between the cluster- and individual-level variables.

Extinction and Non-ignorable Missingness

In many examples it is possible that observations for some stars are either partially (i.e., one or more bands) or fully (i.e., all bands) missing. The missing data mechanism for this missingness can potentially depend on the intrinsic brightness of the stars. That is, brighter stars are more likely to be observed than dimmer ones. Thus, the missing data mechanism can potentially provide information about the model parameters. For a given detector the detection/missingness probabilities are often well understood by careful calibration and testing. In such cases we often have access to a series of functions that express the probability of missingness as a function of the brightness of the star, a functional form that can then be coherently built into our hierarchical model. Again, within the Bayesian framework the extra layer can be added in a relatively straightforward way, although this will entail an additional computational burden. The importance of this missingness mechanism varies depending upon the type of stellar cluster being analyzed, and thus we currently restrict to those datasets where it is unlikely to affect the resulting inference.

Going Beyond Stellar Populations

Computer models and "black-box" likelihoods are increasingly common in many scientific disciplines, and can pose some interesting challenges to traditional com-

putational methods. In the case of analyzing stellar populations, that the likelihood is tabulated is both a blessing and a curse. We benefit in that much of the structure in the model is known *a priori*, and we show how an effective proposal distribution can be pre-computed independent of data. However, the black box proves to be a curse in that understanding and intuition are harder to come by, as are analytic simplifications and approximations. Despite this, one can construct an efficient and effective sampling scheme even for highly nonlinear and degenerate likelihoods that are more robust to the properties of the black box than naive methods.

The frequency of statistical applications involving components of the model that cannot be written down analytically is likely to increase in the coming years. There is much work to be done to better understand the computational and inferential implications of such models, and we hope the strategies and methods explored in this article can contribute to further research in this area.

Acknowledgments We dedicate this article to the memory of Stephen Fienberg, a stellar statistician and academic leader, who spent many decades unveiling the true colors of many black boxes in and outside of statistics.

We thank David van Dyk, Aneta Siemiginowska, and other members of CHASC for many helpful conversations and suggestions. PB and XLM acknowledge partial support from multiple NSF grants, and VLK from the NASA Contract to the Chandra X-ray Center NAS8-03060. AZ acknowledges support from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement n. 617001 (A-BINGOS). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie RISE action, grant agreement No. 691164 (ASTROSTAT).

References

1. J. Anderson, G. Piotto, I. R. King, L. R. Bedin, and P. Guhathakurta. Mixed populations in globular clusters: Et Tu, 47 Tuc? *The Astrophysical Journal*, 697:58–62, 2009.
2. J. Barnard, R. McCulloch, and X.-L. Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10:1281–1311, 2000.
3. M. J. Bayarri, J. O. Berger, J. A. Cafeo, G. Garcia-Donato, F. Liu, J. Palomo, R. J. Parthasarathy, R. Paulo, J. Sacks, and D Walsh. Computer model validation with functional output. *Annals of Statistics*, 35:1874–1906, 2007.
4. M. J. Bayarri, J. O. Berger, R. Paulo, J. Tu, J. Sacks, J. A. Cafeo, J. Cavendish, and C. H. Lin. A framework for validation of computer models. *Technometrics*, 49:138–154, 2007.
5. S. R. Cook, A. Gelman, and D. B. Rubin. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692, 2006.
6. A. Dotter, B. Chaboyer, D. Jevremovic, V. Kostov, E. Baron, and J. W. Ferguson. The Dartmouth stellar evolution database. *The Astrophysical Journal Supplement Series*, **178**:89–101, 2008.
7. D. S. Evans. *Lacaille: astronomer, traveller; with a new translation of his journal*. Pachart, 1992.
8. E. D. Feigelson. Statistics in astronomy. *arXiv*, [astro-ph.IM](0903.0416v1):1–12, 2008.
9. C. J. Geyer. Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163, 1991.

10. C. O. Heinke, J. E. Grindlay, P. D. Edmonds, H. N. Cohn, P. M. Lugger, F. Camilo, S. Bogdanov, and P. C. Freire. A deep CHANDRA survey of the globular cluster 47 Tucanae: Catalog of point sources. *The Astrophysical Journal*, 625:796–824, 2005.
11. S. C. Kou, Q. Zhou, and W. H. Wong. Equi-energy sampler with applications in statistical inference and statistical mechanics. *Annals of Statistics*, **34**-4:1581–1619, 2006.
12. E. Lawrence, D. Bingham, C. Liu, and V. N. N. Nair. Bayesian inference for ordinal data using multivariate probit models. Technical report, Misc, 2006.
13. Thibault Lejeune and Daniel Schaerer. Database of Geneva stellar evolution tracks and isochrones for $(UBV)_J(RI)_C$JHKLL'M, HST-WFPC2, Geneva and Washington photometric systems. *Astronomy & Astrophysics*, 366(2):538–546, 2001.
14. P. Marigo, L. Girardi, A. Bressan, M. A. T. Groenewegen, L. Silva, and G. L. Granato. Evolution of asymptotic giant branch stars. II. Optical to far-infrared isochrones with improved TP-AGB models. *Astronomy and Astrophysics*, **482**:883–905, 2008.
15. X.-L. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical explanation. *Statistica Sinica*, 6:831–860, 1996.
16. R. J. Renka. Algorithm 751: TRIPACK: a constrained two-dimensional Delaunay triangulation package. *ACM Transactions on Mathematical Software*, 22(1):1–8, 1996.
17. J. Rougier, S. Guillas, and A. Maute. Expert knowledge and mulitvariate emulation: The thermosphere-ionosphere electrodynamics general circulation model (TIE-GCM). *Technometrics*, 51(4):414–424, Nov 2009.
18. E. E. Salpeter. The luminosity function and stellar evolution. *The Astrophysical Journal*, 121:161–167, 1955.
19. L. Sbordone, M. Limongi, A. Chieffi, E. Caffau, H.-G. Ludwig, and P. Bonafacio. Sulfur in the globular clusters 47 Tucanae and NGC 6752. *Astronomy and Astrophysics*, 503:121–127, 2009.
20. C. M. Schafer and P. B. Stark. Constructing confidence regions of optimal expected size. *Journal of the American Statistical Association*, 104:1080–1089, 2009.
21. D. A. van Dyk, S. DeGennaro, N. Stein, W. H. Jefferys, and T. von Hipple. Statistical analysis of stellar evolution. *Annals of Applied Statistics*, **3**:117–143, 2009.
22. Y. Yu and X.-L. Meng. To center or not to center: That is not the question – an ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency (with discussion). *Journal of Computational and Graphical Statistics*, 20(3):531–570, 2010.

Chapter 5
The Ways of Bayes (In Memory of Stephen Fienberg)

Thomas A. Louis

5.1 Reminiscence

Steve Fienberg and I were friends and colleagues, from the early 1980s through his death. He was the consummate academic, interested in ideas, concepts and principles, thirsting to teach, and yes, to "profess." But more importantly for science, policy, and public well being, he was action-oriented, applying his skills to important public policy issues, promoting the Bayesian viewpoint (see Fienberg 2011), establishing or helping to establish institutions that activate statistics and statisticians. One of the most recent of these was Steve's creation of the *Annual Review of Statistics and its Application (ARSIA)*.

Steve was a Bayesian with encyclopedic knowledge. He had strong opinions, but liked nothing better than to argue it out, to "disagree agreeably" (approximate quote of Daniel Patrick Moynihan, former U.S. Senator). I recall many enjoyable occasions, including gourmet dining in Leiden NL in 1996 (he was a true gourmet), floating in the Mediterranean off Hersonissos-Heraklion, Crete in 2000 (discussing Bayes and life), participating in the closing session of the International Year of Statistics in 2013, and serving together on the COPSS-CMS profiling project and on the *ARSIA* editorial board. I miss him, we all miss him, statistics and society miss him. All have benefitted from his contributions and example.

T. A. Louis (✉)
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
e-mail: tlouis@jhu.edu
https://www.biostat.jhsph.edu/~tlouis/

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_5

5.2 Introduction

The Bayesian approach to design and analysis is making important contributions to science and policy. While much of the Bayesian advantage comes from success in dealing with highly complex situations, there are also benefits in basic settings. If modeling assumptions are not reasonably correct, an approach, Bayes or otherwise, can perform poorly; and Bayesian modeling relies on additional assumptions. However, as communicated in subsequent sections, its potential benefits are substantial, and in many situations it is worth investing the additional effort to "go Bayes." There are many benefits including: the approach makes assumptions explicit, the laws of probability guide the progression from data to evidence, the approach supports flexible processing to address basic and complex goals, it carries forward all uncertainties, and the posterior distribution can be used to compute expected utilities. Availability of robust Bayesian models, including semi-parametric approaches, and efficient computational approaches broaden the domain of applications and add to the attractiveness of the approach.

I present theory and examples ranging from basic to advanced, showing that Bayesian approaches can confer substantial benefits, but that care is needed.

Thomas Bayes

The Reverend Thomas Bayes 1701–1761, started the whole thing off with his, "…probability that an event has to happen in given circumstances …" (see Fig. 5.1

Fig. 5.1 Steve converting a then-frequentist Rafa Irizarry at Thomas Bayes' grave in the nonconformist section of Bunhill Fields Burial Ground, London (a short walk from the Royal Statistical Society's office)

for Steve "converting" a then-frequentist Rafa Irizarry at Bayes' grave during a lunch break at the wrap-up meeting of the International Year of Statistics, held at the Royal Statistical Society, November 2013).

5.3 Bayes and Frequentist

Why Is Not Everyone a Bayesian?

In Efron (1986) (a must-read) Brad asked, "Why isn't everyone a Bayesian?"[1] He notes that Laplace, Gauss, and others endorsed the Bayesian formulation, but that the twentieth century was frequentist with R. A. Fisher as a leading proponent. Fisher's theory was attractive in part because the overhead relative to Bayesian approaches was low, and methods were relatively straightforward to implement. The frequentist approach solved a specific problem, for example a confidence interval for the population median, without requiring a comprehensive model, which is a far more ambitious goal. The approach was considered "objective," an apparently attractive feature. Quoting Efron,

> "A prime requirement for any statistical theory intended for scientific use is that it reassures oneself *and others* that the data have been interpreted fairly."

A worthy goal, irrespective of your statistical philosophy, but as Lindley states in his discussion:

> "The objective element is the data: interpretation of the data is subjective, ..."

The approach is not without controversy; as Sir Maurice Kendall noted in discussing the use of prior beliefs,

> "If they [Bayesians] would only do as he [Bayes] did and publish posthumously, we should all be saved a lot of trouble."

Many would agree that the Bayesian approach is well-adapted to procedure generation, either to attain frequentist goals (Bayes for frequentist) or Bayesian goals (Bayes for Bayes). This advantage is especially true in complicated settings (e.g., synthesizing data sources), complicated designs (e.g., adaptive protocols), or for complicated goals (e.g., optimizing a non-linear utility). However, irrespective of the way a design and analysis are generated, understanding frequentist properties, including model adequacy, is important (see Box 1980, for the frequentist aspects of a fully Bayesian model).

[1] Also, see Box (1980) for another pivotal article with illuminating discussion.

A Few Comparisons

We provide a few comparisons between Bayesian and frequentist concepts and models with no attempt to be comprehensive.

The (Apparent) Simplicity of Null Hypothesis Significance Testing

Significance testing confers an apparent frequentist advantage, because, as Efron (2005) states, you need only to assume something about the null model. However, it is the case that the test statistic implies assumptions about the underlying statistical model and the alternatives of primary interest (those for which one wants good power). So, in reality, the approach is not all that simple or objective. This point segues into the p-value swamp, which I avoid by citing Wasserstein and Lazar (2016) and Lazar et al. (2019) for discussion of p-values and beyond.

Full Distribution Processing

Full distribution processing is a clear Bayes advantage. It is essential for Bayesians and very effective for frequentists. The posterior distribution consolidates all evidence and can be used for inference and prediction. Full distribution thinking encourages tuning an inference to specific goals rather than using defaults such as the (posterior) mean and variance. This ability is especially, but not uniquely, beneficial in addressing complex and non-standard goals such as estimating ranks, and optimizing complex utilities.

Bayesian processing guided by the laws of probability carries forward full (posterior) distributions in making inferences, a clear Bayesian advantage. Examples include developing a missing data prediction model that incorporates uncertainty in estimated parameters. In using data to provide information about a prior distribution (e.g., a between-unit variance component), Bayes empirical Bayes transmits uncertainties to the posterior distribution. In a context wherein exposures are measured with error, rather than computing exposure/response by operating on a point estimate of exposure, the full posterior distribution of exposure can be processed by the function. Doing so can change the shape of the (measured exposure)/response curve, and produce more appropriate risk assessments.

My use of “infer” rather than “estimate” respects that Bayesian output is a joint, posterior distribution, not a point estimate, and so it delivers uncertainties to the ultimate inference. Of course, only uncertainties identified by the model will be carried forward. For example, if model selection is not incorporated in the data model (incorporation is challenging!), related uncertainties will not be delivered to the posterior distribution. The bootstrap can be used to deliver these uncertainties, and the Bayesian bootstrap (sample from the posterior of parameters, then generate bootstrap data) produces a more “honest” assessment.

Computational Complexity

The additional computational complexity of the Bayesian approach is a decreasing advantage of the frequentist approach. Software such as `rstan`, `BUGS`, and `JAGS` have dramatically reduced the computational challenges, and appropriate frequentist computations are by no means trivial.

Objectivity/Subjectivity, Opinion/Judgment

A great deal has been written on objectivity and subjectivity, on opinion and judgment with Brownstein et al. (2019) providing a recent discussion. Pure objectivity is difficult or impossible to obtain, and statistical studies depend on opinion and judgment in each study phase (conceptualization, design, implementation, and analysis). Subjectivity and judgment are *prima facie* necessary in the design phase; indeed we are all at least informal Bayesians when designing a study. All evaluations are preposterior, integrating over both the data (a frequentist act) and the parameters (a Bayesian act). And, all analysis have subjective elements, including the form of the data model, choice of candidate regressors, of course guided by statistical best practices. Each phase should be conducted in a scientific and reproducible manner; conduct and analysis should be protocol-based, with the analysis protocol allowing/inviting adaption and exploration that go beyond pre-specification. Similarly, as emphasized by O'Hagan (2019), elicitation of expert opinion, including which studies are relevant to the design of the current one, should be scientific, protocol-based, and documented.

Dependence on the Sampling Plan

Some sampling aspects can be ignored in a Bayesian analysis (e.g., stopping rules), others cannot or should not (e.g., survey weights). For example, the posterior distribution does not depend on the stopping rule, the equivalence of Binomial/(Negative Binomial) posterior distributions is a popular example. However, some frequentist and some Bayesian properties of a procedure will depend on the stopping rule, and stopping rules can degrade Bayesian robustness to departures from assumptions (see Rubin 1984, for examples).

A Bayesian is not completely relieved of considering sampling, for example failure to accommodate length-biased sampling will produce biased estimates. In the survey world, Bayesians need to accommodate sampling weights, with Little's design-consistent Bayes (Little 2004, 2012) an attractive approach. And, as discussed in Sect. 5.6, sample size can be both informative and endogenous.

Modeling Flexibility

Modeling flexibility is a clear Bayesian advantage. For example, consider estimating K, unit-specific parameters $(\theta_1, \ldots, \theta_K)$ (e.g., treatment-specific means in an ANOVA, or small-area true disease rates). In the absence of regressors, the "pure" frequentist approach operates at two extremes. Either each of the θs stands on its own (and unit-specific statistical likelihoods provide the information), or all θs are assumed equal in the null model, $\{\theta_1 = \ldots = \theta_K = \theta\}$ (and a single-θ likelihood provides the information).

There are connections with the Bayesian formulation. "Every unit on its own" can be represented by θ-specific priors with no connection among them, and the null model by a single prior on the common θ. Importantly, the Bayesian approach adds considerable flexibility to these extremes by supporting the middle-ground model wherein the unit-specific θ_k are unique, but they are related (are siblings) in that they all come from the same prior distribution.

More generally, the prior can include a regression structure, with model lack of fit (unexplained variation) quantified by the prior variance. The Bayesian formulation substantially expands modeling flexibility, including the hierarchy of hypotheses,

- H_{00}: Prior variance is 0, all unit-specific parameters (residuals) are equal to the prior mean.
- H_0 or H_a: All unit-specific parameters (residuals) come from the same, non-degenerate distribution and so are different, but are related.
- H_{aa}: None of the above ("pure" frequentist).

Bayes and Multiplicity

The prior to posterior mapping does not "know about" multiple comparisons. With additive, component-specific losses, each comparison is conducted separately with no accounting for the number of comparisons (see Sect. 5.3 for an example of non-additive loss). There is indirect attention to multiplicity in that the analyst can choose the number of units (e.g., treatments in an ANOVA) under study. Also, and importantly, shrinkage toward a common mean (or more generally, a regression surface) in a Bayes or empirical Bayes analysis calms multiplicity.

The following, basic example is related to the K-ratio procedure (see Brant et al. 1992), additional examples are provided by Scott and Berger (2006). Using the standard, Gaussian model,

$$\begin{aligned}
\theta_1, \ldots, \theta_K \;& iid \; N(\mu, \tau^2) \\
[Y_{ik} \mid \theta_k] \;& ind \; N(\theta_k, \sigma^2) \\
[\theta_k \mid Y_{.k}] \;& \sim \; N\left(\mu + (1-B)(Y_{.k} - \mu), (1-B)\sigma^2\right) \\
B \;& = \; \frac{\sigma^2}{\sigma^2 + \tau^2},
\end{aligned}$$

compare treatments 1 and 2 via the posterior distribution test statistic,

$$Z^{Bayes} = \frac{(1-B)(Y_{.1}-Y_{.2})}{\sqrt{2(1-B)\sigma^2}} = \frac{(1-B)^{\frac{1}{2}}(Y_{.1}-Y_{.2})}{\sqrt{2}\sigma}$$

$$= (1-B)^{\frac{1}{2}} Z^{freq} = \left\{\frac{(F-1)^+}{F}\right\}^{\frac{1}{2}} Z^{freq},$$

where

$$F = 1/\hat{B} = (\hat{\sigma}^2 + \hat{\tau}^2)/\hat{\sigma}^2$$

is the F-statistic for testing the null hypothesis that all treatment means are equal. The frequentist Z-score is attenuated by a function of the F-statistic; the closer F is to 1.0, the greater the attenuation. Scoping is important because the number and relations among the units in the analysis determine the value of $\hat{\mu}$ and $\hat{B}$.

If the global $H_0: \theta_1 = \theta_2 = \ldots = \theta_K$ (equivalently $\tau^2 = 0, B = 1$) is true, the overall type I error for an ensemble of contrasts is controlled because F will tend to be close to 1.0, and all Z-scores will be substantially attenuated. In fact, for large K, under the global null hypothesis $pr(F \leq 1.0) \approx 0.5$ and so pr(all $Z_{ij} = 0) \approx 0.5$. The family-wise rejection rate is much smaller than 0.5 because rejection requires that an attenuated Z-score exceeds a non-zero testing threshold.

Table 5.1 reports the type I error for a single, one-sided test and for a set of $(K-1)$ independent comparisons for nominal level $\alpha = 0.05$. Even for $K = 5$, type I error control is impressive and converges to complete control with increasing K (the non-monotonicity over the first three rows in the "$K-1$ contrasts" column appears to be correct and not a computational issue). Similar computations show that when $B < 1.0$ ($\tau > 0$), statistical power is maintained, though of course not to the level produced by non-attenuated test statistics.

Table 5.1 Type I error for a one-sided test: for a single comparison, for $(K-1)$ independent comparisons, and the probability that pr($\hat{B} = 1$) (equivalently $\hat{\tau}^2 = 0$). Values are computed by numerical integration with 50,000 grid points set at the quantiles of the appropriate chi-square distribution

K	Single test	$K-1$ contrasts	pr($\hat{B} = 1$) $\times 100$
5	0.00224	0.00893	59.4
10	0.00116	0.01038	56.3
20	0.00050	0.00943	54.3
30	0.00028	0.00796	53.5
50	0.00012	0.00562	52.7
100	0.00003	0.00267	51.9
500	0.00000	0.00009	50.8
1000	0.00000	0.00001	50.6

Non-additive Loss

As in Carlin and Louis (2009), consider a situation wherein making two errors incurs a penalty greater than the sum of the coordinate-specific losses. For example, set the loss equal to 1.0 for a single error, but a loss greater than 2.0 for two errors, specifically,

$$\begin{aligned}
\text{Parameters:} \quad & \theta_1, \theta_2 \in \{0, 1\} \\
\text{Probabilities:} \quad & \pi_{ij} = pr[\theta_1 = i,\ \theta_2 = j] \\
\text{Decisions:} \quad & a_1, a_2 \in \{0,\ 1\} \\
\text{Loss}(\mathbf{a},\ \theta): \quad & a_1(1-\theta_1) + (1-a_1)\theta_1 \\
& + a_2(1-\theta_2) + (1-a_2)\theta_2 \\
& + \boxed{\gamma(1-\theta_1)(1-\theta_2)a_1a_2}
\end{aligned}$$

The optimal decision rule is,

$$\begin{aligned}
\pi_{1+} \le .5, \pi_{+1} \le .5 \quad & a_1 = 0, a_2 = 0 \\
\pi_{1+} \le .5, \pi_{+1} > .5 \quad & a_1 = 0, a_2 = 1 \\
\pi_{1+} > .5, \pi_{+1} \le .5 \quad & a_1 = 1, a_2 = 0 \\
\pi_{1+} > \pi_{+1} > .5 \quad & a_1 = 1 \\
& a_2 = \begin{cases} 0, \text{ if } (2\pi_{+1} - 1) < \gamma\pi_{00} \\ 1, \text{ if } (2\pi_{+1} - 1) \ge \gamma\pi_{00}. \end{cases}
\end{aligned}$$

Frequentist, Bayes, Empirical Bayes, or Bayes Empirical Bayes?

Making inferences solely based on information from a single study/dataset without incorporating external evidence or professional judgment is *prima facie* frequentist. However, as Brownstein et al. (2019) propose, this pure form is difficult if not impossible to achieve. All other inferential activities entail some degree of formal or informal Bayesian evaluation. Compound sampling allows replacing some degree of personal opinion or external data with within-study data. If the number of relevant data sources is large and provides a large amount of information on the prior distribution, plug-in empirical Bayes (EB) performs well. Many genomics examples occupy this domain, and flexible priors are available including using a smoothed, non-parametric, plug-in prior (see Lin et al. 2006). On the other hand, with only one "draw" from the prior, if there is to be Bayes, it needs to be high-church Bayes, driven by personal/expert judgment (see O'Hagan 2019). If the number of relevant

data sources is moderate, then accommodating uncertainty in the inferred prior and posterior distributions via an expansion (Morris 1983), the bootstrap, or hyper-prior Bayes (BayesEB, BEB) is needed. Evaluating whether plug-in is sufficient usually requires comparing it to an approach that brings in uncertainty in estimating the prior, and if it has been implemented, then why not use it?

Example: Exponential/Inverse-Gamma

Carlin and Louis (2009) evaluated preposterior CI length and coverage for the exponential/inverse-gamma model with θ the rate/hazard,

$$\begin{aligned} \theta_1, \ldots, \theta_N &\overset{\text{iid}}{\sim} \text{InvGamma}(\eta, 1) \\ f(y_i \mid \theta_i) &= \frac{1}{\theta_i} e^{-y_i/\theta_i}, \; y_i > 0 \\ \hat{\theta}_i^{mle} &= y_i, \end{aligned}$$

which produces the marginal distribution and marginal mle,

$$m(y_i \mid \eta) = \eta/(y_i + 1)^{\eta+1}, \quad \hat{\eta}^{mmle} = N/\sum_{i=1}^{N} \log(y_i + 1).$$

They compared the frequentist CI to naive EB (plug-in $\hat{\eta}^{mmle}$), Laird and Louis (1987) bootstrap, Carlin and Gelfand (1991) hyper-prior matching using $h_1(\eta) = 1$ and $h_2(\eta) = 1/\eta$, when $N = 5$ and the true $\eta = 2$.

Table 5.2 gives a snippet of their results. The classical interval is well-calibrated, but extremely long relative to all other methods. The naive EB interval substantially under-covers; the Laird/Louis bootstrap and the h_2-based intervals are well-calibrated, each with length substantially shorter than the classical. The h_2-based interval is shorter, but the Laird/Louis does not require specifying h. That the h_1-based interval under-covers shows that for small N situations, choice of hyper-prior matters and care is needed.

Table 5.2 Comparison of simulated unconditional, nominal 95% EB CI length and coverage for the exponential/inverse-gamma model. See Section 5.4.3 and Table 5.4 in Carlin and Louis (2009) for full details

	Method				
Feature ↓	Classical	Naive EB	Laird/Louis	h_1	h_2
Length	38.80	5.22	7.50	4.51	5.66
Coverage	0.952	0.900	0.954	0.930	0.951

Robust Bayes

Chen et al. (2015), Jiang et al. (2011) develop "best predictive small area estimation" that robustifies a core small area estimation model. The idea is as follows. The posterior mean (PM_k) for unit k in a Gaussian/Gaussian model is,

$$\begin{aligned} \text{PM}_k &= B_k \mathbf{X}_k \hat{\beta} + (1 - B_k) Y_k \\ B_k &= \sigma_k^2 / (\sigma_k^2 + \tau^2) \end{aligned}$$

a weighted average of the direct estimate (Y_k) and a regression prediction ($\mathbf{X}_k \beta$) with larger B_k for the relatively unstable direct estimates. On the other hand, $\hat{\beta}$ gives more weight to the units with relatively stable direct estimates and so the high B_k units that "care about" the regression model have less influence. This is fine if the model is correctly specified ($\hat{\beta}$ will be minimum variance and, approximately, unbiased), but if the model is mis-specified, PM_k will be unfair to the high variance (large B_k units); the deck may be stacked against them (see Sect. 5.6 for an example where this unfairness may be in play). The Chen et al. (2015), Jiang et al. (2011) approach increases the weight on relatively unstable direct estimates when estimating β and increases variance, but in many contexts improves MSE. Henderson et al. (2020) expand on this approach, adding a bit more variances, but also increasing robustness. Enriched covariates, including a flexible function of sample size, can make sample size uninformative (see Little 2012) and offers another approach to "robustify" an analysis.

Why Is Not Everyone a Bayesian → Why Not Bayes?

There are risks associated with any design or analysis, but carefully developed and applied, the Bayesian formalism can be very effective for both Bayes and frequentist goals. Guided by the laws of probability, it can handle complex models with careful accounting of uncertainty. As Sect. 5.3 indicates, hybrid approaches with a Bayesian centerpiece that is robustified by philosophy-free adjustments are very effective. While the Efron (1986) question is still germane, for me the burden of proof has switched to "why not Bayes?" or at least "why not eclectic Bayes?"

Purity and Its Discontents

Full Bayes provides a documentary trail, enhancing reproducible research. However, not everything can be formalized and in practice nothing is "pure" anything, so designers and analysts must strike trade-offs, and being eclectic is (almost) always necessary. Indeed, while pure Bayes pairs nicely with Port, when you leave port

and are on the high seas of applications, some degree of impurity is almost always necessary. Consequently, Bayesians who engage in important studies use the paradigm as an (possibly the) aid to navigation, not as a straightjacket. The goal is to do a good job, and one cannot be (too) doctrinaire. Steve most certainly operated in this manner.

5.4 Diagnostics

Bayesian diagnostics rely on frequentist concepts and computations; the usual diagnostics (outlier identification, residual plots, Q-Q plots) are still the lynchpins. Box (1980) notes that the marginal distribution of observed data (a predictive distribution) is model assessment-relevant. Bérubé et al. (2019) note that the commonly used standardized residuals, Residual = (Observed − Expected)/SD, where the Expecteds and SDs come from data point-specific predictive distributions, may be deceptive. Irrespective of model form, if the predictive distributions are well-calibrated, these residuals are mean 0, variance 1. However, unless their distribution is close to Gaussian, they would not be $N(0, 1)$. In this case, outlier detection and model assessments can perform poorly, with poorly calibrated type I error for outlier detection or low power, and misleading residual plots. Therefore, a standardized residual that is not tied to normality has the potential to improve performance, and percentile-based residuals have the potential to provide that improvement.

Let $(\mathbf{Y}_k, \mathbf{X}_k)$ represent all direct data (dependent variable, covariates) for the k^{th} sampling unit, and $(\mathbf{Y}, \mathbf{X})$ all data. We focus on a scalar Y_k, which can be a unit-specific summary statistic. The analyst produces a working model, $[Y_k \mid \mathbf{X}_k, \psi]_{\texttt{xwkng}}$ with covariates $\mathbf{X}_k$ and parameters ψ (all parameters; slopes, variances, variance components, etc.). Examples include linear and logistic regression, CART, random forests, and other machine-learning approaches (for these ψ represents the underlying algorithm's end result). Data analysis produces the working predictive cumulative distribution,

$$D_k(Y_k) = D_k(Y_k \mid \mathbf{X}_k, \text{Analysis}),$$

but the true predictive distribution is,

$$F_k(Y_k) = F(Y_k \mid \mathbf{X}_k, \text{Analysis}).$$

The term "analysis" is quite general. It can mean production of the full posterior distribution of ψ to be used to generate an in or out of sample predictive distribution for Y_k (e.g., the collection of MCMC samples) can be a basic, plug-in approach substituting $\hat{\psi}$ for ψ with no attention to uncertainty in the estimate (consequently, $D_k \neq F_k$), or the end result of a machine-learning algorithm, with or without infusion of uncertainty.

The (O − E)/SD, Standardized Residuals

With $Y_k = y_k$, the (Observed − Expected)/SD standardized residuals are,

$$R_k^* = \frac{y_k - \tilde{\mu}_k}{\tilde{\sigma}_k} \tag{5.1}$$
$$\tilde{\mu}_k = E_D(Y_k \mid \mathbf{X}_k, \text{Analysis})$$
$$\tilde{\sigma}_k^2 = V_D(Y_k \mid \mathbf{X}_k, \text{Analysis}).$$

For model criticism, the empirical distribution of the R_k^* is evaluated relative to the $N(0, 1)$ distribution (e.g., outlier identification, distributional shape); also, plotting R_k^* versus $\tilde{\mu}_k$ can identify the need for model enhancement. If $(\tilde{\mu}_k, \tilde{\sigma}_k) = (\mu_k, \sigma_k)$ (the true values associated with F_k), then the R_k^* have mean 0 and variance 1, but the full distribution can be far from Gaussian unless the F_k are Gaussian or close to it via the central limit theorem (CLT).

Percentile-Based Residuals

To relax dependence of performance on the CLT, consider finding the percentile location of $Y_k = y_k$ in the working predictive distribution D_k and mapping it to the associated quantile of a $N(0, 1)$ distribution. With $Y_k = y_k$ define,

$$R_k^{\ddagger} = \Phi^{-1}\left\{D_k(y_k) - 0.5\text{pr}_D(Y_k = y_k)\right\}. \tag{5.2}$$
$$= \Phi^{-1}\left\{D_k(y_k)\right\} \quad \text{(for continuous } D_k\text{)}.$$

If $D_k = F_k$, irrespective of the form of F_k, $R^{\ddagger} \sim N(0, 1)$. The one-half correction, termed "medialized" by Good (2007) is needed to balance the assessment for a discrete distribution. For example, if D_k puts all mass at a single point and y_k is that point, the uncorrected $R_k^{\ddagger} = \infty$; the corrected (and correct) $R_k^{\ddagger} = 0$. If the direct estimate is equal to the largest value of the predictive distribution, the correction brings $R^{\ddagger}$ from infinity to a finite value. Even for a continuous D_k, either R^* or $R^{\ddagger}$ can be $\pm\infty$, for example if the observed value is beyond the support of the predictive distribution. So, in practice truncate the values at, for example ± 5.0.

5.5 Addressing Non-standard Goals

Bayesian procedure generation is especially effective in addressing non-standard goals, including determining if a parameter is in a complicated, multivariate region (see Sect. 5.5); bioequivalence and non-inferiority relations in clinical trials, esti-

mating ranks and histograms; optimizing non-linear loss functions (see Henderson and Newton 2016; Lin et al. 2006; Paddock et al. 2006; Paddock and Louis 2011, for examples of ranking and histogram/cdf estimation).

Determining If a Parameter Is in a Complex Region ($\mathcal{R}$)

Section 203 of the voting rights act provides an excellent example of determining if a vector parameter is in a complex region. It mandates that a state or political subdivision must provide language assistance to voters,

> If more than 5% of voting age citizens are members of a single language minority group, **and** do not, "speak or understand English adequately enough to participate in the electoral process," **and** if the rate of those citizens who have not completed the fifth grade is higher than the national rate of voting age citizens who have not completed the fifth grade. A political subdivision is **also covered**, if more than 10,000 of the voting age citizens are members of a single language minority group, do not "speak or understand English adequately enough to participate in the electoral process," **and** the rate of those citizens who have not completed the fifth grade is higher than the national rate of voting age citizens who have not completed the fifth grade.

To make this determination, Joyce et al. (2014) and Ashmead and Slud (2017) use Bayes and empirical Bayes methods to stabilize relevant estimates. A worthy alternative is to compute $pr(\theta \in \mathcal{R} \mid \text{data})$ and use a value (e.g., 0.5) as the provide/(no need to provide) decision threshold.

5.6 The COPSS-CMS White Paper

Since 2008, the Center for Medicare and Medicaid Services (CMS) has been publicly reporting condition-specific outcome measures such as risk-adjusted all-cause mortality and readmission measures for Acute MI (AMI), Heart Failure (HF), and Pneumonia. CMS uses random-effects, hierarchical modeling to produce case-mix risk-adjusted hospital-specific measures that compare a hospital's performance to a national standard, and has received considerable push-back from stakeholders as well as research communities regarding use of this methodology. Principal critiques are that the CMS approach,

1. Fails to reveal provider performance variation because the shrinkage effect of hierarchical modeling reduces variation of hospital performance.
2. Masks the performance of small hospitals due to the extreme shrinkage toward the national mean.
3. Generally, there was (and still is) a considerable lack of consensus regarding whether to use hierarchical models, to use a random-effects approach, to stabilize by shrinkage, to use hospital attributes to adjust the risk model and, separately, to set shrinkage targets, and to use procedure volume as a covariate.

To address these concerns the CMS supported a committee appointed by the Committee of Presidents of Statistical Societies (COPSS) to address statistical issues associated with the foregoing criticisms of CMS's approach to modeling hospital quality based on mortality outcomes. Committee members represented the COPSS societies (Arlene Ash, ASA; Stephen Fienberg, IMS; Sharon-Lise Normand, ENAR; Thérèse Stukel, SSC; Jessica Utts, WNAR; Thomas Louis, at large and chair). The review was coordinated by Xihong Lin, and addressed the charge,

> "... provide guidance on statistical approaches ... when estimating performance metrics," and "consider and discuss concerns commonly raised by stakeholders [...] about the use of "hierarchical generalized linear models" in profiling hospital quality."

The committee report (Ash et al. 2012) detailed our evaluations and recommendations; a subsequent article (Normand et al. 2016) is based on the report.

Goal Identification

As a precursor to addressing the technical aspects, there was confusion/disagreement (broadly in the stakeholder community and among committee members) regarding the overall goal of statistical modeling for hospital comparison. Should it be *profiling*,

> "How does this hospital's mortality for a particular procedure compare to that predicted at the national level for the kinds of patients seen for that procedure or condition at this hospital?"

Or, should it be *individual decision-making*,

> "Given my medical status and needs, to which hospital should I go for a particular procedure or treatment?"

The choice between the two matters because, yes, modeling depends on goals. To address the individual decision-making goal (a worthy goal, directly relevant to the individual patient), build a rich model for the probability of death (or other endpoint) that includes patient-specific and hospital-specific characteristics. However, for profiling, hospital characteristics must be omitted from the model, with the resulting hospital effects used for comparisons. That is, profiling depends on an explicit comparison between the target hospital and a counterfactual hospital with the same case-mix that operates at the national quality level, but it is very unlikely that there is another hospital with exactly the same case-mix. It is relevant for quality evaluations, for policy decisions. Resolving this goal issue took considerable discussion, with Steve's input essential for clarifying the issues and coming to consensus.

The profiling goal depends on the Standardized Mortality Ratio (SMR), equivalently, the indirectly standardized, 30-day death rate (SDR). Specifically, for hospital "i'' and each admission condition under consideration, the inferential target is,

$$\text{SMR}_{\mathbf{i}} = \frac{\text{E}_{\mathbf{i}}(\#\text{of deaths}|\text{case-mix})}{\text{E}_{\mathbf{i,national}}(\#\text{of deaths}|\text{same case-mix})},$$

or

$$\text{SDR}_{\text{i}} = \text{SMR}_i \times (\text{the national-level death rate}).$$

We focus on the SMR, with SMR = 1.0 indicating typical performance relative to the national standard for the distribution of patients treated at the hospital. $\text{E}_{\mathbf{i,national}}$ is produced by summing up national-level probabilities that depend on an ensemble of patient-level, pre-admission risk factors for the patients treated in hospital i. $\text{E}_{\mathbf{i}}$ is produced by summing up national-level probabilities that depend on an ensemble of patient-level, pre-admission risk factors for the patients treated in hospital i, and a hospital-specific offset (hospital effect).

A direct estimate,

$$\widehat{SMR}_i^{direct} = \frac{\text{Observed}(\#\text{ of deaths in hospital } i)}{\text{E}i, national(\#\text{of deaths} \mid \text{same case-mix})}$$

is attractive because it requires statistical modeling only for the denominator (Kalbfleisch and Wolfe 2013, advocate this approach), but for many hospitals, especially those with low volume, the estimate will be very noisy and a stabilized estimate,

$$\widehat{SMR}_i^{stabilized} = \frac{\text{Stabilized}(\#\text{ of deaths in hospital } i)}{\text{E}_{i,national}(\#\text{ of deaths}|\text{same case-mix})}$$

is used by the CMS. Stabilization is based on a random-effects, logistic regression model, with stabilization via Bayesian shrinkage of hospital-specific offsets toward 0, consequently shrinking $\widehat{SMR}_i^{direct}$ toward 1.0 with the degree of shrinkage greater for the less stable direct estimates. Much of the controversy relates to this issue, especially the degree of shrinkage for low volume hospitals.

Candidate covariates to the risk model ($\text{E}_{national}$):

Definite: *Pre-admission, patient-level health attributes (i.e., case-mix):* To produce a fair assessment, predictions must account for patient-level, upon-admission characteristics that associate with outcome.

Never: *Post-admission patient attributes including events that coincide with or might be the result of care (e.g., in-hospital infections or patient length-of-stay):* Including this information would adjust-away part of the hospital effect. Features that are on the causal pathway from admission to outcome should not influence the risk adjustment.

Possibly: *Pre- or at-admission, hospital attributes (i.e., presence of cardiac catheterization laboratories):* These can be used to reduce confounding by hospital (cross-level confounding) in the estimated relation between patient attributes and the probability of death, but should not influence shrinkage targets.

The CMS Model and a Generalization

Let $Y_{ij} = 1$ or 0 to indicate whether the event did/(did not) occur for the jth patient treated at the ith hospital, with $\mathbf{x}_{ij}$ a vector of patient-specific characteristics, n_i the number of cases treated at hospital i, and $\mathbf{z}_i$ a vector of hospital-level attributes that are to be used to develop shrinkage targets for the numerator of the SMR. For $i = 1, 2, \cdots, I$,

$$[Y_{ij} \mid \beta_{0i}, \alpha, \mathbf{x}_{ij}] \overset{ind}{\sim} \text{Bern}(p_{ij}) \text{ where } \text{logit}(p_{ij}) = \beta_{0i} + \alpha \mathbf{x}_{ij}$$

$$[\beta_{0i} \mid \mu, \tau^2, \gamma, \mathbf{z_i}] \overset{iid}{\sim} N(\mu + \gamma \mathbf{z_i}, \tau^2).$$

$$SMR_i = \frac{\sum_{j=1}^{n_i} E(Y_{ij} \mid \beta_{0i} + \alpha \mathbf{x}_{ij} + \mu + \gamma \mathbf{z_i}, \tau^2)}{\sum_{j=1}^{n_i} E(Y_{ij} \mid \beta_{0i} + \alpha \mathbf{x}_{ij} + \mu + \gamma \mathbf{z}^*, \tau^2)},$$

with $\mathbf{z}^*$ chosen to satisfy,

$$\bar{Y} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} E(Y_{ij} \mid \beta_{0i} + \alpha \mathbf{x}_{ij} + \mu + \gamma \mathbf{z}^*, \tau^2)$$

$$\bar{Y} = \frac{\sum_{i,j} Y_{ij}}{\sum_i n_i} \text{ (the overall event rate).}$$

The hospital-specific, estimated SMR is based on a counterfactual population of hospitals having the same case-mix as the target hospital ($\mathbf{x}_{ij}$, $j = 1, \ldots, n_i$) with risk effects quantified by the national average. The expectation in the numerator integrates over the posterior distribution of β_{0i} (inducing shrinkage toward 0 from the directly estimated values). Pair-wise comparison of SMR_i to $\text{SMR}_{i'}$ is meaningful only to the extent that the distributions of $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ are similar.

Here, τ^2 represents between-hospital variation after accounting for at-admission patient attributes, allowing true, underlying hospital quality to vary around an overall mean effect denoted by μ. So, if there are no between-hospital differences in the outcome beyond that captured by the $\mathbf{x}_{ij}$, then $\tau^2 = 0$ and $\beta_{01} = \beta_{02} = \cdots = \beta_{0I} = \mu$. An implicit assumption is that n_i is exogenous, conditional on patient characteristics. We consider this assumption subsequently.

Note that the $\mathbf{z}_i$ influence the prior distribution for the β_{0i}, but the hospital non-specific $\mathbf{z}^*$ is used in the denominator to match the overall event rate. The current CMS model ("pure" random effects at the hospital level) is produced by setting $\mathbf{z}_i \equiv 0$, excluding all hospital-level covariates other than the intercept (μ). A saturated fixed-effects model (in the spirit of that advocated by Kalbfleisch and Wolfe 2013) is produced by using $(I-1)$ degrees of freedom in $\mathbf{z}_i$, equivalently replacing $\mu+\gamma\mathbf{z_i}$ by $\mu_i, i = i, \ldots, I$.

While hospital attributes should not be used in risk adjustment, it can be beneficial to use them in developing the national-level risk model to reduce potential confounding induced by correlation between hospital and patient-level attributes. They should also be considered as inputs to shrinkage targets when stabilizing SMRs, but care is needed.

Graphical Example Using USRDS Data

We focus on ranking 3459 dialysis providers using 1998 United States Renal Data System (USRDS). See Lin et al. (2006; 2009), Lockwood et al. (2002), Shen and Louis (2000) for details. We start with directly estimated Standardized Mortality Ratios (SMRs). Figure 5.2 reports substantial variation in the direct estimates and in CI length. Figure 5.3 displays MLEs, posterior means, and standard errors for the SMR. Note the considerable shrinkage of the centers with long whiskers. SMRs estimated as 0 have a reported SE also of 0, but for these the figure shows that

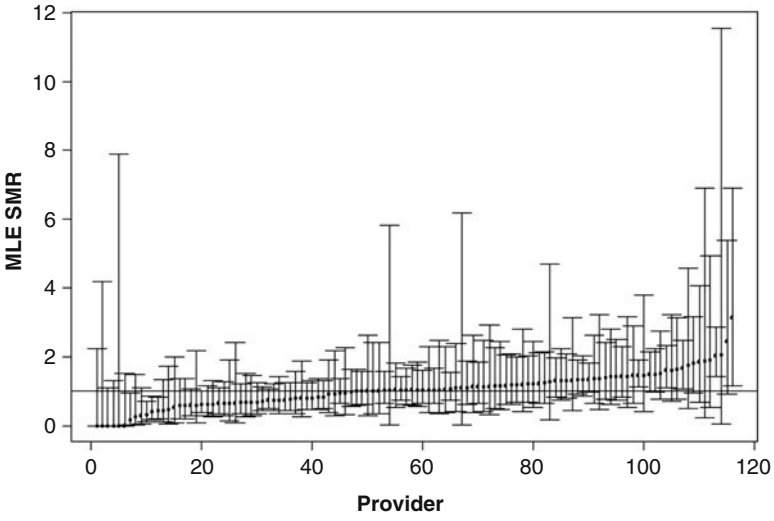


Fig. 5.2 MLEs and exact CIs for the SMR data. Providers are ordered by their MLE, and every 45th value is plotted

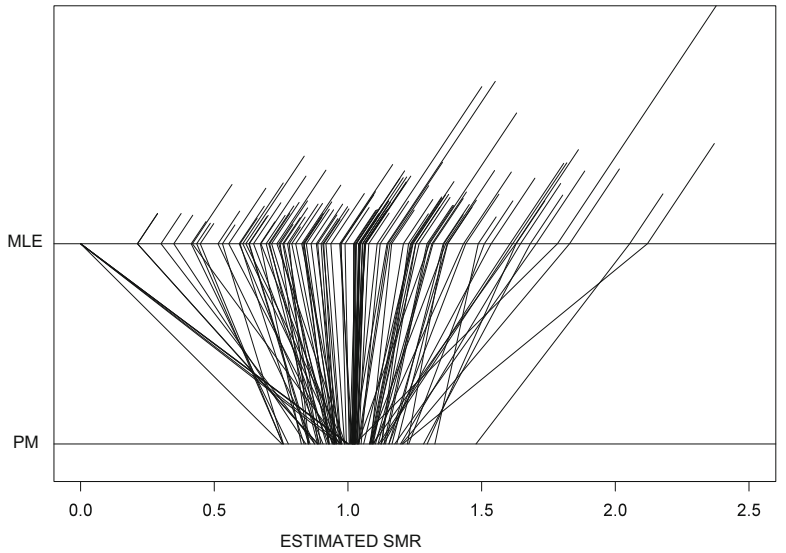


Fig. 5.3 USRDS SMRs: PM is the Posterior Mean, the whisker rising from the MLE line is proportional to the naively estimated SE of the MLE (e.g., estimated SMR = 0 produces an estimated SE = 0)

shrinkage varies, in fact roughly proportional to sample size.[2] This considerable shrinkage is the root of most of the controversy surrounding CMS profiling.

Procedure Volume

Stephen Senn once stated something like, "Mathematicians deal with lemmas, statisticians with dilemmas," but the low volume issue presents a tri-lemma in that there are three principal statistical options with associated issues:

1. Set aside "highly variable estimated SMRs," do not use them in estimating the model and do not report them:

   Implementing this approach requires a definition of "highly variable" and, especially if there are large number of highly variable estimates, sets aside a large amount of information that could be used to estimate the model. Also, these hospitals do require evaluation, and this approach pays no attention to them.

[2] It would be more informative to set whisker length proportional to exact CI length.

2. Report all directly estimated SMRs (no shrinkage) each accompanied by a confidence interval, some very wide:

   The positive is that all hospitals are evaluated and reported; negatives include that all stakeholders (including statisticians!) have a difficult time tempering the point estimate in the context of uncertainty. Furthermore, unstable estimates will bounce around considerably from assessment to assessment.

3. Report stabilized (Bayesian) estimates accompanied by a confidence interval. Shrinkage can either be toward a SMR of 1.0; alternatively toward a regression surface that depends on hospital-level attributes, for example volume:

   The confidence intervals will be relatively narrow, but shrinkage can be substantial (depending on the model used), energizing one of the criticisms that led to convening the COPSS panel. If hospital-level attributes contribute to the shrinkage target, the direct and Bayes estimates can be on different sides of 1.0, likely generating concern.

Shrinkage Targets Based on Hospital-Level Attributes

Including hospital-level attributes in determining the shrinkage target when stabilizing estimated hospital effects is standard practice in other facility assessment settings, and the COPSS committee recommended that the CMS give serious consideration to it. Under the current CMS analysis, volume does have a residual association with quality, possibly in part due to an inadequate risk adjustment that disadvantages small hospitals (see Sect. 5.3). Subject to adequate risk adjustment and other conditions, including it as a regressor to the shrinkage target would validly address the low volume issue. However, volume has a special status, due to it being the one attribute that is intimately associated with the degree of shrinkage. More important, it almost certainly has a combined role as an exogenous attribute that may be independently associated with quality but not "caused" by quality (e.g., practice makes perfect), and an endogenous attribute insofar as today's low volume could be a consequence of previously observed poor quality, and therefore, in the causal pathway to the outcome. This likely dual role generated considerable committee discussion on whether it should be a regressor in a profiling model. There is agreement that using volume in an individual decision-making model is appropriate, but debate continues in the profiling context.

Discussion

Successful preparation of the white paper required deep understanding of the policy issues and goals, the data collection/sampling process, porting these to candidate

statistical models, and evaluating these. There were (and are!) substantial policy implications to our work, and the high consequences generated some disagreements, so there was also some sociology and psychology in play. Without question, Steve was best at melding the policy and technical issues, the best understanding of the full context, including subtle issues of causal analysis, in many ways using Fienberg (2011) as his guide. He clarified the distinction between the exogenous and endogenous roles of patient volume. Specifically, was volume a predictor (exogenous) or a result (endogenous) of performance? The answer is "yes" and "yes." This conjunction of the very attribute that produces large shrinkage also being a predictor of and consequence of performance energizes the debate, with fully appropriate modeling yet to be developed and implemented.

Though the issue is not settled, there is considerable research addressing modeling in these situations with Silber et al. (2010) provide an argument in favor of including volume in determining the shrinkage target. They showed that lower quality is associated with lower volume, and that the shrinkage target for stabilizing the numerator of the estimated SMR is substantially modified using a volume-dependent target. However, the endo/exo issue is still in play. Other relevant research includes that by Varewyck et al. (2014), Varewyck et al. (2015) and Rompaye et al. (2015), with George et al. (2017) providing a comprehensive assessment. Yes, the beat goes on.

5.7 Summary

Carefully developed and applied, the Bayesian formalism adds considerable value, but care is needed. Even committed Bayesians counsel caution. For example, as reported by IJ Good, in 1970 Herman Rubin, in an oral statement at the Symposium on the Foundations of Statistical Inference, Waterloo, Canada, said "A good Bayesian does better than a non-Bayesian, but a bad Bayesian gets clobbered." So, it is important to trade-off some efficiency/optimality for acceptable robustness to departures from assumptions.

The Bayesian future is bright, the benefits of the approach are substantial, and it is frequently worth the bother. But, validity and effectiveness require expertise and care; the approach is by no means a panacea. Computing has enabled accommodating complex data and implementing models, enabling collaboration on challenging and important applications. Continued success will depend on "anchored flexibility" with the Bayesian viewpoint accommodating scientific, policy, and sociological realities. Eclecticism is (almost) always necessary; however, it is essential to have a point of view, a framework. Indeed, in applied statistical life nothing is pure anything; eclecticism, trade-offs and compromises are needed, guided by statistical principles. The domain of Bayesian approaches steadily expands, but there are limits; it is unlikely there will ever be a valid, multi/multi/multi/. . . -level model of a complex system.

Steve was a committed Bayesian, but had a higher commitment to fair and effective statistical design, conduct, analysis, and reporting. Consequently, he was eclectic in practice. I suspect that he would agree with a fair amount of the foregoing, tolerate some, and have a cringe-reaction to some. His cringes would be accompanied by an attempt to set me straight using theory and examples, all buttressed by his encyclopedic knowledge. I cannot predict the outcome and would surely love to have the interaction, but that is not to be.

References

Ash, A. S., Fienberg, S. E., Louis, T A (chair), Normand, S. L., Stukel, T. A. and Utts, J. (2012) Report of the COPSS committee to the Centers for Medicare and Medicaid Services: Statistical issues in assessing hospital performance. *Tech. rep.*, Committee of Presidents of Statistical Societies. URL http://nisla05.niss.org/copss/COPSS_CMS_Report.pdf http://www.cms.gov/HospitalQualityInits/Downloads/.

Ashmead, R. and Slud, E. (2017) Small area diagnostics and validation with applications to the voting rights act section 203. In *Proceedings of the JSM 2017 Survey Research Methods Section*, 1515–1523.

Bérubé, S., Datta, A., Li, Q., Wang, C. and Louis, T. A. (2019) Percentile-based residuals for model assessment. https://arxiv.org/abs/1910.03709.

Box, G. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Ser. A*, **143**, 383–430.

Brant, L. J., Duncan, D. B. and Dixon, D. O. (1992) k-ratio t tests for multiple comparisons involving several treatments and a control. *Statistics in Medicine*, **11**, 863–873.

Brownstein, N., Louis, T., O'Hagan, A. and Pendergast, J. (2019) The role of expert judgment in statistical inference and evidence-based decision-making. *The American Statistician*, **73 sup1**, 56–68.

Carlin, B. P. and Gelfand, A. E. (1991) A sample reuse method for accurate parametric empirical Bayes confidence intervals. *Journal of the Royal Statistical Society, Series B, Methodological*, **53**, 189–200.

Carlin, B. P. and Louis, T. A. (2009) *Bayesian Methods for Data Analysis, 3rd edition*. Boca Raton, FL: Chapman and Hall/CRC Press, 3nd edn.

Chen, S., Jiang, J. and Nguyen, T. (2015) Observed Best prediction for small area counts. *Journal of Survey Statistics and Methodology*, **3**, 136–161.

Efron, B. (1986) Why isn't everyone a Bayesian? (with discussion). *The American Statistician*, **40**, 1–11.

Efron, B. (2005) Bayesians, Frequentists, and Scientists. *Journal of the American Statistical Association*, **100**, 1–5.

Fienberg, S. E. (2011) Bayesian models and methods in public policy and government settings (with discussion). *Statistical Science*, **26**, 212–239.

George, E. I., Ročková, V., Rosenbaum, P. R., Satopaa, V. A. and Silber, J. H. (2017) Mortality rate estimation and standardization for public reporting: Medicare's hospital compare. *Journal of the American Statistical Association*, **112**, 933–947.

Good, I. J. (2007) Medialized p-values. *Journal of Statistical Computation and Simulation*, **31**, 126–127.

Henderson, N. C. and Newton, M. A. (2016) Making the cut: Improved ranking and selection for large-scale inference. *Journal of the Royal Statistical Society, Series B*, **78**, 781–804.

Henderson, N. C., Varadhan, R. and Louis, T. A. (2020) Improved Small Area Estimation via Compromise Regression Weights. *arXiv preprint arXiv:2006.15638*.

Jiang, J., Nguyen, T. and Rao, J. S. (2011) Best Predictive Small Area Estimation. *Journal of the American Statistical Association*, **106**, 732–745.

Joyce, P., Malec, D., Little, R., Gilary, A., Navarro, A. and Asiala, M. (2014) Statistical Modeling Methodology for the Voting Rights Act Section 203 Language Assistance Determinations. *Journal of The American Statistical Association*, **109**, 36–47.

Kalbfleisch, J. D. and Wolfe, R. A. (2013) On monitoring outcomes of medical providers. *Statistics in the Biosciences*, **5**, 286–302.

Laird, N. M. and Louis, T. A. (1987) Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *J. Am. Statist. Assoc.*, **82**, 739–757.

Lazar, N., Schirm, A. and Wasserstein, R. (2019) Special issue: Statistical inference in the 21st century: A world beyond $p < 0.05$. *The American Statistician*, **73**, suppl. Entire Issue.

Lin, R., Louis, T. A., Paddock, S. M. and Ridgeway, G. (2006) Loss function based ranking in two-stage, hierarchical models. *Bayesian Analysis*, **1**, 915–946.

—— (2009) Ranking of USRDS, provider-specific SMRs from 1998–2001. *Health Services and Outcomes Research Methodology*, **8**, 22–38.

Little, R. J. (2004) To Model or Not To Model? Competing Modes of Inference for Finite Population Sampling. *Journal of The American Statistical Association*, **99**, 546–556.

—— (2012) Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics (with discussion). *Journal of Official Statistics*, **28**, 309–372.

Lockwood, J., Louis, T. and McCaffrey, D. (2002) Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics*, **27**, 255–270.

Morris, C. N. (1983) Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*, **78**, 47–55.

Normand, S.-L., Ash, A. S., Fienberg, S. E., Stukel, T., Utts, J. and Louis, T. A. (2016) League Tables for Hospital Comparisons. *Annual Review of Statistics and Its Application*, **3**, 21–50.

O'Hagan, A. (2019) Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, **73 sup1**, 69–81.

Paddock, S. M. and Louis, T. A. (2011) Percentile-based empirical distribution function estimates for performance evaluation of healthcare providers. *J. Roy. Statist. Soc. Ser. C*, **60**, 575–589.

Paddock, S. M., Ridgeway, G., Lin, R. and Louis, T. A. (2006) Flexible prior distributions for triple goal estimates in two-stage, hierarchical linear models. *Computational Statistics and Data Analysis*, **50**, 3243–3262.

Rompaye, B. V., Eriksson, M. and Goetghbeur, E. (2015) Evaluating hospital performance based on excess cause-specific incidence. *Statistics in Medicine*, **34**, 1334–1350.

Rubin, D. (1984) Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *Annals of Statistics*, **12**, 1151–1172.

Scott, J. G. and Berger, J. O. (2006) An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, **136**, 2144–2162.

Shen, W. and Louis, T. A. (2000) Triple-goal estimates for disease mapping. *Statistics in Medicine*, **19**, 2295–2308.

Silber, J. H., Rosenbaum, P. R., Brachet, T. J., Ross, R. N., Bressler, L. J., Even-Shoshan, O., Lorch, S. A. and Volpp, K. G. (2010) The hospital compare mortality model and the volume-outcome relationship. *Health Serv Res*, **45**, 1148–1167.

Varewyck, M., Goetghebeur, E., Eriksson, M. and Vansteelandt, S. (2014) On shrinkage and model extrapolation in the evaluation of clinical center performance. *Biostatistics*, **15**, 651–664.

Varewyck, M., Vansteelandt, S., Eriksson, M. and Goetghebeur, E. (2015) On the practice of ignoring center-patient interactions in evaluating hospital performance. *Statistics in Medicine*, URL https://doi.org/10.1002/sim.6634.

Wasserstein, R. L. and Lazar, N. A. (2016) The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, **70**, 129–133.

Chapter 6
Probability Forecasts and Prediction Markets

Julia Mortera and A. Philip Dawid

6.1 Introduction

In this chapter, we give an overview of various topics tied to probability forecasting, i.e., the use of probability distributions to express uncertainty about future events—a problem area to which Steve Fienberg made seminal contributions. In particular, we consider methods for assisting and assessing a single forecaster; methods for combining the probability forecasts of several forecasters; and prediction markets, where forecasters take turns to announce their current probabilities, taking into account previous announcements.

We start, in Sect. 6.2, by reviewing methods for motivating and assessing a single forecaster. Important tools here include *proper scoring rules*, which motivate the forecaster to give honest predictions; *calibration*, which compares average forecasts to observed frequencies, and *resolution* and *refinement*, which reflect expertise in the subject area.

In Sect. 6.3, we describe methods for *opinion pooling*, where a decision-maker consults a number of experts who give their opinions as probability statements, and needs to combine these somehow. We consider model-based and axiomatic approaches and the application of coherence constraints under a specific definition of what constitutes expertise as seen by you, the decision-maker. These approaches are illustrated with the linear and logarithmic opinion pools.

J. Mortera (✉)
Università Roma Tre, Roma, Italy

University of Bristol, Bristol, UK
e-mail: julia.mortera@uniroma3.it

A. P. Dawid
University of Cambridge, Cambridge, United Kingdom

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_6

Section 6.4 discusses *prediction markets*, which are venues where individuals trade predictions on uncertain future events, and allow participants to stake bets on future events. The individuals take turns to update their probabilities for a future event, taking into account the previously announced probabilities of the other individuals, which may be based on unannounced private information. We show that there will always be convergence to a limiting value, which may or may not be the same as the value they could achieve if they were able to pool all their private information.

Finally, Sect. 6.5 reviews the main contributions of this chapter.

6.2 Evaluating a Single Probability Forecaster

Following Dawid (1986), consider a forecaster F who is required to describe his uncertainty about some unknown event A (coded $A = 1$ if A happens, $A = 0$ if not) by quoting a value $q \in [0, 1]$, intended to be interpreted as his personal probability for the event A. So long as $q \neq 0$ or 1, one might consider that neither outcome of A could discredit F's quote. Nevertheless, a higher value is clearly better when $A = 1$, and a lower value when $A = 0$. Here we consider ways of motivating and evaluating F, both for single and for multiple events.

Scoring Rules

To induce F to give an honest prediction, we might penalise him with a loss $S(a, q)$, depending on his quoted probability forecast q and the eventual outcome a ($= 0$ or 1) of A. Such a loss function S is termed a *scoring rule*. We assume that the forecaster F wishes to minimise his expected loss. Let $p = \Pr(A = 1)$ be F's true subjective probability of A. Then when he quotes probability value q, his expected loss is $S(p, q) := pS(1, q) + (1 - p)S(0, q)$. The forecaster should thus choose q to minimise $S(p, q)$. The scoring rule S is called *[strictly] proper* if, for any true probability p, the expected loss $S(p, q)$ is minimised if [and only if] $q = p$. Under such a scoring rule, honesty is the best policy.

There is a wide variety of proper scoring rules, which can be tailored to emphasise different parts of the probability range. Important examples are the following:

(a) The Brier score or quadratic loss function (Brier 1950; de Finetti 1954):

$$S(1, q) = (1 - q)^2$$
$$S(0, q) = q^2.$$

(b) The logarithmic scoring rule (Good 1952):

$$S(1, q) = -\log q$$
$$S(0, q) = -\log(1 - q).$$

As well as motivating honesty before the event, a proper scoring rule can be used after the event, to quantify the quality of the forecaster's performance, in the light of the observed outcome a, by means of the realised score $S(a, q)$ (a lower value being better). Different forecasters, with their differing q's, can thus be compared.

When a forecaster makes a sequence of probability forecasts, for multiple events, additional evaluation criteria become available. In particular, we can assess the *calibration* and *resolution* of the forecasts issued.

Calibration

Suppose that, over a long sequence, F has issued probability forecast p_i for event A_i. Now choose $\pi \in [0, 1]$, and consider all those occasions i for which $p_i = \pi$ (to a good enough approximation). Supposing there are many such occasions, let $\rho(\pi)$ denote the relative frequency of success ($A_i = 1$) on these occasions. A plot of $\rho(\pi)$ against π is the forecaster's *calibration curve*, and the forecaster is said to be *well-calibrated*, or *probability calibrated*, when he is "getting the relative frequencies right", *i.e.*, $\rho(\pi) \approx \pi$ for all values of π used. In meteorology, calibration is also termed validity or reliability, and a well-calibrated forecaster is called perfectly reliable. It is shown in Dawid (1982) that, when events arise and are predicted in sequence, probability calibration is a necessary (though not sufficient) requirement of a good forecaster.

Resolution

Probability calibration is a fairly weak constraint on a forecaster. It will hold for the "naïve forecaster", who quotes the same value q for every A_i, so long as $q = \pi_0$, the overall relative frequency of success; as well as for the ideal "perfect forecaster", who has a crystal ball and so can always give probability 1 to the outcome a_i of A_i that actually occurs (so $q_i = a_i$). Although both are well-calibrated, the latter is doing a much more useful forecasting job than the former.

More generally, a good forecaster should be able to issue many forecasts close to the extreme values 0 or 1, with few intermediate values, while remaining well-calibrated. The same criterion can be applied to an uncalibrated forecaster, if we first replace each issued probability q_i by its recalibrated version $r_i = \rho(q_i)$. The term "resolution" refers to the extent to which a forecaster's (possibly recalibrated)

forecast probabilities are widely dispersed on the unit interval. Thus, a weather forecaster's resolution is a reflection of his knowledge of, and skill in, forecasting the weather, whereas his calibration addresses his entirely different ability to quantify his uncertainty appropriately.

Resolution can be quantified in various ways, for example by the variance of the recalibrated forecasts. More generally, let S be an arbitrary proper scoring rule. We might assess a forecaster's overall performance, over n events, by his total achieved penalty score, $S_+ := \sum_{i=1}^n S(a_i, q_i)$, where q_i is his quoted probability for A_i, and $a_i = 0$ or 1 is the outcome of A_i. This total score can be used to compare different forecasters. Now introduce the *entropy function* associated with S, $H(p) := S(p, p)$, which is a concave function of p; and the associated *discrepancy function*, $D(p, q) := S(p, q) - H(p)$, which is non-negative and vanishes for $q = p$. (For the Brier score, $H(p) = p(1 - p)$, and $D(p, q) = (p - q)^2$.) Let $r_i = \rho(q_i)$ be the recalibrated version of q_i. Then (DeGroot and Fienberg 1983), we can decompose $S_+ = S_1 + S_2$, where

$$S_1 = \sum_{i=1}^{n} D(r_i, q_i)$$

$$S_2 = \sum_{i=1}^{n} H(r_i).$$

We see that $S_1 \geq 0$, with equality if $r_i = q_i$ for all i: S_1 thus penalises poor calibration. As for S_2, since H is concave, it is smaller when the recalibrated forecasts (r_i) are clustered near 0 and 1, and thus S_2 penalises poor resolution. We can use these components of the overall score to compare forecasters in terms of their calibration and/or their resolution.

We can further decompose $S_2 = nH(\pi_0) - S_3$, where $S_3 = \sum_{i=1}^n D(r_i.\pi_0)$, and $\pi_0 = n^{-1} \sum_{i=1}^n a_i$ is the overall relative frequency of success. Since the first term is fixed, a larger S_3 indicates better resolution. For the Brier score, this delivers the variance criterion.

Refinement

DeGroot and Fienberg (1983) describe a partial ordering between forecasters that is related to resolution. This is based on the theory of sufficiency in the comparison of statistical experiments (Blackwell 1951).

Consider two forecasters, F and F', who issue respective forecasts (q_i) and (q_i') for the same sequence of events (A_i), with outcomes (a_i). We can suppose both forecasters are well-calibrated; if not, we work with their recalibrated forecasts, rather than the raw values. Then, we say that F is *more refined* than F' if there exists a specification of a conditional distribution, $p(q' \mid q)$, of q' given q, such

that, both for $\alpha = 0$ and for $\alpha = 1$, we can generate the (empirical) distribution of q', for those events having $a = \alpha$, by first generating q from its distribution given $a = \alpha$, and then generating q' from $p(q' \mid q)$. In this case, we can consider q' as a noisy version of q, with the noise unrelated to the true outcome—which suggests that F' has poorer performance than F. Indeed, it can be shown that when this property holds, the resolution score S_2 for F will not exceed that for F', for any proper scoring rule S.

6.3 Combining Several Opinions

It is sometimes necessary to construct a single opinion by combining a number of individual opinions. A decision-maker might consult a number of experts (financial, meteorological, medical, etc.) before reaching a final decision.

We can distinguish three types of problems:

(a) When opinions are expressed as probability distributions. For reviews of opinion pooling in this setting, see for example Clemen and Winkler (1999), Dawid et al. (1995), DeGroot and Mortera (1991), Genest and Zidek (1986), Ranjan and Gneiting (2010). Marschak and Radner (1972) developed team theory from an economic perspective.
(b) Group decision-making when opinions are expressed as preferences among alternatives: see for example Arrow (1951), Laffont (1979), Luce and Raiffa (1958).
(c) Meta-analysis where different quantitative methods are used to combine the results of different studies on the same topic.

Here we will be concerned only with problems of type (a). We shall suppose that the experts' opinions are expressed as probability distributions, over a fixed set of events and quantities of interest, but the data underlying those opinions remain undisclosed.

The ideal approach to merging several experts' views would be for each of them to report all the data and background knowledge on which his or her opinions are based, and for you, the decision-maker, to combine all this information with your own prior opinions, and any additional data you may have, using Bayes's theorem; but in the absence of access to the underlying data, you can only work with the experts' opinions, expressed as probability distributions. Their distributions will most probably differ. Your task is to combine these differing opinions, somehow, into a single distribution to use as your own. Let the k experts $E_1, E_2, \ldots E_k$ give their probability predictions $\Pi_1, \Pi_2, \ldots \Pi_k$ for an uncertain quantity, perhaps an event A or the parameter θ of a distribution. You must pool the experts' distributions to form your resulting aggregate or pooled distribution, Π.

Model-Based Approach

In this approach, the experts' opinions are modelled as data (for you), and, on combining the data with your own prior opinions, using Bayes's theorem, you can construct your own posterior distribution. This approach is taken by, among others, Berger and Mortera (1991), French (1986), Lindley (1983), Winkler (1981). For the case of a single event, Winkler (1981) assumes that the various log-odds have a multivariate normal distribution.

However, the process by which probability assessments are generated is not in general very easy to formalise—unlike the mechanisms by which experimental data are typically generated. The model needs to take into account the decision-maker's opinion, the dependence between that and the experts' opinions, the interdependencies among the experts' opinions, and the dependence between all of these and the quantity of interest.

Group with Complete Interaction In the case where all experts exchange information, the problem of consensus of opinions expressed as probability distributions is an example of complete interaction. DeGroot (1974) considers a group of individuals who must act together as a team or committee, each individual in the group having his/her own subjective probability distribution for the unknown quantity. After these are all announced (round 1), each expert updates his distribution to a linear combination of all the distributions. This procedure is repeated over many rounds, the weights varying between experts but being fixed over time. DeGroot (1974) presents a condition under which the group eventually reaches agreement on a common probability distribution. The model can also be applied to problems of reaching a consensus when the opinion of each member of the group is represented as a point estimate, rather than as a probability distribution. Aumann (1976) studied the dynamics of reaching a consensus through Bayesian dialogue, where conflicting opinions in a group are due solely to the fact that the members have different information sets.

Group with Partial Interaction The theory above bears a resemblance to the Delphi technique (Pill 1971), used to reach agreement among a panel of experts. The Delphi technique is a purely empirical procedure and is not based on any underlying mathematical model. Again, it is applied iteratively in a sequence of rounds. At each round, the individuals are informed of the opinions of the others in the group and allowed to reassess their own opinion before proceeding to the next round. Because of the empirical nature of the Delphi technique, it provides no conditions under which the experts can be expected to reach agreement or terminate the iterative process.

Axiomatic Approach

In this approach, a series of axioms are laid down in which an opinion pooling method should satisfy. For example, if all the experts agree on a certain property—e.g., that certain events are independent—then one might require that this should be preserved in the aggregated distribution. Another such property is invariance with respect to marginalisation, i.e., you would attain the same aggregated opinion if you first aggregate overall and then marginalise, or if each expert gives his marginal distribution and you then aggregate those. Depending on the properties assumed, a variety of aggregation methods can be derived. Among these, we will discuss the *linear opinion pool* and the *logarithmic opinion pool*.

Linear Opinion Pool Stone (1961) considered the linear opinion pool:

$$\Pi = \sum_{i=1}^{m} w_i \Pi_i, \tag{6.1}$$

where $w_i \geq 0$ and $\sum_{i=1}^{m} w_i = 1$. He suggested that the opinion pool is democratic if you use equal weights $w_1 = w_2 = \cdots = w_m = 1/m$.

The linear opinion pool has both advantages and disadvantages. McConway (1981) proved that if you require the marginalization property then the rule for aggregation must be linear (if at least three nontrivial events exist). The weights w_i can be interpreted as reflecting the previous performance of the experts. DeGroot and Mortera (1991) derived the optimal weights according to a criterion based on the Brier score.

Ranjan and Gneiting (2010) show that the linear opinion pool is uncalibrated, even when the individual probability forecasts are calibrated.

Logarithmic Opinion Pool The logarithmic opinion pool is given by

$$\log \pi = w_0 + \sum_{i=1}^{m} w_i \log \pi_i,$$

where π [resp., π_i] is the density function of Π [resp., Π_i], and w_0, a function of $(w_1, \ldots, w_m)$, is chosen to ensure that Π is a probability distribution. This was derived by Weerahandi and Zidek (1981). An important property of the logarithmic opinion pool is its consistency under aggregating and updating, i.e., if you first aggregate opinions and then update the pooled opinion when new information is available, or if the experts first update their opinions with the new information and you then aggregate these. However, the weights do not have a simple interpretation, and if a probability given by any expert for an event is zero, then the pooled probability is zero, whatever weight he/she has, and whatever the other experts' opinions are.

Coherent Combination

Dawid et al. (1995) investigate coherent methods for combining experts' opinions, when these are expressed as probabilities for some fixed event A. Neither axiomatic nor modelling assumptions are made. Instead a restricted definition is used of what constitutes *expertise*, as seen by you, the decision-maker: an expert is considered to be someone who "shares your world-view", i.e., if you both had identical information, you would both have identical opinions. However, the expert may know more than you do. It is assumed that the probabilities the expert provides are correctly and coherently computed.

Suppose you have access to k different experts. If you were to obtain a probability for an event A from just one of these, you should adopt it as your own, but the different experts' probabilities will typically differ, since they will be based on differing information. You require a *combination formula* to apply to the full collection $(\Pi_1, \cdots, \Pi_k)$ of expert probabilities, to compute your own probability Π.

Before you consult the experts, their various reports $(\Pi_1, \cdots, \Pi_k)$ will be, for you, uncertain random quantities, jointly distributed together with the uncertain event A of interest. Let P^* denote your overall joint distribution on the random quantities $(\Pi_1, \cdots, \Pi_k, A)$, and let P denote the implied distribution for (Π_i), marginalizing over A.

The laws of coherence require that, on learning all the experts' probabilities, you should assign probability $P^*(A \mid \Pi_1, \cdots, \Pi_k)$ to A. This yields the combination formula

$$\Pi = \Phi(\Pi_1, \cdots, \Pi_k) \equiv P^*(A \mid \Pi_1, \cdots, \Pi_k). \tag{6.2}$$

Note that, if expert i bases her probability on observation of X_i, then $\Pi_i \equiv P^*(A \mid X_i)$, where here P^* is extended to encompass (X_i), but in general the value of X_i will not be fully recoverable from that of Π_i, so that the right-hand side of (6.2) will not usually be the same as $P^*(A \mid X_1, \cdots, X_k)$.

The question addressed is: When will a given combination formula Φ be coherently compatible with some joint distribution P for the experts' reported opinions? i.e., when will there be some overall joint distribution P^* under which $\Pi_i \equiv P^*(A \mid \Pi_i)$, the implied distribution for $(\Pi_1, \cdots, \Pi_k)$ is P, and (6.2) holds?

Compatibility Consider the case of $k = 2$ experts and an event A, where expert E_i observes X_i and reports $\Pi_i \equiv P^*(A \mid X_i)$, for $i = 1, 2$. Then, from the definition of an expert, $\Pi_i \equiv P^*(A \mid \Pi_i)$, and defining $\Phi(\Pi_1, \Pi_2) \equiv P^*(A \mid \Pi_1, \Pi_2)$, Φ must satisfy

$$0 \leq \Phi(\Pi_1, \Pi_2) \leq 1 \tag{6.3}$$

$$\mathrm{E}_P(\Phi \mid \Pi_i) = \Pi_i, \ i = 1, 2. \tag{6.4}$$

Then, $\mathrm{E}_P(\Phi) = \mathrm{E}_P(\Pi_1) = \mathrm{E}_P(\Pi_2) = \pi_0$, say. Thus, by Bayes's theorem,

$$p^*(\pi_1, \pi_2 \mid A) \equiv \pi_0^{-1} \Phi(\pi_1, \pi_2) p(\pi_1, \pi_2),$$

where $\pi_0 = P^*(A)$.

The pair (P, Φ) determines a unique distribution P^* for (Π_1, Π_2, A) with $P^*(A \mid \Pi_i) \equiv \Pi_i$, $P^*(A) = \pi_0$. Conditions (6.3) and (6.4) are necessary and sufficient conditions for logical consistency and the pair (P, Φ) are then termed *compatible*.

Characterizations In general, the problems of characterizing all Φ's compatible with a given P, and *vice versa*, are difficult. This set, defined by (6.3) and (6.4), is convex, but not generally a simplex. It might be empty, or contain just one member, or many.

Let P denote a joint distribution for (Π_1, Π_2) having $\mathrm{E}(\Pi_1) = \mathrm{E}(\Pi_2) = \pi_0$; and let Φ be a combination formula. Define a finite measure Q by $dQ(\pi_1, \pi_2) := \Phi(\pi_1, \pi_2) p(\pi_1, \pi_2)$, and let P_i and Q_i be the marginals for Π_i under P and Q, respectively. Dawid et al. (1995) show that Φ and P are a compatible pair if and only if $dQ_i(\pi_i) \equiv \pi_i dP_i(\pi_i) = dP_i^*(\pi_i)$ for $i = 1, 2$. Given P, this shows that the problem of finding a compatible Φ reduces to that of characterizing measures Q having specified marginals and with $dQ/dP \leq 1$.

As a corollary of the above, for any absolutely continuous coherent joint distribution P for (Π_1, Π_2), there exists a compatible combination formula Φ that takes values 0 and 1 only. This implies that it is logically consistent that the combination of opinions could deliver absolute subjective certainty as to whether the event A holds or not.

Most of the literature on combining opinions uses axiomatic properties or modelling assumptions to derive particular pooling recipes. Compared with these, the assumptions for coherent pooling are less restrictive.

This analysis also offers guidance for assessing pooling formulae that have been suggested from other approaches. Thus, consider combination rules that can be expressed in the generalized linear form

$$g(\Phi) \equiv \alpha_1 g(\Pi_1) + \alpha_2 g(\Pi_2) + c, \tag{6.5}$$

for some monotonic continuous function g. These include those considered in Sect. 6.3.

Linear Opinion Pool Dawid et al. (1995) show that, with (6.3) and (6.4), a generalised linear opinion pool

$$\Phi \equiv \alpha_1 \Pi_1 + \alpha_2 \Pi_2 + c \tag{6.6}$$

has $\alpha_0 + \alpha_1 + \alpha_2 = 1$, where $c = \alpha_0 \pi_0$, with $\pi_0 := P^*(A)$. In particular, if $c \neq 0$, every distribution P^* compatible with Φ must assign the same prior probability

$\pi_0 = c/(1-\alpha_1-\alpha_2)$ to A. When both experts agree, you will adopt their common forecast if and only if it is the exactly same as your prior probability for A.

When α_1, α_2, and $c = \alpha_0\pi_0$ are all non-zero, (6.4) implies that

$$\mathrm{E}_P(\Pi_2 \mid \Pi_1) \equiv \lambda\Pi_1 + (1-\lambda)\pi_0 \tag{6.7}$$

$$\mathrm{E}_P(\Pi_1 \mid \Pi_2) \equiv \mu\Pi_2 + (1-\mu)\pi_0, \tag{6.8}$$

where $\lambda := (1-\alpha_1)/\alpha_2$, $\mu := (1-\alpha_2)/\alpha_1$, i.e., each Π_i has a linear regression on the other.

Conversely, any joint distribution P on $[0,1]^2$ that satisfies (6.7) and (6.8) is compatible with a Φ of form (6.6), for

$$\alpha_1 = \frac{1-\lambda}{1-\lambda\mu}$$

$$\alpha_2 = \frac{1-\mu}{1-\lambda\mu}$$

$$c = \pi_0(1-\alpha_1-\alpha_2)$$

as long as it gives probability 1 to the event $0 \le \alpha_1\Pi_1 + \alpha_2\Pi_2 + c \le 1$, thus satisfying (6.3). This characterizes all distributions P compatible with Φ in (6.6).

Note that not all choices of the α coefficients are coherent. Since, from (6.7) and (6.8), $\lambda\mu$ is the squared correlation ρ^2 between Π_1 and Π_2, we must have $0 \le (1-\alpha_1)(1-\alpha_2)/\alpha_1\alpha_2 < 1$. In contrast with an assumption commonly made for linear opinion pools, it is not coherent for α_1, α_2, and c all to be strictly positive, since then both regression coefficients in (6.7) and (6.8) would exceed 1.

Bradley (2018) claimed to show that a joint distribution P is compatible with a linear opinion pool only if Π_1 and Π_2 are almost surely identical. However, as pointed out by Dawid and Mortera (2020), there is an error in his proof.

Logarithmic Opinion Pool A generalised logarithmic opinion pool can be expressed as

$$\operatorname{logit}\Phi \equiv \alpha_1 \operatorname{logit}\Pi_1 + \alpha_2 \operatorname{logit}\Pi_2 + \alpha_0 \operatorname{logit}\pi_0, \tag{6.9}$$

where $\operatorname{logit} x = \log\{x/(1-x)\}$.

This combination formula can arise as follows. Given A or $\bar{A}$, let (X_1, X_2) be bivariate normal with $\mathrm{var}(X_i \mid A) = \mathrm{var}(X_i \mid \bar{A}) = 1$, and $\mathrm{cov}(X_1, X_2 \mid A) = \mathrm{cov}(X_1, X_2 \mid \bar{A}) = \rho$, with $\rho^2 \neq 1$ and $\mathrm{E}(X_i \mid A) = \delta_i/2$ and $\mathrm{E}(X_i \mid \bar{A}) = -\delta_i/2$ for $i = 1, 2$. Let $\Pi_i := P^*(A \mid X_i)$ and $\Phi := P^*(A \mid X_1, X_2)$. Bayes's theorem yields

$$\operatorname{logit}\Pi_i = \operatorname{logit}\pi_0 + \delta_i X_i,$$

and

$$\text{logit}\, P^*(A \mid X_1, X_2) = \text{logit}\, \pi_0 + (1-\rho^2)^{-1}\{(\delta_1 - \rho\delta_2)X_1 + (\delta_2 - \rho\delta_1)X_2\}.$$

Hence, (6.9) holds, with $\alpha_1 = (1-\rho\eta)/(1-\rho^2)$, and $\alpha_2 = (1-\rho\eta^{-1})/(1-\rho^2)$, where $\eta = \delta_2/\delta_1$, and $\alpha_0 = 1 - \alpha_1 - \alpha_2$.

Again we have $\alpha_0 + \alpha_1 + \alpha_2 = 1$, and if $\alpha_0 \neq 0$, (6.9) determines π_0. Since $\alpha_0 = -(1-\rho^2)\alpha_1\alpha_2$, again α_0, α_1, and α_2 cannot all be strictly positive.

Conditional Independence If the two experts' opinions are conditionally independent given both A and $\bar{A}$, $\Pi_1 \perp\!\!\!\perp \Pi_2 \mid (A, \bar{A})$, then

$$\text{logit}\, \Phi \equiv \text{logit}\, \Pi_1 + \text{logit}\, \Pi_2 + c, \tag{6.10}$$

where $c = -\text{logit}\, \pi_0$.

Dawid et al. (1995) prove the following theorem that characterizes all joint distributions compatible with (6.10).

Theorem 1 *A necessary and sufficient condition for a joint density $f(\pi_1, \pi_2)$ to be compatible with Φ in (6.10) is that*

$$f(\pi_1, \pi_2) \equiv \{[1-\pi_0)\pi_1\pi_2 + \pi_0(1-\pi_1)(1-\pi_2)]/(\pi_0(1-\pi_0))\}g(\pi_1, \pi_2), \tag{6.11}$$

where $\pi_0 = (1+e^c)^{-1}$ and g is a density such that $E_g(\Pi_1 \mid \Pi_2) \equiv E_g(\Pi_2 \mid \Pi_1) \equiv \pi_0$. In this case, π_0 is the common expectation of Π_1 and Π_2 under f and, thus, the prior probability of A.

6.4 Prediction Markets

This part of this chapter revisits some results appearing in the economics literature from a statistical point of view.

A prediction market—also known as a predictive market, an information market, a decision market, or a virtual market—is a venue where experts can trade predictions on uncertain future events and can stake bets on various events occurring. Such events might be, for example, an election result, a terrorist attack, a natural disaster, commodity prices, quarterly sales, or sporting outcomes. Prediction markets also offer trade in possible future outcomes on securities markets, in which case participants who use it are buying something like a futures contract. The Iowa Electronic Markets (http://tippie.uiowa.edu/iem/) of the University of Iowa Henry B. Tippie College of Business is one of the main prediction markets in operation. Also companies like Google have their own internal prediction markets. Prediction markets sometimes operate as an open market like the stock market, or as a closed

market akin to a betting pool. A prediction market translates the wisdom of crowds into predictive probabilities.

For example, suppose that in a prediction market one can bet whether A occurs (before time t), and individuals can trade contracts among each other. Consider a contract that pays 1 if event A occurs, and 0 otherwise. Say the current market price for the contract is 0.58. Offers to buy and sell are fixed at 0.57 and 0.59, respectively. Now you can either pay 0.59 instantly, or post an offer to pay 0.58 and see if anyone is willing to sell at that price. If so, the new market price, 0.58, becomes the consensus probability.

Prediction markets have been discussed by Aldous (2013), Arrow et al. (2008), Hanson (2003), Chen et al. (2010), Hanson et al. (2006), Wolfers and Zitzewitz (2008), and Strähl and Ziegel (2015), among others.

Basic Setup

We shall focus on the opinions of a specific individual, "you", possibly but not necessarily a participant in the market, and how these opinions change in the light of accumulating experience. We suppose that your opinions are expressed as a joint probability distribution, Pr, over all relevant variables. Other individuals may have their own probabilities for various events, but for you these are data. In the sequel, all probabilities are computed under your distribution Pr.

We shall again interpret the term "expert" in the sense of Dawid et al. (1995), DeGroot (1988). That is, an individual E is an expert (for you) if E started with exactly the same joint probability distribution Pr over all relevant variables as you, and has observed everything that you have observed, and possibly more. Then when you learn (just) the probability Π that E assigns to some event A, your updated probability for A will be Π. That is, you will agree with the expert.

In the context of a prediction market, suppose that experts $E_1, E_2, \ldots$ sequentially announce their probability predictions $\Pi_1, \Pi_2, \ldots$ for a future event A. Thus, E_i is the expert that makes the forecast at time i, and we allow that the same expert could make forecasts at different times. At time i, expert E_i has access to all previous forecasts $\Pi_1, \ldots, \Pi_{i-1}$ and possibly additional private information H_i, but E_i will typically not have access to the private information sets $H_1, \ldots, H_{i-1}$ that the previous experts used in formulating their forecasts. However, in some markets, there is an option for forecasters to leave comments, which could give partial information K_i (which might be empty) about H_i. We assume that each forecaster is aware of all such past comments. Thus, $\Pi_i = \Pr(A \mid T_i)$, where $T_i := (K_1, \Pi_1, \ldots, K_{i-1}, \Pi_{i-1}, H_i)$ is the total information available to E_i.

The full public information available just after time i is $S_i := (K_1, \Pi_1, K_2, \Pi_2, \ldots, K_i, \Pi_i)$. Note that S_i and T_i both contain all the information made public up to time $i - 1$. They differ however in the information they contain for time i: T_i incorporates the totality, H_i, of expert E_i's information, both public, K_i, and private, whereas S_i incorporates only E_i's public information, K_i, and her

announced probability forecast, Π_i, for A at time i. The information sets (T_i) are not in general increasing with i, since H_i is included in T_i but need not be in T_{i+1}. The information sets (S_i) are however increasing. The following Lemma and Corollary show that, for you, for the purposes of predicting A both information sets T_i and S_i are equivalent, and your associated prediction is just the most recently announced probability forecast.

Lemma 1 $Pr(A \mid S_i) = Pr(A \mid T_i) = \Pi_i$.

Proof Since $T_i \supseteq S_i \ni \Pi_i$,

$$\begin{aligned} \Pr(A \mid S_i) &= \mathrm{E}\{\Pr(A \mid T_i) \mid S_i\} \\ &= \mathrm{E}(\Pi_i \mid S_i) \\ &= \Pi_i \\ &= \Pr(A \mid T_i). \end{aligned}$$

□

Corollary 2 *If you observe the full public information S_i and have no further private information, your conditional probability for A is just the last announced forecast Π_i.*

Convergence

From Lemma 1 and the fact that the information sequence (S_i) is increasing, we have:

Corollary 3 *The sequence (Π_i) is a martingale with respect to (S_i).*

Then by Corollary 3 and the martingale convergence theorem, we have:

Corollary 4 *As $i \to \infty$, Π_i tends to a limiting value Π_∞.*

The variable Π_∞ is random in the sense that it depends on the initially unknown (to you) information sequence $S_\infty := \lim S_i$ that will materialise but will be a fixed value for any such sequence.

A perhaps surprising implication of Corollary 4 is that, eventually, introduction of new experts will not appreciably change the probability you assign to A—whatever new private information they may bring will be asymptotically negligible compared with the accumulated public information. We term Π_∞ the *consensus probability* of A, and the information S_∞ on which it is based the *consensus information set*. The information S_∞ is *common knowledge* for all experts in the sense of Aumann (1976): see Geanakoplos (1992a;b), McKelvey and Page (1986), Nielsen (1984).

It might be considered that the limiting value Π_∞ has succeeded in integrating all the private knowledge of the infinite sequence of experts. As we shall see below, this is sometimes, but not always, the case.

Two Experts

As a special case, suppose we have a finite set of experts, $E_1, \ldots, E_N$, and we take $E_{N+1} = E_1$ (so $H_{N+1} = H_1$), $E_{N+2} = E_2$, *etc.* Thus, we repeatedly cycle through the experts. Continuing for many such cycles, eventually we will get convergence to some Π_∞, at which point each expert will not be changing her opinion based on the total sequence of publicly announced forecasts, even though she may still have access to additional private information.

At convergence, it will thus make no difference to expert E_i to incorporate (again) her private information H_i. Consequently, we have:

Proposition 5 *For each i, $A \perp\!\!\!\perp H_i \mid S_\infty$.*

In the sequel, we consider in detail the case $N = 2$ of two experts, who alternate $E_1, E_2, E_1, E_2, \ldots$ in updating and announcing their forecasts. Geanakoplos and Polemarchakis (1982) have studied this in the case that there is no side information, and each expert E_i's set of possible private information has finite cardinality, k_i say. They show that exact consensus is reached in at most $k_1 + k_2$ rounds.

Dutta and Polemarchakis (2014) give a simple example, with two experts, that shows that the order in which the experts play can matter. In their example, they show that when one of the experts starts they reach complete consensus (i.e., equivalent to pooling their private information), whereas on changing the starting order they only reach a limited consensus. Dutta and Polemarchakis (2014) further show that if an expert has additional information, this can produce a weaker consensus. They call this "obfuscation".

Vacuous Consensus

We start with some examples where the experts learn nothing from each other's forecasts—although they would learn more if they were able to communicate and pool their private data.

Example 1 (Parity Check) This example is essentially the same as that described by Geanakoplos and Polemarchakis (1982), p. 198.

Let X_1, X_2 be independent fair coin tosses. Expert E_i observes only X_i ($i = 1, 2$). Let A be the event $X_1 = X_2$. This has prior probability 0.5.

On observing his private information X_1, whatever value it may take, E_1's probability of A is unchanged, at 0.5. His announcement of that value is therefore

totally uninformative about the value of X_1. Consequently, E_2 can only condition on her private information about X_2—which similarly has no effect. The sequence of forecasts will thus be 0.5, 0.5, 0.5, ... Convergence is immediate, but to a vacuous state.

However, if the experts could pool their data, they would learn the value of A with certainty. □

Example 2 (Bivariate Normal) With this example, we generalise from predicting an uncertain event to predicting an uncertain quantity.

Suppose that E_1 observes X_1, and E_2 observes X_2, where (X_1, X_2) have a bivariate normal distribution with means $\mathrm{E}(X_i) = 0$, variances $\mathrm{var}(X_i) = 1$, and unknown correlation coefficient ρ—which is what they have to forecast. Let ρ have a prior distribution Π_0. Since X_1 is totally uninformative about ρ, E_1's first forecast is again Π_0 and so is itself uninformative. Again, E_2 has learned nothing relevant to ρ and so outputs forecast Π_0, and so on, leading to immediate convergence to a vacuous state. However, the pooled data (X_1, X_2) is informative about ρ (though does not determine ρ with certainty). □

In the above examples, each expert's private information was marginally independent of the event or variable, generically Y say, being forecast, with the immediate result that the consensus forecast was vacuous, the same as the prior forecast. Conversely, suppose the consensus is vacuous. That is to say,

$$Y \perp\!\!\!\perp S_\infty. \tag{6.12}$$

From Proposition 5 (trivially generalised), we have

$$Y \perp\!\!\!\perp H_i \mid S_\infty. \tag{6.13}$$

Combining (6.12) and (6.13), we obtain

$$Y \perp\!\!\!\perp (H_i, S_\infty)$$

whence, in particular,

$$Y \perp\!\!\!\perp H_i.$$

Hence, the consensus will be vacuous if and only if each expert's private information is, marginally, totally uninformative. The argument extends trivially to any finite number of experts.

Complete Consensus

We use the term *complete consensus* to refer to the case that the consensus forecast will be the same as the forecast based on the totality of the private information available to all the individual forecasters. A simple situation where this will occur is when Π_i is a one-to-one function of H_i, so that, by announcing Π_i, expert E_i fully reveals her private information.

Example 3 (Overlapping Bernoulli Trials) Let θ be a random variable with a distribution over $[0, 1]$ having full support. Given θ, let $Y_0 \sim B(n_0, \theta)$, $Y_1 \sim B(n_1, \theta)$, $Y_2 \sim B(n_2, \theta)$, and $A \sim B(1, \theta)$, all independently.

Suppose E_1 observes $X_1 = Y_0 + Y_1$, and E_2 observes $X_2 = Y_0 + Y_2$. At the first stage, E_1 computes and announces $\Pi_1 = \Pr(A \mid X_1)$—which is a one-to-one function of X_1. For example, under a uniform prior distribution for θ, $\Pi_1 = (X_1 + 1)/(n_0 + n_1 + 2)$. Then at stage 2, E_2 will have learned X_1 and also has private information X_2. Thus, $\Pi_2 = \Pr(A \mid X_1, X_2)$, the correct forecast given the pooled private information of E_1 and E_2 (though different from that based on full knowledge of (Y_1, Y_2, Y_2)). Further cycles will not change this probability, which will be the consensus. □

Example 4 (Linear Prediction) Consider variables $\mathbf{X} = (X_1, \ldots, X_k)$, $\mathbf{Z} = (Z_1, \ldots, Z_h)$, and (scalar) Y, all being jointly normally distributed with non-singular dispersion matrix. Expert 1 observes $H_1 = \mathbf{X}$, Expert 2 observes $H_2 = \mathbf{Z}$, and they have to forecast Y. Each time an expert announces her predictive distribution for Y, she is making known the value of her predictive mean of Y, which will be some linear combination of the predictor variables $(\mathbf{X}, \mathbf{Z})$. So generically, we would expect convergence of the forecasts, after at most $\min\{k, h\}$ rounds, to the full forecast based on the pooled information $(\mathbf{X}, \mathbf{Z})$. This has been shown by Dutta and Polemarchakis (2014). However, they did not give a numerical illustration, which we now supply.

We have made use of the 93CARS dataset (Lock 1993), containing information on new cars for the 1993 model year. There are $n = 82$ complete cases with information on 26 variables, including price, mpg ratings, engine size, body size, and other features. We took $\mathbf{X} = (X_1, \ldots, X_{11})$ to be the variables 7 to 17, $\mathbf{Z} = (Z_1, \ldots, Z_9)$ to be the variables 18 to 26, and Y to be variable 5 (Midrange Price).

Let S denote the uncorrected sum-of-squares-and-products matrix based on the data for these variables. The fictitious model we shall consider for the prediction game has $(\mathbf{X}, \mathbf{Z}, Y)$ multivariate normal, with mean $\mathbf{0}$ and dispersion matrix $\Sigma = S/n$. The predictive distribution of Y, based on any collection of linear transforms of X's and Z's, will then be normal, with a mean formula that can be computed by running the zero-intercept sample linear regression of Y on those variables, and variance that will not depend on the values of the predictors. Note that, although our calculations are based on the sample data, the values computed are not estimates but are the correct values for our fictitious model.

Let U_1 be the variable so obtained from the sample regression Y on $\mathbf{X} \equiv (X_1, \ldots, X_{11})$. Recall that both experts are supposed to know the model, hence Σ, and know which variables each is observing. Consequently, both know the form of U_1, but initially only E_1, who knows the values of $(X_1, \ldots, X_{11})$, can compute its value, u_1 say. Since his round-1 forecast for Y is normal with mean u_1, while its variance is already computable by both experts, the effect of E_1 issuing his forecast is to make the value u_1 of U_1 public knowledge.

It is now E_2's turn to play. At this point, she knows the values of U_1 and $(Z_1, \ldots, Z_9)$, and her forecast is thus obtained from the sample regression of Y on these variables. Let this regression function (computable by both experts) be V_1; then at this round E_2 effectively makes the value v_1 of V_1 public.

Now at round 2, E_1 regresses Y on $(X_1, \ldots, X_{11}, V_1)$ (U_1, which is a linear function of his privately known X's, being redundant) and announces the value u_2 of the computed regression function U_2; and so on.

The relevant computations are easy to conduct using the statistical software package `R` (R Development Core Team 2011). At each stage, we compute the 82 fitted values based on the regression just performed. These can then be used as values for the new predictor variable to be included in the next regression. Moreover, convergence of the forecast sequence will be reflected in convergence of these fitted values.

As a numerical illustration, suppose E_1 has observed

$$\mathbf{X} = \mathbf{x} = (16, 25, 2, 1, 8, 4.6, 295, 6000, 1985, 0, 20.0),$$

and E_2 has observed

$$\mathbf{Z} = \mathbf{z} = (5, 204, 111, 74, 44, 31.0, 14, 3935, 1).$$

Before entering the prediction market, E_1's point forecast for Y, based on his data $\mathbf{X} = \mathbf{x}$, is $u_1 = 40.6163$, and E_2's point forecast for Y, based on her data $\mathbf{Z} = \mathbf{z}$, is $v_0 = 30.6316$. If they could combine their data, the forecast, based on $(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}, \mathbf{z})$, would be 39.73925.

On entering the market, the sequence of their predictions is as given in Table 6.1. We observe convergence, both for the fitted values and the predicted standard deviations, from round 10 onwards. As soon as E_1 has access to the values of $U_1, \ldots, U_9$, he effectively knows $Z_1, \ldots, Z_9$, and his forecast becomes the same as that based on the pooled data. And as soon as E_1 makes that public, E_2 can make the same forecast. The predictions of both experts will remain the same thereafter.

Table 6.1 Sequence of market predictions for Y

i:	1	2	3	4	5	6	7	8	9	10	...
u_i:	40.62	39.49	39.34	39.51	39.55	39.54	39.66	39.75	39.73917	39.73925	...
v_i:	38.28	39.40	39.46	39.54	39.56	39.63	39.67	39.74	39.73924	39.73925	...

As a second illustration, suppose E_1 has observed

$$\mathbf{X} = \mathbf{x} = (22, 30, 1, 0, 4, 3.5, 208, 5700, 2545, 1, 21.1),$$

and E_2 has observed

$$\mathbf{Z} = \mathbf{z} = (4, 186, 109, 69, 39, 27.0, 13, 3640, 0).$$

Before entering the prediction market, E_1's point forecast for Y is 27.80968, and E_2's point forecast is 36.593865. Their market forecasts converge at round 10 to 31.22983, the forecast based on all the data.

These two examples illustrate within-sequence convergence, to a data-dependent limit. □

Limited Consensus

In all the above examples, convergence was either to a vacuous state or to a complete consensus based on the totality of the pooled private information. As the following example shows, it is also possible to converge to an intermediate state.

Example 5 Suppose θ and X_1 have independent $N(0, 1)$ distributions, while, given (θ, X_1), $X_2 \sim N(\theta X_1, 1)$. Expert E_1 observes $H_1 = X_1$, while E_2 observes $H_2 = X_2$. The interest is in predicting θ. A sufficient statistic for θ, based on the combined data (X_1, X_2), is $(X_1 X_2, |X_1|) = (S_1, S_2)$, say. The posterior distribution is

$$\theta \mid (S_1, S_2) = (s_1, s_2) \sim N\left(\frac{s_1}{1 + s_2^2}, \frac{1}{1 + s_2^2}\right).$$

Straightforward computations deliver the joint density of (X_1, X_2), marginalising over θ:

$$f(x_1, x_2) = (2\pi)^{-1}(1 + x_1^2)^{-\frac{1}{2}} \exp -\frac{1}{2}\left(x_1^2 + \frac{x_2^2}{1 + x_1^2}\right). \tag{6.14}$$

Because (6.14) is unchanged if we change the sign of either or both of x_1 and x_2, we deduce (what may be obvious from the symmetry of the whole setup):

Proposition 6 *Conditionally on* $|X_1|$ *and* $|X_2|$, $\text{sign}(X_1)$ *and* $\text{sign}(X_2)$ *behave as independent fair coin flips.*

At the first round, E_1 declares his posterior for θ, based on X_1—but, since $X_1 \perp\!\!\!\perp \theta$ this supplies no information at all about θ. (So we would get the same

answer if E_2 were to go first—the order in which they announce their opinions does not matter.)

Now E_2 goes. Since $X_2 \mid \theta \sim N(0, 1+\theta^2)$, with sufficient statistic $|X_2|$, E_2 is effectively putting $|X_2|$ into the public pot.

At the start of round 2, E_1 knows X_1 and $|X_2|$. By Proposition 6, $\text{sign}(X_2)$ is still equally likely to be 1 or -1. So E_1 knows S_2 but only knows $|S_1|$—for him, S_1 is either $|S_1|$ or $-|S_1|$, each being equally likely. His posterior is thus a 50–50 mixture of the associated posteriors

$$N\left(\frac{|S_1|}{1+S_2^2}, \frac{1}{1+S_2^2}\right)$$

and

$$N\left(\frac{-|S_1|}{1+S_2^2}, \frac{1}{1+S_2^2}\right).$$

On E_1's now announcing this mixture posterior, he is effectively communicating $(|S_1|, S_2) \equiv (|X_1| \times |X_2|, |X_1|)$. The total information in the public pot is thus now equivalent to $(|X_1|, |X_2|)$.

It is now E_2's turn again. At this point, she knows $(|X_1|, X_2)$, so $(|S_1|, S_2)$—but still does not know $\text{sign}(S_1)$, which again behaves as a coin flip. Her forecast distribution is thus exactly the same as E_1's. So we get convergence to the above mixture posterior at round 2. But this limiting forecast is not the same as that based on the pooled data, which would be the relevant single component of the mixture.

Note that, at convergence, the pool of public knowledge is $(|X_1|, |X_2|)$. Since θ has the identical mixture posterior whether conditioned on $(|X_1|, |X_2|)$, on $(X_1, |X_2|)$, or on $(|X_1|, X_2)$, we have both $\theta \perp\!\!\!\perp X_1 \mid (|X_1|, |X_2|)$ and $\theta \perp\!\!\!\perp X_2 \mid (|X_1|, |X_2|)$, in accordance with Proposition 5. □

It might appear that the above behaviour is highly dependent on the symmetry of the problem, but this is not so. As the following analysis shows, the same limited consensus behaviour arises on breaking the symmetry.

Example 6 Consider the same problem as in Example 5 above, with the sole modification that the prior distribution of θ is now $N(\mu, 1)$, where μ is non-zero. The posterior distribution of θ, based on the full data (X_1, X_2) or its sufficient statistic (S_1, S_2), is now

$$\theta \mid (S_1, S_2) = (s_1, s_2) \sim \Pi(s_1, s_2) := N\left(\frac{\mu + s_1}{1+s_2^2}, \frac{1}{1+s_2^2}\right).$$

The following result is immediate.

Proposition 7 *Given only* $|S_1| = m_1, S_2 = m_2$, *the posterior distribution is a mixture:*

$$\theta \sim M(m_1, m_2) = \pi(1)\Pi(m_1, m_2) + \pi(-1)\Pi(-m_1, m_2), \tag{6.15}$$

where

$$\pi(j) = \Pr(\text{sign}(S_1) = j \mid |S_1| = m_1, S_2 = m_2) \quad (j = \pm 1). \tag{6.16}$$

Proposition 8 *Conditionally on* $|X_1|$ *and* $|X_2|$*:*

(i) $\text{sign}(X_1) \perp\!\!\!\perp \text{sign}(X_1X_2)$
(ii) $\text{sign}(X_2) \perp\!\!\!\perp \text{sign}(X_1X_2)$

Proof (i) The joint density of (X_1, X_2), marginalising over θ, is

$$f(x_1, x_2) = (2\pi)^{-1}(1 + x_1^2)^{-\frac{1}{2}} \exp -\frac{1}{2}\left(x_1^2 + \frac{(x_2 - \mu x_1)^2}{1 + x_1^2}\right).$$

This is unchanged if we change the signs of both x_1 and x_2. Consequently, given $|X_1| = m_1, |X_2| = m_2$, $\Pr(X_1 = m_1, X_2 = m_2) = \Pr(X_1 = -m_1, X_2 = -m_2)$, while $\Pr(X_1 = m_1, X_2 = -m_2) = \Pr(X_1 = -m_1, X_2 = m_2)$. But this is equivalent to

$$\Pr(\text{sign}(X_1) = 1, \text{sign}(X_1X_2) = 1) = \Pr(\text{sign}(X_1) = -1, \text{sign}(X_1X_2) = 1)$$

$$\Pr(\text{sign}(X_1) = 1, \text{sign}(X_1X_2) = -1) = \Pr(\text{sign}(X_1) = -1, \text{sign}(X_1X_2) = -1).$$

Thus, $\Pr(\text{sign}(X_1) = 1 \mid \text{sign}(X_1X_2) = 1) = \Pr(\text{sign}(X_1) = 1 \mid \text{sign}(X_1X_2) = -1) = \frac{1}{2}$, which in particular implies $\text{sign}(X_1) \perp\!\!\!\perp \text{sign}(X_1X_2)$.

(ii) We have

$$\Pr(\text{sign}(X_2) = 1 \mid \text{sign}(X_1X_2) = 1) = \Pr(\text{sign}(X_1) = 1 \mid \text{sign}(X_1X_2) = 1)$$

$$\Pr(\text{sign}(X_2) = 1 \mid \text{sign}(X_1X_2) = -1) = \Pr(\text{sign}(X_1) = -1, \text{sign}(X_1X_2) = -1).$$

So from (i), conditional on $|X_1| = m_1, |X_2| = m_2$, $\Pr(\text{sign}(X_2) = 1 \mid \text{sign}(X_1X_2) = 1) = \Pr(\text{sign}(X_2) = 1 \mid \text{sign}(X_1X_2) = -1) = \frac{1}{2}$ so that, in particular, $\text{sign}(X_2) \perp\!\!\!\perp \text{sign}(X_1X_2)$. □

In the first round, E_1 and E_2 behave exactly as before, and again, at the start of round 2, the public pot contains $|X_2|$. So now E_1 knows X_1 and $|X_2|$. In terms of the sufficient statistic, he knows $(|S_1|, S_2)$ but does not know $\text{sign}(S_1)$. Moreover, by Proposition 8(i), his additional knowledge of $\text{sign}(X_1)$ contains no relevant further information about $\text{sign}(S_1)$. Consequently, he will compute and announce

the mixture posterior $M(|S_1|, S_2)$. From this, it is possible to deduce the values of $|S_1|$ and S_2. Hence, at this point, the public pot contains $(|S_1|, S_2)$.

Now E_2 knows $(|S_1|, S_2)$ but is still ignorant of $\text{sign}(S_1)$. And again, although she has the additional knowledge of $\text{sign}(X_2)$, by Proposition 8(ii), this contains no relevant further information about $\text{sign}(S_1)$. Consequently, E_2 will have the same posterior distribution $M(|S_1|, S_2)$, which will be the final (but limited) consensus.

(Note that an essentially identical analysis will hold with any prior distribution for θ.) □

6.5 Discussion

Probability forecasts take explicit account of the uncertainty concerning an unknown quantity or event. We have described three important tools for motivating and assessing the performance of a single forecaster. A proper scoring rule induces the forecaster to give honest predictions and can also be used to evaluate performance after the event. For forecasts made for a sequence of events, calibration measures success in quantifying uncertainty. Resolution measures how close calibrated forecasts come to actual outcomes, and thus reflects expertise in the subject area. Refinement is a relation between the resolutions of different forecasters, which is useful for comparing them.

When there are multiple expert forecasters, you require a method for combining their forecasts. There is strong empirical evidence that probability forecasts suitably combining all the experts' opinions generally result in better predictive performance—this is similar to the case of Bayesian model averaging, a coherent mechanism for accounting for model uncertainty that improves predictive performance. In a Bayesian approach to combining experts' opinions, the decision-maker models the experts' opinions and combines them with his/her own prior opinion, and any additional data he/she may have, using Bayes's theorem. An alternative axiomatic approach imposes constraints that a combination formula is required to satisfy. We have described in detail an approach, based on a specific understanding of "expertise", which imposes only coherence constraints. For the linear opinion pool, the most popular method for combining probability forecasts, coherence requires that not all weights are strictly positive.

We have given a detailed account of prediction markets, with special attention to the case where two experts take it in turn to update their probability of a future event, conditioning only on the revealed probabilities of the other. We have displayed a variety of behaviours for such a process. There will always be convergence to a limiting value, but this may or may not be the same as what could be achieved if the experts were able to pool all their private information.

We have supposed that, although each expert may be unaware of the private information held by the other, he does at least know which variables the other expert knows—though not their values. When even this cannot be assumed, there will be much greater freedom for an expert to update his own probability on the basis of

the revealed probabilities of the other. Nevertheless, this freedom is restricted. The theory of Dawid et al. (1995) relates to combining the announced probabilities of a number of experts without necessarily knowing the private variables on which these are based. It would be challenging, but valuable, to extend this to the sequential case.

References

Aldous, D. J. (2013). Using prediction market data to illustrate undergraduate probability. *American Mathematical Monthly*, 120:583–593.

Arrow, K. J. (1951). *Social Choice and Individual Values*. J. Wiley, New York.

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Ledyard, J. O., Levmore, S., Litan, R., Milgrom, P., Nelson, F. D., Neumann, G. R., Ottaviani, M., Schelling, T. C., Shiller, R. J., Smith, V. L., Snowberg, E., Sunstein, C. R., Tetlock, P. C., Tetlock, P. E., Varian, H. R., Wolfers, J., & Zitzewitz, E. (2008). The promise of prediction markets. *Science*, 320:877–878.

Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics*, 4:1236–1239.

Berger, J. O. & Mortera, J. (1991). Bayesian analysis with limited communication. *Journal of Statistical Planning and Inference*, 28:1–24.

Blackwell, D. (1951). Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (pp. 93–102), Berkeley, Calif. University of California Press.

Bradley, R. (2018). Learning from others: conditioning versus averaging. *Theory and Decision*, 85:5–20.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Chen, Y., Dimitrov, S., Sami, R., Reeves, D., Pennock, D., Hanson, R., Fortnow, L., & Gonen, R. (2010). Gaming prediction markets: Equilibrium strategies with a market maker. *Algorithmica*, 58:930–969.

Clemen, R. T. & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19:187–203.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77:605–610.

Dawid, A. P. (1986). Probability forecasting. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), *Encyclopedia of Statistical Sciences*, volume 7 (pp. 210–218). Wiley-Interscience.

Dawid, A. P., DeGroot, M. H., & Mortera, J. (1995). Coherent combination of experts' opinions (with Discussion). *TEST*, 4:263–314.

Dawid, A. P. & Mortera, J. (2020). Resolving some contradictions in the theory of linear opinion pools. *Theory and Decision*, 88:453–456.

de Finetti, B. (1954). Media di decisioni e media di opinioni. *Bulletin of the International Statistical Institute*, 24:144–157.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, 69:118–121.

DeGroot, M. H. (1988). A Bayesian view of assessing uncertainty and comparing expert opinion. *Journal of Statistical Planning and Inference*, 20:295–306.

DeGroot, M. H. & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32:12–22.

DeGroot, M. H. & Mortera, J. (1991). Optimal linear opinion pools. *Management Science*, 37:546–558.

Dutta, J. & Polemarchakis, H. (2014). Convergence to agreement and the role of public information. *Mathematics and Financial Economics*, 8(4):399–404.

French, S. (1986). Calibration and the expert problem. *Management Science*, 32:315–321.

Geanakoplos, J. D. (1992a). Common knowledge. In *Proceedings of the 4th Conference on Theoretical Aspects of Reasoning About Knowledge* (pp. 254–315), San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Geanakoplos, J. D. (1992b). Common knowledge. *Journal of Economic Perspectives*, 6:53–82.

Geanakoplos, J. D. & Polemarchakis, H. M. (1982). We can't disagree forever. *Journal of Economic Theory*, 32:192–200.

Genest, C. & Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography (with Discussion). *Statistical Science*, 1:114–148.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B*, 14:107–114.

Hanson, R. (2003). Combinatorial information market design. *Information Systems Frontiers*, 5:105–119.

Hanson, R., Oprea, R., & Porter, D. (2006). Information aggregation and manipulation in an experimental market. *Journal of Economic Behavior and Organization*, 60:449–459.

Laffont, J.-J. (1979). *Aggregation and Revelation of Preferences*. Elsevier North-Holland, New York.

Lindley, D. V. (1983). Reconciliation of probability distributions. *Operations Research*, 31:866–880.

Lock, R. H. (1993). The 1993 new car data. *Journal of Statistics Education*, 1. http://www.amstat.org/publications/jse/v1n1/datasets.lock.html.

Luce, R. D. & Raiffa, H. (1958). *Games and Decisions: Introduction and Critical Survey*. Wiley, New York.

Marschak, J. & Radner, R. (1972). *Economic Theory of Teams*. Yale University Press.

McConway, K. J. (1981). Marginalization and linear opinion pools. *Journal of the American Statistical Association*, 76:410–414.

McKelvey, R. D. & Page, T. (1986). Common knowledge, consensus, and aggregate information. *Econometrica*, 54:109–127.

Nielsen, L. T. (1984). Common knowledge, communication, and convergence of beliefs. *Mathematical Social Sciences*, 8:1–14.

Pill, J. (1971). The Delphi method: Substance, context, a critique and an annotated bibliography. *Socio-Economic Planning Sciences*, 5:57–71.

Ranjan, R. & Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:71–91.

Stone, M. (1961). The opinion pool. *The Annals of Mathematical Statistics*, 32:1339–1342.

Strähl, C. & Ziegel, J. F. (2015). Cross-calibration of probabilistic forecasts. arXiv:1505.05314.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Weerahandi, S. & Zidek, J. V. (1981). Multi-Bayesian statistical decision theory. *Journal of the Royal Statistical Society: Series A (General)*, 144:85–93.

Winkler, R. L. (1981). Combining probability distributions from dependent information sources. *Management Science*, 27:479–488.

Wolfers, J. & Zitzewitz, E. (2008). Prediction markets in theory and practice. In L. Blume & S. Durlauf (Eds.), *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, London, 2nd edition.

Chapter 7
Determinantal Priors for Variable Selection

Veronika Ročková and Edward I. George

7.1 Steve the Bayesian

It is with deep appreciation and admiration that we dedicate this contribution to Steve Fienberg. The wide variety of Steve's broad contributions to the theory, methodology, and application of Bayesian analysis were remarkable in how they anticipated so many different areas of flourishing Bayesian research today. His many prescient contributions included the early introduction and development of Bayesian methods for fundamental statistical problems such as latent root analysis and sparse multinomial cell probability estimation, for novel applications such as data confidentiality protection, for disability measurement in elderly populations and legal proceedings frameworks, and for modern machine learning approaches such as mixed membership classification analysis, to name but a few [1]. It is especially notable that in spite of Steve's impressive attention to the foundations and historical evolution of Bayesian analysis, he never let subjective purity get in the way of using whatever kind of Bayesian machinery and thinking would further the statistical goals of the problem at hand. His work exemplified the broad potential of Bayesian analysis at its best. Moved by this spirit, our contribution introduces new Bayesian machinery for tackling the fundamental problem of mitigating unwanted multicollinearity in Bayesian variable selection.

V. Ročková (✉)
Booth School, University of Chicago, Chicago, IL, USA
e-mail: Veronika.Rockova@chicagobooth.edu

E. I. George
Wharton, University of Pennsylvania, Philadelphia, PA, USA
e-mail: edegorge@wharton.upenn.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_7

7.2 Bayesian Variable Selection with Spike-and-Slab Priors

Suppose observations on $\boldsymbol{y}$, an $n \times 1$ response vector, and $\boldsymbol{X} = [x_1, \ldots, x_p]$, an $n \times p$ matrix of p potential standardized predictors, are related by the Gaussian linear model

$$f(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma) = \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathrm{I}_n), \tag{7.1}$$

where $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$ is a $p \times 1$ vector of unknown regression coefficients and σ is an unknown positive scalar. (We assume throughout that $\boldsymbol{y}$ and the x's have been centered at zero to avoid the need for an intercept).

A fundamental Bayesian approach to variable selection for this setup is obtained with a hierarchical spike-and-slab Gaussian mixture prior on $\boldsymbol{\beta}$ [2]. Introducing a latent binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$, $\gamma_i \in \{0, 1\}$, each component of this mixture prior is defined conditionally on σ and $\boldsymbol{\gamma}$ by

$$\pi(\boldsymbol{\beta} \mid \sigma, \boldsymbol{\gamma}) = \mathrm{N}_p(\boldsymbol{0}, \sigma^2 \boldsymbol{D}_{\boldsymbol{\gamma}}), \tag{7.2}$$

where

$$\boldsymbol{D}_{\boldsymbol{\gamma}} = \mathrm{diag}\{[(1-\gamma_1)v_0 + \gamma_1 v_1], \ldots, [(1-\gamma_p)v_0 + \gamma_p v_1]\} \tag{7.3}$$

for $0 \leq v_0 < v_1$. Adding a relatively noninfluential prior on σ^2 such as the inverse gamma prior $\pi(\sigma^2) = \mathrm{IG}(\nu/2, \nu\lambda/2)$ with $\nu = \lambda = 1$, the mixture prior is then completed with a prior distribution $\pi(\boldsymbol{\gamma})$ over the 2^p possible values of $\boldsymbol{\gamma}$.

By suitably setting v_0 small and v_1 large in (7.3), β_i values under $\pi(\boldsymbol{\beta} \mid \sigma, \boldsymbol{\gamma})$ are more likely to be small when $\gamma_i = 0$ and more likely to be large when $\gamma_i = 1$. Thus variable selection inference can be obtained from the posterior $\pi(\boldsymbol{\gamma} \mid y)$ induced by combining this prior with the data $\boldsymbol{y}$. For example, one might select those predictors corresponding to the $\gamma_i = 1$ components of the highest posterior probability $\boldsymbol{\gamma}$.

The explicit introduction of the intermediate latent vector $\boldsymbol{\gamma}$ in the spike-and-slab mixture prior allows for the incorporation of available prior information through the prior specification of $\pi(\boldsymbol{\gamma})$. This can be conveniently done by using hierarchical specifications of the form

$$\pi(\boldsymbol{\gamma}) = \mathsf{E}_{\pi(\boldsymbol{\theta})} \pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta}), \tag{7.4}$$

where $\boldsymbol{\theta}$ is a (possibly vector) hyperparameter with prior $\pi(\boldsymbol{\theta})$.

In the absence of structural information about the predictors, i.e., when their inclusion is a priori exchangeable, a useful default choice for $\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta})$ is the i.i.d. Bernoulli prior form

$$\pi^B(\boldsymbol{\gamma} \mid \theta) = \theta^{q_{\boldsymbol{\gamma}}}(1-\theta)^{p-q_{\boldsymbol{\gamma}}}, \tag{7.5}$$

where $\theta \in [0, 1]$ and $q_{\boldsymbol{\gamma}} = \sum_i \gamma_i$. Because this $\pi(\boldsymbol{\gamma} \mid \theta)$ is a function only of model size $q_{\boldsymbol{\gamma}}$, any marginal $\pi(\boldsymbol{\gamma})$ in (7.4) will be of the form

$$\pi^B(\boldsymbol{\gamma}) = \pi^B_{\pi(\theta)}(q_{\boldsymbol{\gamma}})\, \pi^B(\boldsymbol{\gamma} \mid q_{\boldsymbol{\gamma}}), \quad \pi^B(\boldsymbol{\gamma} \mid q_{\boldsymbol{\gamma}}) = \binom{p}{q_{\boldsymbol{\gamma}}}^{-1} \tag{7.6}$$

where $\pi^B_{\pi(\theta)}(q_{\boldsymbol{\gamma}})$ is the prior on model size induced by $\pi(\theta)$, and $\pi^B(\boldsymbol{\gamma} \mid q_{\boldsymbol{\gamma}})$ is uniform over models of size $q_{\boldsymbol{\gamma}}$.

Of particular interest for this formulation has been the beta prior $\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$, $a, b > 0$, (7.5) which yields model size priors of the form

$$\pi^B_{a,b}(q_{\boldsymbol{\gamma}}) = \frac{\mathrm{Be}(a + q_{\boldsymbol{\gamma}}, b + p - q_{\boldsymbol{\gamma}})}{\mathrm{Be}(a, b)} \binom{p}{q_{\boldsymbol{\gamma}}}, \tag{7.7}$$

where $\mathrm{Be}(\cdot, \cdot)$ is the beta function. For the choice $a = b = 1$, under which $\theta \sim U(0, 1)$, this yields the uniform model size prior

$$\pi^B_{1,1}(q_{\boldsymbol{\gamma}}) \equiv \frac{1}{p+1}. \tag{7.8}$$

An attractive alternative is to choose a small and b large in order to be more effective for targeting sparse models in high-dimensions. For example, [3] shows that the choice $a = 1$ and $b = p$ yields optimal posterior concentration rates in sparse settings with $v_0 = 0$ and heavier-tailed Laplace priors for $\boldsymbol{\beta}$.

7.3 Determinantal Prior Formulations

The main thrust of this contribution is to propose new model space priors $\pi(\boldsymbol{\gamma})$ based on the hierarchical representation (7.4) with the conditional form

$$\pi^D(\boldsymbol{\gamma} \mid \theta) = \frac{\left|c_\theta \boldsymbol{X}_{\boldsymbol{\gamma}}{}' \boldsymbol{X}_{\boldsymbol{\gamma}}\right|}{|c_\theta X'X + I|} \propto |\boldsymbol{X}_{\boldsymbol{\gamma}}{}' \boldsymbol{X}_{\boldsymbol{\gamma}}|\, \theta^{q_\gamma} (1-\theta)^{p-q_\gamma}, \tag{7.9}$$

where $c_\theta = \frac{\theta}{1-\theta}$ and $\boldsymbol{X}_{\boldsymbol{\gamma}}$ is the $n \times q_{\boldsymbol{\gamma}}$ matrix of predictors identified by the active elements in $\boldsymbol{\gamma}$. The first expression for $\pi^D(\boldsymbol{\gamma} \mid \theta)$ reveals it to be a special case of a determinantal prior, as discussed below, while the second expression reveals it to be a reweighted version of the Bernoulli prior (7.5) as in [4]. Thus, this prior downweights the probability of γ for the predictor collinearity measured by the determinant $|\boldsymbol{X}_{\boldsymbol{\gamma}}{}' \boldsymbol{X}_{\boldsymbol{\gamma}}|$, which quantifies the volume of the space spanned by the selected predictors in the γth subset. Intuitively, sets of collinear predictors are less likely to be selected under this prior, due to ill conditioning of the correlation matrix. As will be seen, the use of $\pi^D(\boldsymbol{\gamma} \mid \theta)$ can provide cleaner posterior inference for

variable selection in the presence of multicollinearity, when the correlation between the columns of $\boldsymbol{X}$ makes it difficult to distinguish between predictor effects.

In general, a probability measure $\pi(\boldsymbol{\gamma})$ on the 2^p subsets of a discrete set $\{1, \ldots, p\}$, indexed by the binary indices $\boldsymbol{\gamma}$, is called a *determinantal point process* (DPP) if there exists a positive semidefinite matrix K, such that

$$\pi(\boldsymbol{\gamma}) = \det(K_{\boldsymbol{\gamma}}), \quad \forall \boldsymbol{\gamma}, \tag{7.10}$$

where $K_{\boldsymbol{\gamma}}$ is the restriction of K to the entries indexed by the active elements in $\boldsymbol{\gamma}$. The matrix K is referred to as a marginal kernel as its elements lead to the marginal inclusion probabilities and anti-correlations between the pairs of variables, i.e.

$$P(\gamma_i = 1) = K_{ii}\,; \qquad \mathsf{P}(\gamma_i = 1, \gamma_j = 1) = K_{ii}K_{jj} - K_{ij}K_{ji}.$$

Given any real, symmetric, positive semidefinite $p \times p$ matrix L, a corresponding DPP can be obtained via the L-ensemble construction

$$\pi(\boldsymbol{\gamma}) = \frac{\det(L_{\boldsymbol{\gamma}})}{\det(L + \mathrm{I})}, \tag{7.11}$$

where L_γ is the sub-matrix of L given by the active elements in $\boldsymbol{\gamma}$ and I is an identity matrix. That this is a properly normalized probability distribution follows from the fact that $\sum_{\boldsymbol{\gamma}} \det(L_{\boldsymbol{\gamma}}) = \det(L + \mathrm{I})$. The marginal kernel for the K-ensemble DPP representation (7.10) corresponding to this L-ensemble representation is obtained by letting $K = (L + \mathrm{I})^{-1}L$. The first expression for $\pi^D(\boldsymbol{\gamma} \mid \theta)$ in (7.9) can be now seen as a special case of (7.11) by letting $L = c_\theta \boldsymbol{X}'\boldsymbol{X}$ and $L_{\boldsymbol{\gamma}} = c_\theta \boldsymbol{X}_{\boldsymbol{\gamma}}{}'\boldsymbol{X}_{\boldsymbol{\gamma}}$.

Applying $\pi(\boldsymbol{\gamma}) = \mathsf{E}_{\pi(\boldsymbol{\theta})}\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta})$ to $\pi^D(\boldsymbol{\gamma} \mid \theta)$ with the beta prior $\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$, we obtain

$$\pi^D(\boldsymbol{\gamma}) = h_{a,b}(q_{\boldsymbol{\gamma}})|\boldsymbol{X}_{\boldsymbol{\gamma}}{}'\boldsymbol{X}_{\boldsymbol{\gamma}}|, \tag{7.12}$$

where

$$h_{a,b}(q_{\boldsymbol{\gamma}}) = \frac{1}{\mathrm{Be}(a,b)} \int_0^\infty |c\boldsymbol{X}'\boldsymbol{X} + I|^{-1} \frac{c^{q_{\boldsymbol{\gamma}}+a-1}}{(1+c)^{a+b}}\, dc. \tag{7.13}$$

Although not in closed form, $h_{a,b}(q_{\boldsymbol{\gamma}})$ is an easily computable one dimensional integral.

For comparison with the exchangeable beta-binomial priors $\pi^B(\boldsymbol{\gamma})$, it is useful to reexpress (7.12) as

$$\pi^D(\boldsymbol{\gamma}) = \pi^D_{\pi(\theta)}(q_{\boldsymbol{\gamma}})\, \pi^D(\boldsymbol{\gamma} \mid q_{\boldsymbol{\gamma}}), \tag{7.14}$$

where

$$\pi^D_{\pi(\theta)}(q_{\boldsymbol{\gamma}}) = W(q_{\boldsymbol{\gamma}})\, h_{a,b}(q_{\boldsymbol{\gamma}}), \quad \pi^D(\boldsymbol{\gamma} \mid q_{\boldsymbol{\gamma}}) = \frac{|\boldsymbol{X}_{\boldsymbol{\gamma}}{}' \boldsymbol{X}_{\boldsymbol{\gamma}}|}{W(q_{\boldsymbol{\gamma}})}, \quad W(q) = \sum_{q_{\boldsymbol{\gamma}}=q} |\boldsymbol{X}_{\boldsymbol{\gamma}}{}' \boldsymbol{X}_{\boldsymbol{\gamma}}|. \tag{7.15}$$

Thus, to generate $\boldsymbol{\gamma}$ from $\pi^D(\boldsymbol{\gamma})$ one can proceed by first generating the model size $q_{\boldsymbol{\gamma}} \in \{0, \ldots, p\}$from $\pi^D_{\pi(\theta)}(q_{\boldsymbol{\gamma}})$, and then generating $\boldsymbol{\gamma}$ conditionally from $\pi^D(\boldsymbol{\gamma} \mid q_{\boldsymbol{\gamma}})$. Note that the model size prior $\pi^D_{\pi(\theta)}(q_{\boldsymbol{\gamma}})$ may be very different from the beta-binomial prior $\pi^B_{\pi(\theta)}(q_{\boldsymbol{\gamma}})$. For example, it is not uniform when $a = b = 1$. Therefore, one might instead prefer, as is done in Sect. 7.5 below, to consider the alternative obtained by substituting a prior such as $\pi^B_{\pi(\theta)}(q_{\boldsymbol{\gamma}})$ for the first stage draw of $q_{\boldsymbol{\gamma}}$, but still use $\pi^D(\boldsymbol{\gamma} \mid q_{\boldsymbol{\gamma}})$ for the second stage draw of $\boldsymbol{\gamma}$ to penalize collinearity.

Lastly, note that the computation of the normalizing constant $W(q)$ can be obtained as a solution to Newton's recursive identities for elementary symmetric polynomials [5]. This is better seen from the relation

$$\sum_{q_{\boldsymbol{\gamma}}=q} |\boldsymbol{X}_{\boldsymbol{\gamma}}{}' \boldsymbol{X}_{\boldsymbol{\gamma}}| = e_q(\boldsymbol{\lambda}) := \sum_{q_{\boldsymbol{\gamma}}=q} \prod_{i=1}^{p} \gamma_i \lambda_i, \tag{7.16}$$

where $e_q(\boldsymbol{\lambda})$ is the qth elementary symmetric polynomial evaluated at $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_p\}$, the spectrum of $X'X$. Defining $p_q(\boldsymbol{\lambda}) = \sum_{i=1}^{p} \lambda_i^q$, the qth power sum of the spectrum, we can obtain normalizing constants $e_1(\boldsymbol{\lambda}), \ldots, e_p(\boldsymbol{\lambda})$ as solutions to the recursive system of equations

$$q\, e_q(\boldsymbol{\lambda}) = p_q(\boldsymbol{\lambda}) + \sum_{j=1}^{q-1} (-1)^{j-1} e_{q-j}(\boldsymbol{\lambda}) p_j(\boldsymbol{\lambda}). \tag{7.17}$$

7.4 Implementing Determinantal Priors with EMVS

EMVS [6] is a fast deterministic approach to identifying sparse high posterior models for Bayesian variable selection under spike-and-slab priors. In large high-dimensional problems where exact full posterior inference must be sacrificed for computational feasibility, deployments of EMVS can be used to find subsets of variables associated with the highest posterior modes. We here describe a variant of the EMVS procedure which incorporates the determinantal prior $\pi^D(\boldsymbol{\gamma} \mid \theta)$ in (7.9) to penalize predictor collinearity in variable selection.

At the heart of the EMVS procedure is a fast closed form EM algorithm, which iteratively updates the conditional expectations $\mathsf{E}[\gamma_i \mid \boldsymbol{\psi}^{(k)}]$, where here $\boldsymbol{\psi}^{(k)} = (\boldsymbol{\beta}^{(k)}, \sigma^{(k)}, \boldsymbol{\theta}^{(k)})$ denotes the set of parameter updates at the k^{th} iteration.

The determinantal prior induces dependence between inclusion probabilities so that conditional expectations cannot be obtained by trivially thresholding univariate directions.

With the determinantal prior $\pi^D(\boldsymbol{\gamma}\,|\theta)$, the joint conditional posterior distribution is

$$\pi(\boldsymbol{\gamma} \mid \boldsymbol{\psi}) \propto \exp\left(-\frac{\boldsymbol{\beta}\boldsymbol{D}_{\boldsymbol{\gamma}}\boldsymbol{\beta}}{2\sigma^2}\right)|\boldsymbol{D}_{\boldsymbol{\gamma}}|^{1/2}|c_\theta\, \boldsymbol{X}_{\boldsymbol{\gamma}}{}'\boldsymbol{X}_{\boldsymbol{\gamma}}|, \tag{7.18}$$

where $\boldsymbol{D}_{\boldsymbol{\gamma}} = \mathrm{diag}\{\gamma_i/v_1 + (1-\gamma_i)/v_0\}_{i=1}^p$. We can then write

$$\pi(\boldsymbol{\gamma} \mid \boldsymbol{\psi}) \propto \exp\left[-\frac{1}{2\sigma^2}\left(\frac{1}{v_1}-\frac{1}{v_0}\right)(\boldsymbol{\beta}\circ\boldsymbol{\beta})'\boldsymbol{\gamma}\right]|\boldsymbol{D}_{\boldsymbol{\gamma}}|^{1/2}c_\theta^{q_{\boldsymbol{\gamma}}}\,|\boldsymbol{X}_{\boldsymbol{\gamma}}{}'\boldsymbol{X}_{\boldsymbol{\gamma}}|, \tag{7.19}$$

where $\circ$ denotes the Hadamard product. The determinant $|\boldsymbol{D}_{\boldsymbol{\gamma}}|$ can be written as

$$|\boldsymbol{D}_{\boldsymbol{\gamma}}| = \exp\left\{\left[\log\left(\frac{1}{v_1}\right)-\log\left(\frac{1}{v_0}\right)\right]\boldsymbol{\gamma}'\mathbf{1} + p\log\left(\frac{1}{v_0}\right)\right\},$$

so that the joint distribution in (7.19) can be expressed as

$$\pi(\boldsymbol{\gamma} \mid \boldsymbol{\psi}) \propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\left(\frac{1}{v_1}-\frac{1}{v_0}\right)(\boldsymbol{\beta}\circ\boldsymbol{\beta}) - \log\left(\frac{v_0}{v_1}\right)\mathbf{1} - 2\log(c_\theta)\mathbf{1}\right]'\boldsymbol{\gamma}\right\}|\boldsymbol{X}_{\boldsymbol{\gamma}}{}'\boldsymbol{X}_{\boldsymbol{\gamma}}|.$$

Defining the $p \times p$ diagonal matrix

$$\boldsymbol{A}_{\boldsymbol{\psi}} = \mathrm{diag}\left\{\exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\left(\frac{1}{v_1}-\frac{1}{v_0}\right)\beta_i^2 - \log\left(\frac{v_0}{v_1}\right)\right] - 2\log(c_\theta)\right\}\right\}_{i=1}^p, \tag{7.20}$$

the exponential term above can be regarded as the determinant of $\boldsymbol{A}_{\boldsymbol{\gamma},\boldsymbol{\psi}}$, the $q_{\boldsymbol{\gamma}} \times q_{\boldsymbol{\gamma}}$ diagonal submatrix of $\boldsymbol{A}_{\boldsymbol{\psi}}$ whose diagonal elements correspond to the nonzero elements of γ.

It now follows that the determinantal prior is conjugate in the sense of yielding the updated determinantal form

$$\pi(\boldsymbol{\gamma} \mid \boldsymbol{\psi}) \propto |\boldsymbol{A}_{\boldsymbol{\gamma},\boldsymbol{\psi}}\, \boldsymbol{X}_{\boldsymbol{\gamma}}{}'\boldsymbol{X}_{\boldsymbol{\gamma}}|. \tag{7.21}$$

The marginal quantities from this distribution can be obtained by taking the diagonal of a matrix $\boldsymbol{K}_{\boldsymbol{\psi}} = (\boldsymbol{A}_{\boldsymbol{\psi}}\, \boldsymbol{X}'\boldsymbol{X} + \mathrm{I}_p)^{-1}\boldsymbol{A}_{\boldsymbol{\psi}}\, X'X$, namely

$$\mathsf{P}(\boldsymbol{\gamma}_i = 1 \mid \boldsymbol{\psi}) = [\boldsymbol{K}_{\boldsymbol{\psi}}]_{ii}. \tag{7.22}$$

7.5 Mitigating Multicollinearity with Determinantal Priors

In order to demonstrate the redundancy correction of the determinantal model prior we revisit the collinear example of [7] with $p = 15$ predictors. Under the uniform-on-model-size beta-binomial spike-and-slab prior, the pervasive collinearity here induces severe posterior multimodality, as displayed by the 32 768 posterior model probabilities in the upper plot Fig. 7.1. Models whose design matrix is "ill-conditioned", i.e. with smallest eigenvalue $\lambda_{min}(\boldsymbol{\gamma})$ of the gram matrix $L_{\boldsymbol{\gamma}}$ below 0.1, are designated in red. In contrast, the lower plot of Fig. 7.1 shows how the uniform-on-model-size determinantal spike-and-slab prior has penalized the many multicollinear submodels and put more posterior weight on submodels with less redundant covariate combinations, effectively reducing both posterior multimodality and entropy.

7.6 Discussion

As opposed to traditional beta-binomial spike-and-slab prior formulations that assign equal prior probability to subset models of the same size, determinantal spike-and-slab priors penalize subset models by reducing their prior probabilities

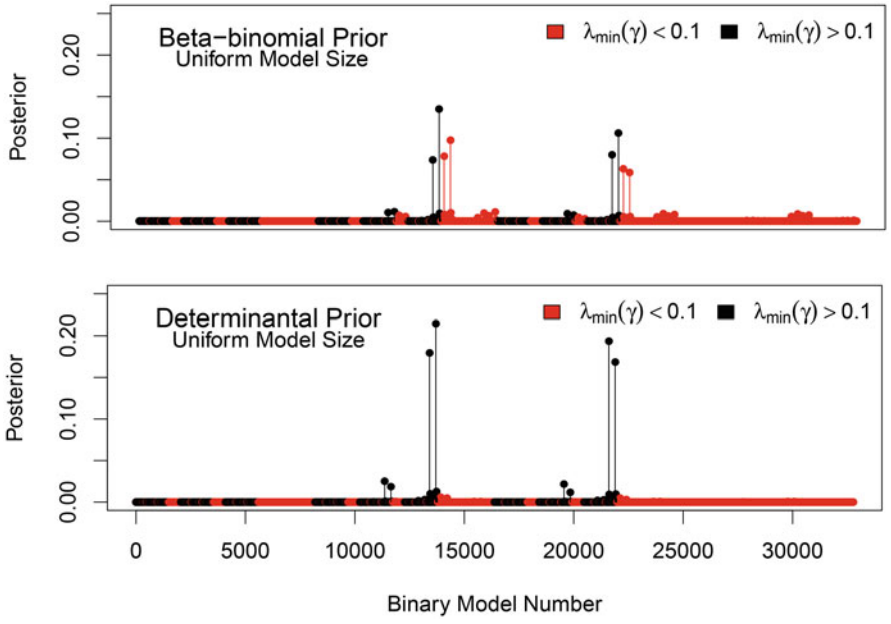


Fig. 7.1 Posteriors arising from beta-binomial and determinantal priors (both uniform-on-model-size)

according to their degree of predictor collinearity. From a practical standpoint, determinantal priors turn attention away from unwanted subset models by allocating more posterior probability to a smaller and more manageable set of interpretable submodels for the statistical analyst to consider. As so clearly demonstrated in Fig. 7.1, determinantal priors also serve to mitigate multimodality due to multicollinearity, thereby facilitating more productive posterior exploration via MCMC or EMVS. Finally, the generality of the determinantal prior formulation allows for its straightforward incorporation into many other Bayesian variable selection methods such as the spike-and-slab lasso and its many variants [8–10].

A preliminary version of this work was presented at the 47th Scientific Meeting of the Italian Statistical Society [11]. Also, an independent related development of these determinant prior ideas can be found in [12].

Acknowledgments This work was supported by the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business and by NSF grants DMS-1916245 and DMS-1944740.

References

1. George, E.I. (2013). Steve the Bayesian. *CHANCE*, 26(4), 16–17.
2. George, E. I. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.
3. Castillo, I. & van der Vaart, A. (2012). Needles and Straw in a Haystack: Posterior Concentration for Possibly Sparse Sequences, *Annals of Statistics*, 40, 2069–2101.
4. George, E.I. (2010). Dilution priors: Compensating for model space redundancy. In: *Borrowing Strength: Theory Powering Applications A Festschrift for Lawrence D. Brown*, IMS Collections, Vol. 6, 158–165.
5. Kulesza, A. & Taskar, B. (2013). Determinantal point processes for machine learning. ArXiv: 1207.6083.
6. Ročková, V. & George, E.I. (2014) EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association*, 109:506, 828–846.
7. George, E. I. & McCulloch, R.E. (1997). Approaches to Bayesian variable selection. *Statistica Sinica* 7 2 339–373.
8. Ročková, V. & George, E.I. (2018). The Spike-and Slab LASSO, *Journal of the American Statistical Association*, 113(521): 431–444.
9. Deshpande S.K., Rockova, V. and George E.I. (2019). Simultaneous Variable and Covariance Selection with the Multivariate Spike-and-Slab Lasso. Journal of Computational and Graphical Statistics, 28:4, 921–931.
10. Moran G.E., Ročková, V. and George E.I. (2019). Variance Prior Forms for High Dimensional Bayesian Variable Selection. *Bayesian Analysis*, Volume 14, Number 4, 1091–1119.
11. Ročková, V. and George, E.I. (2014). Determinantal Priors for Variable Selection. *Proceedings of the 47th Scientific Meeting of the Italian Statistical Society*, CUEC, Cagliari.
12. Kojima1, M. & Komaki, F. (2016). Determinantal point process priors for Bayesian variable selection in linear regression. *Statistica Sinica* 26, 97–117.

Part III
Introduction: Statistics and the Law

William F. Eddy

Decision-making is the heart of statistics and decision-making is the heart of the law. Naturally Steve was drawn to legal issues from his earliest days. He worked on many aspects of the two. One of his last big projects was the creation of the Center for Statistics and Applications in Forensic Evidence (CSAFE). Steve had been involved in the policy side of statistics and the law for a number of years; those efforts will not be reviewed here. In 2014 the NIST (National Institute of Standards and Technology) issued a call for proposals to create a large national research activity centered on the statistical aspects of forensics. Steve was ready. He formed a small group, Alicia Carriquiry of Iowa State, Hal Stern of the University of California Irvine, and Karen Kafadar of the University of Virginia, and developed a large proposal. He asked Alicia to be the principal investigator because of his health, but Steve remained the power behind the throne.

After many months of hard work, the proposal was submitted, and after many more months of review, the project was funded in 2015. The initial relationship between the NIST staff and the CSAFE academics was a little rocky; unlike many similar relationships, the principals were not well known to each other at first, and the form of the relationship (a cooperative agreement) was not well understood by either of the two parties. The kinks were worked out and the two groups have developed a strong relationship. Steve passed away in December 2016, and William Eddy of Carnegie Mellon was asked to replace him. In 2020 the project completed 5 years and CSAFE was renewed for another 5 years. The leadership group evolved, and the activities developed some new foci, but CSAFE has remained true to Steve's original vision.

The four papers in this section all focus on aspects of the legal problems in statistics. Jay Kadane discusses various personal views of particular cases. Hal Stern discusses the use of the likelihood function. That paper overlaps a bit with the work by Ommen and Saunders; they should be read together. Finally, Karen Kafadar discusses the use of eyewitness testimony in legal cases.

Jay Kadane was head of the statistics department when Steve was recruited to Carnegie Mellon in 1980. Jay was also one of those who had an abiding interest in legal issues and statistics. Shortly after Steve moved to Pittsburgh, the book

Statistics and the Law edited by Morris DeGroot, Steve, and Jay was published by Wiley (1986). That book is a very interesting collection of papers written by both statisticians and lawyers that run the gamut of applications.

Kadane's contribution to this volume is his personal view of statistical issues in the law supported by several examples of actual cases. The first of Kadane's examples is a paper by Finkelstein and Farley; the former is a lawyer and the latter is a statistician. It is perhaps the first paper to connect statistics and the law. Kadane decides that their example is "insufficiently Bayesian." Knowing Jay very well, I am not surprised by this decision. The next paper (by Tribe) is seen as a rebuttal to the first paper, and Kadane allows that court procedures should not allow statisticians to interfere with their proper functioning.

The remaining ten papers in Jay's review range in length from a single sentence reaffirming the initial Finkelstein and Fairley paper to several pages that Kadane offers on the Batson "motions." They are all quite interesting especially to statisticians who are not familiar with legal issues. I will only discuss one in detail, the last.

In Batson v. Kentucky, the Supreme Court ruled in 1986 that preemptory challenges could not exclude black members and created a three-step process to be followed by the courts. The steps are (1) a timely objection, identification of the subject of discrimination, and facts to raise inference of discrimination; (2) the other party has to explain why they were not discriminatory challenges; and (3) the court has to decide if step 2 is persuasive. Kadane considers or has participated in the appeals process in several cases. In each case he summarizes the "data" by what is effectively a 2×2 table. I put data in quotes because the relevance of the data to each specific case is variable. Kadane provides an analysis of the data in each case. I note that he chooses in at least one case to provide a Bayesian analysis based on a uniform prior distribution. As a non-Bayesian, I find that a kind of "cop-out." In the paper below (Ommen and Saunders) which received obvious input from Kadane, they argue for a personal Bayesian analysis which, as I understand it, does not allow for "default" priors but requires the fact finder to use their "real" prior.

Hal Stern has contributed a fascinating look at the statistical aspects of using forensic evidence.

Hal was one of the original partners in CSAFE and had decades of experience analyzing forensic evidence. Hal begins with a brief discussion of the Frye standard (the 1923 court decision which in some US states still governs the use of expert witness testimony). In Federal courts and some states, this has been superseded by the 1993 Daubert decision; that lays out in some detail a set of rules that must be followed to properly use such evidence.

In Sect. 9.3 Stern discusses Daubert and the 2016 PCAST report and the use of so-called black box studies. He does note that such studies produce a single overall error rate and cannot produce such a rate for every conceivable study. In the same section, Stern discusses the "two-stage" approach. He does point out the loss of information that results from a yes/no decision and mentions the 2019 report that recommended abandonment of statistical significance tests.

He then begins the discussion on the likelihood ratio, and while he refers to it as the Bayes factor, he does not make a distinction. The next paper in this part of the book, by Ommen and Saunders, is devoted to this distinction. Hal discusses the single-source DNA problem where the likelihood ratio approach works; he goes on to discuss trace evidence, where it might work, and then moves on to pattern evidence, where it hardly works. Towards the end of the section, Stern reviews the pluses and minuses of likelihood ratios. The interested reader should compare this with the discussion in Ommen and Saunders below.

The difficult task of explaining the difference between likelihood ratios and Bayes factors in the context of forensics has been taken on by Ommen and Saunders. This is already a difficult problem for statisticians, restricting the playing field to forensics makes it more difficult. In Sect. 10.2 the authors clarify that Bayes rule is simply a method of calculating probabilities that is a generally accepted statistical tool; it is distinct from the subjective interpretation of probabilities that is generally referred to as Bayesian and is an essential component of the Bayesian paradigm that they are arguing for. In Sect. 10.3 they describe that paradigm, and in Sect. 10.4 they describe an alternative called the likelihood paradigm.

There are a couple of points that are worth a little elaboration. First, there is still not general agreement in the statistics community that the Bayesian paradigm is the "correct" way to go, although the intellectual wars between Bayesian adherents and opponents that went on 40 or 50 years ago have subsided. That subsidence is due in part to the current availability of numerical computational tools (that did not exist back then) that have made it possible for potential users to apply the paradigm. Back then adherents were reduced to saying "This is what you should do (but I recognize you'll have to approximate)." Second, use of the likelihood paradigm seems a bit like a failure of the scientific method: Can't you scientists simply decide? The Bayesian paradigm using a subjective analysis has allowed each analyst to produce her own result; the likelihood paradigm has pushed the final decision past the end of the process.

Sections 10.5 provides more discussion with Sect. 10.6 emphasizing similarities and Sect. 10.7 emphasizing differences. This short paper is not going to resolve any differences and is not going to produce a "final" decision. The authors have written a clear discussion of the issues that will help inform individual readers about their choices. Ultimately there can be no resolution of the "debate" because the solution requires participants to agree on a set of rules that they may not (or may not be allowed to) agree on; essentially "we allow subjective opinions" or "we do not allow subjective opinions."

The contribution of Karen Kafadar to this volume is an interesting and very thorough review of eyewitness testimony as forensic evidence.

Karen was one of Steve's original members of CSAFE, and Karen was a member of the National Academy panel which authored the 2015 report, and she, obviously, participated in the "response to review" and subsequent discussion which led to the revised report (which was ultimately published). This role put her in a unique position to develop a detailed statistical understanding of the process of

eyewitness identification and how serious scientists respond to serious criticism of the identification process.

Karen begins with a discussion of the cognitive scientific models of the eyewitness identification process. Then in Sect. 11.4 she provides a very detailed discussion of the factors that enter a formal statistical model of eyewitness witness identification. The factors are broken into the group she calls signal which a statistician would probably model as "fixed" effects and the group she calls noise which a statistician would probably as "random" effects. The discussion could almost lead the reader to write down the model complete with Greek letters and plus signs between the various terms and perhaps even imagine writing down prior distributions for the unknown parameters.

In Sect. 11.5 she turns to what might be the single most important effect, the distinction between "sequential" presentation of the images and "simultaneous" presentation of the images. After years of neglect, scientists are turning to its study.

Chapter 8
Bayes and the Law

J. B. Kadane

Steve Fienberg and I shared both an interest in the law and an attraction to Bayesian statistics. Just what that relationship is has been discussed over the years by many. The purpose of this paper is to review some of the major contributions to thought on this subject and to raise some attendant questions.

8.1 Making Optimal Decisions: A Review of the Basics

No Uncertainty

When there is no uncertainty, the advice sounds trivial: do your best. What does that mean? First, think creatively about the choices you have. Often in life decisions are framed too narrowly; there may be good options the decision-maker failed to consider. The only options to be considered are those that the decision-maker is prepared to contemplate. Second, think hard about what you are trying to achieve. What are your goals? Often there are many attributes of decisions to consider. How do you trade off those attributes against each other? A formal way of expressing how you feel about the attributes (jointly) and the decisions is in a utility function, a function of both the attributes and the decisions. Such a function is usually scaled so that higher is better. Then formally the recommendation is to maximize the utility function, substituting the attribute values for each decision. Much of microeconomics uses this paradigm. Note that any increasing function of such a utility function results in the same optimal decision.

J. B. Kadane (✉)
Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: kadane@andrew.cmu.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_8

It is important to understand that the decisions made in this way are "good" only in the sense that they reflect what the decision-maker is trying to accomplish. Depending on the taste of the decision-maker, these choices may be ethical or not, may be wise or not, etc. The method is neutral with respect to such considerations.

This theory is intended to be a normative theory, not a descriptive one. That is, it is not proposed that people actually make decisions this way, but rather that if they did, they would make decisions that they would be more satisfied with.

With Uncertainty

Where there is uncertainty that importantly enters the picture, that is, situations in which the best decision would be different were the uncertainty resolved, a different analysis results. Now the opinion of the decision-maker about the uncertain events matters. There are now two inputs to the analysis, a probability distribution reflecting the beliefs of the decision-maker and a utility function reflecting the goals of the decision-maker. The recommendation then is to make a choice that maximizes expected utility, where the expectation reflects the probabilities of the decision-maker. While it goes by the same name, this utility function is different in kind from the utility function described in 8.1 above, as this one is invariant only with respect to positive linear transformations.

Note that again this theory is normative and not descriptive and that it is ethically neutral. Also note that the probabilities and utilities are those of the decision-maker. That other people might disagree with the choices of probability, and utility made by the decision-maker is irrelevant. The only party whose views are reflected are those of the decision-maker.

When Data May Become Available

In this case, the decision-maker is hypothesized to have a joint probability distribution over the data and the other uncertain quantities. Then, Bayes Theorem, which is a simple consequence of the axioms of probability, applies to permit the calculation of the posterior distribution of the remaining uncertain quantities, given what has been learned from the data. (This is where the name "Bayesian" arises from.) Then, the posterior distribution is used to maximize expected utility. The only party whose views are reflected are those of the decision-maker.

Often the joint distribution of the data and the uncertain quantities is expressed as the product of two factors: the distribution of the data given the uncertain quantities and the marginal distribution of the uncertain quantities. The former is called the likelihood function, and the latter the prior distribution. Both reflect the opinion of the decision-maker only. Once again, the theory is ethically neutral and normative.

There is an extensive literature explaining and justifying this treatment of decision-making, particularly under uncertainty and with data. My version can be found in Kadane (2011).

8.2 Relation to the Law

(1) Finkelstein and Fairley (1970) is one of the first expositions to call for the explicit use of Bayesian ideas in court. It came at a time when the full implications of Bayesian ideas were still being worked out. With respect to what probability means, Finkelstein and Fairley write "'Subjective', 'intuitive' or 'personal' probability... have been defined in terms of the odds that a rational person acting after reflection and consistently would regard as fair in betting on the proposition".[1] Although subjective probabilities can be used on this basis, we suggest that in the legal context they are likely to be interpreted as expressing a frequency, just as "the chances of heads is one-half" expresses a frequency. When we say that the defendant is guilty beyond a reasonable doubt, we mean that the evidence has brought us to a state of belief such that if everyone were convicted when we had such a belief the decisions would rarely be wrong" (p. 504). Thus, their interpretation is Bayesian–frequentist, sometimes the one, sometimes the other.

How would Finkelstein and Fairley propose that Bayesian ideas be used in court? They propose a scenario in which the murder weapon is a knife, and suppose that the frequency of the palm mark found on the knife were it left by someone other than the defendant is known. Then, they propose giving the jury a table mapping their prior probabilities into calculated posterior probabilities (their Table 1, page 500). Of course there is an issue about how the required frequency might come to be known. But even if it were, there is no reason that the decision-makers' (i.e., the jurors) subjective probabilities should coincide with that frequency. Therefore, this use of Bayes Theorem relies on a fact not in evidence, namely the jurors' translation of a frequency (applying to a hypothetical infinite sequence of "similar" situations) into the particular situation about which they are called upon to render a verdict.

Hence, the issue that I see with their proposal is not that it is Bayesian, but rather that it is insufficiently Bayesian.

(2) Tribe (1971) is fairly seen as a rebuttal to Finkelstein and Fairley. Before dealing with what it says, I have a few comments on the title "Trial by mathematics: precision and ritual in the legal process." One might read the title as opposing all uses of mathematics in court. Mathematics, as a general matter, is about the conclusions that follow from particular premises. Provided that no mathematical error is involved, if you agree with the premises, you are stuck

[1] Discussed in D. Lindley, Introduction to Probability and Statistics, 32–34 (1965).

with the conclusions.[2] In the particular examples used by both Finkelstein and Fairley and by Tribe, the mathematics involved is Bayes Theorem, which is so easy to derive that it hardly deserves to be called mathematics. There are many other possible uses of mathematics in court, and Professor Tribe does not address nor appear to oppose them.

The issue raised by Professor Tribe is much narrower than the title suggests: the possible use of decision analysis by statisticians (or others) in court to advise jurors how to amalgamate quantitative evidence with other kinds of evidence. Professor Tribe gives reasons why he feels that such use would threaten important values that the ritual of a legal trial embodies, such as that a person is innocent until found guilty, and that conviction must be "beyond a reasonable doubt." I agree with him, but think the proposal is catastrophically misplaced for another reason. It would give entirely too much power to the decision analyst. At present no one is permitted to speak to a juror about the case outside of the formal procedure in court. Indeed, the jurors are not permitted to discuss the case with each other until they have heard closing arguments and the judge's instructions about the law applicable to the case. To permit a decision analyst to ask jurors their opinions about the case, either part way through or at the end, would allow the analyst a powerful position to influence the outcome, if only by their attitude and behavior. Much as I love Bayesian decision theory, I would not contemplate such a power grab on the part of my friends and colleagues.

(3) Finkelstein and Fairley (1971) responded to Tribe, mainly repeating arguments already in their original paper.

(4) Lindley (1977b) is a study of glass fragments found at a crime scene, and glass particles found in the clothing of a suspect. The question he proposes to answer is the extent to which the data support the proposition that the refractive indices of the two sources are similar. To this end, he has data from a sample of glass refractive indices from the two sources. He imposes a joint normal distribution (with distinct means and variances) on these two sources. In addition, he has a reference sample of "true values" of refractive indices of glass. His analysis is to compute a Bayes Factor for the proposition of similar refractive indices. He does so first showing the mathematics under the counter-factual supposition that the reference sample is normal, and then a computational method taking into account the non-normal shape of the reference sample. After a comparison with a frequentist test of significance, he discusses how his analysis might be presented in court, presuming that identity of refractive indices in the two populations is the same as the guilt of the defendant.

There are several important factual matters not discussed in this paper, such as (a) What was the context of the crime scene? For example, was the glass

[2]There is nothing in this statement that should disturb Professor Tribe. By way of disclosure, he and I were classmates in several mathematics courses as undergraduates. I am confident he fully understands the deductive nature of mathematics.

found at the crime scene from a single broken pane of glass, or from some other source? (b) Whatever that source might have been, how variable is the refractive index within such a source? If the sample he has from the crime scene involves nearly all the glass from whatever the source was, then his sample will reflect that variability. (c) How is the reflective index of glass measured? Are the errors in measurement likely to be correlated? (d) Were the measurements from the suspect measured on the same apparatus? In either case, how would the answer influence his analysis? (e) Most critically, how was the reference sample collected? What reasons are there to believe that the reference sample applies to this particular case?

The casual assumption that identity of the refractive indices is equivalent to the guilt of the suspect is unfortunate. The suspect may have a defense that explains why such glass fragments were found in his clothing, and deserves the opportunity to present such a defense. Lindley's use of a Bayes Factor is important because it presages such use in forensic science more generally. However, such foreshadowing loses much of its attention-grabbing panache if it is presented as a Bayes Factor for identity of refractive indices.

(5) Fienberg and Kadane (1983) review the literature up to that point and speculate about a Bayesian understanding of legal terms, like "beyond a reasonable doubt" and "the preponderance of the evidence." Our thought in doing so was not to replace those phrases with numbers (which Tribe would oppose), but simply to understand better, in probabilistic terms, what is being said.

We also review two cases. We took the data from Swaine vs. Alabama (1965) about whether African Americans were being systematically excluded from juries in Alabama. We show that both frequentist and Bayesian treatments lead to the same conclusion, one also reached (without our help) by the US Supreme Court.

Finally, we discuss work that I had done on a procedure used in the USA to choose jurors to hear cases in which the death penalty is a possibility (Kadane 1983). The jury would first render a verdict on guilt. If the defendant were found guilty, a second hearing would be held in which the same jury would decide whether to impose the death penalty. My work (which used Bayesian methods) was aimed at the question of whether the exclusion of jurors who could not fairly and impartially decide whether to impose the death penalty led to a jury more likely to convict. As this case made its way through the California courts, the Supreme Court ruled in Lockhart vs McCree (1986) that the procedure is legal, regardless of social science results that may indicate bias. Consequently, the case I had testified in was moot.

(6) Bright et al. (1988) was the result of a sales-tax audit of a retailer. A sample had been drawn (badly the first time), the result of which was a probability distribution of the amount owed by the retailer. But a probability distribution is a difficult amount to write a check for. We propose a loss function for the state, adoption of which would result in a specified amount being owed.

(7) Kadane (1990) discusses a case of possible employer discrimination on the basis of age. There were four firing waves, resulting in four 2×2 tables

(fired/not fired, over 40, under 40). Reasons are given for treating these as doubly constrained. The (Bayesian) results indicated high probability of discrimination. On a Friday, the jury was chosen. In the judge's chambers, the plaintiff's attorney made a settlement offer, which the defense attorneys rejected. On Monday, the plaintiff and his attorney came ready to start the case, with boxes of evidence. However, the plaintiff himself noticed that the defense attorneys had come to court with no briefcases or boxes of documents. The attorneys were called to the judge's chambers, and the defense attorneys announced that they wanted to accept the settlement offer they had rejected on Friday. The plaintiff's attorney responded that the Friday offer was no longer available. The case settled on grounds substantially more favorable to the plaintiff than the Friday offer. The plaintiff bought the house his wife wanted.

(8) Lindley (1991) (see also Lindley (1977a)) proposes that fact finders, whether judges or juries, should return probabilistic judgments, rather than binary (guilty/not guilty or liable/not liable) verdicts. The court would then apply a legislatively determined utility function to determine the consequence. He also opposes suppressing any evidence, no matter how obtained. He also proposes the abolition of the adversarial system for deciding cases. His proposals are thus directly contrary to the arguments of Tribe (1971). Lindley argues that these changes are consequences of expected utility theory. I argue Kadane (1993) that, depending on whose utility is being discussed and how it is being specified, one can defend current practice about these matters as optimal. Hence, in my view, Lindley was conflating his social views with his epistemological views.

(9) Fienberg and Finkelstein (1996) review studies of mock jurors in five experiments and find that these mock jurors under-respond to proffered testimony, compared to what the use of Bayes Theorem would predict. They conclude from this that jurors need instruction about probability theory, omitting the possibility that the mock jurors did not give full credence to the data they were offered.

After an exposition of how Bayes Theorem might relate to hypothetical cases, they turn to two real cases. The first is a paternity case in which HLA blood typing was used, but not DNA of the modern kind. The appeals court found that the trial record showed much confusion about conditional probabilities and sent the matter back for rehearing. The second was about whether there was fraud in absentee ballots in deciding an election in Eastern Pennsylvania. There was an expert for each side, and a neutral expert appointed by the judge. Fienberg and Finkelstein conclude that "the judge misinterpreted the evidence provided by the experts in the case."

(10) At about the same time, I was hired by the Public Defender's Office in Gloucester, New Jersey to help with a case alleging that the New Jersey State Police were stopping cars on the New Jersey Turnpike driven by African Americans at disproportionate rates. The Defender's Office arranged for a stationary survey, in which counters stationed on a bridge over the Turnpike assessed the race of drivers. About 13.5% of these drivers were African

American. Additionally, we had a rolling survey in which a Public Defender set his cruise control for 5 miles an hour above the speed limit and counted the race of drivers of cars who passed him, and the race of drivers whom he passed. Almost all the cars he encountered were cars that passed him (nearly everybody speeds), and 15% of the cars that passed him were driven by African Americans. Using a random sample of stops by the State Police, 46.2% of the stops in which race was available were of cars driven by African Americans. The problem was that there were many stops that did not have the race of the driver recorded. Using a Bayesian analysis, we showed that making various reasonable assumptions about the recording of race did not upset the overall finding of discrimination (Kadane and Terrin 1977). While the matter was hard fought (the hearing went on 1 day a week for 6 months), my use of Bayesian methods was not challenged. We won. The decision (State v. Soto 1996) upset the New Jersey State Police and the Attorney General, who announced they would appeal. A year and a half later, and a week before the appeal was to be heard, the Governor and the Attorney General held a press conference to announce that they were withdrawing their appeal. They had done their own study, which corroborated our findings. New Jersey then entered into a consent decree with the Civil Rights Division of the US Department of Justice and promised to reform the practices of the State Police. So this is an example in which Bayesian analysis in court actually brought about needed social change.

One of the defense arguments in the Soto case was that perhaps Blacks are heavily represented among the most egregious speeders on the New Jersey Turnpike. This possibility was dismissed by each police officer who testified, saying that the race of egregious speeders mirrored the population.

To find out, I partnered with John Lamberth, a social psychologist who did much of the data collection in the Soto case. We found essentially the same rates of extraordinary speeders among Blacks and Whites on the turnpike (Kadane and Lamberth 2009).

Therefore,

$$\begin{aligned} \mathrm{P}\{\text{black}|\text{speeder}\}/\mathrm{P}\{\text{white}|\text{speeder}\} &= \mathrm{P}\{\text{black \& speeder}\}/\mathrm{P}\{\text{white \& speeder}\} \\ &= \mathrm{P}\{\text{speeder}|\text{black}\}/\mathrm{P}\{\text{speeder}|\text{white}\} \\ &\times \mathrm{P}\{\text{Black}\}/\mathrm{P}\{\text{white}\}. \end{aligned}$$

Showing that the first ratio in the last expression is 1 shows that the proportion of Blacks among speeders mirrors the population numbers.

(11) Lund and Iyer (2017) emphasize the subjective nature of likelihood ratios and consequently of the potential difference between the likelihood ratio of the decision-maker and that of the forensic analyst. Consequently, they urge that likelihood ratios be considered uncertain and propose many alternative assumptions that might be considered.

(12) Batson motions. American juries are chosen from a venire, which is supposed to be a cross-section of the citizens in the community served by the court. The venire is subject to challenges for cause, such as being related to one of the parties, or other clear sources of bias. Then, potential jurors are subject to peremptory challenges. Traditionally, reasons did not have to be given for the use of peremptory challenges. However, it was a common observation that prosectors in criminal cases were systematically using peremptory challenges to exclude Black citizens from juries.

Accordingly, in 1986, the Supreme Court ruled that peremptory challenges could not be used to exclude Blacks. The ruling created a three-step process. The first step is to establish a *prima facie* case consisting of: (1) a timely objection, (2) identification of the racial group alleged to be discriminated against, and (3) facts sufficient to raise an inference of discrimination. The second step requires the non-moving party to give non-discriminatory reasons for its use of peremptory challenges. Finally, in the third step, the court decides whether the reasons offered in the second text are persuasive (Batson v. Kentucky, 476 U.S. 79).

Table 8.1 shows the steps taken by the Supreme Court to extend the rights granted in *Batson*.

Batson motions are notoriously difficult to win at trial, perhaps because the motion in effect asks the judge to find that the opposing attorney is both a bigot and a liar. But appeals against a verdict based on a finding that the trial judge would have found sufficient evidence for a *prima facie* case are another matter. Such appeals are particularly likely in death penalty cases, because it is the responsibility of attorneys arguing appeals in such cases not to omit any plausible ground.

There is data relevant to a Batson challenge, namely the numbers of Black and non-Black potential jurors available to be peremptorialy challenged, and the numbers who were so challenged. However, the situation is complicated by the fact that each state (and sometimes within a state) has its own set of practices for how peremptory challenges are administered.

In one process, used for example in Georgia, each potential juror is considered individually. First the prosecution and then the defense are asked whether they wish to exercise a peremptory challenge. If neither does so, the potential juror is seated and cannot be recalled. A natural model for this process is two independent binomial distributions, one for each racial group. The question is whether the probability of being excluded from the jury is higher if one is Black than if one is not. Using

Table 8.1 Supreme Court cases extending Batson

Case	Holding
Powers v. Ohio (1991) 499 US 400	Defendant's race is irrelevant
Edmonson v. Leesville Concrete (1991) 500 US 614	Both parties in civil cases
Georgia v. McCollum (1992) 505 US 40	Defendant cannot discriminate
J.E.B. v. Alabama (1994) 511 US 127	Covers discrimination based on gender

data from a 2016 Supreme Court case, I discuss various frequentist approaches to the problem. A Bayesian approach with uniform priors shows a probability over 99.9% that Blacks were more likely to be challenged than were non-Blacks (Kadane 2018b).

A second case (Battle), this one from California, used a different procedure to administer peremptory challenges. Here twelve potential jurors are in the box. The sides take turns choosing whom to strike with a peremptory challenge. A challenged juror is replaced before the other side chooses. Again I consider both frequentist and Bayesian solutions to the problem. The Bayesian analysis requires a parameter to measure the degree of bias. The way I chose is to let the probability of a minority person being chosen as $e^d c/(e^d c + m)$, so the probability of a majority person being struck is $m/(e^d c + m)$. Here, m is the number of majority persons, and c the number of non-majority persons. Then, $d > 0$ indicates bias against the minority, and conversely. Applied to the case at hand, the probability of bias ($d > 0$) turns out to be 98.53% (Kadane 2018b).

North Carolina has yet a third procedure for administering peremptory challenges. As in California, twelve potential jurors are seated. The prosecution goes first, and challenges whom it pleases. Then, those potential jurors are replaced, and the prosecution again may challenge whom it pleases. After the prosecution is done, then the twelve potential jurors are passed to the defense, who again can challenge whom they please. Those unchallenged by both parties are seated (and not subject to further challenge). The reduced group of replacement of challenged potential jurors then passes to the prosecution, etc.

In Kadane (2018a), I argue for a doubly constrained 2×2 table for data rising from this procedure, where the variables are race and whether challenged. Under this model, the remaining parameter is the log cross-product ratio. I applied a Bayesian analysis with this model to two North Carolina cases. The first case (Hurd) featured both an unsuccessful Batson challenge of the prosecution and a successful Batson challenge of the defense (called a reverse-Batson challenge). My results suggest that the trial court was correct in both of these decisions: the prosecution was not differentially challenging Black potential jurors, but the defense was differentially challenging Whites.

In the second case (Tucker), my analysis suggests that the prosecution did use peremptory challenges against Blacks at extraordinary rates. Tucker's capital conviction for murder is now under appeal, and one of the grounds for that appeal is his Batson motion, referring to my analysis.

I have not testified in Batson motions yet. Since they are mainly appeal motions, testimony is not pertinent. Yet my reports are parts of capital appeals in both California and North Carolina. Such appeals seem to take forever. Part of the reason may be the incentives of the parties. From the prosecution's viewpoint, the defendant is already in jail. Why expedite an issue that might vacate his conviction? From the defense's viewpoint, the defendant is still alive. Why expedite an issue that might hasten his execution? So it may be a while before appeals courts rule on these motions.

8.3 The Use of Bayes in True Evidence Cases: R v T (2010)

R v T (2010 EWCA Crim. 2439) is a murder case in which a forensic scientist testified about the similarity of the defendant's Nike shoe to the shoe print found at the crime scene. The expert used a database of shoes to buttress a 1 in 100 likelihood ratio.

The appeals court overturned the conviction, writing:

"The use of the FSS' (Forensic Science Service) own database could not have produced reliable figures as it had only 8122 shoes whereas some 42 million are sold every year" (Para 84). Additionally, the expert relied on a likelihood ratio calculation made from this database. This was also found unsatisfactory by the court: "No likelihood ratios or other mathematical formula should be used in reaching [a] judgement."

This decision engendered a vast number of critical comments from various perspectives. The one I want to focus on is Berger et al. (2011), because the basic viewpoint is closest to that expressed in Sect. 8.1 above.

This paper defends the use of subjective probability to express the opinion of the forensic analyst. With this, I can only agree. Furthermore, they write "Overall, the duty of transparency prevails, and the expert should articulate the basis for her probabilistic assignments in a given case." Again I agree.

However, the unspoken assumption here is that the expert's opinion, however expressed, will become those of the finder of fact, whether jury or judge. The very point that probabilities are subjective means that the fact finder's opinions need not be those of the expert. Indeed, the real content of an expert's testimony may revolve around the reasons adduced for their opinion, however expressed. Even with such explanation, a reasonable fact finder might think that the expert is called by the prosecution, who had the choice of which expert to engage and hence may reduce the credence given to the expert's testimony.

Berger et al. (2011) are also vociferous in defending likelihood ratios as a (perhaps the) way to express expert opinion (whether quantitative or qualitative). They write "If the probability of the observations is greater given the prosecution proposition than it is given the defense proposition it is intuitively reasonable and logically justifiable to state that the observations support the former."

But just a minute. The defense is not required to have any proposition at all, other than that the prosecution has not proven its case. What is the correct comparison, say in the R v T case? Is it against a random selection of shoes worn in Britain? Should it be limited to Nike shoes? To Nike shoes of a particular size and vintage? Should the comparison be to that shoe owned by someone other than the defendant whose print would be most similar to that found at the crime scene? And how is the FSS database going to help with this issue?

One can imagine cross-examination on this basis:

Q: Should the jury understand that you were called to testify by the prosecution?
A: Yes.

Q: Do I understand correctly that the likelihood ratio about which you opine is the ratio of the probability of the evidence under the prosecution proposition divided by the probability of the evidence under the defense proposition?
A: Yes.
Q: Did you consult with the defense to ascertain what the defense proposition is?
A: No, I did not.
Q: If you have mistaken the defense proposition, of what use to the court is your opinion about this likelihood ratio?

Likelihood ratios may not be the panacea that Berger et al. (2011) seem to think it is.

8.4 Conclusion

Bayesian ideas and applications to the law are now part of the accepted landscape. Some of the ideas proposed have been more successful than others, of course, and there are many aspects of the conjunction between law and Bayesian methods still to be explored. Perhaps continuing to use Bayesian ideas in court is among the more fruitful ways forward because it provides feedback, sometimes painful, about what courts find helpful.

References

Berger, C., Buckleton, J., Champod, C., Evett, I., and Jackson, G. (2011). "Evidence evaluation: a response to the court of appeal judgment in R v T." *Science and Justice*, 51, 43–49.
Bright, J., Kadane, J., and Nagin, D. (1988). "Statistical Sampling in Tax Audits." *Journal of Law and Social Inquiry*, 13, 305–338.
Fienberg, S. and Finkelstein, M. (1996). "Bayesian Statistics and the Law." In *Bayesian Statistics 5*, eds. J. Bernardo, J. Berger, A. Dawid, and A. Smith, 129–146.
Fienberg, S. and Kadane, J. (1983). "The Presentation of Bayesian Statistical Analyses in legal Proceedings." *Journal of the Royal Statistical Society, Series D (The Statistician)*, 32, 1 / 2, 88–98.
Finkelstein, M. and Fairley, W. (1970). "A Bayesian Approach to identification evidence." *Harvard Law Review*, 83, 489–517.
—— (1971). "The continuing debate over mathematics in the law of evidence." *Harvard Law Review*, 84, 8, 1801–1809.
Kadane, J. (1983). "Juries hearing death penalty cases: statistical analysis of a legal procedure." *Journal of the American Statistical Association*, 78, 544–552.
—— (1990). "A Statistical Analysis of Adverse Impact of Employer Decisions." *Journal of the American Statistical Association*, 85, 925–933.
—— (1993). "Bayesian Decision Theory and the Legal Structure in *Aspects of Uncertainty: A Tribute to Dennis Lindley*." In *A Tribute to Dennis Lindley*, eds. A. Smith and P. Freeman, 261–266. New York: J. Wiley and Sons.
—— (2011). *Principles of Uncertainty*. Boca Raton: Chapman and Hall.

—— (2018a). “Batson and Reverse-Batson Motions in North Carolina: State v Hurd and State v Tucker.” *Law, Probability and Risk*, 17, 263–273.
—— (2018b). “Statistics for Batson Challenges.” *Law, Probability and Risk*, 17, 1, 1–13.
Kadane, J. and Lamberth, J. (2009). “Are Blacks egregious speeding violators at extraordinary rates in New Jersey?” *Law, Probability and Risk*, 8, 139–152.
Kadane, J. and Terrin, N. (1977). “Missing Data in the Forensic Context.” *Journal of the Royal Statistical Society, Series A*, 160, 351–357.
Lindley, D. (1977a). “Probability and the Law.” *The Statistician*, 26, 203–212.
—— (1977b). “A Problem in Forensic Science.” *Biometrika*, 64, 2, 207–213.
—— (1991). “Subjective Probability, Decision Analysis, and the Legal Consequences.” *Journal of the Royal Statistical Society, Series A*, 154, 1, 83–92.
Lockhart vs. McCree (1986). “476 U.S. 162.”
Lund, S. and Iyer, H. (2017). “Likelihood Ratio as Weight of Forensic Evidence: A Closer Look.” *Journal of Research National Institute of Science and Technology*, 122, Article 27.
R v T (2010). Court of Appeal - Criminal Division. ECWA Crim 2439.
State v. Soto (1996). “324 N.J. Super 66.”
Swaine v. Alabama (1965). “380 U.S. 202.”
Tribe, L. (1971). “Trial by mathematics: precision and ritual in the legal process.” *Harvard Law Review*, 84, 1329–1393.

Chapter 9
Statistical Considerations for the Analysis and Interpretation of Forensic Evidence

Hal S. Stern

9.1 Introduction

The application of forensic science in crime solving has a long history with fingerprints being used in the USA since the early 1900s. Forensic science has had a surge in popularity due in part to the success of the television show *CSI: Crime Science Investigation* (and several spinoff shows) which ran for approximately 15 years in the early 2000s. The use of fingerprints and more recently DNA to identify individuals has had a major impact on law enforcement and the legal system. Though there have been many successes, concerns have been raised about whether the claims being made by forensic examiners are supported by sufficient scientific research and whether the claims incorporate appropriate assessments of uncertainty.

Several key events have raised the profile of these concerns. In 2004, the FBI arrested Brandon Mayfield after mistakenly identifying him as the source of a latent fingerprint found on a bag implicated in a terrorist train bombing in Spain. This error received a great deal of attention because it was not uncommon at the time for fingerprint examiners to report while testifying that the fingerprint comparison process had zero error rate. A congressionally mandated study of forensic science in the USA by the National Academies of Science, Engineering and Medicine (National Research Council 2009) examined a number of aspects of the forensic science system and concluded that there was not a strong scientific foundation for many common types of evidence. The 2009 report was followed seven years later by a report of the President's Council of Advisors on Science and Technology (PCAST 2016). The PCAST report described what the Council thought should be expected in

H. S. Stern (✉)
Department of Statistics, University of California, Irvine, CA, USA
e-mail: sternh@uci.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_9

establishing forensic methods as valid and reviewed a number of disciplines, finding again that more study was needed to support testimony about a number of evidence types. The need for improving the practice of forensic science is further reinforced by the work of the Innocence Project (IP); more than 40% of the 350+ wrongfully convicted individuals who have been exonerated by DNA evidence through the efforts of the IP had been convicted in part based on misleading or misapplied forensic evidence.

Stephen Fienberg made substantial contributions to the application of statistics in public policy and his contributions to statistics in the law are especially noteworthy. Indeed, Steve's fingerprints (no forensic pun intended) are on many of the most important contributions to rigorous statistical thinking about scientific evidence. His earliest contributions arose as part of a National Academies panel whose report (Fienberg 1989) addressed the increasing trend for statistical evidence to arrive via expert testimony and some of the issues associated with that trend for the different courtroom participants. He subsequently chaired a panel on the validity of the polygraph exam that found no scientific basis for its use as a lie detector (National Research Council 2003). This turned out to be a precursor of the 2009 report mentioned above (for which Steve served as a reviewer).

Two consequences of the 2009 report were the creation of the joint NIST-Department of Justice National Commission on Forensic Science and the funding of a NIST Center of Excellence on probability and statistics for forensic evidence. Not surprisingly Steve featured prominently in both. He was the only statistician on the Commission, ensuring that the statistical viewpoint was always well represented. In addition, Steve and several colleagues (including the author of this chapter) created the Center for Statistics and Applications in Forensic Evidence (CSAFE), a NIST-funded center of excellence that is exploring a range of research topics aimed at assessing and improving the scientific underpinnings for the analysis of forensic evidence.

Statistical concepts and statisticians have emerged as an important component of the efforts to improve forensic science because of the need for reliable and accurate measurements, rigorous assessments of uncertainty, and decision-making under uncertainty. Through this chapter's review of different approaches to the analysis and interpretation of forensic evidence in criminal cases we hope to build on Steve's legacy of statistical contributions to public policy and the law.

9.2 Forensic Evidence and Expert Testimony

Forensic examinations cover a wide range of questions in criminal investigations. These include determining the timing of events (e.g., time of death), cause and effect (e.g., the sequence of events based on bloodstain pattern analysis), and whether two

items correspond to the same source (e.g., comparing a latent finger mark from a crime scene to the fingerprint of a suspect). This chapter focuses on the last of these, often known as the source identification question or the determination of source. The approaches to evidence described here apply more broadly to the other forensic questions but it is convenient to use a single problem to focus the discussion.

The source identification question is itself quite general in that it is relevant to the comparison of biological samples, trace evidence (glass samples, soil samples), and pattern evidence (bullets/cartridge cases, toolmarks, fingerprints, shoe prints, tire treads, handwriting). In the most common scenario, sometimes called the specific source case, there is one sample with known source and a second sample with questioned or unknown source. Typically one sample comes from the crime scene and another is obtained from a suspect in the investigation. Which sample has known source and which has questioned source can vary depending on the evidence type. For example, in comparing glass samples the fragments at the crime scene may be assumed to have come from the broken window at the scene while the fragments identified on the clothing of the suspect are from an unknown source. By contrast, the finger mark at the crime scene is from an unknown person while the suspect's fingerprint is from a known source. The forensic examiner is called as an expert to assess the evidence and provide their conclusion regarding whether the sample with unknown source could have arisen from the source that provided the sample with known source.

The provision of expert testimony in federal courts and in many states is governed by the Daubert decision of the U.S. Supreme Court (Daubert v. Merrell Dow Pharmaceuticals 1993) and subsequent related decisions. The Daubert standard identifies the judge as the gatekeeper in determining whether proposed testimony should be admitted. The standard identifies a number of factors that are relevant for the judge to consider including whether the technique in question has been tested in actual field conditions, whether the technique has been subject to peer review and publication, the known or potential rate of error of the technique, the existence of standards for application of the method, and whether the method or technique is generally accepted within the relevant scientific community. None of these factors are required but all point the judge in the direction of assuring that the method or technique has a solid scientific foundation. Some state courts are still governed by expert witness guidelines that were in place before the Daubert decision, known as the Frye standard (Frye v. United States 1923). The Frye decision applied only to novel scientific evidence and the criterion for admitting such evidence was general acceptance in the relevant scientific community.

The Daubert decision is codified in the U.S. Federal Rules of Evidence Rule 702. The rule requires a determination that the proffered expert knowledge will help the trier of fact (usually the jury) to understand the evidence. It also requires that the testimony be "based on sufficient facts or data," be "the product of reliable principles and methods," and that the approach was "reliably applied" to the case at hand.

9.3 Evaluation and Interpretation of Forensic Evidence

There are a number of approaches to making a source determination for forensic evidence. All approaches require examining the two samples to identify similarities and differences. The nature of the similarities and differences, including their quantity and quality, are assessed to see if the observations would be expected from two items derived from the same source and also to assess how likely such observations might be for objects from different sources. The three approaches described in detail here are: forensic evidence as expert opinion, a two-stage approach to forensic evidence, and the likelihood ratio.

Forensic Evidence as Expert Opinion

The status quo in many forensic evidence disciplines, especially those known as the pattern evidence disciplines (i.e., fingerprints, shoe prints, firearms, toolmarks, questioned documents), is for an examiner to analyze the evidence based on their experience, training, and accepted methods in the field. The examiner's conclusion regarding the evidence reflects the examiner's expert opinion. The conclusion is often reported as one of a set of categorical conclusions. Fingerprint examiners will, for example, report the result of their analysis as either an identification (the questioned print came from the same finger as a given print of known source), an exclusion (the questioned print did not come from the same finger as the known source print) or as not providing a conclusive result (often described as an inconclusive result). Other forensic disciplines may allow for more categories. For example, handwriting examiners often use a nine-category scale.

To be admissible under the Daubert standard forensic analyses of this type must be "the product of reliable principles and method." The word "reliable" has a both a colloquial meaning and a technical meaning. The colloquial meaning is trustworthy and this is the sense in which legal commentators interpret the word. There are a number of technical concepts that address the trustworthiness of forensic evidence that is provided in the form of an expert opinion. These include the measurement science concepts of reliability and validity. Reliability refers to consistency of measurement. It can be assessed through repeatability studies, in which the same analyst is asked to analyze the same evidence on more than one occasion, and reproducibility studies, in which different analysts are asked to analyze the same evidence. Validity refers to the accuracy of a measurement or decision. Validity is assessed through studies in which evidence samples with known status (as either being from the same source or from different sources) are analyzed by a number of analysts. The application of technical reliability and validity to forensic evidence is described by Stern et al. (2019).

PCAST, in their 2016 report, (PCAST 2016) provided guidance on how the Daubert standards could be addressed. They advocated for "black box" studies in

which the expert is viewed as a decision-making "black box" taking the evidence as input and outputting a decision regarding the same source question. Carrying out such studies on a large number of analysts would allow for a determination of error rates for examiner opinions and an assessment of the overall validity of the forensic science discipline. One example of such a study is the fingerprint study carried out by Ulery et al. (2011). That study included 169 examiners and more than 700 pairs of prints with known true status. Each examiner analyzed a subset of the pairs. They found a false positive rate of 0.15% with 6 nonmated pairs incorrectly identified as mated. The false negative rate was 7.5% indicating that it was relatively common for true mates to be incorrectly identified as nonmated pairs. Examiners were surveyed as part of the study and a large majority were confident that they had never had a false positive or false negative in their casework. Though false positives were indeed rare in the study, more than 80% of examiners in the study had at least one false negative.

There are limitations associated with this (and any) study. Statistics can contribute to strengthening the scientific foundations of using expert opinion as an approach to forensic evidence through its principles of experimental design. For example, the examiners in the Ulery et al. (2011) study knew they were being tested and thus their performance would not necessarily reflect the performance that would be observed in case work or in the standard lab environment. In addition, fingerprint examiners were judged on their analysis without the benefit of a coworker's verification (which is a part of standard practice in the field). Additional research is planned to address concerns such as these.

The same authors also investigated the reliability of fingerprint decisions (Ulery et al. 2012). They found that repeatability (intra-examiner agreement) was 90% for mated pairs and 86% for nonmated pairs. Reproducibility (inter-examiner agreement) was 80% for mated and nonmated pairs. The researchers also carried out additional studies focused on understanding how the examiners reached their conclusions.

One criticism of studies of this type is that they provide only a single overall error rate estimate for the discipline. It is likely that the error rate varies based on various characteristics of the evidence and the examiner. It is natural to expect that errors will be more common for a blurry, low quality latent print or for a low complexity signature (one that is easy to simulate). This suggests that an important future direction is the estimation of error rates that would be relevant to a specific comparison.

It is likely that there will always be comparison questions for which no error rate information is available. One Indiana case involved a comparison of a two plastic garbage bags that attempted to determine if they were produced at roughly the same time by the same machine. Such comparisons may yield useful information but caution is obviously required when there is no empirical evidence that such determinations can be reliably made.

A Two-Stage Approach to Forensic Evidence

The two-stage approach (Parker 1966, 1967; Parker and Holford 1968) to the analysis of forensic evidence addresses the source determination question in two separate steps. Initially a determination is made as to whether the two samples are sufficiently similar based on their measured characteristics to be declared as a "match" or to be identified as "indistinguishable." This is often done via a statistical test or procedure. The second stage of the analysis attempts to address the probative value of the evidence by assessing the probability that two samples from different sources would be found indistinguishable by chance. This is commonly done for measurements of glass fragments. This can include measurements of the refractive index of the glass or measurements of the chemical concentrations of various elements found in the glass fragment. We use glass to describe the approach and to address some of the issues associated with assessing the reliability and validity of the approach.

The first stage can be carried out with a traditional statistical test such as the t-test. The null hypothesis is taken to be that the two sets of measurements (one set from the sample with questioned source and one set from the sample with known source) come from populations with the same mean. The alternative hypothesis is that the two sets come from populations that differ. In practice, examiners appear to often use a variety of alternative statistical procedures that are similar to the t-test. For example, they may compare the mean measurement for the questioned sample to a confidence interval for the mean measurement of the known sample. For purposes of this discussion we assume the first stage is carried out through a traditional statistical test.

There are a number of both technical and conceptual concerns associated with the use of a statistical test to determine matching samples. One technical concern is that in practice the test is carried out based on a small number of measurements. This means that the test procedure will be sensitive to any distributional assumptions that are made. A second concern is that the forensic analysis often involves carrying out a series of tests for the concentration of different elements. This raises the issue of multiple comparisons which must be addressed in the setting of thresholds. A final technical concern when the analysis is carried out through a series of univariate tests of different chemical elements is that this ignores information that is contained in the correlation structure of the multivariate chemical concentration measurements.

In practice, however, the conceptual concerns are arguably more important than the technical concerns. The first of these is the loss of information associated with summarizing a comparison of continuous measurements through a binary decision (distinguishable/indistinguishable). A finding that two samples are indistinguishable, i.e., that the null hypothesis of equal population means could not be rejected, may mean there is excellent agreement or it may mean that the samples differ considerably but not enough to achieve statistical significance. This issue has received a great deal of attention recently with a number of statisticians recommending elimination of the concept of statistical significance (Wasserstein

et al. 2019; Amrhein et al. 2019). The Amrhein et al. comment in *Nature* was signed by more than 800 scientists. The various authors note that a binary decision can ignore potentially meaningful differences because they do not attain a specified p-value and that samples can be found significantly different even when the difference is not practically important.

In the forensic application, the setting of the threshold for declaring two items matched has important implications for justice. Using a generous threshold for statistical significance (e.g., taking $p < .15$ as indicating "distinguishable" samples) risks failing to correctly identify two mated or matching samples. A strict threshold for statistical significance on the other hand (requiring a very low p-value) risks incriminating an innocent person by failing to correctly identify nonmated samples. It is not at all clear how to trade off these two types of errors. It certainly does not seem like the kind of decision that should be determined by an individual examiner.

The challenge of setting a threshold is linked to another key conceptual issue associated with the two-stage approach. As commonly executed the null hypothesis is taken to be that the two samples are indistinguishable. As is well-known, statistical significance tests do not treat the two hypotheses symmetrically. The null hypothesis is assumed true unless the data provide sufficient evidence against that hypothesis. But the null hypothesis incriminates the suspect and thus the statistical approach seems to oppose the traditional position of the justice system in the USA that a suspect is innocent until proven guilty. In fact, somewhat paradoxically, carrying out a "weaker" analysis by taking fewer measurements or using an inferior measurement technique (with larger variance) make it more likely that samples will be found "indistinguishable."

The second stage of the two-stage approach is an assessment of the probability that two samples from different sources would be found indistinguishable by chance. Assessing this probability requires information about the distribution of mean measurements across a population of glass sources. Given information about this distribution and a set of measurements for a given sample from a known source (summarized perhaps by the sample size, the sample mean and the sample standard deviation), one can calculate the probability of a coincidental match as an integral over the population distribution of glass source means of the probability that a sample of size n from a (perhaps normally distributed) sample with the given mean would fail to differ significantly from the known source sample.

There are challenges associated with obtaining and maintaining a database of measurements that can provide relevant information for the second stage. For example, there are many types of glass and each likely requires its own database. As well, one can easily imagine that the distribution of glass source means will vary over time. It is not too surprising then that no such databases exist for most forensic evidence types at the present time. This means that current practice for forensic evidence often stops with the stage one conclusion that two samples are indistinguishable. Future research that would enable examiners to reliably infer the likelihood of a coincidental match is crucial for improving forensic science.

The Likelihood Ratio

The likelihood ratio (LR) or Bayes factor (BF) has long been recognized as a natural approach to the evaluation of evidence (see e.g., Lindley 1977; Aitken and Taroni 2004). Within forensics the LR has been regularly computed for DNA evidence for more than 20 years and this has led many to wonder if the concept can be more generally applied. The justification for the use of the LR or BF can be seen easily through the odds version of Bayes' Theorem

$$\frac{\Pr(H_s|E)}{\Pr(H_d|E)} = \frac{\Pr(E|H_s)}{\Pr(E|H_d)} \times \frac{\Pr(H_s)}{\Pr(H_d)} \tag{9.1}$$

where E represents the evidence (the measurements on the two items), H_s is the hypothesis that the two items share a common source, and H_d is the hypothesis that the two items come from different sources. The notation above is imprecise in that for evidence E comprised of continuous measurements the probabilities would be replaced by probability density functions. Bayes' theorem identifies the LR or BF, $\Pr(E|H_s)/\Pr(E|H_d)$, as the factor by which a trier of fact should update their prior odds of the two hypotheses to obtain their posterior odds of the two hypotheses.

The likelihood ratio addresses the conceptual weaknesses of the two-stage approach. By assessing the likelihood of the evidence under the two competing hypotheses we avoid choosing a single hypothesis as the starting point (e.g., the null hypothesis that two samples are indistinguishable). In addition, the LR avoids separating the analysis into two stages. The LR also eliminates the need for a threshold and associated binary decision. The European Network of Forensic Science Institutes (ENFSI) has endorsed the use of likelihood ratios for evaluation of forensic evidence (ENFSI 2015). Their guideline emphasizes that the evaluation of evidence should be with respect to two or more competing hypotheses, should use probability as a measure of uncertainty, and should be based on the assignment of a likelihood ratio.

A couple of technical points are noteworthy. The first is that assessing the likelihood of the evidence under the same (or especially the different) source hypothesis is likely to depend on a number of unknown parameters. For example, the likelihood of a set of glass measurements is likely to depend on the variability of the measurement process and the variability among sources of glass in the population. Inference for the parameters of the same source distribution may be based on the measurements in the case at hand or may be based on other data regarding the measurement process. Inference for the parameters of the different source distribution usually relies on data from an auxiliary source.

To this point we have referred to the ratio $\Pr(E|H_s)/\Pr(E|H_d)$ as both the likelihood ratio and the Bayes factor. The distinction between the two terms is related to the treatment of the unknown parameters. The likelihood ratio is often used as the preferred term when the unknown parameters are replaced by point estimates and the Bayes factor terminology is used when the unknown parameters

are assigned a prior distribution and averaged over. Naturally if the two hypotheses completely specify the probability distribution of the evidence (rare in forensics though possible in other applications of the LR), then the likelihood ratio and Bayes factor would agree. In the remainder of the chapter we use the likelihood ratio terminology.

Another technical concern is that there may be additional information relevant to the interpretation of the evidence. This information might have to do with the technology used to obtain the measurements or with steps taken during the analysis of the evidence (e.g., the technique used to lift the fingerprint evidence or gather the shoe print evidence). Such task-relevant information can be integrated in the calculation of the LR/BF but is ignored here.

In the remainder of this section, we briefly review how the likelihood ratio works or might work for different types of forensic evidence. Additional details can be found in Stern (2017). We also discuss some of the limitations associated with the use of likelihood ratios to analyze forensic evidence.

Where It Works: DNA

The likelihood ratio approach is best known for its application to DNA evidence. To start we assume there is a crime scene sample that is known to represent a single unknown source and a second sample from a suspect in the case. The evidence E is comprised of two DNA profiles, with each profile identifying the two alleles identified at a set of locations along the genome. Under the hypothesis that the two samples are from the same source, the probability of matching profiles is essentially one. Under the hypothesis that the two samples are from different sources, the probability of matching profiles depends on the frequency of the alleles on which the samples were found to match.

Given estimates of the allele frequencies in the relevant population it is straightforward to compute the probability that a random individual would match the crime scene sample at the specified locations. For a single location each person has two alleles (one from mother and one from father) and the probability of matching the crime scene sample follows from the Hardy-Weinberg principle. By choosing a set of locations that reside on different chromosomes, and hence are inherited independently, it turns out that the probability of matching profiles under the different source hypothesis are often extremely low. This can yield likelihood ratios in the millions or billions.

The single source DNA case is important because it informs about what is required to have the LR approach work. The biology is well-understood, biological theory provides a probability model for the evidence, population databases are available to provide the numbers required by the probability model, and the method has been peer reviewed by the scientific community. Even with these benefits there are still complications that impact our ability to draw reliable inferences from DNA evidence. Increasingly sensitive techniques for capturing DNA evidence can lead to inadvertent contamination. As well, crime scene samples often represent a

mixture of DNA from multiple sources and assessing the probability of observing the evidence (crime scene and suspect sample) under the two hypotheses of interest is much more complicated in such settings.

Where It Might Work: Trace Evidence

The likelihood ratio is not currently applied to trace evidence (e.g., glass fragments) but research is ongoing and there are published examples demonstrating how this might work. Using glass evidence as an example, the evidence E comprises measurements of a number of characteristics on a set of fragments from both the crime scene sample and fragments obtained from the clothing of a suspect. Let $y_{cs} = (y_{cs,1}, \ldots, y_{cs,n_{cs}})$ denote the data for the crime scene sample of n_{cs} fragments with the measurement $y_{cs,i}$ on each fragment possibly multivariate and let y_{sus} denote the data for the suspect sample.

Following the ideas in Section 3.4 of Aitken and Lucy (2004), the data may be modeled as being generated by two probability distributions. The repeated observations of fragments from the same glass source (e.g., window) are modeled as repeated draws from a distribution $f(y|\theta, W)$ where θ is the (unknown) glass source mean and W is a variance matrix for repeated observations within a single homogeneous glass sample. This implicitly assumes the distribution depends only on these two parameters (as is the case for the normal distribution) but other assumptions are possible. The distribution of unknown glass source means across the population of potential sources are modeled as draws from a distribution $f(\theta|\mu, B)$ where μ is the population mean for the relevant type of glass source and B is a variance matrix characterizing the variation of mean vectors across the population. Then the likelihood ratio is

$$LR = \frac{\Pr(E|H_s)}{\Pr(E|H_d)} \tag{9.2}$$

$$= \frac{\int f(y_{cs}|\theta, W) f(y_{sus}|\theta, W) f(\theta|\mu, B) d\theta}{\int f(y_{cs}|\theta_{cs}, W) f(\theta_{cs}|\mu, B) d\theta_{cs} \int f(y_{sus}|\theta_{sus}, W) f(\theta_{sus}|\mu, B) d\theta_{sus}} \tag{9.3}$$

where μ, B, W are to be estimated from available data.

The numerator assumes the two sets of samples come from the same glass source and are thus characterized by a single mean; the denominator assumes the two sets come from different sources characterized by different means. Aitken and Lucy (2004) demonstrate the approach using normal distributions and more flexible kernel density estimates.

Though not currently used in practice, the brief description provided here shows that it is plausible that such methods can be developed. Here the LR approach seems possible because the evidence are characterized by a well-defined and reliably measured set of quantitative characteristics. A major challenge is that the distribution of measurements across the population of relevant glass sources is

not well studied. Indeed, it is not immediately obvious how to define the relevant population of glass sources. Is it all glass manufactured in the area of the crime? in the nation? in the world? Further, a complete assessment of the evidence may also require taking account of the probabilities that glass evidence is transferred from the crime scene to the suspect and is retained on the suspect's clothing (Curran et al. 1998; Curran 2003).

Where It Is Extremely Challenging: Pattern Evidence

Evidence comprised of pattern impressions (e.g., shoe prints) are currently analyzed using methods that rely heavily on the expertise of the examiner. These methods have been criticized for lacking the kind of objective measurements that characterizes DNA and glass evidence. The evidence E typically comprises two images: an image c obtained from the crime scene and an image s obtained from an item/object belonging to the suspect. These may, for example, be a gray-level image of a shoe print found at the crime scene and an analogous image of a test impression taken from a suspect's shoe.

In this setting the LR is defined as $\Pr(c, s|H_s)/\Pr(c, s|H_d)$. The challenges in developing a likelihood ratio are immediately obvious. The data c, s are very high-dimensional, comprised of gray levels for many thousands of pixels in each image. It is nearly impossible to develop probability models for data of this type; a full probability model would require assigning a distribution on the space of possible images. One approach (see, for example, Neumann et al. 2015 for fingerprints) replaces the images by sets of features derived from the images. In the case of fingerprints this might be locations, directions and types of individual minutiae (e.g., ridge endings). As with trace evidence, evaluation of the likelihood ratio requires some information about the variation that would be expected in repeated impressions from the same source (i.e., the distortions that might arise in the deposition of a fingerprint) and about the variation that would be expected in impressions from different items in the population.

As mentioned earlier the ENFSI guideline (ENFSI 2015) endorses the use of likelihood ratios as the appropriate approach to the evaluation of evidence. According to the guideline, probabilities in the likelihood ratio are ideally based on published data but experience and subjective assessments can be used as long as they are justified. The use of experience-based subjective likelihood ratios has been viewed with considerable skepticism in the USA.

Score-Based Likelihood Ratios

The challenges associated with developing likelihood ratios for pattern evidence and other evidence types has led many to consider an approach based on numerical summaries or scores as a summary of the evidence. Though the score-based approach can be applied to a range of forensic disciplines, we develop the idea here

for the pattern evidence setting. Suppose that $S = S(E) = S(c, s)$ is a similarity score derived from the two images c and s with higher values indicating greater similarity of the two images. The definition of a score is quite general. It can be derived directly from the images (e.g., as the sum of squared difference between two aligned images) or through the use of features defined on the images. The key idea is that the scores can then be used in place of the evidence E to reach a conclusion about the hypotheses H_s and H_d.

Scores can be used in two different ways. On the one hand, scores can be used as the basis for a two-stage procedure. A threshold on the score can be used to distinguish between cases that are declared to be "matched" or from a "common source" (analogous to stage one) and then data used to assess the probability of a coincidental match (stage two). This approach has been used, for example, by Hare et al. (2017) to develop an automatic approach to matching bullet land impressions. The score in that case is the estimated probability of a match as output from a random forest prediction model that has been trained to detect whether two bullet land impressions come from the same source. When scores are used in this manner it can be helpful to examine their performance through the receiver operating characteristic curve that displays error probabilities as the threshold used to declare a "match" is varied (Pepe 2004).

Scores can also be used to develop a likelihood ratio. The score-based likelihood ratio, SLR, is defined as $SLR = \Pr(S(c, s)|H_s)/\Pr(S(c, s)|H_d)$ where the probability distributions of the scores under H_s and H_d are usually developed from datasets comprised of known mated pairs (H_s) and known nonmated pairs (H_d). A challenge in this case is that it is often desirable to condition the score-based likelihood ratio on various characteristics of the evidence in the case. For example Swofford et al. (2018) develop a score-based approach to the analysis of fingerprint impressions and show that there is considerable dependence on the number of minutiae marked on the latent print. Hepler et al. (2012) develop score-based likelihood ratios for handwriting evidence and show that different choices for how the denominator of the score-based likelihood ratio is defined have a large impact on the values obtained.

Advantages and Disadvantages of the Likelihood Ratio

The likelihood ratio approach is popular among scientific researchers interested in the analysis of forensic evidence. The LR provides a quantitative summary of the probative value of the evidence and thus eliminates the need for choosing an arbitrary threshold to enable a binary decision when working with continuous data. It also explicitly requires the analyst to compare the relevant hypotheses regarding the source(s) of the evidence. Though challenging to develop, the likelihood ratio can also potentially accommodate a wide range of factors including issues associated with the manufacturing and distribution of some evidence types. Perhaps most important, the LR can be thought of as providing a mapping from a given set of data, assumptions about the data, and a given set of hypotheses to a summary

of the evidence. By being explicit about the data and assumptions the LR approach invites criticism and alternative analyses.

This last point has also emerged as a potential disadvantage of the likelihood ratio approach. There are a number of choices made in developing a statistical model that enables inference for the likelihood ratio. Lund and Iyer (2017) explore the amount of variation that can be observed in the likelihood ratio as a range of plausible models for the data are considered. They demonstrate in the context of glass evidence that the likelihood ratio can vary by several orders of magnitude over a range of plausible models for the within-source and between-source distributions of refractive index measurements. Their example even includes instances where the hypothesis preferred by the likelihood ratio changes across the set of plausible models. One can argue that the set of models considered by Lund and Iyer is too generous, but their work clearly emphasizes the importance of considering the sensitivity of forensic conclusions to the assumptions being made.

A second challenge in applying the LR approach is the difficulty that non-statisticians have in understanding and interpreting the likelihood ratio. As one example, it is common for individuals (lawyers, juries) to misinterpret the denominator of the LR by transposing the arguments of the conditional probability and thinking of $\Pr(E|H_d)$ as $\Pr(H_d|E)$. The former addresses the likelihood of the evidence under the different source hypothesis while the latter directly addresses the likelihood of the different source hypothesis. As made clear in Bayes' Theorem you can only obtain the posterior odds of the hypotheses if you first specify the a priori odds and most experts agree that the examiner of the forensic evidence should not be selecting prior odds for these hypotheses. One interesting line of research is to try and better understand how juries understand different ways of presenting the strength of the evidence (see, e.g., Thompson et al. 2018).

9.4 Conclusions

The summary above describes a variety of approaches to assessing the probative value of forensic evidence. Any approach must address the court's requirement that expert testimony be based on reliable methods. Statistical methods thus have a critical role to play in ensuring the fair administration of justice. Approaches to the analysis and interpretation of forensic evidence should be explicit about the reasoning and assumptions on which the assessment is based. There should be empirical support for the reasoning and assumptions. The empirical support can vary depending on the type of approach used to analyze the evidence. It may be the result of well-designed "black box" studies to assess the performance of analysts' expert opinions or it may be carefully constructed statistical models used to construct likelihood ratios for quantitatively measured evidence. There are significant opportunities for statisticians to contribute to developing methods for this high-impact area of public policy.

References

Aitken CGG and Lucy D. 2004. Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, 53(4), 109–122.

Aitken C and Taroni F. 2004. *Statistics and the Evaluation of Evidence for Forensic Scientists* (2nd edition). Wiley: New York.

Amrhein V, Greenland S, and McShane B. 2019. Scientists rise up against statistical significance. *Nature* 567, 305–307. https://doi.org/10.1038/d41586-019-00857-9

Curran JM. 2003. The statistical interpretation of forensic glass evidence. *International Statistical Review*, 71(3), 497–520.

Curran JM, Triggs CM, Buckleton JS, Walsh KAJ and Hicks T. 1998, Assessing transfer probabilities in a Bayesian interpretation of forensic glass evidence. *Science & Justice*, 38(1), 15–21.

Daubert v. Merrell Dow Pharmaceuticals, Inc. 1993. 509 U.S. 579.

European Network of Forensic Science Institutes (ENFSI). 2015. *ENFSI Guideline for Evaluative Reporting in Forensic Science.*

Fienberg SE (ed.). 1989. *The Evolving Role of Statistical Assessments as Evidence in the Courts.* Springer: New York.

Frye v. United States. 1923. 293 F. 1013.

Hare E, Hofmann H, and Carriquiry A. 2017. Automatic matching of bullet land impressions. *Annals of Applied Statistics*, 11(4), 2332–2356.

Hepler AB, Saunders CP, Davis LJ, and Buscaglia J. 2012. Score-based likelihood ratios for handwriting evidence. *Forensic Science International*, 219(1–3):129–40. https://doi.org/10.1016/j.forsciint.2011.12.009

Lindley, DV. 1977. A problem in forensic science. *Biometrika*, 64(2), 207–213.

Lund SP and Iyer H. 2017. Likelihood ratio as weight of forensic evidence: A closer look. *Journal of Research of National Institute of Standards and Technology*, 122, No. 27. https://doi.org/10.6028/jres.122.027.

National Research Council. 2003. *The Polygraph and Lie Detection*. Committee to Review the Scientific Evidence on the Polygraph. National Academies Press: Washington.

National Research Council. 2009. *Strengthening Forensic Science in the United States: A Path Forward*. Committee on Identifying the Needs of the Forensic Science Community. National Academies Press: Washington.

Neumann C, Champod C, Yoo M, Genessay T, and Langenburg G. 2015. Quantifying the weight of fingerprint evidence through the spatial relationship, directions and types of minutiae observed on fingermarks. *Forensic Science International*, 248, 154–171. https://doi.org/10.1016/j.forsciint.2015.01.007

Parker J. 1966. A statistical treatment of identification problems. *Journal of the Forensic Science Society*, 6, 33–39.

Parker J. 1967. The mathematical evaluation of numerical evidence. *Journal of the Forensic Science Society*, 7, 134–144.

Parker J and Holford A. 1968. Optimum test statistics with particular reference to a forensic science problem. *Journal of the Royal Statistical Society. Series C*, 17(3), 237–251.

President's Council of Advisors on Science and Technology (PCAST). 2016. *Report to the President: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods.* https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

Pepe MS. 2004. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press: Oxford.

Stern HS. 2017. Statistical issues in forensic science. *Annual Review of Statistics and Its Application*, 4(1), 225–244.

Stern HS, Cuellar M, and Kaye DH. 2019. Reliability and validity of forensic science evidence. *Significance*, 16(2), 21–24. https://doi.org/10.1111/j.1740-9713.2019.01250.x.

Swofford HJ, Koertner AJ, Zemp F, Ausdemore M, Liu A, and Salyards MJ. 2018. A method for the statistical interpretation of friction ridge skin impression evidence: Method development and validation. *Forensic Science International*, 287, 113–126. https://doi.org/10.1016/j.forsciint.2018.03.043

Thompson WC, Grady RH, Lai E, and Stern HS. 2018. Perceived strength of forensic scientists' reporting statements about source conclusions. *Law, Probability and Risk*, 17(2), 133–155. https://doi.org/10.1093/lpr/mgy012

Ulery BT, Hicklin RA, Buscaglia J, and Roberts MA. 2011. Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences*, 108, 7733–7738.

Ulery BT, Hicklin RA, Buscaglia J, and Roberts MA. 2012. Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLOS One*, 7, e32800.

Wasserstein RL, Schirm AL, and Lazar NA. 2019. Moving to a world beyond "$p < 0.05$". *The American Statistician*, 73:sup1, 1–19. https://doi.org/10.1080/00031305.2019.1583913.

Chapter 10
Differences between Bayes Factors and Likelihood Ratios for Quantifying the Forensic Value of Evidence

Danica M. Ommen and Christopher P. Saunders

In Memory of Stephen E. Fienberg

Dr. Fienberg paved the way for forensic statisticians through his own work, as well as his support and mentorship of others. He was one of the earliest proponents of using statistics in the courtroom (Fienberg and Straf 1982). He often talked about his ideas for reforming the forensic science system, and later his ideas manifested in the creation of the Organization of Scientific Area Committees (OSAC), the National Commission on Forensic Science (NCFS), and the Center for Statistics and Applications in Forensic Evidence (CSAFE) of which Dr. Fienberg was a founding co-director. More personally, Prof. Fienberg was a major supporter of Chris' early work in forensic science on the FLASH ID system and the early experiments on score-based methods in collaboration with Dr. Amanda Hepler and Dr. Linda Davis (Davis et al. 2012, Hepler et al. 2012, Saunders et al. 2011). Danica fondly remembers Prof. Fienberg for his encouragement to continue pursuing forensic statistics after a particularly disappointing exchange with someone in the audience of one of her first oral presentations on her dissertation work. Prof. Fienberg's work in categorical data analysis has guided our application of categorical tables for handwriting evidence, particularly in advising the work of Danica's first Ph.D. student (Crawford 2020). We are both honored to author this chapter in his memory.

D. M. Ommen (✉)
Department of Statistics, Iowa State University, Ames, IA, USA
e-mail: dmommen@iastate.edu

C. P. Saunders
Department of Mathematics and Statistics, South Dakota State University, Brookings, SD, USA
e-mail: christopher.saunders@sdstate.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_10

10.1 Introduction

> I believe that the basic concept of probability and of weight of evidence should be the same for all rational people and should not depend on whether you are a statistician. – Good (1985)

Unfortunately, this is not the case regarding the weight of evidence in forensic applications where statisticians and forensic practitioners disagree. In statistics, there are a handful of pertinent topics related to "evidence." The first that comes to mind is the *strength of evidence* that we all learn about in our first course on statistics: the p-value. The p-value measures the support that sample data provide against a null hypothesis in a traditional hypothesis (or significance) test. There are a variety of the issues with using p-values alone for scientific conclusions and policy decisions that are summarized in the American Statistician (Wasserstein and Lazar 2016). So, it is no surprise that similar expressions regarding "evidence" mean different things in relation to forensic science and criminal justice where the stakes are high. In forensic statistics, the *value of evidence* is defined as the Bayes Factor (Aitken and Taroni 2004, Good 1991), but often referred to as the likelihood ratio. While statisticians agree that Bayes Factors and likelihood ratios can serve as the value of evidence, statisticians distinguish them as two different statistics, while the two are used interchangeably in forensics. In addition, the *weight of evidence* in forensics is defined as the logarithm of the value of evidence (Aitken and Taroni 2004, Good 1991). Both the value and the weight of evidence are used to make decisions regarding several aspects of the overall fact-finding process in justice proceedings. These aspects range from decisions regarding the source of individual pieces of forensic evidence to the overarching goal of deciding the guilt or innocence of a person in relation to committing a crime. Therefore, methods of quantifying the value of evidence play crucial roles in forensic science.

Let us start with a small snippet of recent history relevant to quantifying evidential value.[1] Most of the methods used to quantify evidential value are broadly called the "likelihood ratio approach." Naming these methods the "likelihood ratio approach" in forensics applications is likely a tradition initiated by discussions that followed Dennis Lindley's paper in Biometrika. In 1977, Lindley's seminal paper "A Problem in Forensic Science" detailed his solution to a problem regarding glass fragments from a broken window (Lindley 1977). This description of that problem was given in an earlier paper by Evett (1977). To summarize, a window was broken to gain entry into a residence for the purpose of committing a crime therein. A suspect was identified in relation to committing the crime, and glass fragments were subsequently found on the suspect. Did the glass fragments found on the suspect come from the window at the crime scene? In his paper, Evett detailed a solution using classical methods of statistical hypothesis testing, which utilize distribution

[1] For a more detailed history of statistical methods for forensic evidence, see Taroni et al. (1998).

functions (i.e. p-values), in a two-stage approach[2] developed by Parker (1966). In contrast, Lindley provided a solution which used Bayesian methods (i.e. the Bayes Factor). Lindley called his solution a "factor" that "uses only density functions, and adheres to the likelihood principle" (Lindley 1977). A short time later, Allan Seheult published a commentary on Lindley's paper, and in this paper Lindley's factor is named a "likelihood ratio" for what we believe is the first time (Seheult 1978). In Evett's follow-up paper regarding the glass problem, he perpetuates the vocabulary by calling his use of Lindley's factor a likelihood ratio (Evett 1986). This seems to be the point where the lines blurred between a Bayes Factor and a likelihood ratio, and the use of "Bayes Factor" seemed to disappear from forensics, being replaced by the misnamed likelihood ratio.

Given its obvious importance, there has been an ongoing debate about how to properly express the forensic value of evidence (NIST 2017). Some advocate for the use of a single number (for example, Taroni et al. (2016)), while others advocate for some sort of interval quantification that would provide the decision-maker with an idea of the uncertainty in the analysis (for example, Lund and Iyer (2017) and Sjerps et al. (2016)). While the debate surrounding the appropriateness of using intervals to quantify the value of evidence is outside the scope of the discussion in this chapter, we believe that the aforementioned arguments are confounded with a debate as to what the *value of evidence* actually means. In response to these debates, one forensic practitioner provided a valuable opinion on the matter of expressing the value of evidence: there is a gap between the sophistication of methods proposed by statisticians and the statistical knowledge of a typical forensic practitioner (Nordgaard 2016). We appreciate this viewpoint and hope to provide information in this chapter to help fill the gap.

At this point, you may be wondering to yourself, "Why does any of this matter?" Our reasoning is this: there have been several reports calling for stronger statistical foundations in forensic evidence interpretation (NRCC 2009, PCAST 2016). In order for that to happen, researchers across disciplines need to work together and that requires good communication. If the language were consistent between statisticians and forensic scientists, then there would be much less confusion regarding quantifying evidential value in the future. As the field of forensic statistics advances, it is becoming increasingly important to make the distinction between a Bayes Factor and a likelihood ratio to reduce the ambiguity surrounding an important and complex area of research. First, we will clarify several terms involving "Bayes." Then, we will discuss some of the current viewpoints regarding the relationships between the Bayes Factor and likelihood ratio. Finally, we will summarize the similarities and differences between a Bayes Factor and a likelihood ratio for the forensic identification of source problems.

[2]The Two-Stage approach is a valid method of forensic evidence interpretation that is currently used in practice. We will not discuss the Two-Stage approach in this chapter because it does not result in a value of evidence (as defined by Good (1991)).

10.2 The "Bayes" Confusion

Following Lindley, a number of researchers (Taroni et al. (2016), for example) and governing bodies (for example, the European Network of Forensic Science Institutes (ENFSI 2015)) have recommended the "Bayesian" approach to forensic evidence interpretation. The Bayesian statistical paradigm is distinct from the classical, or sometimes called the Frequentist, paradigm of statistics by *explicitly* incorporating subjective personal beliefs into the analysis.

It is often misunderstood that any statistical analysis which uses Bayes Theorem is Bayesian. This seems to be a connection that is made solely on the sameness of the names. In reality, *Bayes Theorem* is a tool that can be used for all types of probability. For example, classical statisticians use Bayes Theorem as a way to compute conditional probabilities when the outcome of one event partitions the sample space, i.e. $P(B_1 \cup B_2 \cup \cdots \cup B_n) = 1$ and $P(B_i \cap B_j) = 0$ for all $i \neq j$,

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}.$$

Bayes Theorem is often used in medical diagnostic applications. For example, a doctor examines a patient who presents a set of symptoms, A, and it is known that there are n possible unique causes, $B_1, B_2, \ldots, B_n$, of these symptoms. The doctor will need to find the most likely cause B_k of the symptoms A in order to treat the patient. This can be done via the *Bayes Rule Classifier* which says the most likely cause is the one with the largest conditional probability, $P(B_k|A) > P(B_j|A)$ for **all** $j \neq k$ (Izenman 2013). Under the classical paradigm of statistics, the doctor would use "objective" population data in order to define the probability $P(B_i)$, i.e. the relative frequency of observing the disease B_i in the population (where there is no restriction on the set of symptoms presented with the diseases). In contrast, under the Bayesian paradigm the doctor may also incorporate his/her/their own personal "subjective" beliefs into the definition of $P(B_i)$ based on experience and other relevant information available to the doctor at that time. For example, these probabilities may be altered if the doctor has knowledge (while other doctors may not have the knowledge, making it a personal experience) that a new vaccine has nearly eradicated one of the diseases since the population frequencies were last determined. To be clear, we are using the term "subjective" in a very broad sense, to cover a variety of Bayesian approaches for assigning probabilities. This includes the example of subjective beliefs given above (see Savage (1972) for further details), as well as other approaches utilizing conjugate, default, or non-informative prior probabilities (see Berger (1985) for further details).[3] It is worth noting at this point that the designation of population frequencies as "objective" and

[3] A full treatment of the numerous approaches for assigning probability in the Bayesian paradigm is beyond the scope of this chapter, but the literature on this subject is rich.

beliefs as "subjective" is intended to help clearly separate the two different types of probabilities, and it does not mean that one is superior to the other.

The decision process via the Bayes Rule Classifier is distinct from the recommended Bayesian decision process in forensic science. The confusion between Bayesian statistics and Bayes Theorem is likely caused by the strong reliance on the *odds form of Bayes Theorem* when quantifying the evidential value in forensics. In general, the odds form of Bayes Theorem can be used to compare the conditional probabilities of any two outcomes of the partitioning event, i.e. $P(B_k|A)$ against $P(B_j|A)$ for $j \neq k$ by taking the ratio of the probabilities given by Bayes Theorem:

$$\frac{P(B_k|A)}{P(B_j|A)} = \frac{\left(\frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}\right)}{\left(\frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}\right)} = \frac{P(A|B_k)P(B_k)}{P(A|B_j)P(B_j)}.$$

In the forensic context, the odds form of Bayes Theorem is used as a decision tool for comparing two competing hypotheses after some evidence has been collected (Lindley 1977).

$$\underbrace{\frac{P\left(H_p|e,I\right)}{P\left(H_d|e,I\right)}}_{\text{PosteriorOdds}} = \underbrace{\frac{P\left(e|H_p,I\right)}{P\left(e|H_d,I\right)}}_{\text{BayesFactor}} \times \underbrace{\frac{P\left(H_p|I\right)}{P\left(H_d|I\right)}}_{\text{PriorOdds}} \tag{10.1}$$

Starting with the prior odds,[4] the Bayes Factor is assigned a value and then multiplied by the prior odds to arrive at the posterior odds.[5] Here, P represents a measure of probability, the evidence is denoted by e, the relevant background information by I, and the two competing hypotheses by H_p and H_d which are intended to give competing viewpoints from the "prosecution" and the "defense" that explain how the evidence came to be.

Now, let us take a closer look at each of the components of Eq. (10.1), starting with the measure of probability P. Like the Bayes Rule Classifier, P can be defined in either the Bayesian or the classical way. Under the classical definition of probability, both the Bayes Rule Classifier and the odds form of Bayes Theorem cannot be used as decision tools for quantifying evidential value since they lead to useless prior probabilities for hypotheses (either 0 or 1 depending on the ground truth). Under the Bayesian definition, using the Bayes Rule Classifier as a decision tool would require an exhaustive partitioning of the sample space. This means that

[4]Note that this does not make sense in the classical framework because the probability of a hypothesis is either 0 or 1, depending on the ground truth, leading to a prior odds of either 0 or undefined (infinite).

[5]Again, this does not make any sense in the classical framework because the Bayes Factor becomes irrelevant. Regardless of its value, the prior odds will determine the resulting posterior odds, either 0 or undefined.

the forensic scientist would need to explore **all** possible explanations of the evidence (this is not a feasible method due to many realistic constraints). However, using the Bayes Factor as a decision tool requires only that the hypotheses be mutually exclusive/disjoint (and there is **no** requirement that the hypotheses be exhaustive). Thus far in the discussion, the Bayesian definition of probability is the only viable option for quantifying the value of forensic evidence via the odds form of Bayes Theorem. Therefore, it is the definition of the probability in the Bayesian way that makes this method Bayesian (and not the fact that it uses Bayes Theorem).

Using the style of Larry Wasserman's blog post (Wasserman 2012), we can summarize the main points of this section by:

Bayesian Probabilities $\neq$ using Bayes Theorem
Bayes Theorem $\neq$ Bayes Rule Classifier
Bayes Rule Classifier $\neq$ Bayes Factor

10.3 The Bayesian Paradigm

Now that we have clarified the role of Bayesian probabilities in the odds form of Bayes Theorem, there are several possible options for dealing with the remaining components of Eq. (10.1). In the case that the two competing hypotheses imply a set of parametric modeling assumptions for the evidence, denoted by M_p for the model corresponding to the prosecution hypothesis and M_d for the model corresponding to the defense hypothesis, then the Bayes Factor (BF) can be expressed as follows:

$$BF(e) = \frac{\int f(e|\theta, M_p)\, \pi(\theta|M_p)\, d\theta}{\int f(e|\theta, M_d)\, \pi(\theta|M_d)\, d\theta} \tag{10.2}$$

where θ denotes the collection of all parameters, f denotes the likelihood function for the evidence, and π represents the joint prior density (assuming that the densities exist)[6] for all the parameters. The BF given in Eq. (10.2) is indeed the *value of evidence* in the Bayesian paradigm.

In order to simplify matters going forward, we will only consider situations where the evidence is characterized by features (as opposed to pairwise (dis)similarity scores), and we will assume that the evidence has been generated according to a chosen set of statistical models that result in parametric likelihood functions denoted by f in Eq. (10.2). In addition, we will consider only the two forensic identification of source frameworks detailed in Ommen and Saunders (2018). To briefly summarize, the goal of the identification of *common source* problem is to answer the question,

> Do the two sets of evidence, each with unknown source, originate from the same unspecified source?

[6]Note in Eq. (10.2) the explicit use of prior probabilities for the parameter. This is another distinguishing characteristic of the Bayesian paradigm that is not present in the classical paradigm.

This type of problem is particularly useful in an investigative setting, where the goal is to connect two crime scenes before any person of interest has been identified. In contrast, the identification of *specific source* problem tries to answer the question,

> Did the evidence with unknown source originate from this specific source?

and is relevant to situations where a person of interest has been identified, and you want to link this person to the crime scene via the evidence. Under these two frameworks, the source-level propositions, the corresponding sampling models, and the forms of the likelihood functions are all described in a recent paper by Ommen and Saunders (2021). In addition, the entire collection of evidence can be split into subsets of evidence with unknown source(s) (often called the recovered evidence) and with known source(s) (often called the control evidence) (Ommen and Saunders 2021). The evidence with known sources includes a population of alternative sources, i.e. background database. In the remainder of the chapter, the subset(s) of evidence with unknown source will collectively be denoted e_u and the entire collection of evidence will be denoted e.

10.4 The Likelihood Paradigm

A natural alternative approach to the Bayesian paradigm is the likelihood paradigm[7] because its goals are similar to the Bayesian decision framework, but it does not allow for the use of subjective beliefs. The idea of the likelihood paradigm is to find the strength of support the data provide to the two competing hypotheses and stop short of making any decision regarding which hypothesis to choose (Royall 1997). The main workhorse of the Likelihood paradigm is the likelihood principle:

> If Hypothesis A implies the probability that $X = x$ is $p_A(x)$, while Hypothesis B implies the probability that $X = x$ is $p_B(x)$, then the observation $X = x$ is evidence supporting Hypothesis A over Hypothesis B if and only if $p_A(x) > p_B(x)$ and the Likelihood Ratio $p_A(x)/p_B(x)$ measures the strength of that support. (Royall 1997)

In this sense, the likelihood paradigm can be considered a "data-driven" paradigm as opposed to the "belief-driven" Bayesian paradigm.[8] Another reason that the likelihood paradigm of statistics is being considered for forensic applications is that its overarching theme aligns nicely to the role of the forensic scientist when presenting evidence in the justice system.

> The ultimate posterior probability, of guilt or innocence and their corresponding legal verdicts, is always a question for the fact-finder ... Expert witnesses must not trespass on the province of the jury by commenting directly on the accused's guilt or innocence, and

[7] Davis et al. (2012) use the likelihood paradigm for handwriting evidence.

[8] Again, this distinction is not intended to indicate that one method is better than the other or to say that the Bayesian paradigm does not allow for probabilities that are informed by the data. This is just to emphasize one particular distinction between the paradigms.

> should generally confine their testimony to presenting the likelihood of their evidence under competing propositions. (Aitken et al. 2010)

In the case that the two competing hypotheses imply a set of parametric modeling assumptions for the evidence, then the likelihood ratio (LR) function is defined by

$$LR(\theta; e_u) = \frac{f(e_u|\theta, M_p)}{f(e_u|\theta, M_d)} \tag{10.3}$$

which is a function of the parameters, θ, where e_u is only the evidence with unknown source and M_p and M_d are the models implied by the prosecution and defense hypotheses, respectively. Note that we prefer to denote the likelihood function by f for notational consistency between the two different paradigms. Under the likelihood paradigm, it is this function that forensic scientists should be presenting to a decision-maker upon the conclusion of their analysis (Royall 1997). Precisely how they should visually or graphically present this function, and then what the decision-maker should do with it once they have it, are two topics best left to future research and discussion.

When there is no uncertainty regarding the modeling parameter, then the LR is the value of evidence in the likelihood paradigm, and it is also equivalent to the BF from the Bayesian paradigm. In this case, the true LR takes the simplistic form below,

$$LR(\theta_0; e_u) = \frac{f(e_u|\theta_0, M_p)}{f(e_u|\theta_0, M_d)}. \tag{10.4}$$

Here, the value of θ is known and is the "true parameter value" defined by the sampling models for the evidence. We denote the "true parameter value" by θ_0 and consequently, the true LR represents a single value of the LR function. There are rare circumstances in real life where it is reasonable to assume that the true parameter value is known, such as simple DNA. This is due to the fact that single-source DNA has a strong scientific foundation where population frequencies can be estimated with a high degree of certainty. This special case where the BF and LR are equivalent may be partially to blame for the confusion between the two statistics since advances in DNA evidence interpretation were heavily researched and publicized. However, in most practical applications the value of θ_0 is rarely known with any degree of certainty. Therefore, the value of the true LR is fixed, but unknown. In contrast, when there is uncertainty regarding θ, the LR and the BF are not equivalent.[9] In the following section, we will show the conditions needed for the equivalence under the scenario of uncertainty in θ.

[9]Davis et al. (2012) are working in the situation that the parameters are unknown, and therefore all resulting likelihood ratio-based methods of evidence interpretation are ad-hoc.

10.5 Debates between the BF and LR

Academic debates regarding how to quantify the value of forensic evidence are so influential for future research directions and policy decisions that we believe the debates warrant additional attention, especially since the positions can be made clearer by explicitly stating which statistical paradigm is being recommended. First, consider the debate (as we have interpreted it) between Taroni et al. (2016) and Sjerps et al. (2016) regarding the appropriateness of interval quantifications for the value of evidence. This debate is so important that it sparked an entire special issue devoted to this topic in Science and Justice (Morrison 2016). To summarize, Taroni et al. (2016) advocate for a single number quantification of the value of evidence because they define the BF to be the value of evidence while Sjerps et al. (2016) advocate for interval quantification of the value of evidence because they define the value of evidence to be the LR. The reasoning is that Taroni et al. (2016) very clearly define probability in the subjective Bayesian way, whereas Sjerps et al. (2016) consider the Bayesian approach to dealing with parameter uncertainty as one possible option of many. Sjerps et al. (2016) are also open to other options for parameter estimation that correspond to using the likelihood paradigm.

Consider the following result from Ommen and Saunders (2021): the BF can be expressed by

$$BF(e) = \int LR(\theta; e_u)\, \pi(\theta|e, M_d)\, d\theta \tag{10.5}$$

where $\pi(\theta|e, M_d)$ is the posterior density for θ given the entire set of evidence e where the evidence has been generated according to the defense model M_d (Ommen and Saunders 2021). In other words, the BF is the expected value of the LR function (with respect to the subjective belief about the uncertain parameter). Under the Bayesian paradigm, (Taroni et al. 2016) recommend handling parameter uncertainty by performing the integration given in Eq. (10.5), and then providing the BF as the value of evidence. Under the likelihood paradigm, there are several possible ways of handling parameter uncertainty. For example, Sjerps et al. (2016) suggest providing the "plug-in" LR as an approximate value of evidence

$$LR(\hat{\theta}; e_u) \tag{10.6}$$

where $\hat{\theta}$ is an appropriate estimator of θ in addition to an interval for the LR derived from the interval obtained by estimating θ with $\hat{\theta}$. Therefore, the difference in opinion regarding the use of intervals follows inherently from the fact that the likelihood paradigm and the Bayesian paradigm have different methods of uncertainty quantification.

Another topic surrounding the debates concerning interval quantifications for the value of evidence is whether or not an "ideal value" for quantifying forensic evidence exists. We believe the essence of this debate is the paradigm of statistics

you choose to use for evaluating forensic evidence. If you subscribe to the classical paradigm, it is likely that you will define this ideal value as the one that corresponds to the true LR given by Eq. (10.4) (Morrison and Enzinger 2016). If you subscribe to the Bayesian paradigm, it is likely that you will not believe an ideal value for the BF exists since it represents a state of belief as opposed to a state of nature (Berger and Slooten 2016, Biedermann et al. 2016, Taroni et al. 2016). While the latter opinion is certainly justified within the Bayesian paradigm of statistics, it has caused some discord within the forensics community as to whether or not the Bayesian approach is justified. For example, many researchers and policy-makers agree with the opinion given in Martire et al. (2017):

> Indeed, unless the practitioner's beliefs converge on the truth, their opinions cannot assist the court to reach accurate outcomes.

The following theorem says that under the Bayesian paradigm and given very mild assumptions, an ideal value for the Bayes Factor is guaranteed to exist, and that the practitioner's beliefs will in fact converge on the "truth."

Theorem 1 (Bayes Factor Doob's Consistency) *Given a fixed observation of unknown source evidence e_u, suppose that $LR(\theta; e_u)$ is a bounded random variable with respect to $\Pi(\theta)$. Let the assumptions of Doob's Consistency Theorem be satisfied. Then for every prior probability measure $\Pi(\theta)$ on the parameter space Θ, the sequence of Bayes Factors, $BF(e)$, converges almost surely to the likelihood ratio, $LR(\theta; e_u)$, as $n \to \infty$ for Π-almost every θ and for P_θ^∞-almost every e^∞.*

The proof of this result is given in the Appendix. Note that the result relies on the assumptions that there is a fixed pair of hypotheses that imply a fixed set of parametric modeling assumptions for the evidence, and a fixed method of measuring and structuring the evidence. This result allows the prior distribution for the modeling parameters to vary from person to person provided that those distributions have a non-zero overlapping portion.

Finally, Theorem 1 shows the argument that objective probabilities are better because they lead to an approach where the value of evidence has an ideal value is baseless because an ideal value exists when using subjective probabilities as well. Another criticism of Bayesian beliefs is that many people fear personal experience and "guessing" often serves as a substitute for "objective," or "data-driven" probabilities. There is a good quote from a blog post by Larry Wasserman that nicely summarizes the alternative point of view:

> "Mindless Frequentist Statistical Analysis is Harmful to Science.
> Mindless Bayesian Statistical Analysis is Harmful to Science." (Wasserman 2012)

Therefore, "objective" frequencies and "subjective" beliefs both require a great deal of thought and care in order to be considered good science.

10.6 Similarities between the BF and LR

So far, we have tried to convince you that, within their respective paradigms, both the Bayesian and likelihood approaches to evidence interpretation yield valid results via the BF and the LR, respectively. Several other similarities between the BF and LR can be seen by comparing Eqs. (10.2)–(10.4). The first similarity is that you need to collect evidence in order to quantify the value of either the BF or LR. The second similarity is that you must assume that the data have been generated according to a particular statistical model in order to quantify either the BF or the LR.

Another similarity between the BF and the LR is the system for interpreting the resulting value. There are two broad categories of systems for interpreting the value of evidence: quantitative and categorical/verbal. Under the quantitative system, the value of evidence is interpreted differently depending on whether the value is above, below, or near one. Let V denote the value of evidence in either the Bayesian (i.e. $BF(e) = V$) or the likelihood (i.e. $LR(\theta_0; e_u) = V$) paradigm. When the value of evidence is *greater* than one ($V > 1$), it can be interpreted in the following way (ENFSI 2015):

> It is V times more probable to observe the evidence if H_p is true than if H_d is true.

When the value of evidence is *less* than one ($V < 1$), it can be interpreted as (ENFSI 2015):

> It is V times more probable to observe the evidence if H_d is true than if H_p is true.

When the value of evidence is *equal* to one ($V = 1$), it can be interpreted by (ENFSI 2015):

> The evidence provides no assistance in addressing the issues covered by H_p and H_d.
> (i.e. It is equally probable to observe the evidence if H_p is true or if H_d is true.)

Under the categorical/verbal system, the evidence is interpreted according to Table 10.1 which has been reproduced from the ENFSI Guidelines for Evaluative Reporting (ENFSI 2015). This table can be modified for values of evidence that are less than one by taking the reciprocal of the first column (for example, the range 2–10 becomes the range 0.1–0.5) and switching the order of H_p and H_d in the second column. This table provides one possible system of verbal equivalence statements, although others exist. There are several advantages and disadvantages of verbal equivalent scales. One advantage is that they offer an alternative to providing actual numbers for the value of evidence which has been shown to result in inconsistent results since different people interpret the magnitude of the numbers differently (Thompson and Newman 2015). One disadvantage is that the verbal statements can also mean different things to different people. For example, without looking at Table 10.1, which statement seems stronger to you, "far more probable" or "much more probable" or "appreciably more probable"? When these types of statements are offered in absence of the defining table, the issue of misinterpreting the strength of evidence remains unresolved.

Table 10.1 ENFSI verbal equivalence scale for value of evidence (ENFSI 2015)

Value of evidence	Verbal equivalent expression
1	The forensic findings do not support one proposition over the other.
	The forensic findings provide no assistance in addressing the issue.
2–10	... provide weak support for H_p relative to H_d.
	... are slightly more probable given H_p relative to H_d.
10–100	... provide moderate support for H_p relative to H_d.
	... are more probable given H_p relative to H_d.
100–1000	... provide moderately strong support for H_p relative to H_d.
	... are appreciably more probable given H_p relative to H_d.
1000–10,000	... provide strong support for H_p relative to H_d.
	.. are much more probable given H_p relative to H_d.
10,000–1,000,000	... provide very strong support for H_p relative to H_d.
	... are far more probable given H_p relative to H_d.
1,000,000 and up	... provide extremely strong support for H_p relative to H_d.
	... are exceedingly more probable given H_p relative to H_d.

10.7 Differences between the BF and LR

So far, we have defined the BF to be the value of evidence in the Bayesian paradigm and the true LR to be the value of evidence in the likelihood paradigm. One major difference is the definition of probability, the BF incorporates subjective probability, whereas the LR relies on objective probability. Recalling the debate between Taroni et al. (2016) and Sjerps et al. (2016), one simple explanation of the difference between the Bayesian approach and the likelihood approach to quantifying the forensic value of evidence is whether you choose to find the expected value of the LR function or whether you choose to "plug-in" an estimate of the parameter. Several other differences between the BF and LR can be seen by comparing Eqs. (10.2)–(10.4). The LR differs from the BF because the LR only depends on the unknown source (recovered) evidence, whereas the BF depends on the entire collection of evidence including the background database. Additionally, computing the exact value of the LR requires that you must know the values of the parameters that correspond to the statistical model (or that the parameters can be estimated with a high degree of certainty). In contrast, computing the BF requires you to only characterize your belief about the values of these parameters by a joint prior density. As a consequence, one major difference between the two is that the value of the LR will be the same for everyone, but the value of the BF can be different from person to person.

10.8 Discussion

To paraphrase from Wasserman (2012), Bayesian methods have good Bayesian properties and frequentist methods have good frequentist properties. As noted by Royall (1997), combining the two results in ad-hoc methods without the intrinsic properties of either. These ad-hoc methods will necessarily need further justification before they can be used.

In hopes of making this topic much clearer for your future endeavors, we leave you with the following summary of the similarities and differences between BFs and likelihood ratios.

- The BF is the value of evidence and the log of the BF is the weight of evidence. The BF is a Bayesian statistic because it uses subjective probability.
- The LR is an important statistic in the Likelihood paradigm because it conforms to the likelihood principle and uses objective probability.
- It is important to distinguish between the BF and LR for forensic evidence interpretation. Under standard conditions, the BF and the LR are not equivalent.
- You must have "data" in order to compute both BFs and LRs. The LR depends only on the "data" from the unknown source evidence, whereas the BF depends on the "data" from the entire set of evidence including the background database.
- In order to quantify either the BF or the LR, you must assume that the data have been generated according to a particular statistical model. In order to compute the exact value of the LR, you must know the values of the parameters that correspond to this statistical model. In order to compute the BF, you do not need to know, but your belief about the values of these parameters must be characterized by a joint prior density.
- The value of the LR (known or not) is the ideal value for a BF. This ideal value can only be reached by the BF when sample sizes for the evidence from known sources are very large (or when the value of the parameter is known).

10.9 Conclusion

In an effort to strengthen the statistical foundations of forensic evidence interpretation, it is necessary to stop the common practice of calling Bayes Factors "likelihood ratios," and vice versa. It is important to distinguish between the BF and LR since the two statistics differ in several ways. The importance of this distinction is evident in the current debates surrounding various aspects of the interpretation process for forensic evidence. While the interpretation of the resulting BF or LR is the same, the values (both magnitude and direction of support) themselves can be quite different. These similarities and differences are paramount to good research and policy-making in the future.

Acknowledgments We would like to thank our mentor Dr. JoAnn Buscaglia of the Federal Bureau of Investigation Laboratory Division for all her support throughout the years. We would also like to thank Dr. Peter Vergeer, Dr. Reinoud Stoel, Dr. Jay Kadane for many fruitful discussions which helped shape this manuscript.

Research supported by the National Institute of Justice, Office of Justice Programs, US Department of Justice under Award No. 2014-IJ-CX-K088. The opinions and conclusions or recommendations expressed in this chapter are those of the authors and do not necessarily represent those of the Department of Justice.

Research partially supported by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement #70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California Irvine, and University of Virginia.

Appendix

For ease of reference, Doob's Consistency Theorem is reproduced from van der Vaart (1998):

Theorem 2 (Doob's Consistency Theorem) *Suppose that the sample space* $(\mathcal{X}, \mathcal{A})$ *is a subset of Euclidean space with its Borel sigma-field. Suppose that the random vectors* $X_1, \cdots, X_n$ *are independent and identically distributed according to the probability measure* P_θ*, and that* $P_\theta \neq P_{\theta'}$ *whenever* $\theta \neq \theta'$*. Then for every prior probability measure* Π *on* Θ *the sequence of posterior measures is strongly consistent for* Π*-almost every* θ*.*

Now, the proof of the BF Doob's Consistency Theorem will proceed in two parts: first we will prove the case for the common source setting, and then we will prove the case for the specific source setting.

In order to facilitate the consistency results, we will need to define some additional notation. In the usual setting for the common source problem, the evidence will have a fixed number of sources from the alternative source population, n_a. However, the following result explores the behavior of the BF as the number of sources in the alternative source population increases. Let E_{a,n_a} denote a sequence of random variables corresponding to the generation of hierarchical samples from the alternative source population e_{a,n_a} where n_a is the index that denotes the varying number of sources in the alternative source population with a fixed number of samples from within each source. Also, let $P_{\theta_a}^{n_a}$ denote the joint probability measure on E_{a,n_a} for all $\theta_a \in \Theta_a$. Finally, let $e_{a,\infty}$ denote an infinite sequence of observations from the alternative source population with corresponding probability measure $P_{\theta_a}^{\infty}$ which is the limiting form of $P_{\theta_a}^{n_a}$ as $n_a \to \infty$.

Theorem 3 (Common Source Bayes Factor Consistency) *Given a fixed observation of* e_{u_1} *and* e_{u_2}*, suppose that* $LR_{cs}(\theta_a; e_{u_1}, e_{u_2})$ *is a bounded random variable with respect to* $\Pi(\theta_a)$*. Let the assumptions of Doob's Consistency Theorem be satisfied. Then for every prior probability measure* $\Pi(\theta_a)$ *on* Θ_a*, the sequence of Bayes Factors,* $BF_{cs}(e_{u_1}, e_{u_2}, e_{a,n_a})$*, converges almost surely to the likelihood ratio,*

$LR_{cs}(\theta_a; e_{u_1}, e_{u_2})$, as $n_a \to \infty$ for Π-almost every θ_a and for $P_{\theta_a}^{\infty}$-almost every $e_{a,\infty}$.

Proof For this proof, let $e_{n_a} = \{e_{u_1}, e_{u_2}, e_{a,n_a}\}$ denote an observation of the entire collection of evidence with alternative source population sample size n_a, and let $\Pi_{n_a}(\theta_a'|e_{n_a}, M_d)$ denote the cumulative distribution function corresponding to the posterior measure on the parameter space given the an observation of the entire collection of evidence under the defense model using the standard abuse of notation. We will also let $\delta_{\theta_a}(\theta_a')$ denote the cumulative distribution function corresponding to the probability measure degenerate at θ_a. By Doob's Consistency Theorem (van der Vaart 1998), for Π_a-almost every θ_a and for $P_{\theta_a}^{\infty}$-almost every $e_{a,\infty}$, then as $n_a \to \infty$

$$\Pi_{n_a}(\theta_a'|e_{n_a}, M_d) \to \delta_{\theta_a}(\theta_a') \tag{A.1}$$

for all continuity points θ_a' of δ_{θ_a}. Let D be the class of all Cadlag functions (van der Vaart and Wellner 2000) and let $g : \mathrm{D} \mapsto \mathbb{R}$ be a continuous map such that $g(D) = \int f dD$ for $D \in \mathrm{D}$ and bounded, continuous function f. Then Eq. (A.1) and the Continuous Mapping Theorem, imply that

$$g(\Pi_{n_a}(\theta_a'|e_{n_a}, M_d)) \xrightarrow{as} g(\delta_{\theta_a}(\theta_a'))$$

for all continuity points θ_a' of δ_{θ_a}, for Π_a-almost every θ_a, and for $P_{\theta_a}^{\infty}$-almost every $e_{a,\infty}$. Using alternative notation, this means that as $n_a \to \infty$

$$\int LR_{cs}(\theta_a'|e_{n_a}, M_d) d\Pi_{n_a}(\theta_a'|e_{n_a}) \xrightarrow{as} \int LR_{cs}(\theta_a'|e_{n_a}, M_d) d\delta_{\theta_a}(\theta_a')$$

for Π_a-almost every θ_a and for $P_{\theta_a}^{\infty}$-almost every $e_{a,\infty}$. Therefore, as $n_a \to \infty$

$$BF_{cs}(e_{n_a}) \xrightarrow{as} LR_{cs}(\theta_a; e_{u_1}, e_{u_2})$$

for Π_a-almost every θ_a and for $P_{\theta_a}^{\infty}$-almost every $e_{a,\infty}$. □

In the usual setting for the specific source problem, the evidence will have a fixed number of sources from the alternative source population, n_a, and a fixed number of samples from the specific source, n_s. However, the following result explores the behavior of the BF as n_a and n_s both increase. Let E_{a,n_a} be defined as in the common source problem and let E_{s,n_s} denote a sequence of random variables corresponding to the generation of samples from the specific source e_{s,n_s} where n_s is the index that denotes the varying number of samples. For simplicity, we will assume that $n_a = n_s \equiv n$ although the proofs can be altered to accommodate more flexible relationships between the sample sizes. Also, let P_{θ}^{n} denote the joint

probability measure on E_{a,n_a} and E_{s,n_s} for all $\theta \in \Theta$ where Θ is the joint parameter space for θ_s and θ_a. Next, let $e_{a,\infty}$ be defined as in the common source problem and let $e_{s,\infty}$ denote an infinite sequence of observations of the random variable $E_{s,\infty}$. Finally, let $e_n = \{e_u, e_{s,n_s}, e_{a,n_a}\}$ denote the entire collection of evidence, where e_∞ denotes an infinite sequence of observations from both the specific source and the alternative source population with corresponding probability measure P_θ^∞ which is the limiting form of P_θ^n as $n \to \infty$.

Theorem 4 (Specific Source Bayes Factor Consistency) *Given a fixed observation of e_u, suppose that $LR_{ss}(\theta; e_u)$ is a bounded random variable with respect to $\Pi(\theta)$. Let the assumptions of Doob's Consistency Theorem be satisfied. Then for every joint prior probability measure $\Pi(\theta)$ on Θ, the sequence of Bayes Factors, $BF_{ss}(e_n)$, converges almost surely to the likelihood ratio, $LR_{ss}(\theta; e_u)$, as $n \to \infty$ for Π-almost every θ and for P_θ^∞-almost every e_∞.*

Proof For this proof, we will use the standard abuse of notation and let $\Pi_n(\theta'|e_n, M_d)$ denote the cumulative distribution function corresponding to the posterior measure on Θ given the observation e_n under the defense model. We will also let $\delta_\theta(\theta')$ denote the cumulative distribution function corresponding to the probability measure degenerate at θ. By Doob's Consistency Theorem (van der Vaart 1998), for Π-almost every θ and for P_θ^∞-almost every e_∞, then as $n \to \infty$

$$\Pi_n(\theta'|e_n, M_d) \to \delta_\theta(\theta') \tag{A.2}$$

for all continuity points θ' of δ_θ. Let D and $g : \mathrm{D} \mapsto \mathbb{R}$ be defined as above for the proof of Theorem 3. Then the Continuous Mapping Theorem implies that

$$g(\Pi_n(\theta'|e_n, M_d)) \xrightarrow{as} g(\delta_\theta(\theta'))$$

for all continuity points θ' of δ_θ, for Π-almost every θ, and for P_θ^∞-almost every e_∞. Using alternative notation, this means that as $n \to \infty$

$$\int LR_{ss}(\theta'; e_u) d\Pi_n(\theta'|e_n, M_d) \xrightarrow{as} \int LR_{ss}(\theta'; e_u) d\delta_\theta(\theta').$$

Therefore, for Π-almost every θ and for P_θ^∞-almost every e_∞,

$$BF_{ss}(e_n) \xrightarrow{as} LR_{ss}(\theta; e_u)$$

as $n \to \infty$. □

References

Aitken, C. G. G., Roberts, P. & Jackson, G. (2010), *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings; Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses*, 1st edn, Royal Statistical Society's Working Group on Statistics and the Law, London, UK.

Aitken, C. G. G. & Taroni, F. (2004), *Statistics and the Evaluation of Evidence for Forensic Scientists*, 2nd edn, John Wiley and Sons, Ltd., West Sussex, UK.

Berger, C. E. & Slooten, K. (2016), 'The LR does not exist', *Science and Justice* **56**(5), 388–391.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer Series in Statistics, 2nd edn, Springer, New York, NY, USA.

Biedermann, A., Bozza, S., Taroni, F. & Aitken, C. G. G. (2016), 'Reframing the debate: A question of probability, not of likelihood ratio', *Science and Justice* **56**(5), 392–396.

Crawford, A. (2020), Bayesian hierarchical modeling for the forensic evaluation of handwritten documents, PhD thesis, Iowa State University Graduate Theses and Dissertations. 18078, https://lib.dr.iastate.edu/etd/18078.

Davis, L. J., Saunders, C. P., Hepler, A. & Buscaglia, J. (2012), 'Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios', *Forensic Science International* **216**(1-3), 146–157.

Eur (2015), *ENFSI Guideline for Evaluative Reporting in Forensic Science*.

Evett, I. W. (1977), 'The interpretation of refractive index measurements', *Forensic Science* **9**, 209–217.

Evett, I. W. (1986), 'A Bayesian approach to the problem of interpreting glass evidence in forensic science casework', *Journal of the Forensic Science Society* **26**, 3–18.

Fienberg, S. E. & Straf, M. L. (1982), 'Statistical assessments as evidence', *Journal of the Royal Statistical Society. Series A* **145**(4), 410–421.

Good, I. J. (1985), 'Weight of evidence: A brief survey', *Bayesian Statistics* **2**, 249–270.

Good, I. J. (1991), *"Weight of evidence and the Bayesian likelihood ratio" in The Use of Statistics in Forensic Science*, CRC Press, Boca Raton, FL, USA.

Hepler, A., Saunders, C. P., Davis, L. & Buscaglia, J. (2012), 'Score-based likelihood ratios for handwriting evidence', *Forensic Science International* **219**(1), 129–140.

Izenman, A. J. (2013), *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, Springer Texts in Statistics, Springer, New York, NY, USA.

Lindley, D. V. (1977), 'A problem in forensic science', *Biometrika* **64**(2), 207–213.

Lund, S. P. & Iyer, H. (2017), 'Likelihood ratio as weight of forensic evidence: A closer look', *Journal of the Research of National Institute of Standards and Technology* **122**(27), 1–32.

Martire, K. A., Edmond, G., Navarro, D. J. & Newell, B. R. (2017), 'On the likelihood of "encapsulating all certainty"', *Science and Justice* **57**, 76–79.

Morrison, G. S. (2016), 'Special issue on measuring and reporting precision of forensic likelihood ratios: Introduction to the debate', *Science and Justice* **56**, 371–373.

Morrison, G. S. & Enzinger, E. (2016), 'What should a forensic practitioner's likelihood ratio be?', *Science and Justice* **56**(5), 374–379.

National Research Council Committee on Identifying the Needs of the Forensic Sciences Community (2009), *Strengthening Forensic Science in the United States: A Path Forward*, The National Academies Press, Washington, D.C., USA.

NIST (2017), National Institute of Standards and Technology Technical Colloquium on the Weight of Evidence, U.S. Department of Commerce, https://www.nist.gov/news-events/events/2017/06/technical-colloquium-weight-evidence.

Nordgaard, A. (2016), 'Comment on 'dismissal of the illusion of uncertainty on the assessment of a likelihood ratio' by Taroni F., Bozza S., Biederman A. and Aitken C.', *Law, Probability, and Risk* **15**, 17–22.

Ommen, D. M. & Saunders, C. P. (2018), 'Building a unified statistical framework for the forensic identification of source problems', *Law, Probability, and Risk* **17**, 179–197.

Ommen, D. M. & Saunders, C. P. (2021), 'A Problem in Forensic Science Highlighting the Differences between the Bayes Factor and Likelihood Ratio', *Statist. Sci.* **36**(3), 344–359. https://doi.org/10.1214/20-STS805

Parker, J. B. (1966), 'A statistical treatment of identification problems', *Journal of the Forensic Science Society* **6**(1), 33–39.

President's Council of Advisors on Science and Technology (2016), *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*, Executive Office of the President of the United States.

Royall, R. (1997), *Statistical Evidence: A Likelihood Paradigm*, Vol. 71, CRC Press, Boca Raton, FL, USA.

Saunders, C. P., Davis, L. J., Lamas, A. C., Miller, J. J. & Gantz, D. T. (2011), 'Construction and evaluation of classifiers for forensic document analysis', *The Annals of Applied Statistics* **5**(1), 381–399.

Savage, L. J. (1972), *The Foundations of Statistics*, revised second edition edn, Dover Publications, Inc., New York, NY, USA.

Seheult, A. (1978), 'On a problem in forensic science', *Biometrika* **65**(3), 646–648.

Sjerps, M. J., Alberink, I., Bolck, A., Stoel, R., Vergeer, P. & van Zanten, J. H. (2016), 'Uncertainty and LR; to integrate or not to integrate, that's the question', *Law, Probability, and Risk* **15**(1), 23–29.

Taroni, F., Bozza, S., Biedermann, A. & Aitken, C. G. G. (2016), 'Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio', *Law, Probability, and Risk* **15**(1), 1–16.

Taroni, F., Champod, C. & Margot, P. (1998), 'Forerunners of Bayesianism in Early Forensic Science', *Jurimetrics* **38**(2), 183–200.

Thompson, W. C. & Newman, E. J. (2015), 'Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents', *Law and Human Behavior* **39**(4), 332–349.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge, UK.

van der Vaart, A. W. & Wellner, J. (2000), *Weak Convergence and Empirical Processes*, Springer Series in Statistics, Springer, New York, NY, USA.

Wasserman, L. (2012), 'Normal Deviate Thoughts on Statistics and Machine Learning: What is Bayesian/Frequentist Inference', https://normaldeviate.wordpress.com/2012/11/17/what-is-bayesianfrequentist-inference/.

Wasserstein, R. L. & Lazar, N. A. (2016), 'The ASA Statement on p-Values: Context, Process, and Purpose', *The American Statistician* **70**(2), 129–133.

Chapter 11
Statistical Issues in Assessing the Reliability of Eyewitness Identification

Karen Kafadar

11.1 Introduction

Eyewitness identification (EWI) plays a critical role in criminal cases, from the investigation of the event to the courtroom trial. Almost no other evidence can be as powerful to a jury as a victim who points to a defendant and says, "Yes, he's the one who raped me." Yet, as we all know from our own personal experiences, memory can be fallible. The Innocence Project found that over ∼70% of 375+ DNA exoneration cases involved mistaken eyewitness identifications.[1] Wells et al. [43], Clark et al. [9] as the source of filler error rates (eyewitness mistakenly identifies a known-innocent filler as the perpetrator) is as high as 21.2% when the true perpetrator was present and 34.5% when the true perpetrator was absent. For a procedure commonly used in law enforcement proceedings, these error rates are frighteningly high. What factors lead to such high error rates? Can the levels of any of those factors be controlled by law enforcement to ensure lower error rates?

Errors in EWI can arise from: (a) identifying an innocent suspect from a lineup (*false positive*), or (b) failure to identify the correct perpetrator from a lineup when the perpetrator is present (*false negative*). Either error has severe consequences: the innocent suspect goes to jail and the true perpetrator is free to commit further crimes.

A tragic, yet not rare, example of memory infidelity arose in the case of Jennifer Thompson who, on a July evening in 1984 as a college student in North Carolina, was sexually assaulted. From Thompson's sketch, the police collected a series of potential suspects and asked her to identify her attacker from a photo lineup. She

[1] `innocenceproject.org`, accessed 1 September 2021.

K. Kafadar (✉)
Department of Statistics, University of Virginia, Charlottesville, VA, USA
e-mail: kk3ab@virginia.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_11

tentatively selected Ronald Cotton from the lineup. The police detective asked, "You're sure?" She replied, "Positive. Did I do OK?" The detective reassured her: "You did great." Later, Thompson was asked again to identify her attacker from a second, physical lineup of potential suspects, *in which only Ronald Cotton was included from the first lineup*. Not surprisingly, only Cotton looked familiar to her (from the previous lineup), and she confidently picked the same man. Ronald Cotton spent over 10 years in prison before he was exonerated by DNA evidence in 1995.[2]

The processes of encoding the event, consolidating the information in the brain, and reconstructing it at a later time, are all imperfect, leading to memory degradation and misconstructed events. When those imperfect memories lead to misidentifications, the consequences are serious, both in convicting innocent suspects and in freeing the true culprit to commit further crimes.

Despite much research on memory, relatively few factors have been considered in eyewitness identification (EWI) research, especially in view of its critical role in criminal cases. It can be the most heavily weighted evidence in a case where the available forensic evidence is either lacking (e.g., insufficient biological material for a DNA analysis) or has questionable probative value (e.g., bite marks). In view of law enforcement's reliance on eyewitnesses, the National Academy of Sciences convened a Committee in 2013 to conduct an in-depth study of its procedures and practices, to better assess EWI error rates, and especially to identify the factors that contribute to its reliability or lead to misidentifications. The Committee issued its report in October 2014 ([31] hereafter, "NAS report") and described aspects of vision and memory, environmental factors, and law enforcement procedures, all of which influence the accuracy of eyewitness identification. The report also discussed the use of eyewitnesses in judicial proceedings and evaluated the laboratory studies that had been conducted to date, as well as the statistical methods used to analyze the data from them and to draw conclusions from them. This chapter discusses these aspects, with particular emphasis on the design of experiments and the statistical methods that have been, and can be, used to analyze data from EWI experiments.

I chose this topic for this special volume dedicated to Steve Fienberg because of his decades-long work on statistical methods that he developed for data to ensure the equitable administration of justice, and because of his important role on the National Academy of Sciences' Report Review Committee that led to a far better report than would have been possible without him. I will end this chapter by talking about that role, and all that I learned from him during the course of writing the NAS report.

11.2 Memory and Vision

While one may believe that one's memory can be faithfully recalled like a photograph, in fact, memory is fallible and can be influenced by many factors. Memories of events involve three processes: encoding (placing aspects of the

[2]For the heartbreaking details of this case, see www.thestory.org/stories/2013-06/jennifer-thompson.

event into memory), storage (maintaining the aspects in memory), and retrieval (subsequent recall of aspects from memory). Each stage of memory is subject to degradation:

1. *Encoding*: Events may not be fully captured or encoded at the time of their occurrence due to limited extent of short-term memory, distortion and interference from effects of previous and/or subsequent events (including emotional, behavioral, sensory, observational effects), visual acuity, passage of time, and environmental conditions (e.g., poor lighting), among other factors.
2. *Storage*: The long-term retention of events can be reduced by the mere passage of time, effects of intervening experiences that replace existing memories, advancing age, and degree of associated emotion when the events occurred: "The emotional content of stored memories is a factor that appears to promote long-term retention; memories of highly arousing emotional stimuli, such as those associated with a witnessed crime, tend to be more enduring than memories of non-arousing stimuli" [31, p43].
3. *Retrieval*: Errors arise at this stage from many sources, including similarities to aspects of previously encoded and stored memories (leading to confusion between the events), loss of information about source of memory, and time. Retrieving memory of events from many decades ago may be harder for some people than those that occurred only last year; the opposite may be true for other people. Some aspects of the event may be improperly solidified into memory by the comments of others who unconsciously reinforce one's erroneous statements about the event.

In all phases, environmental conditions and behavioral factors (e.g., stress), as well as the passage of time, can lead to further degradation in the accuracy of a witnessed event.

With all these potential effects leading to memory degradation, why is eyewitness identification even used? For some crimes, it may be the only evidence available. For some crimes, latent fingerprints cannot be captured with sufficient quality to be identifiable nor can DNA, even in trace amounts, be obtained; both can be expected to have more reliability (consistency) and greater accuracy (fewer false positives and false negatives) than EWI. Moreover, eyewitness testimony has been shown to be extremely powerful in the courtroom: jurors are highly affected by the victim who points to a defendant and says, "He is the one who attacked me. I am one hundred percent confident." But can eyewitness identification ever be one hundred percent accurate and routinely trustworthy?

In 2014, the National Academies released a report, *Identifying the Culprit: Assessing Eyewitness Identification* [31]. The report explained features of vision and memory, and summarized much of the research that had been conducted on factors affecting the accuracy and reliability of EWI. The report also discussed the study designs that had been used for EWI experiments and the statistical methods that have been used to analyze the data from these experiments. This chapter describes some of the findings in that report, as well as subsequent studies that have been conducted. This field of inquiry demonstrates the enormous value of

cross-disciplinary research, which Steve Fienberg championed throughout his life and impressed upon his students and collaborators all over the world.

11.3 Challenges in Assessing EWI Reliability

The study that led to the report [31] included psychologists, legal professionals, and law enforcement personnel, as well as one lone statistician. Per the charge from the Laura and John Arnold Foundation (now Arnold Ventures), the report discussed the following issues [31, p.12].

1. Assess the existing body of scientific research related to EWI;
2. Identify research gaps in the current literature, and suggest appropriate research questions to pursue that will further our understanding of EWI and that might offer additional insight into law enforcement and courtroom practice;
3. Identify relevant research from fields outside of existing EWI research;
4. Offer recommendations to law enforcement for best practices and procedures for conducting and reporting eyewitness identifications;
5. Offer recommendations for developing jury instructions;
6. Offer advice regarding the scope of a Phase II consideration of neuroscience research and other areas of research that might affect EWI accuracy and reliability.

Briefly, the report noted these issues:

1. Much of the published literature in this field comes from researchers in university departments of psychology.
2. The existing studies have been conducted using college students or online platforms (e.g., Qualtrics©, or Amazon Mechanical Turk©); neither is likely to recreate the stress of a real-life incident. For example, the college student who must participate in an EWI experiment as a psychology course requirement, or the online participant who views a video of an attempted robbery (for a modest participation fee), is not likely to experience the same degree of anxiety and stress as would occur to a victim or witness of a real-life crime. These online platforms for conducting experiments are popular, however, because many participants can be recruited in a short period of time.
3. Most experiments in this field examine only one factor at a time, thereby preventing the exploration of interaction among factors that influence EWI reliability. Even slightly more sophisticated designs, such as factorial and fractional factorial experiments, were rarely conducted. (Admittedly, complex designs on human subjects can be more difficult to execute than on components in production environments.)
4. More statistical methods could be, but have not been, applied to analyze the data from these experiments, and thus useful data are often ignored, because researchers could not agree on how to incorporate the extra data.

This chapter describes these findings in more detail.

11.4 Factors that Affect EWI Accuracy

Statisticians are accustomed to approaching problems by first identifying the task at hand and then listing possible sources of variation that can affect the outcome. In this case, the task can be described as one of *binary classification*:

		Eyewitness Classification	
		"Guilty"	"Innocent"
True status of suspect	Guilty	True+	False –
	Innocent	False +	True –

In actual practice, a law enforcement officer invites the eyewitness to identify the perpetrator from a lineup, which can be conducted in several ways. Thus, the accuracy of each "binary classifier" (eyewitness) depends on many factors, including both those that can be controlled by the law enforcement officer (e.g., procedures for conducting the lineup) and those that are beyond the control of law enforcement (e.g., distance between eyewitness and the culprit at the time of the crime). In the statistics design literature, the former factors are often called "signal variables" (levels can be adjusted for maximizing the outcome) and the latter factors are "noise variables." In the EWI literature, the terms are "system variables" (signal) and "estimator variables" (noise). Some of these factors include:

Signal (system) variables: Under control of law enforcement

- Protocol for lineup (e.g., live or photo; present simultaneously or sequentially)
- Number of people or photos in lineup (suspect plus "fillers")
- Degree of similarities between suspect and fillers ("fair lineup" if very similar; "biased lineup" if suspect stands out)
- Nature of instructions to eyewitness (e.g., delivered orally or in writing; short or long)
- Presence or absence of feedback to eyewitness
- Request for confidence in identification (e.g., "How sure are you?" or "On a scale of 0–5, how confident are you?")
- Administering official is/is not "blind" (not involved) to the circumstances of the event, including the eyewitness, crime type, location, and potential suspects.

Noise (estimator) variables: Beyond control of law enforcement

- Eyewitness's level of stress or trauma at time of incident
- Conditions affecting visibility (e.g., light or dusk or dark)
- Distance between eyewitness and perpetrator
- Presence/absence of threat (e.g., weapon)
- Presence/absence of distinctive feature (e.g., scar)
- Presence/absence of other distractions (e.g., people, physical structures)
- Common/Different race or ethnicity

- Time between incident & report (*retention interval*)
- Age of eyewitness.

An important goal for law enforcement officers is to conduct EWI procedures using levels of the signal variables that maximize accuracy while minimizing variability, irrespective of the operating levels of noise variables in any given situation. Thus, it is important to design experiments that vary more than one factor at a time, especially because variables may not operate independently (Box, Hunter, Hunter 2005). However, most studies vary only one factor at a time; e.g., lineup format (sequential versus simultaneous: [23, 30]) delay between incident and the eyewitness's identification for the police officers (Deffenbacher et al. [10], Kensinger et al. [20]; Dodson et al. [12]), presence or absence of a weapon [14], same versus different race [28], process instructions to eyewitness (fair versus biased: [8]), and age [35]. Very few studies have varied multiple factors: distinctive feature and lineup format [4, 5]; distinctive feature and weapon focus [6]; lineup format, weapon focus, and distinctive feature [6]; and lineup format, weapon focus, distinctive feature, cross/same race, and target present/absent [12]. The NAS report [31] encouraged more factorial experiments.

Factorial experiments are common in several scientific fields, but they are relatively (and surprisingly) sparse in the eyewitness literature. One reason for their sparsity may lie in the need for advanced statistical methods to analyze the data from them, methods that may be unfamiliar to researchers in the field. The next section describes the most commonly used methods, followed by methods that can take account of multiple factors (which also require adjustment for multiple testing).

11.5 Statistical Methods: "Sequential" versus "Simultaneous"

As noted above, most studies evaluate the effect of a single factor on EWI accuracy, and often by comparing only two levels of that factor. Many of the studies were limited to considering the accuracy of identifications arising from only two levels of a single "system variable": photo lineup format is either "sequential" versus "simultaneous;" i.e., photographs are presented to the eyewitness either one at a time or all together in a "photo array."[3] In a simultaneous lineup, the witness views all individuals at the same time, and either identifies one as the perpetrator or reports that the person seen at the crime scene was not in the lineup. In a sequential procedure, the witness views individuals one at a time and reports whether or not each one is the person from the crime scene. The sequential

[3] See http://theconversation.com/police-photo-lineups-how-background-colours-can-skew-eye-witness-identification-116329 for an example of a simultaneous lineup. The article notes the challenges in constructing a fair lineup, which include the potential bias from different background colors for all photographs and the instructions given to the eyewitness.

procedure had been in practice for decades as a possible way to reduce false identification rates, particularly after the supporting research by Lindsay and Wells [23], who "concluded that sequential presentation of lineups can reduce false identifications of innocent suspects by reducing eyewitnesses' reliance on relative-judgment processes." In the intervening years, an extensive literature has compared witness identification performance using simultaneous and sequential procedures. These comparisons typically have used photo-arrays.

Until recently, most studies compared the performance of simultaneous and sequential procedures in terms of a ratio of correct identifications (i.e., 'hit rate') to false identifications (i.e., 'false alarm rate'), a measure known in the eyewitness identification literature as the "diagnosticity ratio" (*DR*). (The "diagnosticity ratio" is also known in other disciplines by other names; e.g., "positive likelihood ratio" or "$LR+$ = Likelihood Ratio of a Positive Call"; see [22] Section 4.1.) The ratio corresponds roughly to confirming a positive ID; i.e., the diagnosticity ratio is related to the probability that an eyewitness who makes a positive identification correctly selected the true perpetrator. Most laboratory-based studies and systematic reviews showed that, with standard lineup instructions informing the witness that the perpetrator may or may not be present, the sequential procedure produced a higher diagnosticity ratio. That is, when a "witness" (study participant) viewed a video of a crime being committed, and later was asked to identify the culprit from a lineup, the ratio of correct to false identifications was often higher with the sequential than with the simultaneous procedure. Later research (e.g., [30]) proposed that the metric used to assess EWI accuracy was incomplete; using a different metric, simultaneous lineups led to higher accuracy. Which was correct?

A higher diagnosticity ratio (hit rate/false alarm rate) could result from a higher hit rate, a lower false alarm rate, or some combination of the two (or from some factors that affect either rate). Some early papers suggested that sequential procedures lead to fewer false alarms without changing the hit rate, which results in a higher diagnosticity ratio. More recent laboratory-based studies and systematic reviews typically show that sequential procedures are associated with a somewhat reduced hit rate accompanied by a larger reduction in the false alarm rate, thereby yielding a higher diagnosticity ratio than for the simultaneous procedure. (Results from these studies using the diagnosticity ratio have not been uniform: a recent field-based study comparing sequential to simultaneous procedures in a limited number of jurisdictions revealed a modest diagnosticity ratio advantage for the simultaneous procedure; cf. [43].) In light of this laboratory-based evidence showing a higher diagnosticity ratio for sequential lineups, many police agencies and policy makers adopted sequential lineups as their preferred EWI procedures.

However, comparisons based on a single diagnosticity ratio collapsed over all participants in that treatment arm (e.g., DR for sequential and DR for simultaneous) typically do not definitively reveal whether one procedure is "better" than the other. The EWI literature noted that a single diagnosticity ratio is affected by many factors, most importantly (a) how well a witness can *discriminate* the suspect from innocent members of the lineup and (b) by the witness's tendency to pick or not to pick someone from the lineup; this tendency is often termed *response bias*. In general,

given the same hit rate, more conservative responding will tend to lead to fewer false alarms than more liberal responding and hence produce a higher diagnosticity ratio. Thus, a single diagnosticity ratio can be an incomplete performance measure, because it fails to account for an important variable, namely *response bias*. As noted earlier, a higher diagnosticity ratio (hit rate/false alarm rate) could result from a higher hit rate, or from a lower false alarm rate (caused by "conservative" responding), or both [29, 30, 32, 44]. The important point is that another variable (here, *response bias*) affects the measure being used to compare performance, and failure to take this variable into account can lead to naive comparisons. The receiver operating characteristic (ROC) curve was proposed as a way to account for "response bias" by using "expressed confidence level" (ECL) as a proxy for it. The ROC curve here is a plot of the *DR*'s numerator, *hit rate* (*HR*, y-axis), versus the *DR*'s denominator, *false alarm rate* (*FAR*, x-axis), calculated from the responses of participants who expressed different levels of confidence in the identification. (In general, the ROC curve is a plot of *sensitivity* versus (1 – *specificity*), for different levels of a third variable; see; e.g., [2, 21].

Recognized, but often to a much lesser extent, is the fact that many other variables besides a proxy for "response bias" could easily affect sensitivity, specificity, or the diagnosticity ratio (see Sect. 11.6). Moreover, ECLs (expressed confidence levels) may have severe shortcomings as a "proxy" for "response bias" that render it insufficient for comparing two procedures, only some of which are listed below.

1. ECL is likely to be only an imperfect measure of *response bias*. A good measure of *response bias* is essential for ensuring ROC is appropriate for comparing performance of procedures. Little research has explored whether other measures would be more sensitive, and less variable, in capturing the tendency toward "conservative" versus "liberal" responding, nor how variable ECL might be for an individual presented with exactly the same circumstances and the same instructions at a later time.
2. *Discriminability* involves not only making a correct identification but also making a correct exclusion. As a plot of *hit rate* versus *false alarm rate*, the ROC curve captures the first aspect: the slope of this curve is the diagnosticity ratio at different ECLs. The second aspect involves the ratio *specificity*/(1 – *sensitivity*), or, using the terminology in the eyewitness literature, $(1 - FAR)/(1 - HR)$; see Sect. 11.6.
3. The ECL responses in a laboratory experiment are likely to be much different in real-life, highly stressful conditions, which are very difficult (if not impossible) to replicate in an academic setting.
4. An eyewitness's ECL of "50%" might mean something different under one procedure (e.g., "sequential") versus another (e.g., "simultaneous"); that is, the third variable on which the ROC is based could well depend on the procedures that the ROC curves are designed to compare.
5. More than just ECL may affect the diagnosticity ratio; an ROC curve that accounts for other variables may be more useful in evaluating two procedures than an ROC curve based on only ECL (Sect. 11.6).

6. In real life, law enforcement officials recognize the impracticality of asking an eyewitness to quantify his/her stated ECL as "10%", "20%", …, "100%" and rather will invite the eyewitness to express confidence in his/her own words. The translation of a response such as "pretty confident" might indicate a "5" on a 5-point scale for one law enforcement officer but a "4" for another officer, adding to the uncertainty in ECL.
7. The use of ECL to distinguish diagnosticity ratios via an ROC curve is sensible if ECL is related to accuracy. One would hope a more confident witness is more accurate, but recent research acknowledges that the strength of this association between ECL and accuracy likely depends on many factors, such as some of the system and estimator variables already presented in the previous section (and discussed further below); see also [36].
8. The points on an ECL-based ROC curve (false alarm rate, hit rate) for respondents that expressed at least a stated level of confidence are estimates of "true" proportions; a new sample of study participants will yield a different false alarm rate, different hit rate, and different number of participants in that ECL category. The uncertainties in these proportions are rarely displayed, but they should be, and they can indicate different conclusions; see below.
9. Finally, ROC curves are often constructed by plotting the "hit rate" versus "false alarm rate" at different levels of ECLs *across all subjects in the study*. In laboratory experiments where the majority of eyewitness identification studies have been conducted (and sometimes in field studies), the same subject might be asked to view more than one condition, responding "10% confident" in one condition but "50% confident" in another condition. The effects of this lack of independence among subjects on the bias and variability in the estimated ROC curve (which requires estimates of "hit rate" = sensitivity and "false alarm rate" = 1 – specificity) have yet to be quantified. Moreover, the plotted points themselves have uncertainty (see below).

Incidentally, the procedures governing the administration of "sequential lineups" that are used in laboratory settings may differ in substantive ways from those in real life. For example, in a lab setting, the "eyewitness" sees each photo only once and cannot "go back" to previous photos. Conversely, in a real-life setting, a real eyewitness may ask to see a previously viewed photo, and the law enforcement officer (who is administering the lineup) often will allow him/her to go back and view previously shown photographs. This discrepancy in the procedures between lab and real life is one of the reasons why the generalization of inferences from lab experiments to real-life practice may be questionable. The inability to faithfully replicate in a lab the stress of the scene on the eyewitness is another reason for the disconnect between lab experiments and real-life incidents.

Potential Uncertainties in ECL-based ROC Curve

The construction of the ECL-based ROC curve requires study participants to answer after each of their choices, "How confident are you in your decision?" Examples of scales for their responses are (1) 11-point scale "0%, 10%, 20%, …, 100%", (2) 7-point scale of 1 (not confident at all) to 7 (highly confident),

or (3) 6-point scale "0%, 20%, 40%, 60%, 80%, 100%". For example, Brewer and Wells ([3], Table 9, p.24) conducted an experiment with 1200 recruited participants (undergraduates and from the local community) who watched a video of a thief in a restaurant and later were asked to identify the thief from a simultaneous lineup of eight color photographs. Roughly 300 participants viewed a lineup with a "foil" that had either high or low similarity to the culprit, and where the true culprit was either present or absent.[4] They calculated different DRs depending on the ECL category: 3.5 ("0–20%"), 3.3 ("30–40%"), 5.9 ("50–60%"), 13.6 ("70–80%"), 38.3 ("90–100%"), for the 524 participants who chose a suspect from the lineup. (Interestingly, the DRs change little across these five ECL categories for the 676 participants who declined to choose any suspect: 1.0, 0.9, 1.2, 1.8, 2.3, respectively.) Mickes et al. [30, Table 1, p.367] then calculated the DRs using slightly different categories and found that they varied, mostly in the last (90–100%) category: 9.0 ("< 0%"), 9.3 ("< 30"), 10.7 ("< 50%"), 16.0 ("< 70%"), 43.2 ("< 90%"). Similar to the use of receiver operating characteristics (ROCs) used in diagnostic medicine to compare different diagnostic test modalities, these authors argued that, by plotting the hit rate (sensitivity) versus the false alarm rate (1 – specificity) collapsed over participants *who express a given ECL or higher*, the area under the resulting ROC curve (AUC) provides a more appropriate metric for comparison. In doing so, a comparison of AUCs in experiments comparing sequential (seq) and simultaneous (sim) lineups, AUC(sim) > AUC (seq). Notice that the slope of the ROC curve at the plotted point is the DR for that ECL category ("at least x% confident"); hence, a straight line would indicate no effect of ECL on DR.

Figure 11.1 shows data published from a similar (thief) experiment in Mickes et al. ([30] "Experiment 1a" in Table 3) that plots the hit rate (*HR*) versus the false alarm rate (*FAR*) for sequential (Q) versus simultaneous (M) lineups. Their plot (Fig. 11.6a, p.371) has been enhanced here in Fig. 11.1 by the addition of two curves that represent very optimistic standard errors on the data points (binomial variation in the reported *HR* and *FAR*). Indeed, simultaneous (M) appears to have a higher AUC than sequential (Q), especially at higher levels of *HR* and *FAR*. Data from another experiment analyzed in Liu [26] also indicated the possible superiority of simultaneous over sequential. But data from other experiments are far less clear: Fig. 11.2 shows the same type of curve, for data from "Experiment 2" in Mickes et al. [30, Table 3]. Suddenly, now, the advantages of simultaneous (M) over sequential (Q) are not so clear.

Does Higher Confidence Lead to Greater Accuracy?

The underlying premise of the ROC approach over the single collapsed DR is based on the theory that increased accuracy is associated with increased confidence in the eyewitness's identification. (Recall from paragraph 2 of this section that *DR*

[4]The actual counts in Table 2 in Brewer and Wells [3] are: 299 (high similarity, target present); 300 (low similarity, target present); 301 (high similarity, target absent); 300 (low similarity, target absent). In each of these four conditions, ∼150 participants viewed a "biased" lineup and the other half viewed an "unbiased" (sometimes called "fair") lineup.

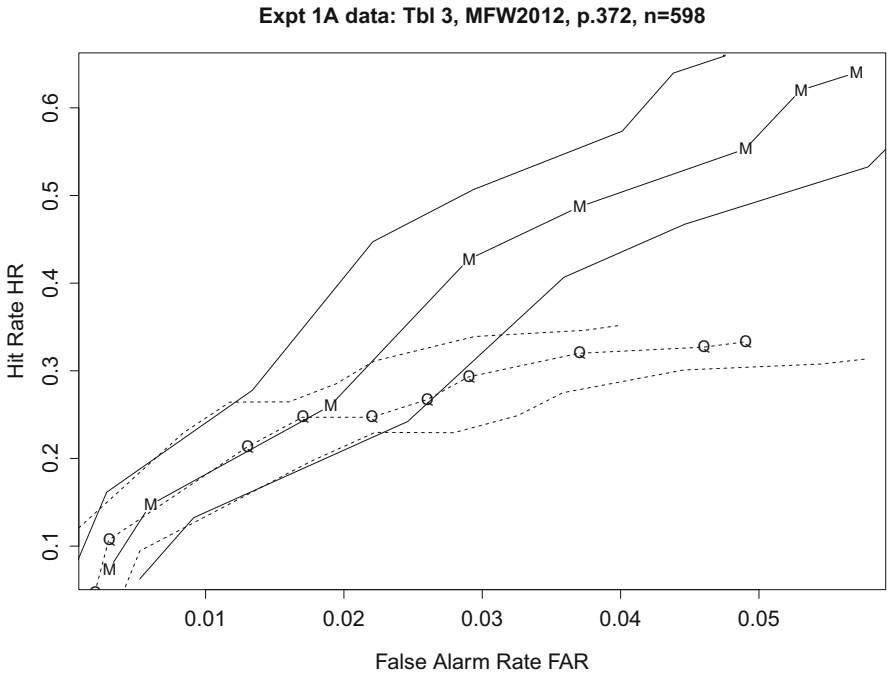


Fig. 11.1 "ROC curve" (hit rate = sensitivity versus false alarm rate = 1 – specificity), using data from Experiment 1A in [30, p.372]. Based on 598 participants who chose a suspect from one of two lineups: Sequential (Q) or Simultaneous (M). Shown with (optimistic) limits of one standard error in the estimated *HR* and *FAR* rates. Points on ROC curve refer to "At least 10% confident," . . . , "At least 90% confident," "At least 100% confident". Figure C-3 from [31], p.148: Reproduced with permission from the National Academy of Sciences, Courtesy of the National Academies Press, Washington, D.C.

is the same as the "positive likelihood ratio" $LR+$.) Data from several experiments suggest that such an association may exist. For example, Figs. 11.3 and 11.4 are plots of DR versus ECL using the same data from [30] as in Figs. 11.1 ("Experiment 1a") and 11.2 ("Experiment 2") above, respectively; see also [11]. However, studies are not consistent in this relationship; e.g., see Juslin et al. [19]. Further, given the multiple sources of variation that can affect ECL, the association is likely to vary greatly depending on the eyewitness, conditions of the event, and other factors yet to be identified.

Thus, even the migration from a single DR to an ECL-based ROC may not adequately capture all the factors that should be taken into account when comparing two procedures (such as "simultaneous" versus "sequential" lineups). Incidentally, few of the articles in the EWI literature take account of multiplicity in hypothesis testing: adjustment for multiplicity, such as the false discovery rate [1], appears to be very rare.

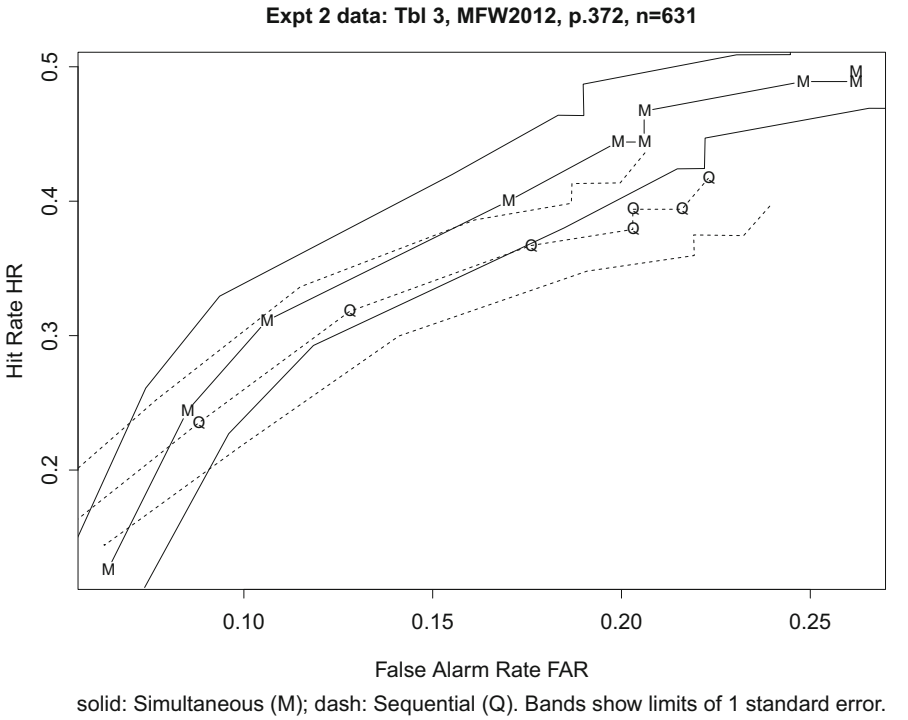


Fig. 11.2 "ROC curve" (hit rate = sensitivity versus false alarm rate = 1 – specificity), using data from Experiment 2 in [30, p.372]. Based on 556 participants who chose a suspect from one of two lineups: Sequential (Q) or Simultaneous (M). Shown with (optimistic) limits of one standard error in the estimated *HR* and *FAR* rates. Points on ROC curve refer to "At least 10% confident," . . . , "At least 90% confident," "At least 100% confident". Figure C-4 from [31], p.140: Reproduced with permission from the National Academy of Sciences, Courtesy of the National Academies Press, Washington, D.C.

11.6 Improved Statistical Approaches: PPV and NPV

The primary goal of an eyewitness procedure is to maximize *both* the *Positive Predictive Value* (*PPV*) as well as the *Negative Predictive Value* (*NPV*). Neither can be observed in real life. We *can* estimate two of the components in the definitions of PPV and NPV, namely *sensitivity* and *specificity*. *Sensitivity* is defined as the probability that an eyewitness, who is shown the *true* culprit, correctly identifies him/her as the culprit. *Specificity* is defined as the probability that an eyewitness, who is shown the *true* innocent suspect, correctly does not identify him/her as the culprit. Both probabilities can be estimated only in a designed study, where the study facilitators *know* the true status of the culprit and innocent suspects. In real life, one does not know the true status. At best, we can only estimate the probability that a positive ID correctly identified the true culprit ("positive predictive

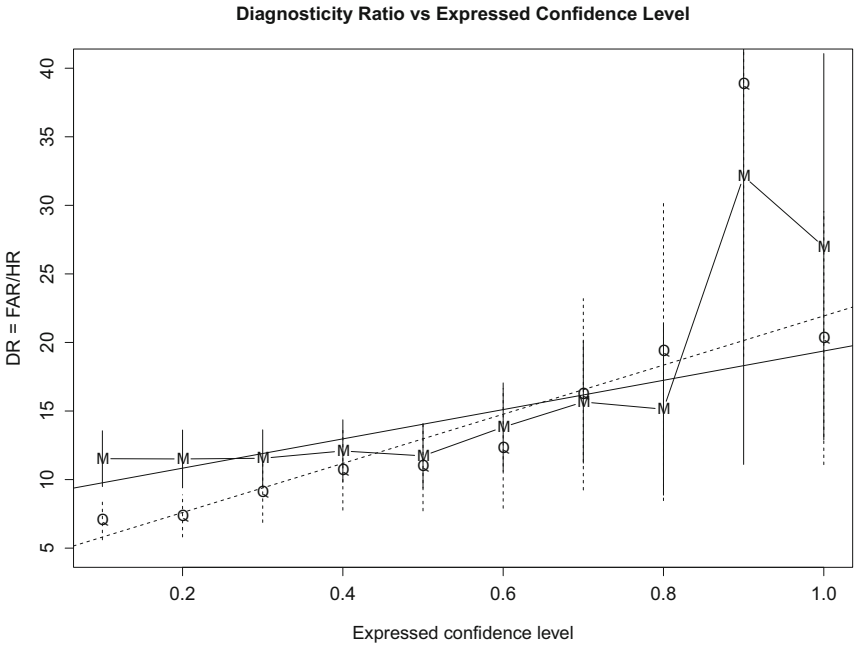


Fig. 11.3 Plot of DR = diagnosticity ratio = hit rate (sensitivity) divided by false alarm rate (1 – specificity), using data from Experiment 1a in [30, p.372]. Based on 598 participants who chose a suspect from one of two lineups: Sequential (Q) or Simultaneous (M). Shown with (optimistic) limits of one standard error in the estimated *HR* and *FAR* rates

value", or PPV) and the probability that no ID correctly recognized an innocent suspect ("negative predictive value", or NPV). Both PPV and NPV depend on three quantities (sensitivity; specificity; and prevalence, or the proportion of individuals that might be the culprit; i.e., 1 out of 6, or 1 out of 100, or ....), but in different ways, as explained below.

Let S denote *sensitivity*, or "hit rate," T denote *specificity*, or "1 – false alarm rate," and p denote the probability that the suspect is the perpetrator (*prevalence*). Let OR denote the "odds ratio" = $(1-p)/p$. Then

$$PPV = (Sp)/[Sp + (1-T)(1-p)] = 1/(1 + OR/d), d = S/(1-T)$$

$$NPV = [T(1-p)]/[T(1-p) + (1-S)p] = 1/[1 + ((1-S)/T)/OR].$$

Recall that $S/(1-T)$ = (hit rate/false alarm rate) is the same as the diagnosticity ratio (DR), or, more generally, $LR+$ = "Likelihood Ratio of a Positive Call;" it measures roughly the probability of a correct ID. So method 1 is preferred over method 2 if $PPV_1 > PPV_2$; i.e., $(OR_1/DR_1) < (OR2/DR_2)$; i.e., $(DR_1/OR_1) > (DR_2/OR_2)$, where DR_1, DR_2 denote the diagnosticity ratios for methods 1 and

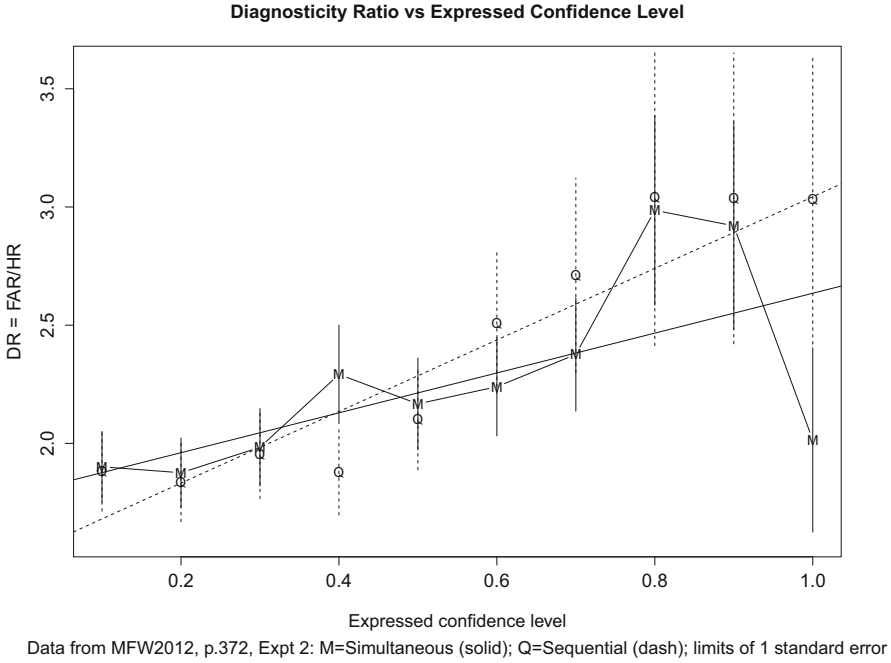


Fig. 11.4 Plot of DR = diagnosticity ratio = hit rate (sensitivity) divided by false alarm rate (1 – specificity), using data from Experiment 2 in [30, p.372]. Based on 556 participants who chose a suspect from one of two lineups: Sequential (Q) or Simultaneous (M). Shown with (optimistic) limits of one standard error in the estimated *HR* and *FAR* rates

2, respectively, under the same circumstances (i.e., $OR_1 = OR_2$, same conditions for both methods, etc.). Thus, a comparison of procedures in terms of criteria based on the diagnosticity ratio ($LR+$) is a comparison in terms of PPV, the probability that an eyewitness's ID really identified the true culprit.

To take into consideration the second aspect, NPV (probability that a non-ID correctly recognized an innocent suspect in the lineup), a second ratio is needed. Corresponding to $LR+ = S/(1-T)$, this ratio, $(1-S)/T$, is known as $LR- =$ "Likelihood Ratio of a Negative Call;" it corresponds roughly to the likelihood of ruling out a suspect ([22] §4.1). A calculation for NPV, similar to the one above for PPV, shows that method 1 is preferred over method 2 if $NPV_1 > NPV_2$; i.e., $T_1/(1-S_1) > T_2/(1-S_2) = 1/(LR-)_1 > 1/(LR-)_2$ – again, assuming that the odds ratio $(1-p)/p$ stays constant in the comparison between the two procedures (e.g., both the simultaneous and the sequential lineups have the same number of "foils"). Both the single diagnosticity ratio and the ROC curve (which shows separate diagnosticity ratios as slopes of the curve at the points corresponding to different levels of expressed confidence) address the PPV criterion; a separate analysis is needed for the NPV criterion. (Note that all calculations have variability,

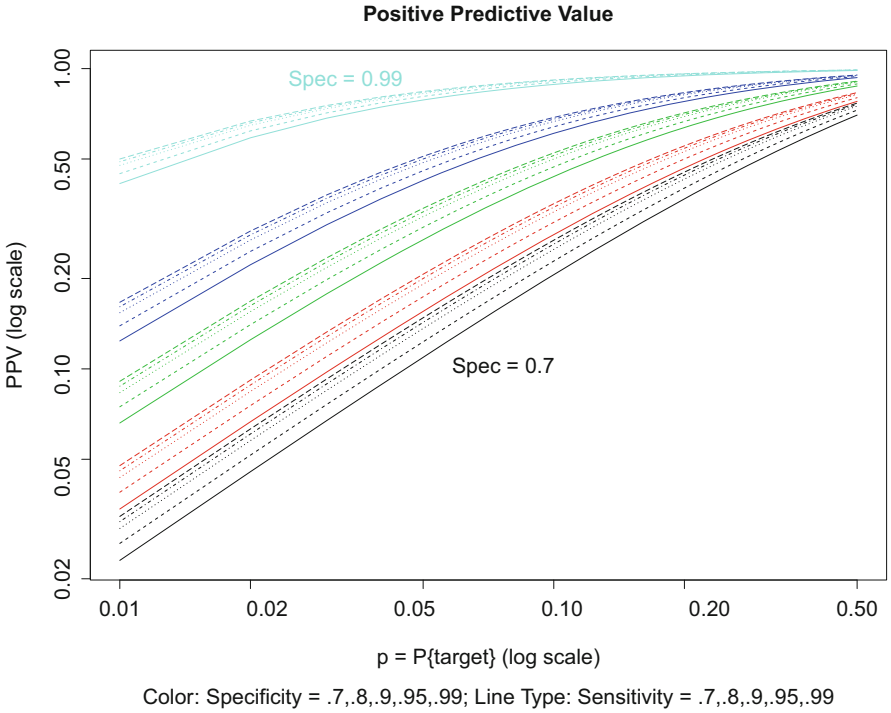


Fig. 11.5 Plot of Positive Predictive Value (PPV) versus prevalence, for levels of specificity (black = 0.7, red = 0.8, green = 0.9, blue = 0.95, turquoise = 0.99) and sensitivity (solid = 0.7, short dash = 0.8, dot = 0.9, dot-dash = 0.95, long dash = 0.99). Both PPV and prevalence are on logarithmic scales. Note that specificity (color) affects PPV much more than sensitivity (line type), and that both high specificity and high prevalence are necessary for high PPV

due to uncertainties in estimating S and T.) The effects of sensitivity, specificity, and prevalence on PPV and NPV are shown in Figs. 11.5 and 11.6, respectively. The figures show that *specificity* affects PPV more than *sensitivity*, that *sensitivity* affects NPV more than *specificity*, and that very high values of both sensitivity and specificity, as well as very high prevalence, are needed for high PPV and NPV.

Thus, although the ECL-based ROC curve is an "improvement" over the single DR collapsed over all subjects, in that it takes into account the eyewitness's expressed confidence in the ID (often seen as a proxy for "response bias" and shown in some studies to be correlated to some degree with accuracy), it focuses on only PPV and ignores NPV as well as other variables that could be affecting accuracy. For sufficiently large experiments, one may be able to construct ROC curves for each participant in the study, characterizing the information in each curve by a summary measure (such as logarithm of the Area Under the Curve; bigger is better) as a function of several factors. For example, [42] propose a hierarchical model for $\log(AUC)$ to distinguish within-participant variability from between-

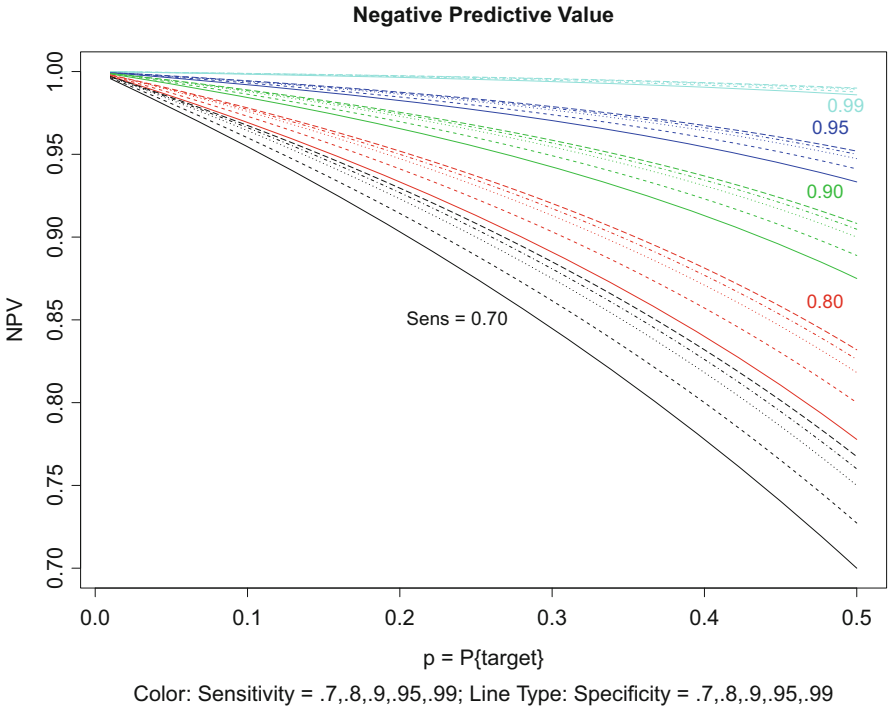


Fig. 11.6 Plot of Negative Predictive Value (NPV) versus prevalence, for levels of sensitivity (black = 0.7, red = 0.8, green = 0.9, blue = 0.95, turquoise = 0.99) and specificity (solid = 0.7, short dash = 0.8, dot = 0.9, dot-dash = 0.95, long dash = 0.99). Both PPV and prevalence are on logarithmic scales. Note that sensitivity (color) affects NPV much more than specificity (line type), and that both high sensitivity and low prevalence are necessary for high NPV

participant variability; other effects can be incorporated into the model.[5] Shiu and Gatonis [37] offer a methodology for incorporating both $LR+$ and $LR-$ through the "predictor receiver operating characteristic" (PROC) curve, by plotting PPV versus 1 – NPV. Other researchers have suggested approaches that would be even more useful, namely bivariate logistic regression models for logit-transformed sensitivity and specificity [34] generalized by Chu and Cole [7]. DuMouchel [13], Junaidi and Stojanovski [18], and Luby [27] also propose hierarchical models that are applicable for data from EWI experiments. See Liu et al. [25, Section 21.4] for further details about using these models in this context.

As noted at the start of Sect. 11.4, the basic task of the eyewitness is to serve as a *binary classifier*. The statistics and computer science literature contains many

[5]Notice that Wang and Gatsonis use AUC, versus pAUC = partial area under the curve; see discussions about AUC versus pAUC in the articles by Pepe [33, p.311] and Walter [40].

proposals for combining the information from binary classifiers; e.g., see [17]. Liu et al. [25, Section 21.4] successfully use random forests and mention other supervised methods that can be applied to the comparison of EWI procedures in the presence of multiple factors, including the "noise" ("estimator") variables that were mentioned above (lighting, distance, race, etc.). Probably the simplest approach would be ordinary logistic regression, which provides a model for the probability of being accurate in light of other manipulated variables [15]; see also Appendix C in the [31] report (pages 150–154) for an analysis of data from an experiment conducted by Carlson and Carlson [6]. The research in the EWI field is yet another example of the benefits of cross-disciplinary research involving statisticians.

11.7 Conclusion and Postscript

This article describes only briefly the state of the experimental designs and statistical methods that have been used in the literature on assessing the reliability of eyewitness identification procedures. The potential for statisticians to contribute to this field is immense, but not without its challenges.

I chose this topic for this chapter because I encountered these challenges while serving as a member of the Committee that authored the [31] report. Steve Fienberg served as co-chair of the Report Review Committee that was in charge of ensuring "that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered" [31, p.xii]. Despite multiple obligations on his plate at the time, Steve was masterful in his role. At one point, the Committee requested a call with him and his co-Chair, Dr. David Korn. With no advanced warning of the Committee's queries to them, Steve responded thoughtfully, eloquently, and precisely to each question raised by Committee members regarding various aspects of the review. His explanations of statistical concepts were crystal clear, emphasizing their complexities and relevance in an unambiguous but non-condescending manner. As I listened on the phone, I was wishing I could have recorded the call; it was impressive, even if he had been forewarned, but nothing short of spectacular in view of the fact that he had no idea what he would be asked. I always had great admiration for Steve from the time I met him, but that call impressed me immeasurably. And so it is, whenever I think of the NRC report, I remember with great fondness the influence that Steve Fienberg had on it, in ways that no one would ever know.

Acknowledgments An early version of this paper was presented at a workshop during the Probability and Statistics in Forensic Science program at the Isaac Newton Institute for Mathematical Sciences supported by EPSRC Grant Number EP/K032208/1. The author thanks the Institute for its hospitality and support during the program. The author also gratefully acknowledges her co-investigators on a grant from Arnold Ventures (Chad Dodson, Brandon Garrett, Joanne Yaffe). This chapter was prepared in part with support from this grant. The views expressed herein are solely those of the author and do not necessarily represent the views of Arnold Ventures or the Isaac Newton Institute.

References

1. Benjamini, Yoav; Hochberg, Yusef (1995): Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B* 57: 289–300.
2. Brown, Christopher D.; Davis, Herbert T. (2006), Receiver operating characterics curves and related decision measures: A Tutorial, *Chemometrics and Intelligent Laboratory Systems* 80: 24–38.
3. Brewer, Neil; Wells, Gary L. (2006). The Confidence-Accuracy Relationship in Eyewitness Identification: Effects of Lineup Instructions, Foil Similarity, and Target-Absent Base Rates," *Journal of Experimental Psychology: Applied* 12(1): 11–30.
4. Carlson C.A.; Gronlund S.D.; Clark S.E. (2008), Lineup composition, suspect position, and the sequential lineup advantage, *Journal of Experimental Psychology-Applied* 14(2): 118–128.
5. Carlson, C.A.; (2011). Influence of a perpetrator's distinctive facial feature on eyewitness identification from simultaneous versus sequential lineups, *Applied Psychology in Criminal Justice* 7:77–92.
6. Carlson, Curt A.; Carlson, Maria A. (2014). An evaluation of lineup presentation, weapon presence, and a distinctive feature using ROC analysis, *Journal of Applied Research in Memory and Cognition* 3(2): 45–53.
7. Chu, Haitao; Cole, Stephen R. (2006), Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach, *Journal of Clinical Epidemiology* 59:1331–1333.
8. Clark, Steven E. (2005), A re-examination of the effects of biased lineup instructions in eyewitness identification, *Law and Human Behavior* 29(4): 395–424.
9. Clark, Steven E.; Howell, Ryan T.; Davey, Sherrie L. (2008), Regularities in eyewitness identification. Law and Human Behavior, 32(3), 187–218. https://doi.org/10.1007/s10979-006-9082-4
10. Deffenbacher K.A.; Bornstein, B.H.; McGorty, E.K.; Penrod, S.D. (2008), Forgetting the once-seen face: estimating the strength of an eyewitness's memory representation, *Journal of Experimental Psychology: Applied* 14(2): 139–150.
11. Dodson, Chad S.; Dobolyi, David G. (2016), Confidence and eyewitness identifications: the cross-race effect, decision time, and accuracy. *Applied Cognitive Psychology*, 30:113–125.
12. Dodson, Chad S.; Garrett, Brandon L.; Kafadar, Karen; Yaffe, Joanne (2020), Eyewitness Identification Speed: Slow identifications from highly confident eyewitnesses hurt perceptions of their testimony, *Journal of Applied Research in Memory and Cognition* 10:259–267.
13. DuMouchel, William (1994), Hierarchical Bayes linear models for meta-analysis. Technical Report 27, National Institute of Statistical Sciences, 19 T.W. Alexander Drive, PO Box 14006, Research Triangle Park, NC 27709, September 1994.
14. Fawcett, J.M.; Russell, E.J.; Peace, K.A.; Christie, J. (2013), Of guns and geese: A meta-analytic review of the 'weapon focus' literature, *Psychology, Crime & Law* 19:35–66. https://doi.org/10.1080/1068316X.2011.599325
15. Garrett, Brandon L.; Liu, Alice J.; Kafadar, Karen; Yaffe, Joanne; Dodson, Chad S. (2020): Factoring the Role of Eyewitness Evidence in the Courtroom. *Journal of Empirical Legal Studies* 17(3): 556–579.
16. GEP Box, WG Hunter, JS Hunter (2005), *Statistics for Experimenters: Design, Innovation, and Discovery, 2nd ed.*, Wiley, New York.
17. Hastie, Trevor; Friedman, Jerome H.; Tibshirani, Robert (2013), *The Elements of Statistical Machine Learning, 2nd edition*, Springer: New York.
18. Junaidi, Darfiana Nur; Stojanovski, Elizabeth (2012), Bayesian estimation of a meta-analysis model using Gibbs sampler. In *Proceedings of the Fifth Annual ASEARC Conference - Looking to the future - Programme and Proceedings*.

19. Juslin, Peter; Olsson, Nils; Winman, Anders (1996), Calibration and diagnosticity of confidence in eyewitness identification: comments on what can be inferred from low confidence-accuracy correlation. *Journal of Experimental Psychology*, 22(5):1304–1316.
20. Kensinger, E.A.; Schacter, D.L. (2008), Neural processes supporting young and older adults' emotional memories, *Journal of Cognitive Neuroscience* 20(7):1161–1173.
21. Lasko, Thomas A.; Bhagwat, Jui G.; Zou, Kelly H; Ohno-Machado, Lucila (2005), The use of receiver operating characteristic curves in biomedical informatics, *Journal of Biomedical Informatics* 38(5), 404–415.
22. Lee, Peter (2012), *Bayesian Statistics: An Introduction, 4th ed.*, Wiley, New York.
23. Lindsay, R.C.L., Wells, Gary L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology*, 70(3), 556–564. https://doi.org/10.1037/0021-9010.70.3.556
24. Lindsay, R.C.L. (1999), Applying applied research: Selling the sequential lineup, *Applied Cognitive Psychology*, 13, 219–225.
25. Liu, Alice; Kafadar, Karen; Garrett, Brandon; Yaffe, Joanne (2020), Bringing new statistical approaches to eyewitness evidence. Chapter 21 in *Handbook of Statistics in Forensic Science* (Eds. David L. Banks, Karen Kafadar, David H. Kaye, Maria Tackett), CRC Press.
26. Liu, Alice Jia (2020). Errors-In-Variables and Random Forests: Theory and Application to Eyewitness Identification Data. Ph.D. Disseration, University of Virginia, Charlottesville, Virginia
27. Luby, Amanda S. (2017), Strengthening analyses of line-up procedures: a log-linear model framework, *Law, Probability and Risk* 16:241–257.
28. Meissner, C.A.; Brigham, J.C. (2001), Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review, *Psychology, Public Policy, and Law* 7(1): 3–35, https://doi.org/10.1037/1076-8971.7.1.3
29. Meissner, C.A.; Tredoux, C.G.; Parker, J.F.; MacLin, O.H. (2005), Eyewitness decisions in simultaneous and sequential lineups: A dual process signal detection theory analysis, *Memory & Cognition* 33, 783–792.
30. Mickes, Laura; Flowe, Heather D.; Wixted, John T. (2012), Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups, *Journal of Experimental Psychology: Applied*, 18(4): 361–376.
31. National Research Council (NRC 2014): *Identifying the Culprit: Assessing Eyewitness Evidence*, National Academies Press (https://doi.org/10.17226/18891).
32. Palmer, M.A., Brewer, N. (2012), Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior*, 36, 247–255.
33. Pepe, Margaret Sullivan (2000), Receiver Operator Characteristic Methodology, *Journal of the American Statistical Association* 95(449), 308–311.
34. Reitsma, Johannes B.; Glas, Afina S.; Rutjes, Anne W.S.; Scholten, Rob J.P.M.; Bossuyt, Patrick M.; Zwinderman, Aeilko H. (2005), Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, 58:982–990, 2005.
35. Rhodes, M.G.; Anastasi, J.S. (2012), The Own-Age Bias in Face Recognition: A Meta-Analytic and Theoretical Review," *Psychological Bulletin* 138(1): 146–174.
36. Roediger, H.L.; Wixted; J.H.; DeSoto, K.A. (2012), The curious complexity between confidence and accuracy in reports from memory, in: L. Nadel & W. Sinnot-Armstrong (Eds.), *Memory and Law*, Oxford, Oxford University Press, 84–118.
37. Shiu, Shang-Ying; Gatsonis, Constantine A. (2008), The predictive receiver operating characteristic curve for the joint assessment of the positive and negative predictive values, *Philosophical Transactions of the Royal Society A* 366:2313–2333.
38. Steblay, Nancy K.; Dysart, Jennifer; Fulero, Solomon; Lindsay, R.C.L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison, *Law and Human Behavior* 25(5): 459–473

39. Steblay, Nancy K.; Dysart, Jennifer; Wells, Gary L. (2011). Seventy-two Tests of the Sequential Lineup Superiority Effect: A Meta-Analysis and Policy Discussion, *Psychology Public Policy and Law* 17(1): 99–139,
40. Walter, S.D. (2005), The partial area under the summary ROC curve, *Statistics in Medicine* 24(13), 2025–2040.
41. Wells, Gary L.; Rydell, S.M.; Seelan, E.P. (1993), The selection of distractors for eyewitness lineups, *Journal of Applied Psychology*, 78, 835–844.
42. Wang, Pei; Gatsonis, Constantine (2008), Hierarchical models for ROC curve summary measures: Design and analysis of multi-reader, multi-modality studies of medical tests, *Statistics in Medicine* 27:243–256 (https://doi.org/10.1002/sim.2828).
43. Wells, Gary L.; Steblay, Nancy K.; Dysart, Jennifer E. (2015), Double-Blind Photo-Lineups Using Actual Eyewitnesses: An Experimental Test of a Sequential versus Simultaneous Lineup Procedure, *Law and Human Behavior* 39(1), 1–14. https://doi.org/10.1037/lhb0000096.
44. Wells, Gary L. (2014) Eyewitness identification: Probative value, criterion shifts, and policy regarding the sequential lineup, *Current Directions in Psychological Science*, 23:11–16.
45. Wells, Gary L.; Kovera, Margaret B.; Douglass, Amy B.; Brewer, Neil, Meissner, Christian A.; Wixted, John T. (2020), Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence, *Law and Human Behavior*, 44(1), 3–36. https://doi.org/10.1037/lhb0000359

Part IV
Introduction: Causal Analysis

Alicia L. Carriquiry

Many questions of interest in medicine, social science, public policy, and several other areas are causal in nature. For example, do smaller class sizes result in better learning outcomes for children? Does living next to a high voltage tower increase the probability of cancer? If speed limits are strictly enforced, will there be fewer fatalities on America's roadways? By understanding the mechanism that generated the data, we can sometimes establish a causal connection between a cause and its effect. Causal reasoning as practiced in statistics and epidemiology is forward looking, in that we wish to understand the effect of applying certain treatment (or factor) to a sample of items. Randomized controlled trials (RCTs) are the gold standard for this type of causal inference at the level of a group.

Steve was no stranger to causal inference research. He was an early proponent of the use of graphical models for representing causal pathways (Spirtes et al. 1999) and argued in favor of the principled use of causal inference in specific problems in public policy (e.g., Fienberg et al. 2003; Fienberg and Haviland 2003). But his real contribution was in what Steve called *individual causation* – how to argue causally from the group to the individual. Can we say that a specific person's cancer was caused by smoking? Or that a particular baby's brain injury was the result of shaking? Assigning responsibility for an outcome is a thorny problem, resulting in challenging technical issues. Yet, this is the question that arises most often in medicine, tort law, and other areas.

A related area of great interest to Steve was about the value of data to guide public policy. Steve was a strong believer and a tireless promoter of data-driven, or more generally, science-driven policy while recognizing the difficulties of devising and implementing such policies in real life. He was a proponent of RCTs for evaluating the effect of new policies and was one of the members of the National Academies of Science, Engineering, and Medicine (NASEM) panel on the use of social science evidence in policy. The panel, in which I also participated, published a report in 2012 entitled *Using Science as Evidence in Public Policy* (Prewitt et al. 2012).

The four contributions in this section are by Maria Cuellar, by Phil Dawid and Monica Musio, by Burt Singer and Ralph Horwitz, and by Norm Bradburn. The first two papers are closely related and focus on the question of individual causality in

the context of the law. Singer and Horwitz also discuss individual causality, but as applied in medicine, for the purposes of medical decision-making. Finally, Bradburn revisits the use of RCTs to evaluate policies and provides some insights on how to make them more effective. The four manuscripts in the section are particularly relevant for this volume; during the last year or two before he passed away, Steve dedicated significant effort to researching technical and practical issues associated to individual causation. Indeed, among the last papers written by Steve are Dawid et al. (2016) and Fienberg (2017), both discussing the *causes of effects (CoE)* problem.

Maria Cuellar was one of Steve's last doctoral students, and the only one to embrace the problem of *CoE* as her research topic. The motivation for Cuellar's work arose from an actual trial, in which a caregiver was accused of causing a baby's brain injury by shaking the baby. Cuellar's contribution focuses on the CoE problem in the context of attribution of blame or responsibility, but helpfully also provides an extensive introduction to the general area. She starts from the 2013 paper (Dawid et al. 2014) in which Dawid, Faigman, and Fienberg argue that causal analysis, as practiced by statisticians and epidemiologists, does not help answer questions posed in Court. This article inspired a lively debate between Judea Pearl, who was a discussant in the publication, and Dawid et al. Cuellar dives into the debate, and her syntheses and interpretations are tremendously valuable contributions to the ongoing discussions about individual causation.

Dawid and Musio synthesize existing work on the G2i (group to individual) causal problem from the last several years, and expand on it. The question they address is how to apply the ideas that underpin the *causes of effects* (CoE) paradigm to assign responsibility once an effect is observed. The probability of causation itself is not identifiable, but Dawid and Musio show that bounds on that probability can be computed and even improved upon when information about covariates or mediators is also available. In this paper, Dawid and Musio include an extensive discussion of the effect of covariates with different attributes, or mediators or both on the lower and upper bounds for the probability of causation.

Singer's and Horwitz's contribution discusses causal reasoning in the context of medical decision-making. The authors make a distinction between causal *reasoning,* useful for making decisions for an individual patient, and causal *inference,* useful for understanding the effect of some factor on the average response in a population. Singer's and Horwitz's argument is that traditional tools for causal inference including randomized controlled trials (RCTs) cannot help a clinician who must make decisions about the course of treatment for an individual patient. RCTs, they argue, provide information about the *average* response in a group of potentially heterogeneous patients, but have little to say about the effect of a treatment on an individual patient who may not look anything like those who participated in the RCT. This is a topic that would have resonated with Steve, who used to refer to himself as a sample of size one when undergoing treatment for his cancer. As Singer and Horwitz do, Steve used to argue that the only useful data to inform his treatment would need to come from other individuals with similar genetic, demographic, and medical history to his. Singer and Horwitz pick up on those ideas and propose the creation of *libraries of matched patients*, or collections of comparable patients from

whom a physician can extract information useful for making decisions about an individual patient.

Finally, the contribution by Norm Bradburn focuses on RCTs and their use in the evaluation of public policies. Bradburn relies on several famous policy failures to introduce the notion that public policies are not static objects and that by their very nature tend to effect a change in society that affects their effectiveness. Bradburn argues that even the most carefully conducted RCT cannot replicate the complexity of the real world and, thus, that one study cannot tell us whether a policy will succeed as intended when implement widely. He proposes that only a battery of RCTs carried out under varying conditions can hope to produce the information that policy makers should have when implementing a new policy in society. While Bradburn does not make a direct mention, his discussion is clearly related to the question of *reproducibility* of scientific studies that has been raised in recent years by, e.g., Ioannidis (2005).

References

Dawid, A.P., Faigman, D., Fienberg, S.E. (2014). Fitting science into legal contexts: Assessing effects of causes or causes of effects? (with discussion and response). *Sociological Methods & Research*, 43(3): 359–431.

Dawid, A.P., Musio, M., Fienberg, S.E. (2016) From statistical evidence to evidence of causality. *Bayesian Analysis*, 11(3): 725–752.

Fienberg, S.E. (2017) Evidence of child abuse: Inferring the causes of effects," Symposium on Child Abuse Evidence: Perspectives from Law, Medicine, Psychology, and Statistics, 50 U Mich. J.L. Reform 773 (2017).

Fienberg, S.E., Haviland, A.M. (2003). Can we measure the causal effect of discrimination? Discussion of Statistics and Causal Inference: A Review by Judea Pearl. *Test*, 12:139–147.

Fienberg, S.E., Glymour, C., Scheines, R. (2003) Expert statistical testimony and epidemiological evidence: The toxic effects of lead exposure in children. *J. Econometrics*, 113: 33–48.

Ioannidis, J.P.A. (2005) Why most published research findings are false. *PLOS Medicine*, 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Prewitt, K., Scwandt, T.A., Straff, ML. (2012) *Using Science as Evidence in Public Policy.* National Academy Press, Washington, DC.

Spirtes, P., Glymour, C., Scheines, C., Meek, C., Fienberg, S.E., Slate, E. (1999) Prediction and experimental design with graphical causal models. Chapter 2 in: *Computation, Causation, and Discovery*, (G. Cooper and C. Glymour, eds.), AAAI Press/MIT Press, pp. 65–93.

Chapter 12
Causes of Effects and Effects of Causes

Maria Cuellar

A Note About Steve Fienberg

Steve Fienberg introduced me to the causes of effects and effects of causes in 2015, when he was my doctoral advisor at Carnegie Mellon University. An attorney contacted Steve and asked him for help with a statistical argument regarding shaken baby syndrome, a brain condition in infants that is caused by shaking or other forms of child abuse and sometimes leads to death. There were numerous trials in which the prosecution would make a statistical argument about why an adult had abused or murdered a child either by shaking the child or through other violent means, but there was very little scrutiny in the courts about the statistics. Steve suggested that I work on this project.

I struggled to write a critique that would be helpful to the attorney and to the individuals who are being tried in court. The usual approaches that I had been taught in my courses seemed inadequate, not only because the data seemed to contain biases that I could not figure out how to eliminate, but also because the reasoning was wrong or incomplete. The statistical arguments in court were trying to determine whether, for a child with a brain condition, the child had been shaken. But even if the child was shaken, what if shaking did not cause the brain condition, and instead it was caused by something else, such as a traumatic birth? In that case, I thought, the adult should not be considered guilty of murder or child abuse.

Steve referred to me some articles he had written about the causes of effects and effects of causes. Once I understood this framework, I was able to vocalize what I thought was wrong with the current arguments in court: they were concerned with whether there had been an exposure, but they were not asking whether the outcome

M. Cuellar (✉)
Department of Criminology, University of Pennsylvania, Philadelphia, PA, USA
e-mail: mcuellar@sas.upenn.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_12

observed (the brain condition) had indeed been *caused* by the exposure in question (shaking).

Questions about the effects of causes address whether there was an exposure, or whether an exposure would, in general, cause that type of outcome. Questions about causes of effects address whether, for a child who was shaken and has a brain condition, the brain condition was caused by shaking, not by something else. In the law, causes of effects questions are related to the "but for" condition, which says that it must be shown that the outcome would not have occurred but for the exposure. This distinction is also discussed in the law as the difference between general and individual causation.

My research on shaken baby syndrome and causes of effects and effects of causes led me to be able to criticize the reasoning used by some researchers and expert witnesses in court. Attorneys have used my research to support arguments about the innocence of individuals. This is precisely how statistics should be used: as a force for the public good. I believe Steve would agree with this, since he strongly supported the use of statistics for informing public policy. This research helped shape my research, and it continues to affect what I do today.

12.1 Introduction

Dewayne Johnson, born in 1972, was a former school groundskeeper in northern California. Between 2012 and 2014, Johnson worked as a school groundskeeper, and during this time, he was exposed to Roundup, an herbicide produced by Monsanto, which is a company owned by Bayer since 2018. In 2014, Johnson was diagnosed with non-Hodgkins lymphoma, a type of cancer in blood cells, as a terminal condition. Johnson then sued Monsanto for having caused his cancer.[1]

What type of argument should Johnson's attorney make? The attorney could argue that when people are exposed to Roundup, they tend to get cancer; or, that when people get this type of cancer, it was because they were exposed to Roundup; or, that, given that Johnson was exposed and got cancer, it is likely that his cancer was caused by his exposure to Roundup and not something else. These three arguments are examples of forecasting, backcasting, and attribution, respectively. In general, the three types of arguments can be defined as the following questions:

1. *Forecasting*. How likely is it that an exposure to A will cause Y? For example, what is the effect of smoking on lung cancer or the effect of lead exposure on crime? This class of questions concerns average effects on individuals in populations and cases in which we have not yet observed the outcome. Mathematically, we would write this as an average causal effect, which, for a binary outcome Y, a binary exposure A, and a potential outcome Y^a, is

[1] Source: https://www.theguardian.com/business/2018/jul/23/monsanto-trial-dewayne-johnson-cancer-roundup-weedkiller.

$P\left(Y^{a=1}\right) - P\left(Y^{a=0}\right)$, or a risk ratio, $P\left(Y^{a=1}\right)/P\left(Y^{a=0}\right)$. This notation will be further discussed later in the chapter.

2. *Backcasting*. Given an outcome Y, how likely is it that the individual had an exposure A? This question deals with uncertainty about whether there was an exposure. For example, for people who have lung cancer, what is the probability that they were smokers, or for people who are violent, what is the probability that they were exposed to lead? Mathematically, we could write this as $P(A = 1|Y = 1)$.
3. *Attribution*. For someone who was exposed to A and had an outcome Y, how likely is it that the outcome was caused by the exposure and not something else? For example, Jane has lung cancer. Was it the five daily cigarettes that she smoked for the past 30 years that caused it, or was it something else? This class of questions concerns average effects on individuals in populations and cases in which we have already observed the outcome. Mathematically, we can write this as the probability of causation, which, for a group of individuals with covariates x, is

$$PC(x) = P\left(Y^{a=0} = 0|Y = 1, A = 1, X = x\right). \tag{12.1}$$

In other words, for an individual who had a positive outcome (e.g. cancer) and a positive exposure (e.g. to Roundup), what is the probability that, had he not been exposed, he would not have had the outcome? Estimating this quantity requires counterfactual reasoning: we are not able to observe the outcome when the individual is exposed and when he is not—unless we can make some further assumptions about the outcome being reversible. Thus, we must make identification assumptions to arrive at observable quantities that can be estimated using statistics or machine learning methods.

Johnson's attorneys should make an argument that combines these three types of questions. It matters whether Roundup is carcinogenic, whether Johnson was actually exposed, and, if he was exposed to Roundup and has cancer, whether his cancer was caused by the exposure. Forecasting and backcasting jointly are called the "effects of causes" (EoC), and attribution is called the "causes of effects" (CoE) (Dawid et al. 2013).

Forecasting receives most of the attention in trials. Indeed, in Johnson's case, the expert witnesses focused primarily on whether Roundup was carcinogenic. In the trial, the jury ruled against Monsanto because it seemed there had been some evidence, undisclosed by Monsanto, that Roundup was harmful to humans.[2] Backcasting is also often discussed, as in cases of shaken baby syndrome, where the most common question is whether a baby was shaken or not. The question of

[2]Source: https://www.baumhedlundlaw.com/toxic-tort-law/monsanto-roundup-lawsuit/dewayne-johnson-v-monsanto-company/#transcripts.

attribution is discussed less often, and the statistical analysis often does not address attribution directly.

This chapter discusses how to formalize questions of attribution, or CoE, by using the probability of causation. It begins with some background on the intersection between statistics and the law in issues of CoE (Sect. 12.2). It continues with a description of a debate about how to define and estimate the probability of causation between A.P. Dawid, D.L. Faigman, S.E. Fienberg and J. Pearl (Sect. 12.3), in which the main points of difference are highlighted. Some questions that were not addressed directly in the debate but are nevertheless important for the application of the CoE framework are then discussed (Sect. 12.4). The debate leads to a comprehensive formulation and identification of the probability of causation (Sect. 12.5). Finally, Sect. 12.6 offers an opinionated summary of the current state of statistical causation in the legal setting and suggests some topics for future research in this area.

12.2 Causal Questions in the Law

Attributions of guilt generally require attributions of causality, as in determining whether an individual's ailment was caused by a specific exposure. In legal proceedings, the two parties (plaintiff and defendant) must follow important pieces of guidance known as burdens of proof and rules of evidence. These rules determine the amount of evidence necessary to accomplish that goal. In most civil cases, the burden of proof that applies is called "a preponderance of the evidence," for which a trier of fact (judge or jury) must return a judgment in favor of the plaintiff if the plaintiff is able to show that a particular event was more likely than not to have occurred. In most criminal cases, the standard is called "beyond a reasonable doubt," and it is the highest standard of proof that may be imposed upon a party in court. This standard requires that the prosecution shows that the most likely explanation was the one proposed.

Thus, in a specific case, we can suppose that if causality is determined beyond the evidentiary standard, the trier of fact will likely side with the plaintiff or prosecution. However, there is a surprising gap between this principle and legal practice. In fact, there is no single standard for making attribution claims in legal settings. For example, the judge or jury must decide not only how to evaluate the evidence with which they are presented but also which standards should be used to determine whether that evidence shows causation of the particular outcome suffered by the harmed party.

In an adversarial system, such as the courts of the United States, the prosecution will make causal attributions that incriminate a defendant, while the defense will usually deny these attributions. Both sides may cite statistical evidence that is admissible in court or may rely on statistical or probabilistic claims made by expert witnesses. But while there are legal standards (e.g. Daubert and Frye) concerning who may be admitted to give expert testimony and what can be said in court,

there is no standard constraining how prosecutors, defense attorneys, or witnesses formulate causal claims. A statistical claim that determines causality in one case might be considered too weak to determine causality in another. Because they are not defined statistically, the general principle that causal claims must be proved "beyond a reasonable doubt" (or a preponderance of evidence) may be applied differently in different settings. Differential application of a legal principle could result in potentially catastrophic miscarriages of justice.

Statisticians have recently worked to improve this situation by formulating a statistically competent theory of legal causation. While causality itself was long considered a topic of philosophy rather than of statistics, since the pioneering work by Holland (1986), Rubin (1974), Neyman (1923), and Pearl (2009) and others, statistical models of causation have come to be a part of mainstream statistical practice. Statistical causal inference is an active field of research, largely focusing on how to estimate the probability that a specific outcome (e.g. a harm) was caused by a specific exposure (e.g. a chemical, a specific event). In the field of statistics, Dawid et al. (2013) and Pearl (2009) define this probability as a causal parameter, while in the field of epidemiology, Robins and Greenland (1989) and Green et al. (2011) define this probability as a statistical parameter. It is curious that under some identifiability assumptions, the causal parameter from statistics is equivalent to the statistical parameter from epidemiology. But, as far as the author is aware, the connection between these two literatures has not been explored in depth.

Questions of attribution can be answered by using the probability of causation: the probability that an outcome was caused by a specific exposure and not something else. Statisticians have been able to calculate the probability of causation in some contexts (Dawid et al. 2016). Given data from a well-designed randomized controlled trial of a particular drug, statisticians are able to estimate the value of the probability of causation that taking the drug relieved patients of a particular symptom, for example. However, there are certain difficulties in applying the concepts of statistical causality in the legal context.

First and perhaps most problematically, standard theories typically calculate causation in a population, in which a control group is compared to a group that received a treatment. In contrast, in the legal context, one is concerned primarily with determining the cause of an individual outcome, for example, a concrete harm. For a statistical theory of causation to be relevant in the legal context, it must offer identification assumptions under which population-level data can be applied to an individual case. This is sometimes referred to as making inferences from "type" causation to "token" causation and sometimes as the g2i (group-to-individual) problem (Dawid et al. 2016, Pearl 2009). Second, there is a question concerning whether it is proper to use a Bayesian framework or to approach calculations of causation from a frequentist perspective. If the probability should be interpreted as subjective belief, then to estimate it, the researcher must choose a prior. But if the probability is a frequency, then it is unclear what evidence should be used in the legal setting since few randomized trials exist. While the disagreement between Bayesians and Frequentists goes far beyond the issue of legal causality, it must also be addressed in this case. Third, there is the question of which specific

quantities should be estimated to make causal attributions, and what assumptions are warranted in making the calculation. What should the probability be conditioned on? How should the counterfactual quantity be stated? What is the type of question that can be answered if this probability were estimated? What happens when there are multiple causes, shorter exposure duration, a long time from exposure to outcome, a very small "dose" in the exposure, and so on? These are all questions that might be relevant in a trial. Can the causal quantity, and its identified version, answer these questions?

Later, identification assumptions will become highly relevant. Regarding identification assumptions, for instance, is it reasonable to make the assumption that a harmful exposure can only hurt, not help, a specific condition? This type of requirement is the "monotonicity" assumption, and it helps identify a closed-form statistical quantity from the causal quantity of the probability of causation. Can it be assumed that for a group of similar individuals, whether they were exposed is independent of what their outcomes would have been under treatment or control? This is the "no unobserved confounders" assumption, and it can be made in situations in which the treatment is randomized. Can it be assumed that anyone in the sample of individuals could have had the exposure? This is the positivity assumption, which can be strong depending on the specific circumstances. And can it be assumed that, if an individual is exposed, the outcome for that individual is the outcome he or she would have had under exposure? Or, as can happen in vaccine trials, can the individual's outcome be affected by another individual's treatment. This type of assumption is called "consistency," and it fails under interference.

12.3 A Debate About the Probability of Causation

Dawid, Faigman, and Fienberg published an article in 2013 in the journal Sociological Methods & Research titled "Fitting Science Into Legal Contexts: Assessing Effects of Causes or Causes of Effects?" Dawid, Faigman, and Fienberg (2013) argue that, on the one hand, statisticians and quantitative social scientists typically study the effects of causes (EoC). On the other hand, attorneys and the courts are more concerned with understanding the causes of effects (CoE). Thus, the evidence that is cited in court is often useful to answer EoC questions. But, because it does not focus on CoE, the evidence is often not relevant to assign blame to a specific individual. The authors provide a substantive survey of the differences between how causation is framed by courts and how it is framed by philosophers and scientists. They "examine how law and science might better align their approaches to causation so that, in particular, courts can take better advantage of scientific expertise." The audience for the article is presumably social scientists and attorneys, given that the motivational examples take up most of the article, that very few formulas

are provided, and that the formulas provided are simplified. This simplification, although it might have been meant to clarify the issue for researchers outside statistics, was noted by J. Pearl (2015) as a flaw in their analysis.

In his response article, Pearl (2015) focuses on three issues in his critique of Dawid et al. (2013). First, Pearl claims that their analysis only applies to determinations of causation in a population, which is irrelevant to the courtroom context, where the question is whether an individual caused a certain outcome. Second, Pearl argues that Dawid and his co-authors are only able to calculate the probability of causation for a population that has gone through a randomized experiment. Since the defendants in court cases are not randomly selected, this is an unwarranted assumption in the legal context. Finally, it seems as though Pearl is interpreting the probability of causation as frequentist (although this is not said explicitly or tested with data), while Dawid, Faigman, and Fienberg claim that it should be interpreted as a Bayesian quantity. Dawid, Faigman, and Fienberg respond to Pearl in Dawid et al. (2014). They argue that because it oversimplifies and idealizes legal causality, Pearl's model is not applicable in the legal context. They claim that there is no clear way that a judge or jury could implement Pearl's procedure to obtain an actual estimate of whether an individual caused an event to happen. Moreover, they criticize the specific method Pearl uses to make inferences from population data to cases of individual causality.

The debate between Pearl and Dawid, Faigman, and Fienberg is important because it allows us to dissect the estimation problem into four pieces: (1) How should we *define* the correct probability of interest in Causes of Effects questions? (2) How should we *identify* this probability? (i.e. What assumptions are we willing to make?) (3) How should we *estimate* this probability? And (4) How should we *use* the results from a Causes of Effects analysis? In the next section, we describe the authors' arguments from each article, and we summarize the debate to shed light on using statistics to speak about causal questions in the law.

Arguments

Dawid, Faigman, and Fienberg (2013) argue that statisticians can estimate the probability of causation, which is the probability that, after conditioning on the observed information, the individual was exposed and had a specific outcome, the individual would not have had the outcome had she not been exposed. The authors begin by citing an example of the distinction between effects of causes (EoC) and causes of effects (CoE). An EoC question is, for example, "Ann has a headache. She is wondering whether to take aspirin. Will that cause her headache to disappear?" A CoE question is, "Ann had a headache and took aspirin. Her headache went away. Was that caused by the aspirin?"

To argue that an individual had some damage done to him or her by a specific exposure to a treatment, Dawid et al. (2013) introduce the probability of causation as

$$PC = P(Y^0 = 0|Y^1 = 1), \tag{12.2}$$

where (Y^0, Y^1) be the potential outcomes that will eventuate if the exposure A equals zero or one. The potential outcome framework, developed by Neyman (1923) and Rubin (2005) is used to denote an outcome that exists before A is determined for each individual. This probability can be read in words as, "the probability that, given that the individual had the outcome when he/she was exposed, the individual would not have had the outcome if he/she had not been exposed." It contains counterfactuals since it is not possible, in general, to know the outcome under exposure and no exposure for an individual. Note that the definition in (12.2) is somewhat simplified compared to the definition given by Dawid et al. (2016), an aspect that is addressed in Dawid et al. (2014).

Writing PC in this way is useful because quantities from a randomized trial can easily be placed in the form of a contingency table. In the authors' example, the disappearance of Ann's headache is caused by taking the aspirin only if $Y^1 = 1$ and $Y^0 = 0$. In other words, Ann's headache disappears if she takes the aspirin but does not disappear if she does not take it. This requirement eliminates the possibility that another cause, say a different medicine, caused Ann's headache to go away. They note that Ann taking the aspirin ($A = 1$) and her headache going away ($Y^1 = 1$) are only causally connected if $Y^0 = 0$.

To estimate PC, the joint distribution of the potential outcomes (Y^0, Y^1) is required, which can never be observed for a single individual. Therefore, the closest one can get to PC is to place a bound on it. The bounds, which are derived under assumptions not mentioned in the article, are

$$1 \geq PC \geq 1 - 1/RR, \tag{12.3}$$

where RR is the risk ratio $P(Y^1 = 1)/P(Y^0 = 0)$. The lower bound is sometimes called the "excess risk ratio" or ERR. The additional assumptions, not mentioned in the article, are consistency and positivity.

The authors provide an example. Suppose a randomized trial was performed in which a sample of 100 individuals with headaches were divided into taking the control (chalk) or the treatment (aspirin). Some of the individuals recovered and some did not. Fictional results are presented in Table 12.1.

In an Effects of Causes analysis, we would need to use only the information from Table 12.1. But in a CoE analysis, we need to know the potential outcome values. So, the authors propose creating a new table shown in Table 12.2. To estimate the probability of causation, they first note that $PC = P(Y^0 = 0|Y^1 = 1) = x/30$. Then, they use the fact that the cells in a table cannot be smaller than zero, and

Table 12.1 Results from a fictional randomized experiment

	No recovery $Y = 0$	Recovery $Y = 1$	Total
Placebo (chalk), $A = 0$	88	12	100
Aspirin, $A = 1$	70	30	100

The study consists of 200 individuals with a headache, 100 of whom are given aspirin and 100 of whom are given chalk (a placebo). Some individuals recovered (i.e. did not have a headache anymore) and some did not

Table 12.2 Potential outcomes of the randomized experiment from Table 12.1

	$Y^0 = 0$	$Y^0 = 1$	Totals
$Y^1 = 0$	$88 - x$	$x - 18$	70
$Y^1 = 1$	x	$30 - x$	30
Totals	88	12	100

This table was generated considering the cell with x in it as the cell of interest, since it denotes $P(Y^0 = 0|Y^1 = 1)$. The remaining cells in the inside are filled in as differences between the marginal totals and x

therefore $x \geq 18$. By combining these two facts, they find that

$$1 \geq PC \geq 0.6. \tag{12.4}$$

In a civil trial, in which the preponderance of evidence is in question, this PC calculation shows that the evidence leans toward causation, since PC > 0.5. In other words, it is more likely than not that the aspirin did indeed cure a headache.

Pearl defines causes of effects as situations in which we observe both the effect, $Y = y$, and the putative cause $A = a$ and we are asked to assess, counterfactually, whether the former would have occurred absent the latter. Pearl defines the quantity of interest as the probability of necessity (although he uses different notation, it is replicated here with the chapter's notation for consistency),

$$PN = P(Y^0 = 0|Y = 1, A = 1). \tag{12.5}$$

The difference is that Pearl conditions on a positive outcome. Pearl notes that the quantity in Dawid et al. (2013) represents the probability that the drug was the cause of death of a subject who died in the experimental setup. However, very few court cases deal with deaths under experimental circumstances. Therefore, their formulation is a specific instance of the more general formulation he has derived. The more general form is needed, since the injury was not necessarily suffered as the result of a controlled experiment.

In Pearl's notation, Dawid, Faigman, and Fienberg's lower bound is the *experimental* excess risk ratio,

$$PC = \frac{P(Y^1) - P(Y^0)}{P(Y^1)}. \tag{12.6}$$

In Pearl's text, he writes do$(A = a)$ instead of $Y^{A=a}$. His "do" notation denotes that the value of the variable A has been set to a (in this case either zero or one) by the experimenter. Therefore, Pearl notes, this denotes experimental data, in which individuals are assigned to take aspirin or not, and not observational data, where individuals happen to take aspirin because of a headache. Note that this does not assume exchangeability, which would hold in a randomized trial.

To identify the probability of necessity, Pearl bounds PN by

$$\max\left\{0, \frac{P(Y=1) - P(Y^0=1)}{P(Y, A)}\right\} \leq PN \leq \min\left\{1, \frac{P(Y^0=0) - P(Y=0, A=0)}{P(Y=1, A=1)}\right\}, \tag{12.7}$$

which follows a Frechet inequality. Pearl then makes the assumption of monotonicity to avoid having bounds and obtain one quantity. If Y is monotonic relative to X, that is, $Y^1(u) \geq Y^0(u)$ for every unit, or individual, u, then PN is identifiable whenever $P(Y^1 = 1)$ is identifiable. Moreover,

$$PN = \frac{P(Y=1) - P(Y^0=1)}{P(Y, A)}, \tag{12.8}$$

which is the lower bound in (12.7). It is useful to note that in Dawid et al. (2013), the lower bound would also be the point estimate under monotonicity. This point estimate can be stated as a sum of two ratios,

$$\frac{P(Y=1|A=1) - P(Y=1|A=0)}{P(Y=1|A=1)} + \frac{P(Y=1|A=0) - P(Y^0=1)}{P(Y, A)}, \tag{12.9}$$

where the first term is the observational excess risk ratio and the second term is the correction for confounding bias. Under the assumption of exogeneity (that is, no unobserved confounders) and monotonicity, Pearl finds that the probability of necessity can be formulated as

$$PN = \frac{P(Y=1|A=1) - P(Y=1|A=0)}{P(Y=1|A=1)}, \tag{12.10}$$

which is the observational excess risk ratio from the first term in (12.9).

Table 12.3 Experimental and nonexperimental (i.e. observational) data used to illustrate the estimation of PN, the probability that drug A was responsible for a person's death $Y = 1$

	Experimental		Nonexperimental	
	$do(A = 1)$	$do(A = 0)$	$A = 1$	$A = 0$
Deaths ($Y = 1$)	16	14	2	28
Survivals ($Y = 0$)	984	986	998	972

Note that in an experimental setting, this quantity equals the one provided in Dawid et al. (2013) because the only change is that $P(Y^a)$ becomes $P(Y = 1|A = 1)$, or in other words, $do(A = a)$ becomes $A = a$,

$$\frac{P(Y^1) - P(Y^0)}{P(Y^1)} = \frac{P(Y = 1|A = 1) - P(Y = 1|A = 0)}{P(Y = 1|A = 1)}. \tag{12.11}$$

Thus, under Pearl's monotonicity and exogeneity,

$$PC = PN, \tag{12.12}$$

so Dawid, Faigman, and Fienberg's probability of causation (PC) is just a special case of Pearl's probability of necessity (PN).

Pearl demonstrates this by example. Suppose you have experimental and nonexperimental (i.e. observational) data as shown in Table 12.3. Assume monotonicity (recalling that without monotonicity this result is the lower bound). Then,

$$\begin{aligned} PN &= \frac{P(Y = 1|A = 1) - P(Y = 1|A = 0)}{P(Y = 1|A = 0)} + \frac{P(Y = 1|A = 0) - P(Y^0 = 1)}{P(A = 1, Y = 1)} \\ &= \frac{0.002 - 0.028}{0.002} + \frac{0.028 - 0.014}{0.001} = -13 + 14 = 1. \end{aligned}$$

Thus, under monotonicity, the observational ERR is -13, but the correction for confounding makes $PN = 1$. Thus, the drug A was indeed responsible for the person's death with probability one. Pearl then shows that estimating PC from Dawid et al. (2013),

$$PC = \frac{P(Y^1 = 1) - P(Y^0 = 1)}{P(Y^1 = 1)} = \frac{0.016 - 0.014}{0.016} = 0.125, \tag{12.13}$$

had they used their excess risk ratio, for which they assumed exogeneity in their formulation. Thus, using the experimental ERR, we get $PC = 0.125$, which implies that drug A was not responsible for the person's death in terms of the preponderance of evidence. Pearl summarizes his results in a table from Tian and Pearl (2000), replicated here in Table 12.4.

Table 12.4 PN as a function of assumptions and available data

Assumptions			Data available		
Exogeneity	Monotonicity	Additional	Experimental	Observational	Combined
+	+		ERR	ERR	ERR
+	–		Bounds	Bounds	Bounds
–	+	Covariates	–	Corrected ERR	Corrected ERR
–	+		–	–	Corrected ERR
–	–		–	–	Bounds

ERR stands for excess risk ratio, and corrected ERR refers to (12.9)

Finally, Pearl states that there is no way to directly estimate a probability for an individual from population-level data. The implications for the legal setting are that

> By using the wording "more probable than not," lawmakers have instructed us to ignore specific features that are either irrelevant or for which data are not likely to be available, and to base our determination on the most specific yet essential features for which data are expected to be available... knowing in advance that we will never be able to match all the idiosyncratic properties of Mr. A, the lawmakers' intent must be interpreted relative to the probability bounds provided by PN.

Thus, Pearl believes that (1) researchers must ignore the observed data about the individual for which there is no available data in the population and (2) the lawmakers' decision must be based on the estimated probability of necessity.

But what are the assumptions used by Dawid et al. (2013) to arrive at their bounds in (12.3)? And are there other assumptions made by Pearl that allow him to derive different quantities from his more general formulation? Additional questions regarding assumptions are answered in two more extensive articles (Dawid et al. 2016, Tian and Pearl 2000) where the authors describe their arguments, which we summarize after we discuss the final response by Dawid et al. (2014).

Dawid et al. (2014) respond to Pearl first by noting that their assumptions are different from Pearl's, and therefore, his analysis differs even though they arrive at the same bounds after the exogeneity assumption. The authors view probability as personalist and subjective, and therefore not only is the probability different (even if it looks the same) but the assumptions for identifiability are different as well. They entertain the idea of making their analysis be more similar to Pearl's by using a new variable D, the event that the individual desires to be treated.

The authors do not say this explicitly in their response, but they refer to their previous article (Dawid et al. 2013), in which they defined the probability of causation before making assumptions as

$$PC_i = P_i(Y^0 = 0 | Y^1 = 1, A = 1, X), \tag{12.14}$$

where X are observed covariates. It is interesting to note that it does not look so different from Pearl's probability of necessity

$$PN = P(Y^0 = 0 | Y = 1, A = 1, X). \tag{12.15}$$

The only differences are the conditioning on Y^1 vs. on Y and the subscript i. The former is not so relevant since by assuming consistency, which is not a strong assumption in this case, we get $A = a \implies Y^a = Y$. But the latter is relevant because it denotes that Dawid, Faigman, and Fienberg are using a subjective Bayesian probability. This makes not only their analyses different but also their philosophies.

They also respond by saying that even though his contributions to the problem are useful, Pearl has made a mistake. He oversimplifies the legal problem. His "use of the legal burden of proof in civil cases is misleading and simplistic." Dawid, Faigman, and Fienberg object to Pearl "aligning the CoE statistical issue with the trier of fact's ultimate determination regarding causation." The authors believe that Pearl is missing a crucial aspect of the analysis, which is how to communicate it to a court, and how a court should use it.

A complication arises if we are not certain that Ann took the aspirin. So far, we have assumed we know that $A = 1$ and $Y = 1$ and do not know whether there is a causal link between the two. What about the situations in which we observe the response but are not sure whether the individual was exposed at all? In these cases, we need to multiply the probability of causation by the probability of exposure, conditioned on the (known) fact that there was a positive response. Thus, we write the modified probability of causation for Ann as

$$\mathrm{PC}^*_A = \mathrm{PC}_A \cdot \Pr_A(A = 1 | Y = 1, X) \tag{12.16}$$

and the bounds as

$$\Pr(A = 1 | Y = 1, X) \geq \mathrm{PC}^* \geq \max\left\{0, 1 - \frac{\Pr(A = 0 | Y = 1, X)}{\Pr(A = 0 | X)}\right\}. \tag{12.17}$$

These new bounds are useful when accounting for uncertain exposure, which could be a crucial issue in a legal trial.

Discussion About the Debate

The debate between Pearl and Dawid, Faigman, and Fienberg addresses some of the main disagreements between the authors on using causes of effects to answer legal questions. The quantities, assumptions, and data types used by each side are displayed in Table 12.5.

It seems that Pearl is more interested in providing a general mathematical formulation and in making weaker identifiability assumptions. Dawid, Faigman, and Fienberg are more interested in what the law asks, which is a question about an individual, and how to address this group-to-individual question in a real-life setting.

Table 12.5 Comparison of the causes of effects formulations by Pearl and Dawid, Faigman, and Fienberg (DFF), based on their debate in the journal Sociological Methods and Research

Issue	Dawid, Faigman, Fienberg	Pearl
Quantity of interest	$PC_i = P_i(Y^0 = 0 \mid Y^1 = 1, A = 1, X)$	$PN = P(Y^0 = 0 \mid Y = 1, A = 1, X)$
Philosophy	Personalist Bayesian	Frequentist
Identifiability	No unobs. confounders ($\sim$)	–
Assumptions	Exchangeability of Y^a	
	Exogeneity	
	$\rightarrow$ Experimental ERR	$\rightarrow$ Observational ERR + Correction
Data required	Experimental	Experimental + Observational
Further assumptions	–	No unobs. confounders
	$\rightarrow$ Experimental ERR	$\rightarrow$ Experimental ERR
Group-to-individual	Posterior expectation	Ignore specific features
	PC specific to individual	PN conditions on X
Uncertain exposure	$PC_i \times P_i(Y = 1 \mid A = 1)$	–
Other quantities	–	PS, PNS

Although both DFF and Pearl start with different probabilities and assumptions, they arrive at the same observable quantity: the excess risk ratio (ERR). (Note: $\rightarrow$ denotes the lower bound (or the single quantity under monotonicity)

It is curious that Dawid, Faigman, and Fienberg start by defining an individual probability, which is what the law requires, but to identify it, they make assumptions that lead to a group-based probability. They do not explicitly say how one gets from the group probability back to an individual probability, but in Dawid et al. (2016), the authors display an example of an application. In this application, they use a Bayesian method to simulate the experimental RR from (12.3), and then they take the mean of the posterior. This indeed yields an individual value, but does it mean that it is the individual's probability of causation? It is unclear how this resolves the problem of group-to-individual causality (sometimes referred to as G2i).

The authors address the sensitivity to the choice of prior by saying that none of their priors capture zero or one, which are the most important values. But it is still not clear which values between zero and one are relevant. Perhaps the preponderance of evidence is useful, for which a value above 0.5 means there is a causal link and below 0.5 means there is not one. But it is unclear how to define quantities for "beyond a reasonable doubt." Should it be 0.5 or 0.8 or 0.99? There is no consensus about what the answer should be here. This is a common problem in applying statistics in the law, which has been addressed by numerous researchers (see, for example, Green et al. (2011)).

The debate sheds light on how best to answer the four questions posed earlier (and repeated here): (1) How should we *define* the correct probability of interest in Causes of Effects questions? (2) How should we *identify* this probability? (i.e. What assumptions are we willing to make?) (3) How should we *estimate* this probability?

And (4) How should we *use* the results from a Causes of Effects analysis? In this section, we describe the authors' arguments from each article, and we summarize the debate to shed light on using statistics to speak about causal question in the law.

Dawid, Faigman, and Fienberg's definition of the probability of causation directly from the requirements by law is useful. Pearl's derivations under different assumptions and types of data are also useful. In addition, there are extensions that each side developed, but the other did not. Pearl included the definition of the probability of sufficiency (PS) and the probability of necessity and sufficiency (PNS). This could be used to incorporate stricter requirements in making statistical claims in court. Dawid, Faigman, and Fienberg provide an adjustment for uncertain exposure (e.g. for cases in which we are not sure whether the individual took the aspirin). Perhaps it is possible to incorporate the insight from both sides to develop a methodology that is stronger than any single one.

Below, we replicate Pearl's and Dawid, Faigman, and Fienberg's complete arguments, which they did not include in their entirety in Dawid et al. (2013) but which can be useful to understand the frameworks.

12.4 Further Questions

The topics of mediators and other possible causes are not addressed in the debate in the articles; nevertheless, they deserve some attention.

Group-to-Individual Problem

The question of how to use the probability from a Causes of Effects analysis requires understanding how to interpret the probability for an individual given that you have population-level data. Dawid et al. (2014) call this the problem of group-to-individual (or G2i) probability. The authors state that Pearl (2015) does not answer the question about the individual when "he makes clear that his statistical analysis does not, indeed cannot, take into account all the... variables... But it is exactly these variables that jurors will be considering alongside any statistical proof of CoE." Thus, the authors state that the trier of fact will learn individual information, and somehow the trier must interpret the estimated probability for the individual, because the courts do expect to hear about "all the anatomical and psychological variables that determine an individual's (or bus's) behavior." The authors add that this simplification to the point of caricature is reminiscent of the famous Blue Bus Company hypothetical, a problem that has been addressed by Fienberg (1986), Dawid (2015), and Pearl (2009), among many others, as well as unpublished work by K. Steele and M. Colyvan at the London School of Economics.

The Blue Bus Company hypothetical asks whether a plaintiff can recover for damages in an accident with a bus when the Blue Bus Company operates 80% of the

buses on the street where the accident occurred, but no other evidence is proffered by either side. The authors state that legal scholars generally agree that the proportion of blue buses is relevant evidence (and likely admissible), but it is not sufficient to sustain a verdict for the plaintiff (in a civil case, or prosecution in a criminal case).

A similar hypothetical not mentioned by Dawid, Faigman, and Fienberg is the problem of the 100 prisoners. Suppose there are 100 prisoners in a courtyard prison, and 99 of them attack a prison guard and murder him. Then, the prisoners are tried individually. For any random individual, this individual has a 99% chance of having attacked and murdered the guard, according to the evidence. Therefore, they will all be considered guilty, even though there was one prisoner who is innocent and not involved in the attack. According to Dawid et al., legal scholars would say the 99% probability is not sufficient to sustain a verdict for the plaintiff.

The question of group-to-individual probability is one that has not been settled in statistics. The literature in precision medicine (Krakow et al. 2017) has addressed some of these issues about how to interpret a probability that has been estimated for a population for an individual. Some of these issues include what additional assumptions must be made, and on which populations? Although we do not discuss this issue in depth in this chapter, we mention it here because it is relevant to interpreting probability for legal trials.

Lagakos and Mosteller (1986) have an interesting view of the group-to-individual, or G2i, problem. They argue that instead of calling it the probability of causation or necessity, the quantity discussed by Pearl, which is conditioned on a group of individuals with observed characteristics X, it should be called the "assigned share." The assigned share or AS addresses the issue that researchers have limited data, and thus they must make simplifying assumptions about how the treatment affects the individual. In addition, it is possible that the individual in question was not in an experiment, but data from an experiment is used. The external validity of this experiment must then be evaluated.

Lagakos and Mosteller (1986) argue that it might be possible to gather additional information about the individual and further condition the probability based on the information. It might be possible to further condition the probability until it reaches a stable value (in the sense that it stops varying with further conditioning), and conditioning it on further information does not alter its value. However, it is possible that the order in which new variables are conditioned on can change whether the probability increases or decreases. So, this idea should only be carried out when there is monotonicity in the sense of checking whether the probability changes values always in the same direction (increasing or decreasing) with specified additional information. However, the authors also warn about possibly not wanting to condition on certain types of information. They state,

> In discussing these matters, we tend to get caught up in the science side and forget that the purpose of the enterprise is to make decisions that are seen to be equitable, if only because they are evenhanded and uniform. One can visualize an individual with a certain type of cancer, whose AS is lower than what it would be if he had another form of cancer, arguing that the distinction between the types of cancer should not be made. One can even imagine society and the courts agreeing with such a position in spite of the legislation and in spite

of the use made of this distinction in constructing the tables. An important point in policy work with populations is that society may choose not to take account of all or even any of the distinctions that matter.

Lagakos and Mosteller (1986) also argue that it might be possible to use AS tables within the tort (or personal injury) legal system to calculate the compensation that should be given to the individual. The authors derive several compensation rules, which include information about the effects of the outcome on quality or quantity of life after diagnosis. For a cancer patient, the authors speculate that "some amount of compensation could be based on the type of cancer, age at diagnosis, prognosis for life-shortening, prognosis for restricted activity, etc., and then this compensation could be adjusted in proportion to the AS values." This is perhaps the most concrete policy suggestion in the literature about causes of effects.

Mediators

Suppose the treatment A is taking a pill and the outcome Y is dying. Suppose a man has taken the pill and died. What happens if the reason he died is not because the pill poisoned him and made his body fail, but instead when he took the pill, he decided to drive a car, and the pill caused him to become debilitated and thus get into a fatal car crash? In that case, the probability of causation would be positive: taking the pill is the cause of his death. But in reality, the car crash caused his death. What do we do in this case? The law sometimes requires the "but for" requirement: it requires that experts demonstrate that the person would not have died but for taking this pill. In this case, however, the man would not have died had he not taken the pill. Without the pill he does not die, but with the pill he does die. This scenario still satisfies the but for criterion. But we cannot assume that the man's decision to drive his car after taking the pill is the same as it would be for a sample of individuals from the population because the exchangeability assumption would fail. If everyone in the sample also decided to drive after taking the pill, then the man would be exchangeable and the analysis could proceed. If not, then one of the main assumptions would be violated. Perhaps a study performed by statisticians and legal scholars about proximate causes could help inform this discussion. Finally, Dawid et al. (2019) have recently written about how to include mediators in identification and in deriving bounds for the causes of effects.

Multiple Causes

What happens when the question of interest is not just "Was Y caused by A?", but instead, "Was Y caused by A or B or C?" The probability of causation only answers the first question. To adapt it to answer the second question, the probability of causation could be applied to each event A, B, and C separately. If events A,

B, and C are mutually exclusive, and they are the only possible causes, we would expect their three probabilities of causation to sum to one. However, if they are not mutually exclusive, then it is possible that their probabilities of causation add up to more than one. What does it mean when to say that the probability that it was A or B or C that caused Y is, for example, two? Since it is still a probability of causation, should it not be restricted to be between zero and one? This topic has not been addressed in the literature, but it is of great importance when comparing different causes. It might also be possible to express the cause being A, B, or C, in a different way using potential outcomes, but that is outside the scope of this chapter. For now, the question of comparing probabilities of causation of different possible exposures is unexplored.

Overdetermination is the case in which the outcome was caused by several different causes, say by A and by B and by C (e.g. a man died because he ingested poison, was shot, and drowned). In other words, each cause was sufficient, but not necessary. In that case, if we write

$$P(Y^{A=0} = 0|Y = 1, A = 1, B = 1, C = 1, X = x), \tag{12.18}$$

then $PC = 0$ because the outcome was also caused by both B and C. In that case, the probability of causation fails to capture the fact that A caused Y. We could repeat this by evaluating the case for B, $PC = P(Y^{B=0} = 0|Y = 1, A = 1, B = 1, C = 1, X = x) = 0$, and the case for C, $PC = P(Y^{C=0} = 0|Y = 1, A = 1, B = 1, C = 1, X = x) = 0$. So, we see that A was not the cause, B was not the cause, and C was not the cause. But in fact, each one was a cause. It just was not *the* (only) cause. Overdetermination breaks this formulation of the probability of causation. However,

$$P(Y^{A=0} = 0, Y^{B=0} = 0, Y^{C=0} = 0|Y = 1, A = 1, B = 1, C = 1, X = x) = 0, \tag{12.19}$$

and

$$P(Y^{A=1} = 0, Y^{B=0} = 0, Y^{C=0} = 0|Y = 1, A = 1, B = 1, C = 1, X = x) = 1. \tag{12.20}$$

Etiology (as in the case of etiologic fractions) is the case in which the cases A, B, and C together cause Y. In other words, each A, B, and C is a partial cause of Y. Writing the usual probability of causation, we get that A was not the cause of Y:

$$P(Y^{A=0} = 0|Y = 1, A = 1, X = x) = 0. \tag{12.21}$$

We also get that B and C are not the cause of Y. The but for requirement applies because Y would not have happened but for A happening (this is true for B and C as well). But if we combine them, we can write

$$P(Y^{A=0,B=0,C=0} = 0|Y = 1, A = 1, B = 1, C = 1X = x) = 1. \tag{12.22}$$

Also, if you set any of A, B, and C to zero, then PC = 1:

$$P(Y^{A=0,B=1,C=1} = 0|Y = 1, A = 1, B = 1, C = 1X = x) = 1. \quad (12.23)$$

In other words, blocking any of them makes the counterfactual false, and therefore PC = 1.

In conclusion, one must be wary when estimating the probability of causation in the case of overdetermination (A or B or C) and etiologic fractions (A and B and C). Pearl has some discussion of this in Tian and Pearl (2000), where he talks about the probability of necessity and sufficiency. It could be interesting to apply Pearl's definitions to real-world cases to see if any new arguments could be made to support either the prosecution or the defense.

Bayesian versus Frequentist Interpretation

If one's goal is to quantify one's subjective beliefs about the probability of causation for a specific case, then one should use Bayesian statistics. In this case, an expert witness can state a likelihood from population-level data and a prior from information specific to the individual. Therefore, the exchangeability assumption is necessary only for the likelihood, and the prior can be specific to the individual at hand. Subsequently, the expert provides his results (as a posterior distribution), and the trier of fact can incorporate that into his or her (or their) prior information (hopefully the judge or jury is unbiased and will have a flat prior before learning about the case, but the prior could include other evidence presented in the case) to make the final conclusion about guilt. More precisely, an expert witness is allowed to give a personal opinion in court. It is subjective, and there is nothing problematic about that. The question is whether, and to what extent, does the trier of facts (judge or jury) change their subjective opinions given the expert's testimony. That is a matter of how persuasive the expert is.

If one's goal is to provide a methodology that whenever it is repeated in the long run it provides 95% confidence intervals for the probability of causation, the one should use frequentist statistics. In this case, the probability of causation should be estimated by using population-level data, for instance, from a randomized trial. PC is a function of x, so there will be an estimate for each observation, and in fact for any new x by using predictions from a model. A way to derive a probability specific to an individual is by conditioning on observed covariates of interest and by assuming that the individual of interest is exchangeable with the people from the sample with his or her same covariates. The more covariates one conditions on, the narrower the group of individuals that fit into that group, and the easier it is to satisfy the exchangeability assumption, and the harder it is to obtain good estimates.

12.5 A Comprehensive Treatment of the Probability of Causation

What can we conclude from the debate? Or in other words, What definition and identification should a researcher use if they are interested in estimating the probability of causation? The formulation in Cuellar and Kennedy (2018) was informed by the debate, and it is replicated here. Let the probability of causation be defined as

$$PC = P(Y^0 = 0|Y = 1, A = 1, X = x), \tag{12.24}$$

where Y is the binary outcome, A is the binary exposure, $Y^{A=a}$ is the binary potential outcome, and $X = x$ is the vector of covariates that can be observed about the individual or group in question. The probability of causation is the causal parameter of interest.

The probability of causation includes the counterfactual relationship between the potential outcomes under exposure and under no exposure, and these two will never be observed for an individual or even for a group. This is the fundamental problem of causal inference, as noted by Holland (1986). To arrive at a statistical quantity that can be estimated from data, we must make some identification assumptions. We start with a causal parameter, make identification assumptions to arrive at a statistical parameter, and then make estimating assumptions to arrive at an estimator. Here, a causal parameter means a quantity of interest that contains potential outcomes and thus can rarely be estimated directly. A statistical parameter is a quantity that does not have potential outcomes and thus could be estimated from data. An estimator is a rule for calculating an estimate of a given statistical parameter based on observed data; thus, the rule (the estimator), the quantity of interest (the estimand, here the statistical parameter), and its result (the estimate) are distinguished.

Assume outcome Y is binary. Start by using the Bonferroni–Fréchet–Hoeffding bounds:

$$\min\left\{1, \frac{P(Y^0=0|X,A=1)}{P(Y^1=1|X,A=1)}\right\} \geq \text{PC} \geq \max\left\{0, 1 - \frac{P(Y^0=1|X,A=1)}{P(Y^1=1|X,A=1)}\right\}. \tag{12.25}$$

Assume no unobserved confounders, otherwise called exchangeability, i.e. $(Y^0, Y^1) \perp\!\!\!\perp A|X$. A is a random variable. Then, conditioning on X and $A = 1$ or on X and $A = 0$ are interchangeable, so we can write

$$\min\left\{1, \frac{P(Y^0=0|X,A=0)}{P(Y^1=1|X,A=1)}\right\} \geq \text{PC} \geq \max\left\{0, 1 - \frac{P(Y^0=1|X,A=0)}{P(Y^1=1|X,A=1)}\right\}. \tag{12.26}$$

Assume consistency, i.e. $A = a \implies Y^a = Y$. In this step, we shift from potential (sometimes unobservable) outcomes to observable outcomes. Then,

$$\min\left\{1, \frac{P(Y=0|X,A=0)}{P(Y=1|X,A=1)}\right\} \geq \text{PC} \geq \max\left\{0, 1 - \frac{P(Y=1|X,A=0)}{P(Y=1|X,A=1)}\right\}. \tag{12.27}$$

Assume monotonicity, i.e. $P(Y^1 \geq Y^0) = 1$. Monotonicity implies that the treatment (e.g. aspirin) can only make a person better (e.g. cure the person's headache), but it cannot make the person worse. In summary, under these causal assumptions:

1. Binary outcome and treatment: $Y \in \{0, 1\}$, $A \in \{0, 1\}$,
2. Consistency: $A = a \implies Y^a = Y$,
3. No unobserved confounders: $Y^a \perp\!\!\!\perp A|X$, and
4. Monotonicity: $Y^1 \geq Y^0$,

the probability of causation is identified as

$$PC \underbrace{=}_{\text{l.t.p.}} \frac{P(Y^0=0, Y^1=1|A=1,X)}{P(Y^1=1|A=1,X)} \underbrace{=}_{1,4} \frac{P(Y^1-Y^0=1|A=1,X)}{P(Y^1=1|A=1,X)} \underbrace{=}_{2,3} 1-\frac{P(Y=1|A=0,X)}{P(Y=1|A=1,X)} \underbrace{=}_{1} 1-\frac{\mathbb{E}(Y|A=0,X)}{\mathbb{E}(Y|A=1,X)},$$

where l.t.p. stands for the law of total probability, and therefore,

$$PC = 1 - ERR. \tag{12.28}$$

Finally, PC under monotonicity is exactly the lower bound in (12.27). A researcher can use this strategy, after verifying that the causal assumptions hold in the specific substantive area and dataset of interest, to estimate the probability of causation.

12.6 Discussion

Attorneys and courts often need to understand whether a specific exposure, or something else, caused a defendant's outcome. In other words, they need to determine whether an outcome can be attributed to a specific cause. To answer questions of causality in which the exposure and outcome have already been observed, researchers should use the framework of causes of effects, not effects of causes. They can do this by using the probability of causation (PC). In the statistics literature, the questions of how to define this probability properly and what identification assumptions to make in a specific application are still debated.

Dawid, Faigman, and Fienberg have one way of answering these questions. They argue that the probability of causation should be defined for a specific individual, as denoted by subscript i in PC_i. That it should condition on Y^1 instead of Y, since in an experimental trial the treatment A was set to one. For Y^1 to equal Y, the authors must assume consistency, which Pearl does not necessarily assume. And that, it should be interpreted as a subjective Bayesian quantity that is stated by a researcher with specific prior knowledge of the situation. They argue that to identify the probability of causation, the researcher must assume that there are no unobserved confounders, and thus essentially only experimental data can be used to answer questions about the probability of causation. "Identify" here means to derive a statistical quantity from a causal quantity by making identification assumptions.

The no-unobserved-confounders assumption is almost never true in observational data, but it is true in experimental data. Finally, they argue that the individual must be exchangeable with the individuals in the sample, in order to make inferences about the individual in question from data (Dawid et al. 2013; 2016).

Pearl has a different way of answering these questions. He argues that the probability of causation (which he has claimed to have already defined as his probability of necessity) should condition on Y, not Y^1 since the researcher must condition on what was observed, which was the real outcome, not the potential outcome. Under consistency, the definition by Dawid, Faigman, and Fienberg is almost the same as Pearl's, but again, Pearl does not necessarily assume consistency. Pearl argues that the probability should be interpreted as a frequentist quantity that could be estimated by anyone. Pearl does not interpret the probability of causation for an individual in the sense that he accepts that, since the data used to estimate PC comes from a population, the closest one can get to finding PC for an individual is by conditioning further by using the covariates X. To identify the quantity, Pearl says there are no necessary assumptions. If one assumes no unobserved confounders, then one can get the experimental excess risk ratio, and with monotonicity, the quantity is a point estimate (not a bound). In general, Pearl makes fewer assumptions.

If one is interested in quantifying subjective beliefs, one should use Dawid, Faigman, and Fienberg's subjective Bayesian approach. If one is interested in frequency guarantees, one should use Pearl's approach. Each has to make assumptions to draw inferences from data about an individual. The assumptions made will depend on the application in question and particularly on the data available. For example, in a vaccine trial consistency might not hold, and in an observational setting, there might be unobserved confounders. Thus, it is up to the researcher to justify each assumption carefully. Work remains to be done to incorporate some additional issues into the formulation of the probability of causation that might be of interest for legal or epidemiological questions. Furthermore, connecting the views from epidemiology, the law, and statistical causal inference on this topic is a serious gap in the literature that if filled could better inform legal and medical decisions.

Further questions about the probability of causation that were not addressed in the debate are important to understanding how to answer questions of attribution and causes of effects. First is the group-to-individual problem, which deals with how to interpret the probability of causation for an individual if what the researcher has access to is data about a group. Second is a discussion about mediators, whether the probability of causation should be defined the same way if there is a sequence of possible causes instead of a single cause. Third is a discussion about multiple causes, whether the definition of PC should change if the question of interest is about whether an outcome was caused by one cause or another, or another. Finally, they include a discussion about the differences between a Bayesian treatment and a frequency treatment of the analysis.

When answering questions about causality in the law, from the point of view of statistics, it is important to clarify what the question of interest is (e.g. is it about forecasting, backcasting, or attribution?), how the causal parameter is defined,

and what causal assumptions are made to estimate the quantity of interest. As statisticians, we are used to addressing questions of forecasting, and sometimes backcasting (effects of causes), but we are less familiar with questions of attribution (causes of effects). It is important to answer the questions we are actually interested in, and not just the ones we have the tools to answer. Therefore, the causes of effects deserve some attention such that questions in court and in investigations can be answered appropriately.

References

Cuellar, M. and Kennedy, E. H. (2018), "A nonparametric projection-based estimator for the probability of causation, with application to water sanitation in Kenya," *arXiv:1810.00767 [stat.AP]*.

Dawid, A. P. (2015), "On Individual Risk," *Synthese*, 1–30.

Dawid, A. P., Faigman, D. L., and Fienberg, S. E. (2013), "Fitting Science Into Legal Contexts: Assessing Effects of Causes or Causes of Effects? (with Discussion)," *Sociological Methods and Research*, 43, 359–390.

Dawid, A. P., Faigman, D. L., and Fienberg, S. E. (2014), "On the Causes of Effects: Response to Pearl," *Sociological Methods and Research*, 44, 165–174.

Dawid, A. P., Musio, M., and Fienberg, S. E. (2016), "From Statistical Evidence to Evidence of Causality," *Bayesian Analysis*, 11, 725–752.

Dawid, P., Humphreys, M., and Musio, M. (2019), "Bounding Causes of Effects with Mediators," *arXiv:1907.00399 [math.ST]*.

Fienberg, S. E. (1986), "Gatecrashers, blue buses, and the Bayesian representation of legal evidence," *Boston University Law Review*, 66, 693.

Green, M. D., Freedman, M. D., and Gordis, L. (2011), "Reference Guide on Epidemiology," in *Reference Manual on Scientific Evidence: Third Edition*, Washington, D.C.: The National Academies Press, pp. 549–632.

Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.

Krakow, E., Hemmer, M., Wang, T., Logan, B., Arora, M., S, S. S., Couriel, D., Alousi, A., M, P. J. L., Lachance, S., and Moodie, E. (2017), "Tools for the Precision Medicine Era: How to Develop Highly Personalized Treatment Recommendations From Cohort and Registry Data Using Q-Learning," *American Journal of Epidemiology*, 186, 160–172.

Lagakos, S. W. and Mosteller, F. (1986), "Assigned Shares in Compensation for Radiation Related Cancers," *Risk Analysis*, 6, 345–357.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science*, 5, 463–472.

Pearl, J. (2009), *Causality: Models, Reasoning and Inference*, Cambridge University Press.

Pearl, J. (2015), "Causes of Effects and Effects of Causes," *Sociological Methods & Research*, 44, 149–164. MR3313822. https:doi.org/10.1177/0049124114562614

Robins, J. and Greenland, S. (1989), "The Probability of Causation under a Stochastic Model for Individual Risk," *Biometrics*, 45, 1125–1138.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Rubin, D. B. (2005), "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *Journal of the American Statistical Association*, 100, 322.

Tian, J. and Pearl, J. (2000), "Probabilities of causation: Bounds and identification," *Annals of Mathematics and Artificial Intelligence*, 28, 287–313.

Chapter 13
What Can Group-Level Data Tell Us About Individual Causality?

A. Philip Dawid and Monica Musio

13.1 Introduction

"Causal inference" subsumes two related but distinct enterprises: inference about the expected Effects of applied Causes (EoC) and inference about the likely Causes of observed Effects (CoE). The former is the bread and butter of most scientific research, while the latter is important in the law for assigning blame or responsibility. While EoC questions can be phrased at the level of individuals or of groups, CoE questions are fundamentally individualistic. In this chapter, we shall be specifically interested in the possibilities for making CoE inferences, on the basis of data collected on groups.

As an example, in a legal "toxic tort" case, an individual might sue a pharmaceutical company, claiming that taking their drug was responsible for her developing a certain adverse outcome. Evidence is presented by expert epidemiologists of the incidence rate of this outcome, both among those exposed and among those not exposed, to the drug. This evidence is clearly directly relevant to EoC inference: would we expect the outcome, were an individual to take, or not take, the drug? But this is not the question before the court, which wishes to assess whether there was a causal link between taking the drug and the outcome in this particular case—a CoE question. The epidemiological evidence is at best indirectly relevant to this issue.

So how are we to use such group-level epidemiological data to assist in addressing the CoE task for the individual case at hand? This has been termed the G2i (Group to individual) problem (Faigman et al. 2014). In a number of

A. P. Dawid
University of Cambridge, Cambridge, United Kingdom
e-mail: apd@statslab.cam.ac.uk

M. Musio (✉)
Università degli Studi di Cagliari, Cagliari, Italy
e-mail: mmusio@unica.it

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_13

contributions (Dawid 2011, Dawid et al. 2014; 2021; 2016a;b; 2017, Murtas et al. 2017), we have addressed various aspects of this question. We have shown that group-level data can in general only supply, at best, interval bounds for the CoE target, the "probability of causation" (PC). These intervals can however be improved if we also have data on additional variables, such as covariates or mediators (with or without the associated information for the individual case).

Here, we summarise and build on our previous collaborative work in this area with Stephen Fienberg, Rossella Murtas, and Macartan Humphreys. We give more detailed analysis, including proofs for some results asserted in Dawid et al. (2017)—which, like the present work, was dedicated to the memory of Stephen Fienberg, a consummate statistician, a stimulating collaborator, and a dear friend. We also present some new results on the use of mediating variables.

Plan of Chapter

In Sect. 13.2, we describe and contrast the problems of EoC and CoE inference. Whereas the former can be formulated and solved using only the standard tools of probability, statistics, and decision analysis (Dawid 2000; 2015), the latter necessitates some form of "counterfactual reasoning": for example, in the above toxic tort example, one would want to consider whether the outcome might have been different if—contrary to actual fact—the plaintiff had not taken the drug.

In Sect. 13.3, we describe a formalisation of the CoE problem in terms of "potential responses", which supports direct formulation of a counterfactual query and the associated probability of causation, PC. But Sect. 13.4 shows that, even with perfect information about the dependence of the response on the exposure, typically one cannot identify PC exactly. Nevertheless, under certain assumptions, one can use this information to determine interval bounds for PC.

The remainder of this chapter shows how such bounds can be varied or improved by including further information on additional variables. Section 13.5 discusses the use of information in the data on a "sufficient covariate", which may affect both exposure and outcome but conditional on which there is no residual confounding. We derive the associated bounds on PC, both (a) when the covariate is observed for the case at hand and (b) when it is not. A special case considered is where the covariate does not affect exposure, in which case the bounds for situation (b) are at least as good as the simple bounds of Sect. 13.4, which also apply in this case. They are also at least as good as the bounds of Tian and Pearl (2000) applied to this case.

In Sect. 13.6, we suppose instead that the additional information in the data relates to a variable that acts as a complete mediator between exposure and response. When this is unobserved in the case at hand, the lower bound on PC is unchanged from the simple lower bound, but the upper bound is improved. We also give results for a chain of complete mediator variables. Section 13.7 derives bounds for cases combining both covariate and mediator variables. Finally, Sect. 13.8 summarises the work and suggests further extensions.

13.2 Effects of Causes and Causes of Effects

The overwhelming majority of the literature on statistical causality, as in most scientific concerns about causality, focuses on understanding the *effects of causes* (EoC). That is to say, we are interested in understanding what would happen to an individual if subjected to a certain intervention or exposure, $E \leftarrow e$. Here, E is the *cause variable* and e its contemplated value, and the expression $E \leftarrow e$ indicates the state of affairs where an external intervention is applied to make E take value e. There will typically be a specific response variable R of interest—the *effect variable*. Our query would then be essentially answered if we could assess the conditional probability distribution for R, consequent on the intervention $E \leftarrow e$. The major purpose of EoC analysis is to find ways of estimating these interventional probabilities, when the available data are in some way deficient—for example, because they are purely observational and we suspect confounding, where an unmeasured variable (such as a patient's general health) is a common cause of both exposure and response. Such estimation will require assumptions to be made, and justified, in order to transfer properties from the observed but deficient data to the hypothetical interventional situation of interest (Dawid 2021).

A somewhat different set of causal questions arises in, for example, legal proceedings to assign liability for an already observed response. This requires inference as to the *causes of effects* (CoE). Thus, suppose Ann took the drug benfluorex as an appetite suppressant ($E = 1$) and sometime later developed valvular heart disease ($R = 1$). She sues the manufacturer of the drug for misconduct, an essential component of her case being that it was because she took the drug that she developed the condition. Typically in such a toxic tort case, there will be epidemiological evidence ("group level" data) presented to argue that the exposure is indeed associated with the response, eg., because the response rate is higher in the exposed than in the unexposed. This is information about the effects of causes. But even full knowledge of the probabilities of the response, with and without exposure, is not obviously relevant to the individual case at hand: we face non-trivial problems of "group to individual" (G2i) inference. In particular, in the individual case, both the exposure and the response have been observed—so where does uncertainty about their causal relationship reside?

The CoE question might be reformulated as:

> "If Ann had not taken benfluorex, would she still have developed valvular heart disease?"

For if the answer to this is "Yes", and the same response would have occurred anyway, one could not say that it was only because of the actual exposure that the actual response occurred. On the contrary, a negative answer means that the actual exposure did make a difference to the outcome, so can be regarded as a causative factor.

However, the above question is *counterfactual*, in referring to a state of affairs (where Ann's exposure was different) that is contrary to the known facts. This raises delicate questions of how we could ever address such a question on the

basis of empirical observations. More specifically, suppose we have a complete epidemiological understanding of how the response depends (probabilistically) on an applied exposure. What does this tell us that might be helpful for addressing the above counterfactual CoE query about Ann?

13.3 Potential Responses

The above counterfactual query is not clear-cut if there are more than two possible values for the exposure, since the answer could depend on just which alternative value we consider might have been realised. To avoid this complication, we restrict attention to the case of a binary exposure variable E. The response variable R could be of any nature, but for simplicity we will also require R to be binary. The CoE question is then (where suffix A refers to Ann)

> "We have observed $E_A = 1$, $R_A = 1$. If (counterfactually) it had been the case that $E_A = 0$, would it have been the case that $R_A = 0$?"

The way in which such a question has generally been formalised statistically is in terms of *potential responses*. Instead of a single response variable R, we posit a *pair* of response variables, $\mathbf{R} = (R(0), R(1))$, where $R(e)$ $(e = 0, 1)$ is conceived of as the response that would occur under an actual or hypothetical intervention $E \leftarrow e$. Both these variables are supposed to exist even prior to the determination of the value of E, and indeed to be independent of whether E is set by intervention or arises "naturally". After the value of E has been determined, the actual response R will be $R(0)$ if in fact $E = 0$, or $R(1)$ if in fact $E = 1$. That is to say, $R = R(E)$. However, while we can observe the single component of the pair $\mathbf{R}$ corresponding to the realised value of E, the other component will remain forever unknown.

With this notational and philosophical elaboration, the CoE question becomes:

> "We have observed $E_A = 1$, $R_A(1) = 1$. Is it the case that $R_A(0) = 0$?"

If we had a full joint distribution for the three quantities $(E_A, R_A(0), R_A(1))$, we could compute the associated *probability of causation* (what Pearl (1999) terms the *probability of necessity, PN*):

$$\mathrm{PC}_A := \Pr\left(R_A(0) = 0 \mid E_A = 1, R_A(1) = 1\right). \tag{13.1}$$

The problem is that the information we have, even if obtained from individuals who can be regarded as similar to (exchangeable with) Ann, will typically not allow us to determine this joint distribution. In particular, since we can never observe both $R(0)$ and $R(1)$ simultaneously, it is problematic to estimate a probability for $R(0)$ conditioned on $R(1)$.

13.4 Simple Analysis

Suppose we have access to extensive data supplying values for the population probabilities $\Pr(R = r \mid E \leftarrow e)$ $(e, r = 0, 1)$. We can express these probabilities in terms of parameters τ and ρ, where

$$\tau := \Pr(R = 1 \mid E \leftarrow 1) - \Pr(R = 1 \mid E \leftarrow 0)$$
$$\rho := \Pr(R = 1 \mid E \leftarrow 1) - \Pr(R = 0 \mid E \leftarrow 0).$$

Then, τ is the *average causal effect* of E on R, while ρ is a measure of how common the outcome is.

The transition matrix from E to R is thus

$$P = P(\tau, \rho) := \begin{pmatrix} \frac{1}{2}(1 + \tau - \rho) & \frac{1}{2}(1 - \tau + \rho) \\ \frac{1}{2}(1 - \tau - \rho) & \frac{1}{2}(1 + \tau + \rho) \end{pmatrix}, \tag{13.2}$$

where the row and column labels are implicitly 0 and 1 in that order. The necessary and sufficient condition for all the transition probabilities to be non-negative is

$$|\tau| + |\rho| \leq 1. \tag{13.3}$$

We have equality in (13.3) only in the *degenerate* case that one of the entries of P is 0. Henceforth, we suppose this is not so.

The probabilities in P are clearly relevant to EoC analysis. But how can they assist CoE analysis?

In terms of potential responses, we can equate

$$\Pr(R(e) = r) = \Pr(R = r \mid E \leftarrow e).$$

The data thus supply the marginal distributions of each of $R(0)$ and $R(1)$ but can give no further insight into their joint distribution.

We can express the joint distribution for $\mathbf{R}$ as in Table 13.1, where the margins are determined by the data, but the internal entries are indeterminate, having one degree of freedom crystallised in the unspecified "slack variable" ξ. The only constraint on

Table 13.1 Joint probability distribution of $(R(0), R(1))$

	$R(1) = 0$	$R(1) = 1$	
$R(0) = 0$	$\frac{1}{2}(1 - \rho - \xi)$	$\frac{1}{2}(\xi + \tau)$	$\frac{1}{2}(1 + \tau - \rho)$
$R(0) = 1$	$\frac{1}{2}(\xi - \tau)$	$\frac{1}{2}(1 + \rho - \xi)$	$\frac{1}{2}(1 - \tau + \rho)$
	$\frac{1}{2}(1 - \tau - \rho)$	$\frac{1}{2}(1 + \tau + \rho)$	1

ξ is that all internal entries of Table 13.1 must be non-negative, which holds if and only if

$$|\tau| \leq \xi \leq 1 - |\rho|. \tag{13.4}$$

Since we are assuming that P is non-degenerate, this will be a non-trivial interval.

We now attempt to apply the group-level probabilities to the individual case of Ann. In order to begin to do this, we make two assumptions:

Exchangeability Ann is similar to the population from which the probabilities have been computed, so those probabilities apply to her. So we can drop the suffix A on probabilities.

No confounding The fact that Ann chose to take benfluorex is not informative about her response to it, either factually or counterfactually. Formally, we require independence between E and $\mathbf{R}$: $\mathbf{R} \perp\!\!\!\perp \mathrm{E}$.

The probability of causation (13.1) now becomes

$$\mathrm{PC} = \Pr\left(R(0) = 0 \mid R(1) = 1\right).$$

In terms of Table 13.1, this is

$$\mathrm{PC} = \frac{\xi + \tau}{1 + \tau + \rho}.$$

In view of (13.4), we have the following interval bounds for PC:

$$l := \max\left\{0, \frac{2\tau}{1 + \tau + \rho}\right\} \leq \mathrm{PC} \leq \min\left\{1, \frac{1 + \tau - \rho}{1 + \tau + \rho}\right\} =: u, \tag{13.5}$$

or equivalently

$$l = \max\left\{0, 1 - \frac{1}{\mathrm{RR}}\right\} \leq \mathrm{PC} \leq \min\left\{1, \frac{\Pr(R = 0 \mid E \leftarrow 0)}{\Pr(R = 1 \mid E \leftarrow 1)}\right\} = u, \tag{13.6}$$

where

$$\mathrm{RR} = \frac{\Pr(R = 1 \mid E \leftarrow 1)}{\Pr(R = 1 \mid E \leftarrow 0)} \tag{13.7}$$

is the *risk ratio*.

Without further information or assumptions, this interval for PC is the best information we can extract from the group-level data.

It may however be that we do have additional information, either for the population, or for Ann, or for both. When this is the case, we can generally use that information to modify the simple interval bounds of (13.5). In this chapter, we will focus on the use of additional information about covariates and/or mediators.

13.5 Using Covariates

Consider the diagram of Fig. 13.1, with exposure E, response R, and *sufficient covariate* S. That is, S is an additional variable that can be measured before exposure is determined and that can have an effect on both E and R. Again, we suppose variables E and R are binary, and (for simplicity) we suppose S is discrete, with values in $\mathcal{S}$ say, and $\Pr(S = s) > 0$ for all $s \in \mathcal{S}$. We suppose that, from the study data, we can identify $\Pr(S = s)$, $\Pr(E = e \mid S \leftarrow s)$, and $\Pr(R = r \mid S \leftarrow s, E \leftarrow e)$. For Ann, we have observed $E_A = 1$ and $R_A = 1$ but may or may not have observed S_A. We want to bound the probability that $E_A = 1$ caused $R_A = 1$.

The relevant potential responses are now:

- $E(s)$, the exposure when $S \leftarrow s$
- $R(s, e)$, the response when $S \leftarrow s$, $E \leftarrow e$

We define $\mathbf{E} := (E(s) : s \in \mathcal{S})$, $\mathbf{R} := (R(s, e) : s \in \mathcal{S}, e = 0 \text{ or } 1)$. The relationship between the potential and actual responses is $E = E(S)$, $R = R(S, E)$.

We again assume exchangeability and no unobserved confounding, the latter now being formalised as mutual independence between S, $\mathbf{E}$, and $\mathbf{R}$. In particular, we have

$$\Pr(E = e \mid S = s) = \Pr(E(s) = e) = \Pr(E = e \mid S \leftarrow s)$$

and similarly

$$\Pr(R = r \mid E = e, S = s) = \Pr(R = r \mid E \leftarrow e, S \leftarrow s).$$

Thus, we have joint distribution

$$\Pr(S = s, E = e, R = r) = \Pr(S = s) \Pr(E = e \mid S \leftarrow s) \Pr(R = r \mid E \leftarrow e, S \leftarrow s).$$

Bounds

Consider first the case that, for Ann, we have observed $E = 1$, $R = 1$, $S = s$. What now is the relevant probability of causation, $\mathrm{PC}(s)$ say?

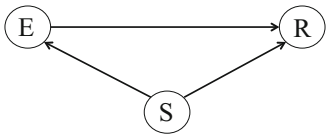


Fig. 13.1 Sufficient covariate. E is the exposure variable, R is the response variable, and S is the covariate, which may affect both E and R. We assume no residual confounding

We have

$$\begin{aligned} \mathrm{PC}(s) &= \Pr\left(R(S,0)=0 \mid S=s, E(s)=1, R(s,1)=1\right) \\ &= \Pr\left(R(s,0)=0 \mid R(s,1)=1\right) \end{aligned} \tag{13.8}$$

by the no-confounding assumption. We can also set

$$\Pr(R(s,e)=r) = \Pr(R=r \mid S \leftarrow s, E \leftarrow e).$$

An analysis parallel to that of Sect. 13.4 can now be performed, with the sole modification that all probabilities are further conditioned on $S=s$. So we obtain

$$l(s) := \max\left\{0, 1-\frac{1}{\mathrm{RR}(s)}\right\} \le \mathrm{PC} \le \min\left\{1, \frac{\Pr(R=0 \mid E \leftarrow 0, S \leftarrow s)}{\Pr(R=1 \mid E \leftarrow 1, S \leftarrow s)}\right\} =: u(s), \tag{13.9}$$

with

$$\mathrm{RR}(s) = \frac{\Pr(R=1 \mid E \leftarrow 1, S \leftarrow s)}{\Pr(R=1 \mid E \leftarrow 0, S \leftarrow s)}.$$

Alternatively, suppose that we cannot measure S for Ann, so we only know $E(S)=1$, $R(S,1)=1$. We have to consider what would have been the response if, counterfactually, Ann's exposure had been $E=0$. We further assume that this is the minimal change made between the factual and counterfactual world, so that, in particular, there is no change to the value or distribution of S.

The probability of causation is now

$$\begin{aligned} \mathrm{PC} &= \Pr\{R(S,0)=0 \mid E(S)=1, R(S,1)=1\} \\ &= \mathrm{E}\{\Pr(R(S,0)=0 \mid S, E(S)=1, R(S,1)=1) \mid E(S)=1, R(S,1)=1\} \\ &= \mathrm{E}\{\mathrm{PC}(S) \mid E(S)=1, R(S,1)=1\} \\ &= \sum_s \mathrm{PC}(s) \times \Pr(S=s \mid E=1, R=1). \end{aligned} \tag{13.10}$$

Now there are no logical relationships between the distributions of $(R(s,0), R(s,1))$ for different values of S. So by independently varying the values taken by the slack variables in the joint distribution of these potential responses, all the lower bounds $l(s)$ for $\mathrm{PC}(s)$ given by (13.8) can be achieved simultaneously. This leads to an achievable lower bound for PC:

$$\mathrm{PC} \ge L := \sum_s l(s) \times \Pr(S=s \mid E=1, R=1). \tag{13.11}$$

Using

$$\frac{\Pr(S = s \mid E = 1, R = 1)}{\Pr(R = 1 \mid E = 1, S = s)} = \frac{\Pr(S = s \mid E = 1)}{\Pr(R = 1 \mid E = 1)},$$

we can re-express the lower bound as

$$\begin{aligned} L = & \frac{1}{\Pr(R = 1 \mid E = 1)} \\ & \times \sum_s \max\{0,\ \Pr(R = 1 \mid E \leftarrow 1, S \leftarrow s) - \Pr(R = 1 \mid E \leftarrow 0, S \leftarrow s)\} \\ & \times \Pr(S = s \mid E = 1), \end{aligned} \tag{13.12}$$

where we compute

$$\Pr(S = s \mid E = 1) \propto \Pr(E = 1 \mid S \leftarrow s)\Pr(S = s), \tag{13.13}$$

and then

$$\Pr(R = 1 \mid E = 1) = \sum_s \Pr(R = 1 \mid E \leftarrow 1, S \leftarrow s) \Pr(S = s \mid E = 1). \tag{13.14}$$

Similarly, we can re-express the upper bound $U = \sum_s u(s) \times \Pr(S = s \mid E = 1, R = 1)$ as

$$\begin{aligned} U = 1 - & \frac{1}{\Pr(R = 1 \mid E = 1)} \\ & \times \sum_s \max\{0,\ \Pr(R = 1 \mid E \leftarrow 1, S \leftarrow s) - \Pr(R = 0 \mid E \leftarrow 0, S \leftarrow s)\} \\ & \times \Pr(S = s \mid E = 1). \end{aligned} \tag{13.15}$$

The results (13.12) and (13.15) agree with the bounds asserted in formula (11) of Dawid et al. (2017). They can also be shown to agree, under the conditions assumed, with formula (4) of Kuroki and Cai (2011).

Note that it is not appropriate to compare the above bounds with the simple bounds of Sect. 13.4. Those bounds only apply when we can assume the simple "no confounding" property, which in particular requires $\Pr(R = 1 \mid E = 1) = \Pr(R = 1 \mid E \leftarrow 1)$. However, in the presence of the unobserved confounder variable S, this will typically not be so.

Exposure Independent of Covariate

A similar situation was considered in Dawid (2011), with the difference that there it was assumed that, for both Ann and the study data, exposure E is not affected by S—e.g., because E is set externally. The response R may however still depend on E and S jointly.

This can be treated as a special case of our story above, where now $E(s)$ does not in fact depend on s and thus is just E. Otherwise, the assumptions are the same. This change does not affect the derivation of (13.10), nor the property that we can vary $\mathrm{PC}(s)$, as given by (13.8), independently for different s.

We now have $E \perp\!\!\!\perp S$, so that (13.12) becomes

$$L = \frac{1}{\Pr(R = 1 \mid E = 1)} \times \sum_s \max\{0,\ \Pr(R = 1 \mid E \leftarrow 1, S \leftarrow s) - \Pr(R = 1 \mid E \leftarrow 0, S \leftarrow s)\} \times \Pr(S = s),$$

which agrees with the formula given in Dawid (2011). Likewise, formula (13.15), with $\Pr(S = s)$ replacing $\Pr(S = s \mid E = 1)$, reduces to that given for the upper bound in Dawid (2011).

In this case, the assumptions underlying the simple bounds of Sect. 13.4 do hold, so it is appropriate to compare the above bounds with the simple bounds. It can easily be shown that $L \geq l$, with equality if and only if all the conditional risk ratios

$$\frac{\Pr(R = 1 \mid E \leftarrow 1, S \leftarrow s)}{\Pr(R = 1 \mid E \leftarrow 0, S \leftarrow s)} \qquad (s \in \mathcal{S})$$

lie on the same side of 1. Commonly, we might expect all these risk ratios to be bigger than 1, in which case access to the covariate S does not offer an improvement. Similarly, $U \leq u$, with equality if and only if all the ratios

$$\frac{\Pr(R = 1 \mid E \leftarrow 1, S \leftarrow s)}{\Pr(R = 0 \mid E \leftarrow 0, S \leftarrow s)} \qquad (s \in \mathcal{S})$$

lie on the same side of 1.

Comparison with Tian and Pearl (2000)

Tian and Pearl (2000) consider a case where both observational and experimental data are available on E and R, but we do not have information on the sufficient

covariate. In our notation, their equation (17) gives the following lower bound for PC:

$$\max\left\{0, \frac{\Pr(R=1) - \Pr(R=1 \mid E \leftarrow 0)}{\Pr(E=1, R=1)}\right\}. \tag{13.16}$$

This can in fact be derived as a special case of our expression (13.12), if we take as S the binary variable $D =$ "desired exposure", which behaves as a sufficient covariate (this follows from the consistency condition assumed by Tian and Pearl (2000)—see Dawid et al. (2015).) The introduction of such a variable has been found useful in a number of contexts (Corradi and Musio 2020, Dawid 2021, Geneletti and Dawid 2011, Richardson and Robins 2013). The actual exposure E will be identical with D in an observational setting but need not be so in an experimental setting (where D may not be observable).

Since $\Pr(D=1 \mid E=1) = 1$, $\Pr(D=0 \mid E=1) = 0$, (13.12) reduces to

$$L = \max\left\{0, \frac{\Pr(R=1 \mid E=1, D=1) - \Pr(R=1 \mid E=0, D=1)}{\Pr(R=1 \mid E=1)}\right\}. \tag{13.17}$$

Now since in the observational regime $E = D$, we have

$$\Pr(R=1 \mid E=1, D=1) = \Pr(R=1 \mid E=1), \tag{13.18}$$

which is estimable from observational data. But we cannot so identify $\Pr(R = 1 \mid E = 0, D = 1)$, since the conditioning event does not occur in observational circumstances. However, we can estimate $\Pr(R = 1 \mid E \leftarrow 0)$ from experimental data; and we have

$$\begin{aligned}\Pr(R=1 \mid E \leftarrow 0) = {} & \Pr(R=1 \mid E=0, D=0) \times \Pr(D=0 \mid E \leftarrow 0) + \\ & \Pr(R=1 \mid E=0, D=1) \times \Pr(D=1 \mid E \leftarrow 0),\end{aligned}$$

where $\Pr(D=0 \mid E \leftarrow 0) = \Pr(D=0) = \Pr(E=0)$, $\Pr(D=1 \mid E \leftarrow 0) = \Pr(D=1) = \Pr(E=1)$, while $\Pr(R=1 \mid E=0, D=0) = \Pr(R=1 \mid E=0)$. So we can now identify

$$\Pr(R=1 \mid E=0, D=1) = \frac{\Pr(R=1 \mid E \leftarrow 0) - \Pr(R=1, E=0)}{\Pr(E=1)}. \tag{13.19}$$

Inserting (13.18) and (13.19) in (13.17), we recover (13.16). In similar fashion, we can recover the upper bound in equation (17) of Tian and Pearl (2000).

Now consider again the case that we do have access to a sufficient covariate S. Then, we can compute $\Pr(R = 1 \mid E \leftarrow 0)$ by the "back-door formula" (§3.3.1 Pearl (2009)):

$$\Pr(R = 1 \mid E \leftarrow 0) = \sum_s \Pr(R = 1 \mid E \leftarrow 0, S \leftarrow s) \times \Pr(S = s). \tag{13.20}$$

So we can also compute the Tian–Pearl lower bound (13.16). Since we have additional information, our lower bound should be better (larger) than theirs. This claim of Dawid et al. (2017) is proved as follows.

We have

$$\begin{aligned} L = & \frac{1}{\Pr(R = 1, E = 1)} \\ & \times \sum_s \max\{0, \Pr(R = 1, E = 1, S = s) - \Pr(R = 1 \mid E \leftarrow 0, S \leftarrow s)\} \\ & \times \Pr(S = s, E = 1). \end{aligned}$$

Using

$$\sum_s \max\{0, a_s\} \geq \max\left\{0, \sum_s a_s\right\}, \tag{13.21}$$

we find $L \geq \max\{0, A/\Pr(R = 1, E = 1)\}$, where

$$\begin{aligned} A = & \Pr(R = 1, E = 1) - \sum_s \Pr(R = 1 \mid E \leftarrow 0, S \leftarrow s) \times \Pr(E = 1, S = s) \\ = & \Pr(R = 1, E = 1) \\ & - \sum_s \Pr(R = 1 \mid E \leftarrow 0, S \leftarrow s) \times \{\Pr(S = s) - \Pr(E = 0, S = s)\} \\ = & \Pr(R = 1) - \sum_s \Pr(R = 1 \mid E \leftarrow 0, S \leftarrow s) \times \Pr(S = s) \\ = & \Pr(R = 1) - \Pr(R = 1 \mid E \leftarrow 0) \end{aligned}$$

on using (13.20). This shows that our lower bound is at least as good as that of Tian and Pearl (2000), as given by (13.16).

We note that there is equality in (13.21) if and only all the $\{a_s\}$ have the same sign. Consequently, our lower bound will be the same as that of Tian and Pearl (2000) when all the conditional risk ratios

$$\frac{\Pr(R = 1 \mid E \leftarrow 1, S \leftarrow s)}{\Pr(R = 1 \mid E \leftarrow 0, S \leftarrow s)} \qquad (s \in \mathcal{S})$$

lie on the same side of 1, in which case the Tian and Pearl bound is just as good as ours—our additional information about S has not been helpful (except in enabling us to compute $\Pr(R = 1 \mid E \leftarrow 0)$ without actually running an experiment).

In similar fashion, we can show that our upper bound U cannot exceed the upper bound derived by Tian and Pearl, namely

$$\min\left\{1, \frac{\Pr(R = 0 \mid E \leftarrow 0) - \Pr(E = 0, R = 0)}{\Pr(E = 1, R = 1)}\right\},$$

with equality if and only if all the ratios

$$\frac{\Pr(R = 1 \mid E \leftarrow 1, S \leftarrow s)}{\Pr(R = 0 \mid E \leftarrow 0, S \leftarrow s)} \qquad (s \in \mathcal{S})$$

lie on the same side of 1.

We illustrate the above with the following example from Dawid et al. (2017).

Example 1 Suppose S is binary, and from the data, we obtain the following probabilities:

$$\begin{aligned}
\Pr(S = 1) &= 0.5 \\
\Pr(E = 1 \mid S \leftarrow 0) &= 0.8 \\
\Pr(E = 1 \mid S \leftarrow 1) &= 0.2 \\
\Pr(R = 1 \mid E \leftarrow 1, S \leftarrow 1) &= 0.2 \\
\Pr(R = 1 \mid E \leftarrow 0, S \leftarrow 1) &= 0.8 \\
\Pr(R = 1 \mid E \leftarrow 1, S \leftarrow 0) &= 0.8 \\
\Pr(R = 1 \mid E \leftarrow 0, S \leftarrow 0) &= 0.2.
\end{aligned}$$

Then, we obtain the following lower bounds for the probability of causation (the upper bound being 1 in all cases):

When Ann is observed to have $S = 1$:	$\text{PC} \geq 0$
When Ann is observed to have $S = 0$:	$\text{PC} \geq 0.75$
When Ann's value for S is not observed:	$\text{PC} \geq 0.71$
Ignoring S and using the Tian and Pearl (2000) bounds:	$\text{PC} \geq 0.53$.

□

13.6 Complete Mediator

Consider now a problem illustrated by the diagram of Fig. 13.2, where binary variable M acts as a complete mediator of the effect of E on R. For CoE analysis, we

introduce the potential responses $M(e)$ and $R(m)$, with $M = M(E)$ and $R = R(M)$. We again assume exchangeability and no confounding, expressed as the mutual independence of E, $\mathbf{M} := (M(0), M(1))$, and $\mathbf{R} := (R(0), R(1))$. We then have the observational conditional independence

$$R \perp\!\!\!\perp E \mid M.$$

We assume that we have data supplying values for $\Pr(M = m \mid E \leftarrow e)$ and $\Pr(R = r \mid M \leftarrow m)$. Then, we can compute

$$\Pr(R = r \mid E \leftarrow e) = \sum_m \Pr(R = r \mid M \leftarrow m) \Pr(M = m \mid E \leftarrow e). \tag{13.22}$$

In Dawid et al. (2016a), it was shown that when we do not observe M for Ann, the lower bound on PC is obtained when the slack variables in the distributions of $\mathbf{M}$ and $\mathbf{R}$ are both set at their lower bounds. And this yields exactly the same value as the simple lower bound l given by (13.6) and (13.7), which entirely ignores the existence of M (except that we use M for computing $\Pr(R = r \mid E \leftarrow e)$ in (13.22)).

However, knowledge of the mediation process does improve the upper bound, obtained when both slack variables are set at their upper bounds. This gives upper bound

$$\frac{\min\{ac + (1-d)(1-b),\ bc + (1-d)(1-a),\ ad + (1-c)(1-b),\ bd + (1-c)(1-a)\}}{\Pr(R = 1 \mid E \leftarrow 1)}, \tag{13.23}$$

where

$$\begin{aligned} a &= \Pr(M = 0 \mid E \leftarrow 0) \\ b &= \Pr(M = 1 \mid E \leftarrow 1) \\ c &= \Pr(R = 0 \mid M \leftarrow 0) \\ d &= \Pr(R = 1 \mid M \leftarrow 1) \end{aligned}$$

(so that the denominator of (13.23) is $\Pr(R = 1 \mid E \leftarrow 1) = (1-b)(1-c) + bd$).

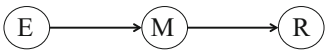


Fig. 13.2 Complete mediator. E is the exposure variable, R is the response variable, and M is the mediator. E directly affects M, M directly affects R, and there is no further direct effect of E on R

For example, suppose we obtain the following values from the data:

$$\Pr(M = 1 \mid E \leftarrow 1) = 0.25$$
$$\Pr(M = 1 \mid E \leftarrow 0) = 0.025$$
$$\Pr(R = 1 \mid M \leftarrow 1) = 0.9$$
$$\Pr(R = 1 \mid M \leftarrow 0) = 0.1.$$

On applying (13.23), we get upper bound PC $\leq$ 0.76, whereas without taking account of the mediator M the upper bound is 1.

Multiple Mediators

The above analysis was extended by Dawid et al. (2021) to the case where we have a complete mediation sequence $E = M_0 \rightarrow M_1 \rightarrow \cdots \rightarrow M_{n-1} \rightarrow M_n = R$ and know the probabilistic structure of each link in the chain. It is assumed there is no confounding at any stage. They consider cases where some of the M's may be observed for Ann, and others unobserved.

Let the transition matrix from M_{i-1} to M_i be $P(\tau_i, \rho_i)$, defined as in (13.2). It can be shown that

$$\tau = \prod_{i=1}^{n} \tau_i$$
$$\rho = \sum_{i=1}^{n} \rho_i \prod_{j=i+1}^{n} \tau_j.$$

Mediators Observed

Suppose we observe each $M_i = m_i$ (with $m_0 = m_n = 1$). Let PC_i be the probability that $M_{i-1} = m_{i-1}$ causes $M_i = m_i$. Then,

$$\mathrm{PC} = \prod_{i=1}^{n} \mathrm{PC}_i,$$

and the same product form will hold for the upper and lower bounds on these probabilities. For the case that all $\tau_i \geq 0$ and all $m_i = 1$, this yields bounds

$$\prod_{i=1}^{n} \left(\frac{2\tau_i}{1 + \tau_i + \rho_i} \right) \leq \mathrm{PC} \leq \prod_{i=1}^{n} \left(\frac{1 + \tau_i - |\rho_i|}{1 + \tau_i + \rho_i} \right).$$

While the lower bound is never smaller than l of (13.5), the upper bound may be smaller or larger than u.

Mediators Unobserved

When $\tau \geq 0$ and all mediators are unobserved, we obtain bounds

$$\frac{2\tau}{1+\tau+\rho} \leq \mathrm{PC} \leq \frac{\tau + \prod_{i=1}^{n}(1-|\rho_i|)}{1+\tau+\rho}, \tag{13.24}$$

where the lower bound is the same as l in (13.5), while the upper bound does not exceed (and is generally less than) u. For $n = 2$, (13.24) agrees with (13.23).

13.7 Mediator with Covariate

Figure 13.3 illustrates a combination of the problems of Sects. 13.5 and 13.6, involving both a complete mediator M and a sufficient covariate S. We allow that S might affect all of E, M, and R but exclude any unmediated (direct) effect of E on R. We thus have potential responses $E(s)$, $M(e,s)$, and $R(m,s)$. From data, we have identified $\Pr(S=s)$, $\Pr(E=e \mid S \leftarrow s)$, $\Pr(M=m \mid E \leftarrow e, S \leftarrow s)$, and $\Pr(R=r \mid M \leftarrow m, S \leftarrow s)$.

We again assume exchangeability and no confounding. We then have the observational conditional independence

$$R \perp\!\!\!\perp E \mid (M,S),$$

and thus the joint distribution is given by

$$\Pr(S=s, E=e, M=m, R=r) =$$
$$\Pr(S=s)\Pr(E=e \mid S \leftarrow s)\Pr(M=m \mid E \leftarrow e, S \leftarrow s)\Pr(R=r \mid M \leftarrow m, S \leftarrow s),$$

so that

$$\Pr(R=r \mid E \leftarrow e, S \leftarrow s) = \sum_m \Pr(R=r \mid M \leftarrow m, S \leftarrow s)\Pr(M=m \mid E \leftarrow e, S \leftarrow s). \tag{13.25}$$

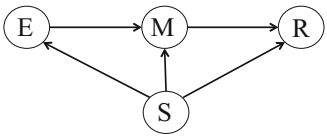


Fig. 13.3 Complete mediator (M) with covariate (S). S affects all the other variables

Conditional on $S = s$, the problem reduces to that considered in Sect. 13.6, except that all probabilities are to be conditioned on $S = s$.

Lower Bound

Knowledge of a complete, but unobserved, mediator does not affect the lower bound on PC. Consequently, if $S = s$ is observed, but not M, the lower bound is given by $l(s)$ of (13.9) (on using (13.25)). When S is also unobserved, we can independently set the slack variables in the distributions of $(M(s,0), M(s,1))$ and of $(R(s,0), R(s,1))$ $(s \in \mathcal{S})$ to their lower bounds, so achieving the lower bound $l(s)$ for PC(s) simultaneously for all s. We thus recover the overall lower bound $L = \mathrm{E}\{l(S) \mid E = 1, R = 1\}$ (as is also given by (13.12), on applying (13.25), (13.13), and (13.14)).

Upper Bound

From (13.23), conditionally on $S = s$, we have upper bound

$$\begin{aligned} u(s) = \frac{1}{\Pr(R = 1 \mid E \leftarrow 1, S \leftarrow s)} \times \\ \min\{\, a(s)\, c(s) + (1 - d(s))\, (1 - b(s)), \\ b(s)\, c(s) + (1 - d(s))\, (1 - a(s)), \\ a(s)\, d(s) + (1 - c(s))\, (1 - b(s)), \\ b(s)\, d(s) + (1 - c(s))\, (1 - a(s)) \quad \} \end{aligned}$$

with

$$\begin{aligned} a(s) &= \Pr(M = 0 \mid E \leftarrow 0, S \leftarrow s) \\ b(s) &= \Pr(M = 1 \mid E \leftarrow 1, S \leftarrow s) \\ c(s) &= \Pr(R = 0 \mid M \leftarrow 0, S \leftarrow s) \\ d(s) &= \Pr(R = 1 \mid M \leftarrow 1, S \leftarrow s). \end{aligned}$$

Again, for S unobserved, we can vary the slack variables in the distributions of the relevant potential responses, independently for different s. So all upper bounds $u(s)$ can be achieved simultaneously. Hence, the overall upper bound is $\mathrm{E}\{u(S) \mid E = 1, R = 1\}$.

Example 2 Suppose we have binary covariate S and obtain the following values from the data:

$$\begin{aligned}
\Pr(S=0) &= 0.5\\
\Pr(E=1 \mid S \leftarrow 0) &= 0.8\\
\Pr(E=1 \mid S \leftarrow 1) &= 0.2\\
a(0) = \Pr(M=0 \mid E \leftarrow 0, S \leftarrow 0) &= 0.98\\
a(1) = \Pr(M=0 \mid E \leftarrow 0, S \leftarrow 1) &= 0.975\\
b(0) = \Pr(M=1 \mid E \leftarrow 1, S \leftarrow 0) &= 0.75\\
b(1) = \Pr(M=1 \mid E \leftarrow 1, S \leftarrow 1) &= 0.24\\
c(0) = \Pr(R=0 \mid M \leftarrow 0, S \leftarrow 0) &= 0.95\\
c(1) = \Pr(R=0 \mid M \leftarrow 0, S \leftarrow 1) &= 0.90\\
d(0) = \Pr(R=1 \mid M \leftarrow 1, S \leftarrow 0) &= 0.65\\
d(1) = \Pr(R=1 \mid M \leftarrow 1, S \leftarrow 1) &= 0.905.
\end{aligned}$$

We find, using (13.25),

$$\begin{aligned}
&\Pr(R=1 \mid E \leftarrow 0, S \leftarrow 0) = 0.062\\
&\Pr(R=1 \mid E \leftarrow 1, S \leftarrow 0) = 0.5\\
&\Pr(R=1 \mid E \leftarrow 0, S \leftarrow 1) = 0.120\\
&\Pr(R=1 \mid E \leftarrow 1, S \leftarrow 1) = 0.293.
\end{aligned}$$

Then, we obtain the following lower bounds for the probability of causation:

When M is unobserved and $S=0$: $l(0) = 0.876$
When M is unobserved and $S=1$: $l(1) = 0.590$.

Also,

$$\Pr(S=s \mid E=1, R=1) \propto \Pr(S=s)\ \Pr(E=1 \mid S \leftarrow s) \times \{(1-b(s))(1-c(s)) + b(s)d(s)\}$$

from which we find $\Pr(S=0 \mid E=1, R=1) = 0.872$. So, on using this with (13.11), when S too is unobserved, we have lower bound $L = 0.840$.

For the upper bounds, we have

$$\begin{aligned}
u(0) &= \frac{0.4885}{0.5} = 0.977\\
u(1) &= \frac{0.21837}{0.2932} = 0.745
\end{aligned}$$

and, for S unobserved, $U = \mathrm{E}\{u(S) \mid E=1, R=1\} = 0.947$.

Note that if we were to ignore entirely the existence of M, on using (13.6), we would then have both $u(0)$ and $u(1) = 1$ (and so $U = 1$). □

Variation

As a variation of the above problem, we might assume that there is no dependence of M on S, as illustrated in the diagram of Fig. 13.4. The effect of this is to replace the potential variable $M(e, s)$ by $M(e)$.

The "no-confounding" assumptions imply that the observational joint distribution of (S, E, M, R) obeys the conditional independencies encoded by Fig. 13.4, namely:

$$M \perp\!\!\!\perp S \mid E$$
$$R \perp\!\!\!\perp E \mid (M, S),$$

yielding the joint distribution

$$\Pr(S = s, E = e, M = m, R = r) =$$
$$\Pr(S = s)\ \Pr(E = e \mid S \leftarrow s)\ \Pr(M = m \mid E \leftarrow e)\ \Pr(R = r \mid M \leftarrow m, S \leftarrow s), \tag{13.26}$$

so that

$$\Pr(R = r \mid E \leftarrow e, S \leftarrow s) = \sum_m \Pr(R = r \mid M \leftarrow m, S \leftarrow s)\ \Pr(M = m \mid E \leftarrow e). \tag{13.27}$$

We assume that, from data, we can evaluate the factors on the right-hand side of (13.26).

However, none of these modifications changes the essential logic of our previous analysis. We obtain the same lower and upper bounds as in Sects. 13.7 and 13.7, subject only to computing $\Pr(R = r \mid E \leftarrow e, S \leftarrow s)$ using (13.27), and replacing $a(s)$ with $a = \Pr(M = 0 \mid E \leftarrow 0)$, and $b(s)$ with $b = \Pr(M = 1 \mid E \leftarrow 1)$.

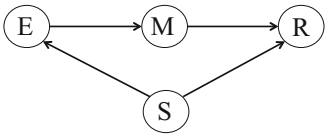


Fig. 13.4 Complete mediator, with covariate S affecting exposure E and response R, but not mediator M

Example 3 We modify Example 2 by replacing $\Pr(M = m \mid E \leftarrow e, S \leftarrow s)$ by

$$\Pr(M = m \mid E \leftarrow e) = \sum_s \Pr(M = m \mid E \leftarrow e, S \leftarrow s)\ \Pr(S = s)$$

(this application of the back-door formula can be regarded as supplying the closest approximation to Example 2 under the additional assumption embodied in Fig. 13.4). The effect of this is to replace $a(s)$ by $a = 0.978$ and $b(s)$ by $b = 0.495$, all other input values being unchanged.

We now find, using (13.27),

$$\Pr(R = 1 \mid E \leftarrow 0, S \leftarrow 0) = 0.064$$
$$\Pr(R = 1 \mid E \leftarrow 1, S \leftarrow 0) = 0.347$$
$$\Pr(R = 1 \mid E \leftarrow 0, S \leftarrow 1) = 0.118$$
$$\Pr(R = 1 \mid E \leftarrow 1, S \leftarrow 1) = 0.498$$

so that we have

When M is unobserved and $S = 0$: $l(0) = 0.817$
When M is unobserved and $S = 1$: $l(1) = 0.763$.

Also $\Pr(S = 0 \mid E = 1, R = 1) = 0.736$, which gives, when S too is unobserved, the following value for the lower bound: $L = 0.803$.

For the upper bounds, we have

$$u(0) = \frac{0.323}{0.347} = 0.930$$
$$u(1) = \frac{0.448}{0.498} = 0.898$$

and, for S unobserved, $U = \mathrm{E}\{u(S) \mid E = 1, R = 1\} = 0.922$.

If we were to ignore M altogether, we would once again obtain $u(0) = u(1) = U = 1$. □

13.8 Discussion

We have highlighted the important distinction between two varieties of causal inference: EoC inferences, about the effects of contemplated interventions, and CoE inferences, about causation or liability in a fully observed individual case. Correspondingly, different mathematical formalisms are required for these two varieties, with CoE inferences expressed in terms of "potential response" variables.

Data sources will typically be available at group level. Under suitable conditions, such as an experimental or ignorable observational setting, such data can be used directly to determine desired EoC-type probabilities. However, they are not directly relevant to CoE-type queries. This chapter has investigated just how, and how well, group-level data can be used to make CoE inference about the "probability of causation", PC, in an individual case. Typically, even with ideal group-level data, PC cannot be identified exactly but only confined to an interval.

The simplest case is when the data relate to exposure and response in an experiment. Here, we have also considered what can be inferred for PC in an individual case when we have some access to the inner workings of the "black box" relating response to exposure, by observing additional variables in the data. Specifically, we have considered the addition of information on covariates and/or complete mediators. In all cases, we get interval bounds on PC, which—when the problem allows the comparison—will be at least as good as the bounds for the simple case.

The structure of each of the problems we have considered has been described by means of a simple directed acyclic graph (DAG). Other problems could involve other types of additional variables, structured into more complicated DAG representations of the black box. It would be highly desirable to develop a more general approach that could handle any such problem.

References

Corradi, F. and Musio, M. (2020). Causes of effects via a Bayesian model selection procedure. *Journal of the Royal Statistical Society, Series A*, **183**, 1777–1792. https://doi.org/10.1111/rssa.12560.

Dawid, A. P. (2000). Causal inference without counterfactuals (with Discussion). *Journal of the American Statistical Association*, **95**, 407–48.

Dawid, A. P. (2011). The role of scientific and statistical evidence in assessing causality. In *Perspectives on Causation*, (ed. R. Goldberg), pp. 133–47. Hart Publishing, Oxford.

Dawid, A. P. (2015). Statistical causality from a decision-theoretic perspective. *Annual Review of Statistics and its Application*, **2**, 273–303. https://doi.org/10.1146/annurev-statistics-010814-020105.

Dawid, A. P. (2021). Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, **9**, 39–77.

Dawid, A. P., Faigman, D. L., and Fienberg, S. E. (2014). Fitting science into legal contexts: Assessing effects of causes or causes of effects? (with Discussion and authors' rejoinder). *Sociological Methods and Research*, **43**, 359–421.

Dawid, A. P., Faigman, D. L., and Fienberg, S. E. (2015). On the causes of effects: Response to Pearl. *Sociological Methods and Research*, **44**, 165–74.

Dawid, A. P., Humphreys, M., and Musio, M. (2021). Bounding causes of effects with mediators. *Sociological Methods and Research* (in Press).

Dawid, A. P., Murtas, R., and Musio, M. (2016a). Bounding the probability of causation in mediation analysis. In *Topics on Methodological and Applied Statistical Inference*, (ed. T. D. Battista, E. Moreno, and W. Racugno), pp. 75–84. Springer.

Dawid, A. P., Musio, M., and Fienberg, S. E. (2016b). From statistical evidence to evidence of causality. *Bayesian Analysis*, **11**, 725–52.

Dawid, A. P., Musio, M., and Murtas, R. (2017). The probability of causation. *Law, Probability and Risk*, **16**, 163–79.

Faigman, D. L., Monahan, J., and Slobogin, C. (2014). Group to individual (G2i) inference in scientific expert testimony. *University of Chicago Law Review*, **81**, 417–80.

Geneletti, S. and Dawid, A. P. (2011). Defining and identifying the effect of treatment on the treated. In *Causality in the Sciences*, (ed. P. M. Illari, F. Russo, and J. Williamson), pp. 728–49. Oxford University Press.

Kuroki, M. and Cai, Z. (2011). Statistical analysis of 'probabilities of causation' using co-variate information. *Scandinavian Journal of Statistics*, **38**, 564–77.

Murtas, R., Dawid, A. P., and Musio, M. (2017). New bounds for the probability of causation in mediation analysis. arXiv:1706.04857.

Pearl, J. (1999). Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, **121**, 93–149.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*, (Second edn). Cambridge University Press, Cambridge.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs: A primer. Second UAI Workshop on Causal Structure Learning, Bellevue, Washington, July 15 2013.

Tian, J. and Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, **28**, 287–313.

Chapter 14
Evidence to Guide Decision Making in Clinical Medicine

Burton Singer and Ralph I. Horwitz

This chapter is dedicated to the memory of Steve Fienberg, who could well be called "Mr. Statistics" of the last several decades. Beyond his many statistical contributions, Steve was a scientific generalist of the highest order. This was exemplified by his chairmanship of the National Research Council's Report Review Committee, which basically covered all of science. Within Steve's large corpus of publications, many were targeted on issues in medicine and epidemiology. His focus, however, was on statistical issues, which dealt with responses of populations and their responses on average. Steve, nevertheless, had an interest in clinical medicine per se from both personal and scientific perspectives. This brings us much closer to the question of evidence needed to guide clinical practice. There, the focus is never the population; it is the individual patient. The present chapter can be considered an initial exegesis on reasoning by the physician about an individual patient. Much of it is case-based reasoning from the experience of the physician and literature focused on similar individual cases. Embedded in this discussion are issues of causality. Here we are faced with the fact that clinical medicine deals with causality from outcomes to candidate causes. This is precisely the opposite of most of the statistics literature, which reasons from candidate causes to outcomes. Although Steve situated most of his writing on the topic in the standard statistics perspective, towards the end of his life, he was just beginning to consider in some detail the real problems of clinical practice which start from the individual patient and of necessity must reason about the patient in his or her own terms. We hope

B. Singer (✉)
Emerging Pathogens Institute, University of Florida, Gainesville, FL, USA
e-mail: bhsinger@epi.ufl.edu

R. I. Horwitz
Lewis Katz School of Medicine, Temple University, Philadelphia, PA, USA
e-mail: ralph.horwitz@yale.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_14

that this chapter can serve to not only honor Steve and his many contributions in the medical domain but also point the way to a vast area of research that the statistics community can add to its corpus. What we are presenting here is only a start but one that we hope would have given Steve great satisfaction about a new domain for scientific investigation.

14.1 Introduction

Any discussion of evidence to guide decision making in clinical practice requires prior explication of the process of reasoning by the clinician starting from initial contact with a given patient. The reasoning process begins with the presentation of symptoms and signs signifying some kind of deviation from normal biological or physiological functioning for the patient at hand, as well as a limited set of sociodemographic facts – e.g., age, sex, ethnic background, and limited work and/or family history. This limited set of information serves to trigger one or more initial hypotheses in the mind of the clinician about mechanisms (dysfunction in one or more biological systems) that might have produced the given symptoms and signs. The basis for initial hypothesis generation is usually the expert knowledge of the physician as well as clinical histories and diagnoses of patients, the prior experience of the attending physician, or case reports from the extant literature, where the information set of these cases approximately matches the presenting conditions of the patient at hand. Without necessarily formalizing the process, indeed, in many instances relying solely on pattern recognition, the clinician is developing a mental database of approximate matches to the patient at hand which can serve as an initial guide to formulating a diagnosis and to suggesting one or more treatments that have a track record of benefitting persons in what is regarded as the relevant comparison population [1].

Iteration of the above process continues with the clinician sequentially examining more and more of the patient's profile of biological and biographical information. New hypotheses are proposed about the mechanisms that might have produced the total examined information at each stage, while also rejecting earlier hypotheses that do not continue to receive support. Often one or more responses to therapeutic treatments are assessed and factored into the formulation of a diagnosis. This strategy ultimately leads to a well-defended diagnosis, a report of several possible diagnoses each with some probability attached to it, or no clear diagnosis or set of them at all. Along the way, new approximate match populations are considered, still based on the prior experience of the attending physician, but also possibly from archives of patient profiles that are called up using diverse combinations of matching conditions suited to the specific question(s) being addressed about the manner in which observed effects for the patient at hand were generated.

The above outline might almost be viewed as an expanded definition of abduction, as put forth by Charles Sanders Peirce [2]. It is essentially a process of learning from data where models of the observed phenomena are generated and

revised, new hypotheses are generated, and new data is assembled with the overall objective of reaching a diagnosis with the strongest available defense. For a closely related overview from the vantage point of economics, see Heckman and Singer [3]. Abduction is a process of detective work where the methodology of Sherlock Holmes is the salient mode of operation. As put forth by Arthur Conan Doyle [4]:

> All knowledge comes useful to the detective . . . . The temptation to form premature theories on insufficient data is the bane of our profession. I can see only two things for certain at present — a great brain in London and a dead man in Sussex. It's the chain between that we are going to trace.
>
> —Sherlock Holmes
> [4, p. 33 and p. 43]

Currently, when decisions about clinical care are considered, it is all too customary to see reference to the randomized controlled trial (RCT) as the "gold standard" that guides decision making. Yet a close reading of the above paragraphs makes clear how little application the RCT has to clinical reasoning. It has long been acknowledged that RCTs are useful for estimating average effects in populations, a result that is most useful to pharmaceutical companies developing drugs and regulators who license them [5]. A more extended set of recent critiques of the method has identified other limitations of RCTs that are especially pertinent to this consideration of clinical reasoning for the patient at hand. For instance, Deaton and Cartwright have emphasized the transportability problem, that if the average result of the RCT is by chance true, it is true for those who were in the trial and not necessarily the patient being treated by the clinician [6].

Other recent critiques are also pertinent when considering the evidential base clinicians require to support clinical reasoning for the individual patient. RCT results may tell the clinician how to start treatment, but not what to do when the patient fails to respond to therapy, or the disease progresses and treatment modifications are needed. Clinical reasoning requires an iterative process that incorporates new data as the patient's clinical trajectory changes over time [7]. The dynamic nature of the patient's illness journey demands a similarly dynamic process of evidence generation that is tuned to clinical reality [8]. Case-based reasoning, starting from the clinician's experience and augmented by examination of similar cases to a patient at hand using extensive literature and patient record archives, facilitated by contemporary computer technology and analytical methods for interfacing with large data sets, is at the heart of clinical decision making.

With this background, the purposes of this paper are to (i) clarify the evidential base that is relevant for decision making in clinical medicine, (ii) describe causal reasoning tuned to clinical practice and contrast it with causal inference as considered in the statistics and epidemiology literatures, and (iii) briefly explicate what we regard as important research directions where advances would enhance clinical practice in medicine.

14.2 Approximate Match Populations

Case-based reasoning, reflecting the attending physician's prior experience, has always been at the heart of specification of diagnoses and decision making about therapeutic treatments for a patient at hand. Although a clinician can only consider small blocks/modules of information at any one point in time [9] from what one hopes is a rich biology-biography profile of the patient at hand [10], the sequential examination of multiple modules triggers recall of different approximate match cases for comparison with the patient's profile.

In principle, case-based reasoning pertaining to a particular patient could be dramatically expanded if the clinician could access a library of detailed patient profiles and select a subset of them whose characteristics are matched to those of the patient at hand. Criteria that define a match are, of necessity, idiosyncratic to a question(s) being asked by the clinician about the particular patient. For example, when contemplating use of a particular pharmacological agent for a patient already receiving two or more co-therapies, matching criteria would include the co-therapies and other clinical course descriptors deemed essential for declaring that comparison cases were an adequate match to the patient at hand. The matched cases would be used for evaluating responses to treatment as well as response to no treatment or alternatives to the contemplated treatment as likely scenarios for what would happen to the patient at hand under the various options [11]. A key point about this process is that it requires availability of large libraries of patient profiles if there is to be assurance of identifying approximate matches for rather idiosyncratic sets of conditions.

Elaborating further on this point, consider molecular diagnostics-guided targeted therapies that have become standard treatments for patients with lung cancer [12]. Molecular analyses of various biomarkers in tumor tissue or cytology specimens are part of standard laboratory tests for the clinical management of lung cancers. Identifying approximate match groups for the test results of a lung cancer patient at hand and examining responses to alternative treatments within this group is an increasingly common practice. However, doing such approximate matching while ignoring basic demographic information, smoking history, and especially comorbidity status can ultimately mean that the extent of overlap between the profile of the patient at hand and the comparison group is quite limited. Taking the more refined information into account leads to tradeoffs between richness of detail and approximate match group size. An example of this tradeoff is illustrated by the recent development of targeted therapies for patients with Stage 3 lung cancer. Previous therapy for this disease was limited to surgery for a small minority of eligible patients and chemoradiation for patients ineligible for surgery. Many patients with Stage 3 lung cancer are too frail for these treatments, with the result that as many as half or more of older patients are left untreated [13]. Immune checkpoint inhibitors are now available with a better side effect profile and can be offered to frail older patients previously left untreated.

Referring back to our remarks about RCTs in the introductory section, it is important to emphasize that no RCT or group of RCTs have sufficient numbers of relevant patients to guide clinical decisions for such lung cancer patients. In this setting, only real-world cases can generate the approximate matches that clinicians and patients need [14]. In fact, this is a generic weakness of results from RCTs in which, as Austin Bradford Hill pointed out in the early days of RCTs in medicine, "they cannot answer the clinician's question about the effectiveness of a particular treatment for a given patient" [15]. It is the approximate match group, of whatever size attainable, that is the relevant comparison population for decision making about the patient at hand. It is also important to emphasize that the members of an approximate match group can derive from hospital records, one or more RCTs, observational studies, or clinical practice archives as long as they are subject to quality control standards. Although difficult to do at the present time because of privacy restrictions, you could build up approximate match groups of considerable size were it possible to draw cancer profiles from a multiplicity of sites – e.g., M.D. Anderson, Dana Farber, Sloan Kettering, Mayo Clinic, etc. Some commercial entities are doing that now for patients with cancer (e.g., Flatiron: https://flatiron.com/).

Here it is useful to comment on the recent development of a predictive approach to treatment effect heterogeneity referred to by the acronym PATH [16]. The authors refer to this approach as personalized evidence-based medicine (EBM). The stated goal is to use RCT data to identify optimal subgrouping schemes based on all relevant patient characteristics that yield more individualized estimates of treatment effects for each patient than the average results from the overall trial. From our perspective there are several key problems with this approach, which is currently receiving wide attention in the medical community. These are the following:

(i) The insistence on using data from RCTs virtually guarantees that broad classes of patients – e.g., those with rare conditions, those on multiple co-therapies, those with conditions deemed ineligible to be enrolled in a trial, etc. – will not be in any trial that might be the target of the subgroup formation being considered. This ensures that there will be no approximate matches, here interpreted as members of an appropriate subgroup in an RCT, for comparison in response to a candidate treatment, with a patient at hand.

(ii) Virtually all RCTs lack nuanced biographical information about the enrolled patients. The following case summary from Arthur Kleinman [17] portrays in glaring terms what you could not identify into any "optimal subgroup" formation in an RCT:

> "A medical anthropologist is asked by a pediatrician in California to consult in the care of a Mexican man who is HIV positive. The man's wife had died of AIDS 1 year ago. He has a 4-year-old son who is HIV positive, but he has not been bringing the child in regularly for care. The explanation given by the clinicians assumed that the problem turned on a radically different cultural understanding. What the anthropologist found, though, was to the contrary. This man had a near complete understanding of HIV/AIDS and its

treatment – largely through the support of a local nonprofit organization aimed at supporting Mexican-American patients with HIV. However, he was a very-low-paid bus driver, often working late-night shifts, and he had no time to take his son to the clinic to receive care for him as regularly as his doctors requested. His failure to attend was not because of cultural differences, but rather his practical, socioeconomic situation. Talking with him and taking into account his 'local world' were more useful than positing radically different Mexican health beliefs." This example highlights the critical importance of including detailed biographical information in patient profiles for the large libraries of such profiles that we have proposed above.

(iii) If you were to start with a profile of a patient at hand and seek guidance about the patient's likely response to a treatment assessed in a given RCT, you might, and you might not, be able to form an "optimal subgroup" for comparison with your patient where the subgroup was a good match – as deemed by the attending physician – to the patient. Even if the subgroup formed a good approximate match, which we view as unlikely given the heterogeneity of patients enrolled in RCTs, and the simultaneous exclusion of classes of patients who are candidates for the treatment, the resulting analyses would be average responses across the members of the subgroup, which is standard RCT reporting. However, the clinician would like to know the clinical course over time of the patients in the subgroup, whether or not complications or adverse effects occurred and how soon after treatment administration, and whether or not a patient's treatment needed to be changed at some relatively short time after initial administration. In a word, much more flexibility in analysis strategies than those that accompany RCT reporting would be desirable for the clinician whose instincts might demand a focus on special details which are triggered by knowledge of her particular patient.

Point (i) in the above listing is pertinent to the general point we have been making about the need for large archives of patient profiles whether the patients are enrolled in an RCT or simply having their detailed health records as part of a library. As a concrete illustration, when rare cancers are at issue, and the need to identify approximate match groups is particularly pressing, library size can be a major limiting factor. In this regard, we recently considered a patient at hand who had diagnosed cholangiocarcinoma (CC), a relatively rare, but dangerous, bile duct cancer. Some years prior to the CC diagnosis, the same patient had, and was treated successfully for, Hodgkin's lymphoma (HL). Having access to a large and diverse patient database at a single medical center from which to do approximate matching, we found at the coarsest level for matching that there were 4284 HL cases available, 1125 CC cases available, but only 2 cases where HL was followed by a CC diagnosis at any point later in life. Without the option of accessing profile data across multiple medical centers, the approximate matching effort was stopped before it ever got started. This phenomenon is by no means limited to cancers. It is a major motivation for a national, and international, movement to develop large

libraries accessible to clinicians broadly and subject to nuanced ethical constraints that await full delineation.

Approximate Matching and N-of-1 Trials

The increasing availability of sensory devices that can continuously monitor a diversity of physiological parameters, as well as patient responses to treatments, has stimulated a renewed focus on N-of-1 clinical trials [18]. This is an interesting option for treatment of chronic conditions in which the patient acts alternately over time as a recipient of a treatment and as a control without it. The earliest careful treatment of an N-of-1 trial and associated design issues is by Hogben and Sim [19], a paper that received very little attention until it was resurrected by George Davey Smith and republished in his *International Journal of Epidemiology* in 2011 [20]. Although a given N-of-1 trial may benefit the patient who is its subject, our purpose in raising this topic here is to point out that results of N-of-1 trials in approximate match populations for a particular patient at hand can provide more highly controlled guidance about performance of a particular therapy than results from a single administration of the treatment under consideration. Being able to access a library of patient profiles that contain N-of-1 trials has the potential to enhance the evidential base for clinical practice. In particular, this is a route into being able to characterize for whom a particular treatment is effective and for whom it is not efficacious. The importance of such characterizations has been apparent in clinical practice for over a century. However, the widely employed evaluative instrument for diverse therapies, the randomized controlled trial (RCT), does not provide answers to this important question.

14.3 Causal Reasoning

Two places in clinical practice where issues of causality arise are the processes of developing a diagnosis and formulation of a prognosis or estimate regarding the efficacy of a treatment(s) administered to a particular patient at hand. These two activities are not disjoint. Indeed response to an intended therapeutic treatment may be part of the evidence leading to a diagnosis.

Elaborating on these points, we first emphasize that most causal reasoning in clinical practice is reasoning from effects to causes. This stands in stark contrast to the direction of reasoning in the statistics literature [21], where we find "causal inference is ultimately concerned with the effects of causes on specific units" [21, page 947]. There has been an extensive development of methods of causal inference reasoning from causes to effects in the 34 years since [21] was published [22–25]. Particularly interesting are the publications focusing on causal inference with interference [26, 27]. Unfortunately, this literature is not relevant to the form of

causal reasoning that is central to clinical practice, particularly in the development of diagnoses, with the exception of the use of therapeutic interventions as part of the process of developing a diagnosis. Since a focus on causal reasoning in clinical practice heavily emphasizes what goes on in the mind of the attending physician, the cognitive psychology/neuroscience interface is of central importance for providing an ever more refined view of knowledge representation and causal claims when the reasoning is from effects to causes.

It is worth explaining this point further. In making a diagnosis, physicians are seeking to know a cause that led to an effect and, in evaluating treatment, how a therapy led subsequently to a treatment response. In both circumstances, the causal chain proceeds forward from cause to effect. But clinicians are always reasoning backwards from the observed effects to the proximate causes. Patients come with signs and symptoms from a cause (disease) that already exists. And clinicians understand treatment effects better retrospectively after they occur than before they are observed. Effects of treatment that occur soon after a treatment, such as when a medication to lower blood pressure has its effects soon after it is begun, are easy for physicians to appreciate. What is not easy to appreciate from direct observations are the long-term effects of lowering blood pressure such as reductions in later heart failure, stroke, or kidney disease. For these outcomes, collections of patients with longitudinal follow-up provide the basis for identifying the causes by studying the effects.

We also emphasize that the measures of uncertainty customarily required in causal inference in the statistics literature are not applicable to clinical practice. Clinical reasoning seeks to identify the best option for the single patient at hand, recognizing that judgments are made under conditions of considerable uncertainty [28]. This further supports the position that abduction is a preferred inferential framework that better conforms to clinical reality.

It is important to indicate at the outset that detailed rules for causal attribution in medicine are not available. Nevertheless, physicians intensively seek causal explanations of the symptoms, signs, and the broader range of phenomena they observe. Such searches for a cause not only are the basis for understanding pathogenesis of clinical manifestations in individual patients, but, in some cases also may be the modality by which new hypotheses about the mechanisms of disease are first identified.

To clarify our position, we explicate first in general and then, by concrete analysis of a particular patient, the process of reasoning by the clinician starting from initial contact with a given patient. The reasoning process begins with the presentation of symptoms and signs signifying some kind of deviation from normal biological, physiological, or immunological functioning for the patient, as well as a limited set of sociodemographic facts – e.g., age, sex, ethnic background, and limited work and/or family history. This limited set of information serves to trigger one or more initial hypotheses in the mind of the clinician about conditions and mechanisms that might have produced the given symptoms and signs. Such mechanistic formulations represent candidate causal models to explain the observed symptoms and signs, i.e., the effects. The basis for this initial hypothesis generation is usually the

knowledge of the physician on clinical histories and diagnoses of patients in her prior experience, or case reports from the extant literature, where the information set of these cases approximately matches the presenting conditions of the patient at hand. Without necessarily formalizing the process, the clinician is developing a database of approximate matches to the patient at hand which can serve as an initial guide to formulating a diagnosis and to suggesting one or more treatments that have a track record of benefitting persons in what we regard as the relevant comparison population [1].

There is frequently a considerable set of nonequivalent mechanisms – candidate explanatory models – that can produce the initial set of symptoms and signs. They also provide a basis for the clinician asking to see the results of particular additional assessments – e.g., laboratory assays, imaging, gene expression profiles [29], responses to prior treatments, etc. – that can facilitate selection of a proper subset of the original candidate diagnoses for further consideration, temporarily ruling out those that are discarded at this stage. With this more extensive set of data from the patient's profile, the clinician again proposes more refined candidate mechanistic descriptions that are capable of producing the full set of so far examined data from the patient at hand. Again, these hypotheses may be based on cases from prior clinical experience of the attending physician. However, this limited set of recalled cases can be augmented by identifying a new set of approximate matches to the patient at hand from a large archive of patient profiles.

At this stage, development of a diagnosis from the more refined set of data would be based on delineation and verification of a proposed mechanism as being consistent with the current and past conditions of the patient at hand. Prior to presentation of a diagnosis, the clinician, in consultation with the patient, may want to ascertain the response of the patient to one or more therapies, suggested by the patient profile data and/or examine the results of additional tests that may serve to discriminate among candidate mechanisms. Under both scenarios, different approximate match populations should be assembled from a library of cases where, for example, a contemplated treatment is applied to some of the cases and where it is not for others, or an alternative treatment is applied. Responses that are short term and long term to the treatment regimens can provide further guidance about a diagnosis.

As the many cases in [1] attest, the iteration process with increasing information about the patient at hand may continue through many rounds, often with discarding of the entire set of initial diagnoses and proposal of new ones. In our opinion, there is no simple straightforward taxonomy of the possibilities. There is no substitute for simply building up a library of many detailed cases, as illustrated by the set of 68 examples in [1].

Example 1 – Adaptation of CASE 35 from [1]. Here we illustrate the reasoning by the clinician for a particular patient at hand. Information from the patient's profile, *PP,* is indicated in *bold face* and designated by *PPx,* where *x* is a label for a particular aliquot of information. The clinician's response to *PPx* and *PPx'* for x', having been examined prior to x, is labeled *CRx.* Although an extensive amount of information may be in the profile, no clinician can grasp it all in parallel

[5]. The sequential examination of information and corresponding generation of alternative hypotheses by the clinician is the actual mode of reasoning involved in clinical practice. We give this feature particular prominence as it seems to be absent from virtually all discussions of causal inference in clinical medicine of which we are aware. The most famous forums to display clinical reasoning are the case studies of the Massachusetts General Hospital where the sequential presentation and discussion of the case has been the method employed for nearly 100 years. We also hasten to add that reasoning of the kind we are illustrating can only be carried out by someone with clinical experience who is also expert in the relevant disciplines of clinical medicine.

The case presentation from *PP1* thru *CR10* and *PP11* is taken from [1, pp. 180–183].

PP1: A 44-year-old man was transferred to the hospital for recurrent episodes of fever, malaise, headache, and confusion.

CR1: This could be a serious situation. In anyone with fevers, headaches, and confusion, I would be concerned about central nervous system infection. Since he was transferred from another institution, I'd wonder if he has already been studied, and if no cause was found, he might have an unusual central nervous system infection. If he has had these symptoms for some time, then bacterial infections such as meningococcus infection is unlikely.

PP2: The patient was in good health until 6 weeks prior to admission when he was admitted to a local hospital with gastrointestinal bleeding from gastritis that was attributed to aspirin and nonsteroidal anti-inflammatory drugs that he had been taking for chronic back pain. He also had a long history of depression.

CR2: It's hard to know if this history is relevant to his current symptoms. Certainly if he were continuing to take nonsteroidal anti-inflammatory drugs, then aseptic meningitis due to these medicines would be a possibility. The combination of gastrointestinal bleeding and back pain raises the possibility of some type of inflammatory bowel disease and a concomitant spondyloarthropathy. Possibly while in the hospital, he acquired a nosocomial infection, perhaps related to the endoscopic procedure, and the complication is only now becoming clinically apparent.

PP3: His current symptoms began 1 month later when over the course of 2 days he became increasingly lethargic, confused, and ataxic. At work he was found sitting at his computer not knowing what to do. His temperature was 39 °C and he was taken to the hospital.

CR3: If the current symptoms were related to his hospitalization a month earlier, then we are dealing with an indolent process. On the other hand, it may be totally unrelated. The confusion raises the question of some type of encephalopathy. The ataxia makes me think of something going on in his cerebellum or perhaps a peripheral neuropathy or perhaps some type of ear infection that has led to a brain abscess. I'm thinking about the possibility that some drug could have caused these symptoms. I don't think that proton-pump inhibitors could be responsible. If he were achlorhydric from the proton-pump inhibitor, he might be more susceptible to an infection like tuberculosis. My main concern at this time is some type of basilar meningitis.

PP4: He was very lethargic but awake. He had difficulty saying "Methodist Episcopal." There was mild left-right confusion and difficulty in repeating. He was diffusely weak with no other focal findings. The remainder of his physical examination was not revealing. Complete blood count, blood glucose, electrolytes, and serum calcium were normal. Lumbar puncture was normal and a toxic screen was negative. The neurologist was concerned about encephalitis and initiated acyclovir therapy.

CR4: He appears to have a diffuse encephalopathy with signs of bilateral cortical dysfunction. He could have a vocal apraxia or something more serious like a loss of language itself. The diffuse weakness makes me think of some type of toxin with systemic effects. Could he have one of the viral encephalopathies? Does he live in an endemic area? Is it the right time of year? He does not appear to have any metabolic abnormalities. Given the potential seriousness of herpes encephalitis, the initiation of acyclovir seems reasonable since the potential benefits outweigh the risks at this point. Finally, could he have some type of a paraneoplastic syndrome?

PP5: *Head CT, MRI, and EEG were unremarkable. Chest X-ray showed a possible right basilar infiltrate. Blood and CSF cultures were negative. Acyclovir was stopped and ceftriaxone therapy was initiated. His symptoms gradually improved and he was discharged. The final diagnosis was probable viral syndrome and depression.*

CR5: The presence of a possible pulmonary infiltrate and neurologic dysfunction makes me wonder about the possibility of a Legionella infection, although the ataxia would be unusual. The improvement of his symptoms in the hospital could be related to the antibiotics or could just be coincidental. PP6: He returned 1 week later with similar symptoms. Blood and urine cultures were negative. Mono spot and a TSH were normal. Lyme titers and a PPD were negative. Other routine lab studies were unremarkable. Chest CT raised the question of a right middle lobe consolidation and he was given clarithromycin. By the third hospital day, he had improved remarkably and was discharged.

PP6: He returned 1 week later with similar symptoms. Blood and urine cultures were negative. Mono spot and a TSH were normal. Lyme titers and a PPD were negative. Other routine lab studies were unremarkable. Chest CT raised the question of a right middle lobe consolidation and he was given clarithromycin. By the third hospital day, he had improved remarkably and was discharged.

CR6: Does he have some type of relapsing disease or could he have a partially treated infection? Could he have some type of occult abscess which is causing intermittent symptoms? Is he being exposed to some kind of toxin? Still's disease could be associated with intermittent fevers and back pains, but not central nervous system symptoms, and vasculitis must always be on the differential of intermittent febrile illnesses; it can involve the central nervous system.

PP7: 2 days later he returned with the same symptom complex including slurred speech, ataxia, and confusion. Meanwhile he had been given paroxetine 30 mg qd for depression and pantoprazole 40 mg qd for his stomach symptoms.

CR7: I'm still wondering about undiagnosed basilar meningitis, central nervous system tuberculosis, and fungal infections. Prior to HIV disease, cryptococcal

disease often occurred in immunocompetent hosts. Is the patient taking any illicit drugs?

PP8: The patient lived with his wife and worked as a sewer inspector. He had a distant history of cigarette smoking. He denied recent travel, sick contacts or pets. He had abused alcohol in the past but had been sober for 6 years. He exercised regularly before his recent illness.

CR8: Given his work as a sewage inspector, leptospirosis immediately comes to mind. This can be a relapsing illness associated with confusion but also hepatic and renal disease. Has he been bitten by a rat and acquired rat bite fever? I doubt it. The regular exercise makes one wonder if he were using performance-enhancing drugs such as anabolic steroids or the like.

PP9: He was in no distress. Blood pressure 100/65 mm Hg, heart rate 54 per minute with no orthostatic changes. He was afebrile. General examination was unremarkable. He was oriented but slow to respond. He could not remember his phone number. He was ataxic and unsteady on his feet. The rest of the neurological examination was normal.

CR9: His ataxia appears to be central in origin. CNS infection still is on my list, but could he have one of the avitaminoses such as thiamine deficiency or another nutritionally related illness?

PP10:-The following laboratory studies were normal: Electrolytes, glucose, liver function tests, serum calcium, serum magnesium, and CK. White cell count was 7400 per with a normal differential. CBC was unchanged. Sedimentation rate was 53 mm/hr. Hepatitis serologies and HIV test were negative. Brucella titers were negative.

CR10: If his outside studies have been reviewed and are normal, then repeat imaging studies are unlikely to be very helpful. Is he taking any health foods of any kind?

PP11: On further questioning, the patient's wife said that 2 months previously the patient had started taking kava kava and valerian root for his depression. Each time he was admitted to the hospital, he stopped the herbal medications and his symptoms subsided. No further tests were performed, and except for drugs for depression, no other medications were given. As before, within 2 days the neurological findings disappeared. He was advised not to take the herbal products. The patient subsequently had no recurrence of his symptoms.

Analysis of Causal Reasoning in the Present Case

Three times the patient presented with confusion and ataxia, and twice he improved during a short hospital stay. Nevertheless, the physicians caring for him did not recognize the nature of his illness until the third time he recovered. The clinical responder (*CR*) seemed wiser; even after she heard the information from the patient's first admission, she raised the possibility of "some type of toxin," and she repeated this concern when she heard about the information from the second

and third admissions. Rapid improvement in symptoms during hospitalization on repeated occasions is a signal to invoke a classic causal hypothesis, namely, "post hoc, ergo propter hoc" (after this, therefore because of this).

There are not many disorders that fit a relapsing and recovering pattern, but exposure to a toxin (and subsequent elimination of the toxin) is one. Examination of the available literature on the toxicity of these herbal substances is frustrating, principally because so little research has been done on the side effects of these materials. Most sites list no side effects of valerian root and only liver toxicity for kava kava. Yet the case for kava kava and/or valerian root as the toxin in this case is quite convincing, based on the return of symptoms after exposure and improvement after cessation of the herbals. In essence, this is the nature of the "challenge-dechallenge-rechallenge" concept, the idea that the strength of a causal attribution increases if the response disappears when a stimulus is removed and reappears when the stimulus is reapplied. In this patient, dechallenge and rechallenge occurred twice after the initial event, providing convincing evidence of a cause-and-effect relationship between the herbals and the patient's illness. Given the multiple recurrent episodes, it seems safe to make the causal connection here.

In terms of the general framework for causal reasoning that we presented above, in the present case, repeated relapse and recovery sets up a cause-and-effect hypothesis, a causal field, or the context, which then demanded satisfaction and explanation. The causal field in this instance could be construed as a chain consisting of a simple explanation, namely, could a toxin at home be the cause of symptoms. The causal field also is important in setting the number of alternative explanations for an event or finding. Here, numerous possible diagnoses were raised from serious infections to vasculitis.

Regarding cues to causality, the intensity of the stimulus and the result are probably concordant even though data on the toxicity of the herbals was difficult to come by. Necessarily, this relation is a probabilistic one, with probability interpreted by experienced physicians as degree of belief. On the basis of the cues alone, we would be justified only in being suspicious that the herbals caused the central nervous system manifestations.

Many measures of strength of a causal linkage were satisfied for the particular patient, but the relation in time and space between the putative stimulus and the response was the most potent. Such relationships are, of course, only correlations.

Finally, in considering counterfactuals, no other explanation than the patient's ingestion of herbal products having some toxicity better explained the patient's clinical pattern or gained any serious credibility.

Mechanistic Models and Clinical Reasoning

In the above clinical responses, *CRx*, the physician's causal assertions are simply presented as claims that a particular cause, X, could have produced an effect, Y. There is an extensive amount of physiology, frequently accompanied by

mathematical models, which underlies each claim. There is also case experience of the physician with multiple patients where the given Y was an effect of X and where X may have been part of a diagnosis. In the present example, the physician was confident about the candidate causal claims and the reasoning that took place with the ten sets of information considered from the patient's profile. No attempt was made at any stage in the iteration process to identify an approximate match population as a source of additional evidence to guide the reasoning toward a diagnosis.

The brevity of the physician's responses masked some complex details that underlie causal reasoning in clinical practice and that we summarize here. A fundamental point is that it is important to understand the representation of causal knowledge in the minds of expert clinicians. Kuipers and Kassirer [30] studied these representations by acquiring verbal transcripts of clinicians describing their causal reasoning processes in detail for phenomena where mechanistic theory in the form of differential equation models characterized some physiological processes that were the causes of effects observed in patients.

When soliciting clinicians' more detailed explication of mechanisms for how a potential cause, X, produced an effect, Y, what would be described as the description of the mechanism was a sequence of qualitative statements such as "a given parameter increased between two consecutive observation times," "another parameter was found to be persistently lower than what is expected when the system is in equilibrium," "two intermediate variables are simultaneously moving in opposite directions," etc. The fundamental point here is that the physician typically lacks precise numerical values for many parameters characterizing a patient's state. Further, some parameters may be difficult or impossible to measure. The physiological system about which she is making a causal claim in the context of a patient at hand frequently has a textbook description via, for example, a system of ordinary differential equations (ODE). The clinician may understand this and be conversant with the behavior of solutions of the ODEs subject to a variety of initial conditions and parameter ranges. However, this is a far more precise level of description than can be put forth, measured, and validated for any given patient.

The key point is that an expert physician reasoning about a case uses only those factors she considers particularly relevant and thus is able to restrict her attention to a much smaller model. To make up for the lack of detail, the expert must then have many different small models, each with its own assumptions and thus expressing different "points of view." The causal model representation is intended to express this highly modularized knowledge structure, so its models will typically be relatively small. Indeed, it appears that there is a match between the limited working memory and processing capacity of the human and the inability of the causal model representation to handle very large models.

The discrete modular nature of information about the dynamics of a physiological system in the clinician's mind, despite the fact that the system may be more completely characterizable by ODEs, raises the question of the extent to which sets of qualitative statements, augmented by numerical constraints can closely approximate the more precise characterization. This question was taken up by

Kuipers and Kassirer [30], who provided a detailed representation of nephrotic syndrome and the pathway from impaired kidney function to edema observed in a patient via both ODEs and their qualitative, but less precisely specified, counterpart QDEs (qualitative differential equations) [31–34]. As a point of fact, the clinician makes detailed assessments of candidate causal mechanisms in a patient at hand via sparse qualitative claims. This then raises the mathematical question of characterizing the physiologically meaningful systems of dynamical equations that are consistent with a given QDE specification of inequalities, initial conditions, and simple algebraic relationships.

In a bit more detail, the allowed conditions in a QDE specification represent the results of an attempt to develop a knowledge representation capable of describing human commonsense reasoning and explanation about physical causality. *Commonsense* causal reasoning is qualitative reasoning about the behavior of a mechanism that can be done without external memory or calculation aids, although it may draw, and indeed does in clinical medicine, on knowledge learned from the advanced study of a particular domain such as medical biology and physiology. In order to be useful for modeling human commonsense knowledge, the computational primitives of the QDE formulation must not require excessive memory or processing resources [31].

While commonsense causal reasoning characterizes the knowledge representations for clinician's consulting the information base in their minds, some of which is derived from approximate matches to patients at hand from their prior clinical practice experience, much more intricate knowledge representations can be introduced if an AI/Expert System is to assist the reasoning process. To this end, Kuipers and Kassirer [20] introduced qualitative physiological system descriptions that match the level and kind of detail that the clinician may have available in a patient profile, but that can also be mapped onto families of dynamical system models (here, ODEs) that represent a more fine-grained set of causal explanations for observed effects in the patient at hand. Detailed examples of such mappings are exhibited in Kuipers et al. [33].

A central point is that the clinician's qualitative assessments cannot be used to identify an ODE representation of physiological dynamics. They can be used to identify a class of QDE models that have the same qualitative behavior as the ODE system, and this is the sense in which causal mechanism is delineated in clinical practice. For a rare example with all the details, the reader should study Kuipers and Kassirer [30] and their thorough and, of necessity, intricate analysis of nephrotic syndrome. Here the problem is to provide a causal explanation of edema in a patient at hand based on sparse measurements and qualitative approximations to dynamics of a physiological system.

14.4 Large Libraries of Patient Profiles

The word "large" in the title of this section has two meanings for our purposes. First, considering the patient at hand, the individual profile consisting of biological, clinical, and biographical information assembled longitudinally has the potential to

grow to enormous size, far beyond what is currently part of the electronic health record (EHR). Second, we have repeatedly pointed out the need to have many profiles available in a library that can be searched for approximate matches to a given patient at one or more points in time and with varying criteria as the basis for matching. We elaborate on these two points of view beginning with the content of profiles for the patient at hand.

In an initial effort to delineate the content of profiles for patients with systemic lupus erythematosus (SLE) [10], profile specification was based on data from randomized controlled trials (RCTs) of belimumab, noting that most of the information in the patient record was not used in published analyses of the results of the trials. This is a generic situation for patient records in RCT files. However, our interest was not in the trials per se, but in assembling profiles that would be part of a library to be used for approximate match population specification for given women (and less commonly men) who are SLE patients. The profiles included longitudinal representations of clinical courses in accordance with a taxonomy of such histories tuned to SLE. They also contained standard clinical and biological/biomarker data, and social/functional information derived from a 36-item short form health survey (SF36) [35].

An important limitation of these profiles is the lack of biographical information beyond basic demographics. This is a common feature of most medical records, whether they be EHRs, hospital patient files, RCTs or observational study files, or medical insurance records. The need for much more extensive biographical information is documented in patient records – but not systematically – in the many instances where family, work, community engagement, and psychological features have played a critical role in recovery from illnesses. It is possible that some of this information is contained in parts of the medical record not commonly accessed, such as social worker and nurse notes. New tools of natural language processing are available to explore these sources of biographical information. This perspective is further developed in Lobitz et al. [36].

Substantially expanding on the profile structure in the SLE population is the integrative personal omics profile (iPOP) [29], which includes genomic, transcriptomic, proteomic, metabolic, and autoantibody profiles for a single patient followed over 14 months. While iPOP files are extensive relative to currently routine EHRs, they should properly be viewed as prototypes for what will surely become a standard part of patient profiles in the near future. This kind of profiling should be viewed in parallel with the more strictly metabolic formulation of the patient journey [37]. Another important multi-omics profiling example is the wellness study of 108 individuals [38] which also includes some ambulatory measurements. As lucidly pointed out by Torkami/Topol and colleagues [39], data from sensing devices of diverse types, worn routinely even by people who are not patients with particular illnesses of concern, will lead to a further explosion in data that will become a routine part of a person's health record.

The challenge with the forthcoming vast expansion of information in patient profiles is effective utilization of it in the iterative process between details about a patient at hand and the clinician's focusing on diagnoses and treatment decisions.

The detailed example of this iterative process in Part III can be viewed as a small-scale approximation to the nuance and subtlety of analyses in the future that incorporate multi-omics and multi-sensory device information. Many different approximate match populations can be envisioned for a given case even at one point in the evolving clinical course, each depending on one or more contemplated explanations/models for how a subset of features in the profile came about.

Given the growing emphasis on Expert Systems to aid in the process of developing diagnoses and selecting treatments likely to benefit a patient at hand, and their success in disciplines that involve imaging or pattern recognition, we feel it imperative to point out a current and likely long-term limitation of automated systems in clinical practice. In particular, the inadequately understood phenomenon of judgment by the physician is an integral feature of hypothesis generation and interpretation of analyses as the diagnostic/treatment decision making process moves along. Automating this critical feature of clinical practice would, it seems to us, require a far deeper understanding of the working of the human mind than is currently at hand. An informative discussion of this point is given in [40].

Shifting attention to the second notion of "large" in the title of this section, we come face-to-face with the question of the size of the library of patient profiles that a given clinician might access in any attempt to assemble a population of approximate matches to a patient at hand. The need for such an assembly becomes particularly acute for patients with rare single disorders or with unusual sets of comorbid conditions. The current balkanization of patient records among medical centers, insurance providers, and the pharmaceutical industry presently inhibits the creation of a much-needed national library that could be responsive to the need to construct approximate match populations for patients where even the specification of a defensible diagnosis is problematic based on case-based reasoning by an attending physician(s). There is, a priori, no guarantee that approximate match populations will supply the requisite discrimination among alternative diagnoses and treatment choices. However, they open an avenue for evidence that has not been part of the information bases of clinical practice heretofore. Increasing the scope for approximate matching using data from multiple centers seems to us to be a pressing issue. Many complex ethical, political, and financial issues remain to be addressed to make large libraries of patient profiles a reality in the United States. This is a major topic in its own right that lies beyond the scope of this paper.

14.5 Discussion

Two facets of clinical practice that are in need of much further development formed the basis for our consideration of evidence to guide patient management in clinical medicine. These are (i) the centrality of approximate match populations as natural comparison groups for a patient at hand and (ii) a focus on causal reasoning and in

clinical practice as a sequential process of modular evidence interpretation, hypothesis generation, and increasingly refined and supported attribution of particular mechanisms as the source of symptoms, signs, and technologically sophisticated assessments of aberrant functioning of a patient at hand. Point (ii) is an instance of abduction, as emphasized in Part I. That the entire process of developing diagnoses, integrated with testing of treatments, should properly be viewed as abduction, and an act of detective work in the spirit of Sherlock Holmes has been emphasized in {1, 179–180; 247–248] and other writers trying to characterize the essence of clinical reasoning. We also reiterate our emphasis that this process stands in stark contrast to causal inference as developed in the statistics and epidemiological literatures.

One aspect of causal reasoning in clinical practice that deserves more attention is the correspondence between qualitative statements from the clinician about mechanistic models, qualitative dynamical models – particularly, qualitative differential equations and qualitative simulations – and continuum models as in differential equations modeling of physiological systems. At the coarse level of causal claims provided by the clinician reasoning about a patient at hand, the more tightly reasoned causal modeling in physics, physiology, immunology, and even economics may seem rather far removed from clinical practice. However, the mapping from a clinician's qualitative statements in the context of nephrotic syndrome to a qualitative dynamical system and then to underlying differential equations, demonstrated in [30], provides proof of principle that such systematic linkage is feasible. The topic has been further developed in a series of papers by Kuipers and colleagues [31–34], some of which parallels qualitative modeling studies by De Kleer and Forbus [41–43]. This topic needs further clinical medicine development in the present environment of molecular-level modeling. The correspondence between clinician's qualitative statements and qualitative molecular models is where much emphasis belongs. Developing this area will require close attention to the thought processes of clinicians reasoning about individual patients. This is a highly labor-intensive activity, but it is necessary if the analogue of the Kuipers and Kassirer [30] analysis from the 1980s is to be brought into twenty-first-century clinical medicine.

Finally, it should be noted that the issues discussed herein are part of a larger enterprise evolving under the name medicine-based evidence (MBE) [44–46], meaning evidence development and utilization tuned to the needs of clinical practice. MBE should be, and has been, contrasted [44] with evidence-based medicine (EBM) where a hierarchy of evidence dominated by the RCT is a primary focus of attention. That the RCT is ill-suited to provide guidance about management of a patient at hand has been well known at least since Austin Bradford Hill's Heberden Oration 53 years ago. We have refrained from entering into lengthy critiques of RCT evidence – much of it already being available in standard sources – as our primary objective has been to stimulate further research on issues that are directly germane to clinical practice looking ahead.

References

1. Kassirer J, Wong J, Kopelman R. *Learning Clinical Reasoning, 2nd Edition.* Philadelphia: Lippincott, Williams and Wilkins. 2010
2. Peirce CS. 'Pragmatism and Pragmaticism' in *Collected Papers of Charles Sanders Peirce,* Vol. V, edited by Charles Hartshorne and Paul Weiss, 117. Cambridge, MA: Harvard University Press. 1934
3. Heckman J, Singer B. Abducting Economics. *American Economic Review*, 107(5): 298 -302. 2017
4. Doyle, AC. *The Valley of Fear*, New York: George H. Doran Co. 1914
5. Horwitz RI, Singer B, Viscoli C, Makuch R. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. *J. Clinical Epidemiology*, 49(4): 395–400. 1996
6. Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*. 210C: 2–21. 2018
7. Concato J, Horwitz RI. Randomized trials and evidence in medicine: A commentary on Deaton and Cartwright. *Social Science & Medicine*. 210C: 32–36. 2018
8. Horwitz RI, Singer BH. What works? And for whom? *Social Science & Medicine*. 210C: 22–25. 2018
9. Miller GA. The Magical Number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*. 63(8): 81–97. 1956
10. Wivel AE, Lapane K, Kleoudis C, Singer BH, Horwitz RI. Medicine Based Evidence for Individualized Decision Making: Case Study of Systemic Lupus Erythematosus. *American Journal of Medicine*, 130(11): 1290–1297; 1297 e1–e6. 2017
11. Johnson T, Libeler D, Chen JL. Opportunities for Patient Matching Algorithms to Improve Patient Care in Oncology. *Clinical Cancer Informatics*. ascopubs.org 70.185.104.94, 8 pp. 2017
12. McMillan EA, Ryu M-J, Diep CH, Mendiratta S, Clemenceau JR, Vaden RM, Kim J-H, Motoyaji T, Covington KR, Peyton M, Huffman K, Wu X, Girard L, Sung Y, Chen P-H, Mallipeddi PL, Lee JY, Hanson J, Vortuganti S, Yu Y, Park S, Sudderth J, DeSevo C, Muzny DM, Doddapaneni H, Gaidar A, Gibbs RA, Hwang T-H, Heimbach JV, Wistuba I, Coombes KR, Williams NS, Wheeler DA, MacMillan JB, Deberardinis RJ, Roth MG, Posner BA, Minna JD, Kim HS, White MA. Chemistry-First Approach for Nomination of Personalized Treatment in Lung Cancer. *Cell*. 173: 864–878. 2018
13. Casaluce F, Sgambato A, Maine P, Spagnolo A, Gridelli C. Lung cancer, elderly and immune checkpoint inhibitors. *J. Thoracic Diseases*. 10(Suppl 13): S1474–S1481. 2018
14. Passer A, Spitaleri G, Gyawali B, de Marinis F. Immunotherapy in Non-small-cell Lung CancerPatients with Performance Status 2: Clinical Decision making with Scant Evidence. *J. Clinical Oncology*. 37(22): 1863–1867. 2019
15. Hill AB. Reflections on Controlled Trials. *Annals of Rheumatic Diseases*. 25(2): 107–113. 1966
16. Kent DM, Paulus JM, van Klavern D, D'Agostino R, Goodman S, Hayward R, Ioannides JPA, Patrick-Lake B, Morton S, Pencina M, Raman G, Ross JS, Selker HP, Varadhan R, Vickers A, Wong JG, Steyerberg EW. The Predictive Approaches to Treatment Heterogeneity (PATH) Statement. *Annals of Internal Medicine*. 172(1): 35–45. 2020
17. Kleinman A, Benson P. Anthropology in the Clinic: The problem of cultural competency and how to fix it. *PLoS Medicine*. 3(10) e294: 1673–1676. 2006
18. Mirza RD, Punja S, Vohra S, Guyatt G. The History and Development of N-of-1 trials. *J. Royal Society of Medicine*, 110(8): 330–340. 2017
19. Hogben L, Sim M. The Self-Controlled and Self-Recorded Clinical Trial for Low-Grade Morbidity. *British Journal of Preventive and Social Medicine*, 7: 163–179. 1953
20. Ebrahim S. Hogben on speed, paradox and strain. *International Journal of Epidemiology*. 40(6): 1429–1430. 2011

21. Holland PW. Statistics and Causal Inference. *Journal of the American Statistical Association*. 81(396): 945–960. 1986
22. Pearl J. Causal inference in statistics: An Overview. *Statistics Surveys*. vol. 3: 96–146. 2009
23. Capri G, Wang W, Chan PH, Austin PC. A review and empirical comparison of causal inference methods for clustered observational data with application to the evaluation of the effectiveness of medical devices. *Statistical Methods in Medical Research*. 28(10–11): 3142–3162. 2019
24. Rubin DB. Essential concepts of causal inference: a remarkable history and an intriguing future. *Biostatistics and Epidemiology*. 3(1): 140–155. 2019
25. Heckman JJ, Pinto R. Causal Analysis after Havvelmo. *Economic Theory* 31(1): 115–151. 2015
26. Hudgens M, Halloran ME. Toward Causal Inference with Interference. *Journal of American Statistical Association.* 103(482): 832–842. 2008
27. Halloran ME, Hudgens M. Estimating population level effects of vaccination using large, routinely collected data. *Statistics in Medicine.* 37: 294–301. 2017
28. Kahneman D, Tversky A. Judgment under uncertainty: Heuristics and Biases. pp. 203–221 in A. Tversky, *Preference, Belief, and Similarity* (Edited by Eldar Shafir). Cambridge, MA: MIT Press. 2004
29. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, Miriam E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lachroute P, Bettinger K, Boyle AP, Kasowski M, Gruber F, Sei S, Garcia M, Whirl-Carillo M, Gallardo M, Blanco MA, Greenberg PL, Snyder P, Klein TE, Altman RB, Butte AJ, Ashley EA, Gerstein M, Nadeau KC, Tang H, Snyder M. Personal Omits Profiling Reveals Dynamic Molecular and Medical Phenotypes. *Cell*, 148: 1293–1307. 2012
30. Kuipers B, Kassirer JP. Causal Reasoning in Medicine: Analysis of a Protocol. *Cognitive Science.* 8: 363–385. 1984
31. Kuipers B. Commonsense reasoning about causality: Deriving behavior from structure. *Artificial Intelligence.* 24: 169–203. 1984
32. Kuipers B. Reasoning with Qualitative Models. *Artificial Intelligence*. 59: 125–132. 1993
33. Kuipers B. Qualitative Simulation then and now. *Artificial Intelligence.* 59: 133–140. 1993
34. Kuipers B. Qualitative Simulation. *Artificial Intelligence*. 29: 289–338. 1986
35. RAND Health Care. https://www.rand.org/health-care/surveys_tools/mos/36-item-short-form.html (accessed 7-29-2019).
36. Lobitz G, Armstrong K, Concato J, Singer BH, Horwitz RI. The Biological and Biographical Basis of Precision Medicine. *Psychotherapy and Psychosomatics* (in press). 2019
37. Holmes E, Nicholson JK, Li J, Darzi AW. Phenotyping the Patient Journey. pp. 49–74 in E. Holmes, JK Nicholson, AW Daezi, JC Lindon. *Metabolic Phenotyping in Personalized and Public Health Care.* Amsterdam: Elsevier. 2016
38. Price ND, Magis AT, Earls, JC, Glusman G, Levy R, Lausted C, McDonald DT, Kusebauch U, Moss CL, Zhou Y, Qin S, Moritz RL, Brogaard K, Omenn GS, Lovejoy JC, Hood L. A wellness study of 1o8 individuals using personal, dense, dynamic data clouds. *Nature Biotechnology*, 35(8): 747–756. 2017
39. Torkamani A, Andersen KG, Steinhubl SR, Topol EJ. High-Definition Medicine, *Cell*, 170: 828–843. 2017
40. Quer G. Muse ED, Nikzad N, Tool EJ, Steinhubs SR. Digital Medicine: Augmenting diagnostic vision with AI, *The Lancet*, 390: 221. 2017
41. De Kleer J, Brown S. A Qualitative Physics Based on Confluences. *Artificial Intelligence.* 24: 7-83. 1984.
42. Forbus KD. Qualitative Process Theory. *Artificial Intelligence.* 24: 85–168. 1984
43. Forbus KD. *Qualitative Representations: How People Reason and Learn about the Continuous World*. Cambridge MA: MIT Press. 2019

44. Horwitz RI, Hayes-Conroy A, Caricchio R, Singer BH. From Evidence Based Medicine to Medicine Based Evidence. *American Journal of Medicine.* 130(11): 1246–1250. 2017
45. Horwitz RI, Charlson ME, Singer BH. Medicine Based Evidence and Personalized Care of Patients. *European Journal of Clinical Investigation.* 48(7): e12945. 2018
46. Horwitz RI, Singer BH. Introduction. What works? And for Whom? *Social Science and Medicine.* (1982) 210: 22–25. 2018

Chapter 15
What Can Policies Do and How Can Social Science Help?

Norman M. Bradburn

Steve Feinberg dedicated his life to improving the use of data in public policy. He served on many National Academies of Science committees and panels that advocated the use of evidence both in formulating and evaluating government programs. He was a leader in the application of statistical methods to public policy issues and was a strong advocate of random controlled trials (RCTs) for evaluating policy effectiveness. His legacy is seen in the current widespread support, including legislation, for evidence-based policies.

In this chapter I would like to step back from technical considerations and reflect on the interplay between social science theory and research and the formulation and implementation of public policies. What can public policies realistically do about producing changes in society and what can the social sciences contribute to the formulation and successful implementation of these policies? These are extremely complex questions, and my reflections are molded by own experiences with these issues.

My career has been largely spent as an empirical social scientist involved partly in the evaluation of US governmental programs during the 1960s and 70s, and I began my career as a colleague of Steve's at the University of Chicago. This was a period in the United States of great experimentation with new policies to solve social problems. It was the time of the War on Poverty and the coming of age of program evaluation. Only a small portion of my career has been involved in policy

I wish to thank Jim Granato, Ian David Moss, Miron Straf and Judy Tanur for their insightful comments and help in preparing this chapter.

N. M. Bradburn (✉)
NORC at the University of Chicago, Chicago, IL, USA
e-mail: bradburn-norman@norc.org

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_15

making, but I am familiar with the implementation of many of the most important social policies in the United States over the past five decades.

15.1 Limitations of Social Policies

My experiences have led me to two main conclusions about public policies that I think are not sufficiently appreciated by policy makers or social scientists who do policy-related research. The first is that policies are formulated as essentially limited static instruments; that is, they assume a set of conditions and change one or two of the parameters of that set. Frequently the perception of the initial set of conditions is faulty due to lack of data or good analysis and may be dominated by views of a modal case, which might not, in fact, be reflective of the underlying distribution of conditions. Governmental policies are blunt instruments to bring about social change. They are designed to intervene in complex systems that are changing all the time. They almost never consider the dynamics put in motion by those changes. Thus, they inevitably suffer from unintended consequences. These unintended consequences are often large enough to nullify the positive effects of the policies or, even, to produce the opposite effect from that intended. Several examples that illustrate this process are discussed below.

Second, the implementation of governmental policies often falls far short of that envisioned by the designers of the policies either because they are underfunded or, in order to be successfully implemented, they require bureaucratic changes that do not occur or both. In the United States at least, the costs of policies are projected through economic models used by budget planners in the administration and in the Congress. These models often do not agree and may severely under- or overestimate the true cost of the programs, and larger budget considerations may affect the actual funding levels for programs. In program evaluation, the first, and too frequently the only thing we can ascertain, is whether the program ever got implemented at all, and, if it did, how closely it resembled its original design. In order to make these points more vivid, I will give two brief examples. I start with a small policy change implemented in the 1970s after the oil embargo. The problem was an acute shortage of oil, and a number of policies to deal with the situation were implemented quickly. One of them was changing to year-round daylight savings time (YRDST) in order to save electricity. This policy was enacted and the country stayed on DST in the fall rather than returning to standard time. The underlying logic was that making it light later in the morning in the winter and longer into the evening when presumably more people would be active, would decrease the demand for energy. It is not clear how well documented this assumption was, but such a policy had been implemented during World War II and presumably was effective.

The United States is a large country spanning 4 time zones. The variance in hours of daylight over the year is also sensitive to how far north one is, with much greater variance for the northern than for the southern parts of the country. In addition,

for economic reasons, some areas near the edge of a time zone opted years ago to join the time zone adjacent to them, so that they were in effect already on perpetual summer time. Demographic analysis easily showed that the conditions that gave rise to the original setting of the times had changed—the time zones were established when the country was largely agricultural—and that the present geographic distribution of the population and time distribution of work hours made it sensible to adjust the clocks by 1 hour throughout the year to increase the hours of daylight, on average, when the population was awake and needed them. Positive popular support for the policy was confirmed by public opinion surveys, although it was noted that there was a minority of the population that was adamantly opposed to the policy.

While the policy may have achieved what it was designed to do in the aggregate, several areas of the country were adversely affected by the policy for geographical reasons or, because they had already opted for the policy (but under a different name), the policy in their areas did not have the intended effect. In some areas in winter, the consequence of the new policy was that rural children who had to take school buses to school had to wait in the morning darkness for the bus to arrive. In one of the areas, southwestern Georgia, parents complained that the children were in danger of being hit by cars in the dark, and, a few weeks into the new policy, indeed a child waiting for a bus in the early morning was hit and killed by a car. This event was given great play by the mass media and support for YRDST evaporated almost overnight. YRDST was repealed and the country returned to the usual policy of daylight time and standard time the next spring.

This example illustrates my point about the bluntness of policy instruments and unintended consequences. While the policy may have produced the intended effect of saving energy and, for the vast majority of the population, did increase the useful hours of daylight in the winter, the policy could not, because of geographical reasons, and did not, because of historical reasons, produce the same effect for everyone. (As an aside, later analysis of data on traffic accidents showed that the very slight increase in deaths in the early morning was offset by a decrease in deaths from traffic accidents in the afternoon when it stayed light longer.)

My second example is one with which social science research has been heavily involved. This is the policy related to racial segregation of public schools in the United States. Social science research on the effects of racial segregation on individuals had been a potent force in the Supreme Court decision declaring the "separate but equal" doctrine unconstitutional and led to the dismantling of legally backed segregation of public facilities including schools. The implementation of this decision, which challenged long standing mores, habits, and social organization, particularly in the South, was openly resisted and, on occasion, required the use of force to implement. School integration has been one of the most contentious issues in the United States. Social science analysis had indicated that school integration was one of the most important instruments for long-term change in social relations between the races as well as for improving the economic condition of African-Americans.

Even after the end of legal segregation, the historical and economic effects of residential segregation meant that many schools, which had local attendance areas, continued to be de facto segregated. A famous and influential study by a team of sociologists led by James Coleman (1966) found that educational achievement was as much a consequence of familial conditions as it was of the schools. One of the most important findings was that in classrooms that were racially mixed (and in effect economically mixed), the African-American children did better than comparable children in segregated schools and the white children did not do any worse. This study was interpreted as a call for intervention by the government into the principle of local attendance areas and led in many cities to a court-ordered program of busing children out of their local attendance areas to schools in other areas in order to achieve racial balance and thus the presumed effects on achievement of racial integration.

This policy has proven to be one of the most politically sensitive policies pursued by the US government in the past few decades. It has been attacked and defended using social science analysis and empirical studies of the effects of busing. For the purposes of this discussion, I would like to point out a few features of the issue that illustrate my main points. First, the policy of busing students was a policy that changed one element of a complex system, one that arguably did not follow from the Coleman Report. The Coleman Report was based on comparison of segregated classrooms with classrooms that had been voluntarily or naturally integrated as a result of a large set of circumstances, many of which were not measured in the study. To generalize, the finding to conditions in which the integration was produced by external policies took a leap of faith and had to be based on social theories coming from more general social science analysis.

Second, the policy was static in that it failed to take into account the reaction of parents of many middle-class white families whose children or local schools were subject to the busing policy. The consequence was the acceleration of the movement of white families from the central cities to the suburbs, a phenomenon that was referred to as "white flight." Over a decade or so, there was a vast change in the racial composition of urban schools and greater segregation than before the policy was instituted because the proportion of white children in the cities declined markedly. While Coleman initially supported busing, he quickly saw the unintended consequences of the policy and became one of its leading opponents.

Third, the policy was also never fully or thoroughly implemented. For many years, busing was restricted to the schools in one school district, and there was no busing across district lines. Busing is costly and, in many cases, school districts lacked funds to implement the policy unless specifically forced to do so by suits brought by parents who supported busing. After the effects of "white flight" became clear so that it was impossible to achieve integration within a single district, courts began to order busing across district lines. This resulted in considerable litigation, but perhaps more importantly required longer bus rides for the children and support for the policy declined among parents of all races. Attention shifted to efforts to improve the schools regardless of their racial composition.

15.2 Limitations of Current Social Science Practice

Social science theories view society as a dynamic system with many interrelated parts, but empirical studies in the social sciences most often take only one or two parts of the system to study, leaving the impression that these parts can change without having implications for other parts of the system. When policy makers do look to the social sciences for guidance in formulating policies to bring about desired changes, they often will find studies that have focused on discrete elements in the system and have ignored the larger system implications of the changes. If they then take up these partial elements and make them the basis of policies, they are likely to discover other interrelated elements in the system that affect the outcome of the policies.

For example, if one takes a simplified look at the primary educational system, it is made up of schools that have principals with a degree of autonomy, but also may be organized into larger units like districts that have rules and a structure that constrains the autonomy of the individual school. Schools are organized in classrooms with individual teachers of varying degrees of experience and ability. Classrooms are of varying sizes and made up of students with varying backgrounds, motivations, and abilities. The system is constrained by a curriculum, resources, parental involvement (or lack of), physical structures, the socioeconomic characteristics of the families in the attendance area, and a host of other factors. The performance of the system is a product of these factors working together over time, and, while individual elements of the system may change, as when there is a new principal or a curricular change, the performance of the system as a whole is difficult to change more than incrementally.

As it is difficult to study all the elements in the entire system at once, it is natural to focus on one or two parts of the system to study what might be important leverage points to bring about desired changes in performance. For example, class size seems a likely constraint on learning, particularly for less advantaged or less able children, leading to the hypothesis that smaller class size would give teachers more time to devote to individual students and improve learning. Such a hypothesis led to the well-known Tennessee STAR experiment in class size (Ritter and Boruch 1999) which, indeed, did appear to show positive effects on learning.

But focusing on this one element of the system ignored the fact that class size in the Tennessee schools studied was embedded in a larger system that included the distribution of experienced teachers, student expectations, family backgrounds, and resources. When the results of the experiment were later implemented on a large scale in California, where a large number of new, relatively inexperienced teachers had to be hired to decrease class size, where there was a greater diversity of racial and linguistic minority student population and larger class sizes, the positive results were not replicated. Indeed, when the STAR experiment were applied to a larger program in Tennessee that focused first on poverty areas and later on the entire state, they failed to replicate the STAR results (von Hippel and Wagner 2018).

15.3 How Can Social Science Help Address These Issues?

First, it can build models of social systems that make clear the important elements of the system and their interrelationships. Too often one or two elements that are thought to be important drivers of the system are singled out for consideration as part of a policy change without making it clear how they function as part of the whole. Elements, such as class size, are then made the subject of policies that are implemented with little regard for the impact of the implementation on other parts of the system. This pattern is supported by the use of random controlled trials (RCTs) as the mode of evaluation for the effectiveness of the policy produced intervention. RCTs, by the logic of their design, treat the intervention as a discrete element to be contrasted with the control group that did not get the intervention. While the randomization of the assignment to the treatment and comparison groups controls statistically the other (static) elements of the system, it does not take into account the context of the experiment which may contain elements which are an essential condition for the experiment to succeed. It also does not consider the dynamics that may be put into play by the interventions that produce unintended consequences. The Tennessee class size experiment STAR was a well-designed and executed RCT, but failed to replicate when implemented at a larger scale where the conditions of the original RCT could not be reproduced. The larger point is that experiments must be replicated many times in different populations and conditions before the results can be relied upon. This is rarely done in testing policy programs; indeed, the literature on randomized controlled trials says little about replication. That omission implies to a lot of people that the randomization takes care of everything.

As is probably obvious from the examples I have chosen, I approach these issues from the perspective of a social systems theorist and fault applications of social science analysis and research that fail to think through the dynamics of social systems and to pursue research that enables us to model more completely the effects of policy changes. Theories help us understand the systems and guide us toward the relevant relations that might be most susceptible to policy interventions. As Ronald Coase (1994) has pointed out: "..., a theory is not like an airline or bus timetable. We are not interested simply in the accuracy of its predictions. A theory also serves as a base for thinking. It helps us to understand what is going on by enabling us to organize our thoughts."

I do not underestimate the difficulty of this task, but it is the direction that I think social sciences must go. This dynamic reorientation requires new theories, new tools, and new data. The development of economic theory and modeling has improved our understanding of the dynamics of economic systems and the probable effects of economic policies, although we obviously have a long way to go before we understand things fully. I believe that there are promising new techniques, such as stochastic modeling of social interactions and network analysis, which can lead to better understanding of the dynamics of social systems at both the micro and macro levels. We also have powerful new computational tools that enable us to build more

realistic models. The data to support these models, however, are often insufficient. I am less sure how much progress we can make on that front in the near future.

In sum, I see a reinvented social science as primarily concerned with formal theory building using dynamic models, closely coupled with empirical testing of the theories and requiring larger and better data bases to provide adequate data for testing those theories. Such a future also requires changes in our training of social scientists so that they will have the skills necessary to carry out this program. Whether we are up to the challenge remains to be seen. If Steve were here to help, I would be more optimistic.

References

Coase, R. H. 1994. "How Should Economists Choose?" In *Essays on Economics and Economists*, ed. R. H. Coase. Chicago: University of Chicago Press.

Coleman, James S., et al. 1966. "Equality of Educational Opportunity." Washington: U.S. Dept. of Health, Education, and Welfare, Office of Education.

Ritter, G. W., & Boruch, R. F. (1999). The Political and Institutional Origins of a Randomized Controlled Trial on Elementary School Class Size: Tennessee's Project STAR. Educational Evaluation and Policy Analysis, 21(2), 111–125. https://doi.org/10.3102/01623737021002111

von Hippel, Paul and Wagner, Chandi, (2018) Does a Successful Randomized Experiment Lead to Successful Policy? Project Challenge and What Happened in Tennessee After Project STAR (March 31, 2018). Available at SSRN: https://ssrn.com/abstract=3153503 or https://doi.org/10.2139/ssrn.3153503

Part V
Introduction: Surveys and Censuses

Judith M. Tanur

Steve gave credit for his initial interest in surveys to Bill Kruskal. Shortly after Steve went to the University of Chicago as an assistant professor, Bill started giving him clippings from the Chicago Sun Times about their straw polls for the 1968 election. As Bill had undoubtedly hoped, Steve became interested in the way the polling was done and the accuracy of the predictions. Steve's interest culminated in an article in JASA entitled "The Sun-Times Straw Poll, 1968 and 1970: A Statistical Appraisal" in which he partitioned the prediction error into components using an additive linear model. He then compared the straw poll results with simple persistence prediction and found that they were not much better. He suggested using previous data on the sampling units via regression adjustments, as was being done in election night national television forecasts. He cautions that "[t]he significant realignment of ward boundaries in Cook County following the census every ten years clearly limits the amount of prior information available." The implications of a census count were already part of Steve's world view.

And this interest in the straw poll led to his appearance on an early morning television program along with Norman Bradburn and Kenneth Prewitt where they explained the process of doing a survey—explaining sampling, putting a questionnaire together, interviewing, data processing, and how the results were interpreted and used.

Later Steve would spend a good deal of his career thinking and writing about surveys, both in terms of "hard" issues of sampling and analysis and of "softer" issues such as the cognitive issues involved in respondents' understanding and responding to questions. Some of these harder issues were addressed in the work he and I did jointly considering the parallels between experiments and surveys; some of the softer ones guided his initiation of and participation in the work of the Committee on National Statistics to sponsor the Advanced Research Seminar on Cognitive Aspects of Survey Methodology and its many follow-up activities. The chapter on Multiple Imputation for Nonignorable Item Nonresponse in Complex Surveys Using Auxiliary Margins by Olanrewaju Akande and Jerome Reiter addresses some of those technical issues.

Steve was concerned about the US Census for almost half a century. In preparing a talk to be delivered in celebration of his 70th birthday, I searched his 52-page cv for the word "census" and found it 82 times—and the cv was already 6 months old when I was working on it. These mentions included the following:

- 1 book *Who Counts: The Politics of Census-Taking in Contemporary America*, written with Margo J. Anderson and published in 1999
- 28 papers + 9 *Chance* articles
- 2 brief papers and editorials
- 4 technical discussions
- 1 book review
- 8 occasions of Congressional testimony
- And one PhD student Yi Ding, Department of Statistics, Carnegie Mellon University, "Capture-Recapture Census with Uncertain Matching," 1991

Steve was very much in favor of adjustment for the 1990 Census to correct the differential undercount and served in many capacities during that controversy. These included a CNSTAT panel, congressional testimony, technical original research on capture/recapture or dual system estimation, and a series of expository pieces in *Chance* about the legal battle over whether to adjust. These Chance articles, read in retrospect, give a blow-by-blow explanation of the complicated proceedings in which the statistical issues were overshadowed by constitutional ones, with the Supreme Court having the final say. One can almost tell the stories from the titles. Steve's account is a model of sorting out the legalities and explaining them to statisticians and lay people, not necessarily versed in the law. His own position only very faintly colors his reportage.

"An adjusted Census in 1990?" (in nine parts):

(a) "An adjusted Census in 1990?" *Chance*, **2** (No. 3), (1989), 23–25
(b) "An interim report," *Chance*, **3** (No. 1), (1990), 19–21
(c) "Back to court again," *Chance*, **3** (No. 2), (1990), 32–35
(d) "The judge rules and the PES begins," *Chance*, **3** (No. 3), (1990), 33–36
(e) "Commerce says 'no'," *Chance*, **4** (No. 3), (1991), 44–52
(f) "A full-scale judicial review approaches," *Chance*, **4** (No. 4), (1991), 22–24, 29
(g) "The trial," *Chance*, **5** (No. 3–4), (1992), 28–38
(h) "Trial judgment set aside," *Chance*, 7, (No. 4), (1994), 31–32
(i) "The Supreme Court decides," *Chance*, **9**, (No. 2) (1996), 4–9 (with M. Anderson)

Controversies about the Census continue and two former Census Directors discuss the issues faced by two recent Censuses. First John Thompson contributes "Insights into the Decision on Whether to Statistically Adjust the 2000 Census for Coverage Errors," and then Kenneth Prewitt in "A Sensible Census" addresses issues raised by the 2020 Census and likely to be relevant to future Censuses as well.

Chapter 16
Multiple Imputation for Nonignorable Item Nonresponse in Complex Surveys Using Auxiliary Margins

Olanrewaju Akande and Jerome P. Reiter

16.1 Introduction

Many surveys suffer from item nonresponse that may be nonignorable. This can complicate analysis or dissemination of survey data. In some settings, we can leverage auxiliary information from other data sources to help adjust for the effects of nonignorable nonresponse. For example, suppose that in a simple random sample, a question on sex suffers from item nonresponse, so that 70% of the respondents are women. Suppose we know that the target population includes 50% of men and 50% of women. This implies that respondents with missing values of sex are more likely to be men than women. Thus, if we impute values for the missing sexes, we should impute more "male" than "female."

Generalizing this example, we desire to leverage reliable estimates of low-dimensional margins for variables with item nonresponse—available, for example, from high-quality surveys or administrative databases—when imputing missing items. However, we do not want to use solely these population margins to inform the imputations. We should also take advantage of observed information in other variables, so as to preserve multivariate relationships as best as possible. In the case where the data are from a complex survey, we also need to somehow account for the survey design weights in the imputations (Reiter et al. 2006, Zhou et al. 2016). We are not aware of any principled ways to do all this simultaneously when performing multiple imputation for item nonresponse.

O. Akande
Social Science Research Institute, Duke University, Durham, NC, USA
e-mail: olanrewaju.akande@duke.edu

J. P. Reiter (✉)
Department of Statistical Science, Duke University, Durham, NC, USA
e-mail: jreiter@duke.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_16

In this chapter, we propose a framework for multiple imputation of missing items in complex surveys that leverages auxiliary margins. Our approach is to use the auxiliary margins to identify additive nonignorable (AN) models (Hirano et al. 1998; 2001), with an additional requirement that the completed datasets result in plausible design-based estimates of the known margins. We do so by fusing the AN model with large sample results under frequentist (survey-weighted) paradigms. In this way, we ensure that imputations are influenced by relationships in the data and the auxiliary information, while being faithful to the survey design through survey weights.

Our work connects to several areas of research in which Stephen Fienberg made key contributions. In particular, the methods are examples of using marginal information (Chen and Fienberg 1976, Fienberg 1970), and of course handling missing values (Bishop and Fienberg 1969, Chen and Fienberg 1974, Fienberg 1972), in the analysis of contingency tables. Our approach also uses Bayesian techniques for official statistics and survey sampling, a perspective that he championed for many areas including disclosure limitation, record linkage, and the analysis of categorical data.

The remainder of this chapter is organized as follows. In Sect. 16.2, we review the AN model. In Sect. 16.3, we present our approach. In Sect. 16.4, we illustrate the performance of the approach using simulation studies with stratified sampling. In Sect. 16.5, we conclude and discuss possible extensions. For clarity, we present the methodology for data that does not have unit nonrespondents. We discuss extensions to scenarios including unit nonrespondents in Sect. 16.5.

16.2 Review of the AN Model

Our review of the AN model closely follows the review in Akande (2019, Chapter 4). For additional discussion of the AN model, see Bhattacharya (2008), Das et al. (2013), Deng et al. (2013), Nevo (2003), Sadinle and Reiter (2019), Schifeling et al. (2015), Si et al. (2015). Although the AN model was developed originally for handling nonignorable attrition in longitudinal studies with refreshment samples (Deng et al. 2013, Hirano et al. 1998), it can be applied to our setting by viewing the data from the refreshment samples as auxiliary information, as we now describe.

Notation

Let $\mathcal{D}$ comprise data from the survey of $i = 1, \ldots, n$ individuals, and $\mathcal{A}$ comprise data from the auxiliary database. Let $X = (X_1, \ldots, X_p)$ represent the p variables in both $\mathcal{A}$ and $\mathcal{D}$, where each $X_k = (X_{1k}, \ldots, X_{nk})^T$ for $k = 1, \ldots, p$. Let $Y = (Y_1, \ldots, Y_q)$ represent the q variables in $\mathcal{D}$ but not in $\mathcal{A}$, where each $Y_k = (Y_{1k}, \ldots, Y_{nk})^T$ for $k = 1, \ldots, q$. We assume that $\mathcal{A}$ only contains sets of

marginal distributions for variables in X, summarized from some external database and measured with negligible error. We disregard variables in $\mathcal{A}$ but not $\mathcal{D}$, as the margins for these variables generally do not provide much information about the missing values in $\mathcal{D}$.

We also introduce variables to account for item nonresponse. For each $k = 1, \ldots, p$, let $R_k^x = (R_{1k}^x, \ldots, R_{nk}^x)^T$, where each $R_{ik}^x = 1$ if individual i would not respond to the question on X_k in $\mathcal{D}$, and $R_{ik}^x = 0$ otherwise. Similarly, for each $k = 1, \ldots, q$, let $R_k^y = (R_{1k}^y, \ldots, R_{nk}^y)^T$, where each $R_{ik}^y = 1$ if individual i would not respond to the question on Y_k in $\mathcal{D}$ and $R_{ik}^y = 0$ otherwise.

Finally, for simplicity, we use generic notations such as f and η for technically different functions and parameters, respectively, although their actual meanings should be clear within each context. For example, f, η_0, and η_1 need not be the same in the conditional probability mass functions $\Pr(X_1 = 1|Y_1) = f(\eta_0 + \eta_1 Y_1)$ and $\Pr(Y_1 = 1|X_1) = f(\eta_0 + \eta_1 X_1)$.

AN Model Specification

To make the AN model specification easy to follow, we work with an example where $\mathcal{D}$ comprises only two binary variables, X_1 and Y_1. Following our notation, $\mathcal{A}$ contains the auxiliary marginal distribution for X_1 but no auxiliary marginal distribution for Y_1. For simplicity, we also suppose X_1 suffers from item nonresponse but Y_1 is fully observed. Thus, we need a model for R_1^x, the fully observed vector of item nonresponse indicators for X_1. We assume that we do not need to include a model for R_1^y, since there is no nonresponse in Y_1. The observed and auxiliary data take the form shown in Table 16.1a. The incomplete contingency table representing the joint distribution of (X_1, Y_1, R_1^x), with observed and auxiliary marginal probabilities excluded, is shown in Table 16.1b.

Due to the empty cells in the contingency table in Table 16.1b, we cannot fit a fully saturated model to these data. To see this, we use a pattern mixture model factorization (Glynn et al. 1986, Little 1993) to characterize the joint distribution of (X_1, Y_1, R_1^x). The factorization, which we write as

$$\begin{aligned}\Pr(X_1 = x, Y_1 = y, R_1^x = r) = \Pr(X_1 = x|Y_1 = y, R_1^x = r) \\ \times \Pr(Y_1 = y|R_1^x = r) \Pr(R_1^x = r),\end{aligned} \tag{16.1}$$

can be fully parameterized using seven parameters: the four values of $\theta_{yr} = \Pr(X_1 = 1|Y_1 = y, R_1^x = r)$, $\pi_r = \Pr(Y_1 = 1|R_1^x = r)$, and $q = \Pr(R_1^x = 1)$.

Five of the seven parameters, that is, q, π_0, π_1, θ_{00}, and θ_{10}, can be directly estimated from the observed data alone, as long as the sample data is representative of the target population. Unfortunately, the observed data contain no information about θ_{01} and θ_{11}. We need to make assumptions about the missingness mechanism to estimate the full joint distribution. For example, we could set $\theta_{01} = \theta_{00}$ and $\theta_{11} = \theta_{10}$, resulting in a missing at random (MAR) mechanism.

Table 16.1 Setup for the AN model with two binary variables Y_1 and X_1. Y_1 is fully observed and X_1 suffers from item nonresponse. We know the population margin for X_1. Here, "✓" represents observed components and "**?**" represents missing components

(a) Data

	X_1	Y_1	R_1^x
Original data	✓	✓	0
	?		1
Auxiliary margin →	✓	**?**	**?**

(b) Contingency table

	$R_1^x = 0$		$R_1^x = 1$	
	$X_1 = 0$	$X_1 = 1$	$X_1 = 0$	$X_1 = 1$
$Y_1 = 0$	✓	✓	**?**	**?**
$Y_1 = 1$	✓	✓	**?**	**?**

The auxiliary marginal distributions provide information that we can use to specify such identifying assumptions. In our two-variable example, the auxiliary marginal distribution of X_1 provides one linear constraint about θ_{01} and θ_{11}. We write this constraint as

$$\Pr(X_1 = 1) - \Pr(X_1 = 1, Y_1 = y, R_1^x = 0) = q\,[\theta_{01}(1 - \pi_1) + \theta_{11}\pi_1]\,. \tag{16.2}$$

Although (16.2) does not provide enough information to identify both θ_{01} and θ_{11}, it does increase the number of estimable parameters from five to six.

The AN model takes advantage of this additional constraint. In particular, the AN model assumes that the reason for item nonresponse in X_1 depends on X_1 and Y_1 through a function that is additive in X_1 and Y_1. We have

$$(X_1, Y_1) \sim f(X_1, Y_1|\Theta) \tag{16.3}$$

$$\Pr(R_1^x = 1|X_1, Y_1) = h(\eta_0 + \eta_1 X_1 + \eta_2 Y_1), \tag{16.4}$$

where Θ, η_0, η_1, and η_2 represent the parameters in f and h. Here, $h(a)$ should be a strictly increasing function satisfying $\lim_{a\to-\infty} h(a) = 0$ and $\lim_{a\to\infty} h(a) = 1$. The models in (16.3) and (16.4) represent a selection model factorization (Little 1995) of the joint distribution of (Y_1, X_1, R_1^x), instead of the pattern mixture factorization in (16.1). Hirano et al. (2001) prove that the AN model is likelihood-identified for general distributions, such as probit and logistic regression models. The interaction term between X_1 and Y_1 is not allowed, as additivity is necessary to enable identification of the model parameters.

The AN model is appealing in that it includes ignorable and nonignorable models as special cases. For example, $(\eta_1 = 0, \eta_2 = 0)$ results in a missing completely at random (MCAR) mechanism, $(\eta_1 \neq 0, \eta_2 = 0)$ results in a MAR mechanism, and $\eta_2 \neq 0$ results in a missing not at random (MNAR) mechanism. In particular, $(\eta_1 = 0, \eta_2 \neq 0)$ results in the nonignorable model of Hausman and Wise (1979). This allows the data to determine an appropriate mechanism from among these possibilities. The AN model does rely on the assumption of additivity of the response model in X_1 and Y_1, which may be reasonable in practice. Deng et al. (2013) describe sensitivity analysis for non-zero interaction effects. Hirano et al. (2001) suggest results are not overly sensitive to the choice of h.

It is possible to use mechanisms other than the AN model to estimate up to six unique parameters (in our example here). For example, one can set either θ_{01} or θ_{11} equal to zero. Setting $\theta_{01} = 0$ but $\theta_{11} \neq 0$ implies that all nonrespondents cannot have $X_1 = 1$ whenever $Y_1 = 0$. On the other hand, setting $\theta_{11} = 0$ but $\theta_{01} \neq 0$ implies that all nonrespondents cannot have $X_1 = 1$ whenever $Y_1 = 1$. Both assumptions seem more restrictive than setting an interaction effect in the model for R_1^x to zero, and we do not recommend adopting them unless the specific application at hand justifies such strong assumptions. As another example, one can set $\theta_{01} = \theta_{11} + b$ for some constant b. With $b = 0$, this equates to $\theta^\star = \theta_{01} = \theta_{11}$, which then simplifies (16.2) to

$$\theta^\star = \frac{\Pr(X_1 = 1) - (1 - q)\left[\theta_{00}(1 - \pi_0) + \theta_{10}\pi_0\right]}{q}. \tag{16.5}$$

This constraint implies conditional independence between Y_1 and X_1 for nonrespondents. This seems a strong assumption in general.

These two options, as well as other mechanisms that we do not cover here, are seemingly more restrictive than the AN assumptions or do not maximize all available information. The AN model does not force analysts to make as many untestable assumptions as most of the other mechanisms do, while allowing analysts to estimate as many parameters as possible with auxiliary data. However, the AN model as developed by Hirano et al. (1998; 2001) does not incorporate complex survey designs directly. We now extend the model to do so.

16.3 Extending the AN Model to Account for Complex Surveys

Let N represent the number of units in the population from which the n survey units in $\mathcal{D}$ are sampled. Let $W = (w_1, \ldots, w_n)$, where each w_i is the base weight for the ith unit in the sample $\mathcal{D}$. Here, we let $w_i = 1/\pi_i$, where π_i is the probability of selection of the ith unit. We present methods where weights are not subject to calibration or nonresponse adjustments, although one could use the approach for adjusted weights as well. Let the superscript "pop" represent the population counterparts of the survey variables. For example, X^{pop} and Y^{pop} represent the population-based counterparts of X and Y, respectively, where each $X_i \in X^{pop}$ and $Y_i \in Y^{pop}$. We do not observe values of X^{pop} or Y^{pop} for all non-sampled units in the population.

To present the methodology, we continue to work with the two-variable example in Sect. 16.2, with one minor modification. We now let Y_1 be a categorical variable with three levels, that is, $Y_1 \in \{1, 2, 3\}$. We do so to show that our approach can extend to non-binary variables. The data and incomplete contingency table take similar forms to Table 16.1, with weights now included and Y_1 having three levels.

Following our discussions in Sect. 16.2, we once again cannot fit a fully saturated model to the data. However, we can uniquely estimate seven of the nine parameters in a fully saturated model. Without any auxiliary information, we can fit the following model to the observed data as a default option within the missing data literature. We have

$$Y_1 \sim f(\theta) \tag{16.6}$$

$$\Pr(X_1 = 1|Y_1) = g(\alpha_0 + \alpha_{1j} 1[Y_1 = j]) \tag{16.7}$$

$$\Pr(R_1^x = 1|X_1, Y_1) = h(\gamma_0 + \gamma_{1j} 1[Y_1 = j]), \tag{16.8}$$

resulting in a MAR mechanism, where $j = 1, 2, 3$. We set $\alpha_{11} = 0$ and $\gamma_{11} = 0$ to ensure the model is identifiable; the model then only contains seven parameters as desired. For more flexibility however, we seek to fit a nonignorable model that includes $\gamma_2 X_1$ in (16.8), so that (16.8) becomes the AN model

$$\Pr(R_1^x = 1|X_1, Y_1) = h(\gamma_0 + \gamma_{1j} 1[Y_1 = j] + \gamma_2 X_1). \tag{16.9}$$

To do so, we need to incorporate at least one constraint on the remaining parameters. When the survey design is complex, it may not be sufficient to use the auxiliary margin to force an extra constraint on the remaining parameters as we did in Sect. 16.2, since that approach does not incorporate the survey weights directly. To account for the survey weights, we take a different approach.

In practice, the most common marginal information is the population total (or mean) of some of the variables. For example, for totals, we know that

$$T_X = \sum_{i=1}^{N} X_{i1}^{pop} = N \times \Pr(X_1^{pop} = 1), \tag{16.10}$$

where $\Pr(X_1^{pop} = 1)$ is the true auxiliary marginal probability. A classical design-unbiased estimator of T_X in this case is the Horvitz–Thompson estimator (Horvitz and Thompson 1952), henceforth referred to as HT estimator, which is

$$\hat{T}_X = \sum_{i \in \mathcal{D}} \frac{X_{i1}}{\pi_i} = \sum_{i \in \mathcal{D}} w_i X_{i1}. \tag{16.11}$$

In large enough samples, finite population central limit theorems ensure that $\hat{T}_X$ is approximately normally distributed around T_X, with a variance V_X that is estimated using design-based principles (Fuller 2009). Thus, for fully observed data, we have

$$\sum_{i \in \mathcal{D}} w_i X_{i1} \sim N(T_X, V_X). \tag{16.12}$$

When the data contain nonresponse, we cannot compute $\hat{T}_X$ directly. However, it is reasonable to expect this unobserved value of $\hat{T}_X$ to be distributed around T_X as governed by (16.12). Thus, when we impute the missing values for X_1, it is reasonable to require any completed dataset to produce a value of $\hat{T}_X$ that is plausible under (16.12) as well. We operationalize this logic as follows. For all $i \in \mathcal{D}$, let $X_{i1}^{\star} = X_{i1}$ when $R_{i1}^{x} = 0$, and let $X_{i1}^{\star}$ be an imputed value when $R_{i1}^{x} = 1$. We impose the probabilistic constraint,

$$\sum_{i \in \mathcal{D}} w_i X_{i1}^{\star} \sim N(T_X, V_X). \tag{16.13}$$

In this way, we favor imputations consistent with (16.13) when generating imputed values for X under the posterior predictive distribution implied by (16.6), (16.7), and (16.9). Using a probabilistic constraint, as opposed to a deterministic constraint that $\hat{T}_X$ be as close to T_X as possible, reflects uncertainty about $\hat{T}_X$ more appropriately. Here, we assume V_X is pre-specified and treated as known; for example, it could be based on previous knowledge or an average of estimates from preliminary sets of completed data. We discuss considerations with unknown V_X further in Sect. 16.5.

We incorporate (16.13) into a Markov chain Monte Carlo (MCMC) sampler for the model parameters through a Metropolis algorithm. At each MCMC iteration t, let the current draw of each $X_{i1}^{\star}$ be $X_{i1}^{\star(t)}$ and let $\hat{T}_X^{\star(t)} = \sum_{i \in \mathcal{D}} w_i X_{i1}^{\star(t)}$. We use the following sampler at iteration $t + 1$.

S1. For all $i \in \mathcal{D}$, i.e., $i = 1, \ldots, n$, set $X_{i1}^{\star} = X_{i1}$ when $R_{i1}^{x} = 0$. When $R_{i1}^{x} = 1$, generate a candidate $X_{i1}^{\star}$ for the missing X_{i1} from the following posterior predictive distribution implied by (16.7) and (16.9). We have

$$\Pr(X_{i1}^{\star} = 1 | \ldots) \propto g(\alpha_0 + \alpha_{1j} 1[Y_{i1} = j])\, h(\gamma_0 + \gamma_{1j} 1[Y_{i1} = j] + \gamma_2 X_{i1}^{\star}), \tag{16.14}$$

using the current posterior draws of the parameters at iteration $t + 1$, where "..." represents conditioning on all other variables and posterior draws of all parameters in the model.

S2. Let $\hat{T}_X^{\star} = \sum_{i \in \mathcal{D}} w_i X_{i1}^{\star}$. Calculate the acceptance ratio,

$$p = \frac{N(\hat{T}_X^{\star}; T_X, V_X)}{N(\hat{T}_X^{\star(t)}; T_X, V_X)}. \tag{16.15}$$

S3. Draw a value u from $u \sim Unif(0, 1)$. If $u \leq p$, accept the proposed candidate $(X_{i1}^{\star}, \ldots, X_{in}^{\star})$, and set $X_{i1}^{\star(t+1)} = X_{i1}^{\star}$ for $i = 1, \ldots, n$. Otherwise, reject the proposed candidate, and set $X_{i1}^{\star(t+1)} = X_{i1}^{\star(t)}$ for $i = 1, \ldots, n$.

Intuitively, these steps reject completed datasets that yield highly improbable design-based estimates of T_X, while simultaneously allowing us to estimate $\gamma_2 X_1$

in (16.9). Although (16.13) provides a stochastic constraint, whereas using the auxiliary margins as in Sect. 16.2 forms linear constraints, $\gamma_2 X_1$ is still estimable when using (16.13), as we show using the simulations in Sect. 16.4.

We recommend that analysts monitor the acceptance ratio of the missing data sampler in Steps S1–S3, as with any Metropolis sampler. In cases where the acceptance ratio is considerably low, analysts can inflate or tune V_X or consider other methods of generating more realistic imputations from the implied posterior predictive distribution. In our simulation scenarios in Sect. 16.4, there is no need to do so as the samplers mix adequately. We do not worry about cases where the acceptance ratio is high because we view (16.13) as a constraint rather than a target distribution. Therefore, we interpret a high acceptance ratio as the sampler doing a good job of generating imputations that respect the survey design, as desired.

16.4 Simulations with Stratified Sampling

In this section, we illustrate the approach described in Sect. 16.3 via simulation studies with stratified sampling. We create ten populations, each of size $N = 50{,}000$ split into two strata: 70% of the units are in stratum 1 ($N_1 = 35{,}000$), and 30% of the units are in stratum 2 ($N_2 = 15{,}000$). For each observation in each population, we generate values of a three-valued Y_1 and binary X_1 using

$$Y_{i1} \sim \text{Discrete}(\theta_1, \theta_2, \theta_3) \tag{16.16}$$

$$X_{i1}|Y_{i1} \sim \text{Bernoulli}(\pi_{X_{i1}}); \quad \Phi^{-1}(\pi_{X_{i1}}) = \alpha_0 + \alpha_{1j} 1[Y_{i1} = j], \tag{16.17}$$

for $j \in \{2, 3\}$, where $\pi_{X_{i1}} = \Pr[X_{i1} = 1|Y_{i1}]$. Here, the discrete distribution refers to the multinomial distribution with sample size equal to one, and Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution. We set $\theta = (\theta_1, \theta_2, \theta_3) = (0.5, 0.15, 0.35)$ in stratum 1 and $\theta = (0.1, 0.45, 0.45)$ in stratum 2. This ensures that the joint distributions of Y_1 and X_1 differ across strata. We set different values for α_0, α_{12}, and α_{13} to explore how the strength of the relationship between X_1 and Y_1 affects results.

For each of the ten simulation runs, we randomly select $n = 5000$ observations from the corresponding population using stratified simple random sampling. We sample $n_1 = 1500$ units from stratum 1 and $n_2 = 3500$ units from stratum 2. This disproportionate sampling allocation ensures that the base weights matter in the estimation of finite population quantities. The survey weights $w_i = N_1/n_1 = 35000/1500 = 23.33$ for all units in stratum 1 and $w_i = N_2/n_2 = 15000/3500 = 4.29$ for all units in stratum 2.

We introduce item nonresponse in X_1 for each of the simulation runs by generating missingness indicators from an AN model. For each $i \in \mathcal{D}$ in each

population, we sample the missingness indicator from a Bernoulli distribution with probability

$$\Phi^{-1}(\Pr[R_{i1}^{x} = 1|Y_{i1}, X_{i1}]) = \gamma_0 + \gamma_{1j}1[Y_{i1} = j] + \gamma_2 X_{i1}, \tag{16.18}$$

where $j \in \{2, 3\}$. We set different values for γ_0, γ_{12}, γ_{13}, and γ_2 to investigate how departures from an ignorable missing mechanism affect the performance of the imputation strategies. All sets result in approximately 30% missing values in X_1.

After making the missing values, we use several approaches to impute the item nonresponse in X_1. For each approach, we use (16.16) and (16.17) as the models for the survey variables. We use different methods for specifying and estimating the selection model, in particular for incorporating the weights and auxiliary information. The approaches include the following.

1. MAR+Weight: We incorporate the survey weights by including w_i as a covariate in (16.17). Since there is a one-to-one mapping between weights and strata in our simulation setup, we incorporate w_i by adding an indicator variable S_i for strata, so that we have

$$X_{i1}|Y_{i1} \sim \text{Bernoulli}(\pi_{X_{i1}}); \quad \Phi^{-1}(\pi_{X_{i1}}) = \alpha_0 + \alpha_{1j}1[Y_{i1} = j] + \alpha_2 1[S_i = 2] \tag{16.19}$$

   as the model for X_1 instead of (16.17). We exclude the parameter for $1[S_i = 1]$ in (16.19) to ensure identifiability. Additionally, since $\gamma_2 X_{i1}$ in (16.18) cannot be identified from the observed data alone, we exclude $\gamma_2 X_{i1}$ in (16.18), so that we have

$$\Phi^{-1}(\Pr[R_{i1}^{x} = 1|Y_{i1}, X_{i1}]) = \gamma_0 + \gamma_{1j}1[Y_{i1} = j]. \tag{16.20}$$

   This is a MAR model for the item nonresponse. This approach represents a default approach analysts might use in this scenario. It does not use auxiliary information about the margin of X_1.
2. AN+Weight: We use (16.19) to incorporate the weights and fit the AN model in (16.18). However, we do so without using any auxiliary information. Although $\gamma_2 X_{i1}$ in (16.18) is not identifiable as previously discussed, the model can be estimated (albeit not accurately) under the Bayesian paradigm because of the prior distribution. This represents a naive application of a nonignorable modeling strategy.
3. AN+Constraint: We fit the AN model in (16.18), using the method in Sect. 16.3 to incorporate the auxiliary information and survey design. We incorporate the auxiliary total T_{X_1} and survey weights through the constraint in (16.13). We set V_X equal to approximately the theoretical variance of $\hat{T}_X$ without any missing values.

4. AN+Constraint+Weight: We combine the AN+Weight and AN+Constraint approaches. Specifically, we follow the AN+Constraint method but use (16.19) instead of (16.17) to further control for the weights.

We use non-informative priors for all parameters. Specifically, we use the Dirichlet(1, 1, 1) distribution as the prior distribution for $(\theta_1, \theta_2, \theta_3)$ and a standard multivariate normal distribution as the prior distribution for the set of parameters in each probit model in (16.17)–(16.20). We fit all models using MCMC sampling. We run each MCMC sampler for 10,000 iterations, discarding the first 5000 as burn-in, resulting in 5000 posterior samples. We create $L = 50$ multiply-imputed datasets, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(50)})$, from every 100th posterior sample. From each completed dataset $\mathbf{Z}^{(l)}$, we compute the design-based estimates of T_X, α_0, α_{12}, and α_{13}, along with the corresponding standard errors, using the survey-weighted generalized linear models option in the R package, "survey." Although there are differing opinions associated with using survey weights in regression modeling (Gelman 2007, Pfeffermann 1993), we use them to ensure all analyses account for the selection effects in the survey design. We also compute estimates of γ_0, γ_1, and γ_2 (which do not depend on the weights by design), along with the corresponding standard errors, from each completed dataset, using the generalized linear models option in the R package, "stats."

Within any simulation run, we combine all the estimates across all multiply-imputed datasets using multiple imputation (MI) rules (Rubin 1987). As a brief review of MI, let q be the point estimator of some estimand of interest Q in a completed dataset, and let u be the estimator of its variance. For $l = 1, \ldots, L$, let q_l and u_l be the values of q and u in completed dataset $\mathbf{Z}^{(l)}$. The MI point estimate of Q is $\bar{q}_L = \sum_{l=1}^{L} q_l/L$, and the corresponding MI estimate of the variance of $\bar{q}_L$ is given by $T_L = (1 + 1/L)b_L + \bar{u}_L$, where $b_L = \sum_{l=1}^{L}(q_l - \bar{q}_L)^2/(L-1)$ and $\bar{u}_L = \sum_{l=1}^{L} u_l/L$. We write $\bar{q}_L^m$ and T_L^m to represent the values of $\bar{q}_L$ and T_L in the simulation run indexed by m, where $m = 1, \ldots, 10$.

We consider eight simulation scenarios resulting from a $2 \times 2 \times 2$ factorial design. The factors include strong and weak associations among X_1 and Y_1; large and small departures from ignorable missingness mechanisms; and, margins for X_1 known either for the entire population only (T_X) or for each of the two strata. In the interest of space, we report detailed results only for the four scenarios described in Table 16.2. In each scenario, we report averages of MI estimates across the 10 runs, including $\sum_{m=1}^{10} \bar{q}_L^m/10$ for the point estimate of each estimand Q, and

Table 16.2 Simulation scenarios presented in Sect. 16.4

Scenario	Association (X_1, Y_1)	Departure from ignorable missingness	Margins
1	Strong	Large	Population only
2	Weak	Small	Population only
3	Strong	Large	Both strata
4	Weak	Small	Both strata

$\sqrt{\sum_{m=1}^{10} T_L^m/10}$ as a measure of the corresponding standard error. For comparison, we also report results before introduction of missing data, using the average of the ten point estimates and the square root of the average of the variance estimates.

Results for Scenario 1 and Scenario 2

In scenario 1, we set $\alpha_0 = 0.5$, $(\alpha_{12}, \alpha_{13}) = (-0.5, -1)$, $\gamma_0 = -0.25$, $(\gamma_{12}, \gamma_{13}) = (0.1, 0.3)$, and $\gamma_2 = -1.1$. This represents a strong relationship between Y_1 and X_1 and a nonresponse mechanism that deviates substantially from an ignorable mechanism. Here, T_X is known only for the entire population and not for the individual strata.

For each method, Table 16.3a displays the average of the ten HT estimates for T_X and the square root of the average of the variances of these estimates in scenario 1. AN+Constraint and AN+Constraint+Weight offer the most accurate estimates, whereas AN+Weight and MAR+Weight offer the least accurate estimates. Controlling for the weights in the model for X_1 as in the AN+Constraint+Weight method apparently decreases the standard error in comparison to AN+Constraint. It also increases the acceptance ratios in the MCMC samplers. The standard error associated with AN+Weight is much higher than all other methods. This is due primarily to the weak identification issues associated with using the AN model without any auxiliary information, resulting in greater uncertainty from the nonresponse mechanism.

Table 16.3b also shows survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13}, and γ_2, along with the corresponding standard errors, again combined across all ten simulation runs. Here, both AN+Constraint and AN+Constraint+Weight give nearly identical results and closely estimate the true parameter estimates. The AN+Constraint and AN+Constraint+Weight approaches outperform the other choices in this scenario. AN+Weight and MAR+Weight again give the least accurate results.

In scenario 2, we weaken both the relationship between the variables of interest and the nonignorable nonresponse. We set $\alpha_0 = 0.15$ and $(\alpha_{12}, \alpha_{13}) = (-0.45, -0.15)$ to reflect a weak relationship between Y_1 and X_1, and we set $\gamma_0 = -1$, $(\gamma_{12}, \gamma_{13}) = (-0.6, 1.4)$, and $\gamma_2 = -0.2$ to reflect a small departure from an ignorable nonresponse mechanism. T_X is known only for the entire population.

Table 16.4a,b presents results of 10 simulation runs of scenario 2. Once again, the AN+Constraint and AN+Constraint+Weight outperform the other methods. AN+Constraint has a slightly smaller standard error for T_X in scenario 2 than AN+Constraint+Weight. Also, MAR+Weight performs much better in scenario 2 than in scenario 1. In the presence of a weakly nonignorable nonresponse mechanism, there appears to be little degradation when using a MAR model. In addition, whatever degradation or bias that should have been attributed to the survey design appears to be taken care of by including the strata indicator in the model

Table 16.3 Results for scenario 1: overall auxiliary margin for X_1, strong relationship between Y_1 and X_1, and strong nonignorable nonresponse

(a) HT estimates for T_X under each method, the corresponding standard errors, and acceptance ratios. "Population" is the value of T_X in the population of $N = 50{,}000$ individuals. "No Missing Data" is the weighted estimate based on the sampled data before introducing item nonresponse. For AN+Constraint and AN+Constraint+Weight, the estimated Monte Carlo standard errors of $\sum_{m=1}^{10} \bar{q}_L^m/10$ are less than 150, ruling out chance error as explanation for the improved performance of these two models over AN+Weight and MAR+Weight.

	T_X		Acceptance Ratio	
Method	Mean	SE	Mean	Range
Population	25,026	–	–	–
Mo Missing Data	25,275	582	–	–
MAR+Weight	30,579	670	–	–
AN+Weight	28,222	2789	–	–
AN+Constraint	24,993	741	0.82	[0.79, 0.84]
AN+Constraint+Weight	25,019	718	0.83	[0.80, 0.86]

(b) Survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13}, and γ_2, along with the corresponding standard errors. "MAR+W" is MAR+Weight, "AN+W" is AN+Weights, "AN+C" is AN+Constraint, and "AN+C+W" is "AN+Constraint+Weight." Standard errors of the averaged point estimates are small enough to rule out chance error as explanations for the improved performance of AN+C and AN+C+W over MAR+W and AN+W.

		MAR+W		AN+W		AN+C		AN+C+W	
Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE
α_0	0.50	0.74	0.05	0.63	0.13	0.49	0.05	0.49	0.05
α_{12}	−0.50	−0.45	0.07	−0.47	0.07	−0.49	0.07	−0.49	0.06
α_{13}	−1.00	−0.88	0.07	−0.92	0.10	−0.98	0.06	−0.98	0.06
γ_0	−0.25	−0.88	0.04	−0.63	0.35	−0.22	0.07	−0.23	0.07
γ_{12}	0.10	0.29	0.05	0.21	0.11	0.10	0.06	0.10	0.06
γ_{13}	0.30	0.63	0.05	0.48	0.17	0.27	0.07	0.27	0.07
γ_2	−1.10	–	–	−0.48	0.57	−1.15	0.14	−1.15	0.13

for X_1. AN+Weight performs worse than the other three methods. Unlike before, AN+Weight actually underestimates rather than overestimates T_X in this scenario. Overall, the range of acceptance ratios has decreased slightly from the previous scenario.

We note that we find similar overall conclusions in the two other scenarios where we know the margin of T_X only for the whole population.

Results for Scenario 3 and Scenario 4

We next investigate the performance of the approaches when we know the auxiliary margin of X_1 in each stratum. In this case, it is possible to implement the constraint in (16.13) for each stratum. For each stratum $s \in \{1, 2\}$, we require that

Table 16.4 Results for scenario 2: overall auxiliary margin for X_1, weak relationship between Y_1 and X_1, and weak nonignorable nonresponse

(a) HT estimates for T_X under each method, the corresponding standard errors, and acceptance ratios. "Population" is the value of T_X in the population of $N = 50{,}000$ individuals. "No Missing Data" is the weighted estimate based on the sampled data before introducing item nonresponse. For AN+Constraint and AN+Constraint+Weight, the estimated Monte Carlo standard errors of $\sum_{m=1}^{10} \bar{q}_L^m/10$ are less than 150, ruling out chance error as explanation for the improved performance of these two models over AN+Weight and MAR+Weight.

	T_X		Acceptance Ratio	
Method	Mean	SE	Mean	Range
Population	24,677	–	–	–
Mo Missing Data	24,742	570	–	–
MAR+Weight	26,098	662	–	–
AN+Weight	23,705	2519	–	–
AN+Constraint	24,666	698	0.79	[0.77, 0.81]
AN+Constraint+Weight	24,653	705	0.79	[0.77, 0.81]

(b) Survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13}, and γ_2, along with the corresponding standard errors. "MAR+W" is MAR+Weight, "AN+W" is AN+Weights, "AN+C" is AN+Constraint, and "AN+C+W" is "AN+Constraint+Weight." Standard errors of the averaged point estimates are small enough to rule out chance error as explanations for the improved performance of AN+C and AN+C+W over MAR+W and AN+W.

		MAR+W		AN+W		AN+C		AN+C+W	
Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE
α_0	0.15	0.19	0.05	0.12	0.08	0.15	0.05	0.15	0.05
α_{12}	−0.45	−0.48	0.06	−0.43	0.08	−0.45	0.06	−0.45	0.06
α_{13}	−0.15	−0.04	0.07	−0.23	0.21	−0.15	0.07	−0.16	0.07
γ_0	−1.00	−1.12	0.05	−0.97	0.21	−1.00	0.06	−1.00	0.06
γ_{12}	−0.60	−0.57	0.07	−0.64	0.10	−0.61	0.07	−0.61	0.07
γ_{13}	1.40	1.42	0.06	1.41	0.06	1.42	0.06	1.42	0.06
γ_2	−0.20	–	–	−0.44	0.47	−0.23	0.08	−0.23	0.08

$$\sum_{\substack{S_i=s;\\ i\in\mathcal{D}}} w_i X_{i1}^{\star} = \frac{N_s}{n_s} \sum_{\substack{S_i=s;\\ i\in\mathcal{D}}} X_{i1}^{\star} \sim N(T_X^{(s)}, V_X^{(s)}), \tag{16.21}$$

where $T_X^{(s)}$ is the auxiliary total of X_1^{pop} for stratum s, and $V_X^{(s)}$ is the corresponding variance associated with it. For the AN+Constraint and AN+Constraint+Weight methods, we implement this constraint by applying the Metropolis steps S1–S3 in Sect. 16.3 within each stratum.

We first set the parameters as in Sect. 16.4 to reflect a strong relationship between Y_1 and X_1, and strong nonignorable nonresponse mechanism. Table 16.5a shows the average HT estimates for T_X, the standard error under each method, and the acceptance ratios by strata. Table 16.5b shows survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13}, and γ_2, and the corresponding standard errors. The

Table 16.5 Results for scenario 3: auxiliary margin for X_1 within each stratum, strong relationship between Y_1 and X_1, and strong nonignorable nonresponse

(a) HT estimates for T_X under each method, the corresponding standard errors, and acceptance ratios by strata. "Population" is the value of T_X in the population of $N = 50{,}000$ individuals. "No Missing Data" is the weighted estimate based on the sampled data before introducing item nonresponse. For AN+Constraint and AN+Constraint+Weight, the estimated Monte Carlo standard errors of $\sum_{m=1}^{10} \bar{q}_L^m / 10$ are less than 250, ruling out chance error as explanation for the improved performance of these two models over AN+Weight and MAR+Weight.

	T_X		Acceptance Ratio			
			Stratum 1		Stratum 2	
Method	Mean	SE	Mean	Range	Mean	Range
Population	24,994	–	–	–	–	–
Mo Missing Data	25,043	580	–	–	–	–
MAR+Weight	30,447	668	–	–	–	–
AN+Weight	28,488	3034	–	–	–	–
AN+Constraint	25,062	665	0.81	[0.66, 0.91]	0.80	[0.74, 0.83]
AN+Constraint+Weight	25,070	667	0.80	[0.61, 0.90]	0.79	[0.74, 0.84]

(b) Survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13}, and γ_2, along with the corresponding standard errors. "MAR+W" is MAR+Weight, "AN+W" is AN+Weight, "AN+C" is AN+Constraint, and "AN+C+W" is "AN+Constraint+Weight." Standard errors of the averaged point estimates are small enough to rule out chance error as explanations for the improved performance of AN+C and AN+C+W over MAR+W and AN+W.

		MAR+W		AN+W		AN+C		AN+C+W	
Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE
α_0	0.50	0.74	0.05	0.64	0.13	0.50	0.05	0.50	0.05
α_{12}	−0.50	−0.45	0.07	−0.46	0.08	−0.50	0.07	−0.50	0.07
α_{13}	−1.00	−0.89	0.07	−0.90	0.12	−1.00	0.06	−1.00	0.07
γ_0	−0.25	−0.89	0.04	−0.73	0.44	−0.27	0.06	−0.27	0.06
γ_{12}	0.10	0.30	0.05	0.22	0.11	0.12	0.06	0.12	0.06
γ_{13}	0.30	0.65	0.05	0.52	0.19	0.31	0.06	0.31	0.06
γ_2	−1.10	–	–	−0.41	0.69	−1.08	0.09	−1.08	0.09

overall conclusions are qualitatively similar to those in Sect. 16.4. Incorporating the auxiliary margin by strata in AN+Constraint and AN+Constraint+Weight reduces the standard errors. AN+Weight and MAR+Weight again yield the least accurate results. The range of acceptance ratios is much wider suggesting that there is a smaller set of combinations of imputed values that fulfill the constraints within each stratum, than with the combined constraint.

We also set the parameters as in Sect. 16.4 to reflect a weak relationship between Y_1 and X_1, and a weakly nonignorable nonresponse mechanism. Table 16.6a,b display the results. The conclusions are qualitatively similar to those in previous simulations. The primary difference is that implementing the constraint by strata reduces the standard errors for AN+Constraint and AN+Constraint+Weight.

Table 16.6 Results for scenario 4: auxiliary margin for X_1 within each stratum, weak relationship between Y_1 and X_1, and weak nonignorable nonresponse

(a) HT estimates for T_X under each method, the corresponding standard errors, and acceptance ratios by strata. "Population" is the value of T_X in the population of $N = 50{,}000$ individuals. "No Missing Data" is the weighted estimate based on the sampled data before introducing item nonresponse. For AN+Constraint and AN+Constraint+Weight, the estimated Monte Carlo standard errors of $\sum_{m=1}^{10} \bar{q}_L^m/10$ are less than 350.

	T_X		Acceptance Ratio			
			Stratum 1		Stratum 2	
Method	Mean	SE	Mean	Range	Mean	Range
Population	24,724	–	–	–	–	–
Mo Missing Data	24,613	569	–	–	–	–
MAR+Weight	25,969	669	–	–	–	–
AN+Weight	23,551	3038	–	–	–	–
AN+Constraint	24,710	651	0.79	[0.60, 0.87]	0.76	[0.68, 0.79]
AN+Constraint+Weight	24,689	672	0.77	[0.59, 0.86]	0.75	[0.66, 0.78]

(b) Survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13}, and γ_2, along with the corresponding standard errors. "MAR+W" is MAR+Weight, "AN+W" is AN+Weight, "AN+C" is AN+Constraint, and "AN+C+W" is "AN+Constraint+Weight."

		MAR+W		AN+W		AN+C		AN+C+W	
Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE
α_0	0.15	0.18	0.05	0.12	0.09	0.15	0.05	0.15	0.05
α_{12}	−0.45	−0.48	0.06	−0.44	0.08	−0.46	0.06	−0.46	0.06
α_{13}	−0.15	−0.05	0.07	−0.24	0.25	−0.14	0.07	−0.14	0.07
γ_0	−1.00	−1.09	0.05	−0.97	0.26	−0.98	0.06	−0.98	0.06
γ_{12}	−0.60	−0.60	0.07	−0.68	0.12	−0.64	0.07	−0.64	0.07
γ_{13}	1.40	1.38	0.05	1.37	0.06	1.38	0.06	1.38	0.06
γ_2	−0.20	–	–	0.50	0.62	−0.21	0.07	−0.21	0.07

Results for the remaining two scenarios with known population totals per stratum are qualitatively similar to those presented here.

16.5 Discussion

The results suggest that the approach in Sect. 16.3 can allow survey analysts to incorporate survey weights and auxiliary information when imputing nonresponse in complex surveys. In particular, AN+Constraint and AN+Constraint+Weight appear to outperform the default option of controlling for the weights in the joint model for the variables in $\mathcal{D}$. The MAR+Weight approach offers good results when the nonresponse mechanism is only weakly nonignorable; we expect that this method should perform even better for fully ignorable nonresponse mechanisms. However, the results based on AN+Constraint and AN+Constraint+Weight are the most

consistently best across the different scenarios. Of course, these results are based on a limited set of simulation scenarios, and the methods could perform differently in other scenarios.

Opportunities for extensions of this approach exist as future research topics. First, future work could explore extensions of the approach to other sampling designs, in particular when weights have many unique values. Preliminary simulations, not shown here, suggest that our approach can also work well for many valued, unequal weights. However, generating plausible imputations that satisfy the constraint can be challenging whenever the set of combinations of imputed values that result in completed datasets that satisfy the constraint is small compared to the set of all possible combinations. When this is the case, we have found that one needs efficient samplers for generating proposals for the imputations. Finding general strategies for such proposals is an important topic for future work.

Second, as we suggested in Sect. 16.3, there are opportunities to investigate different approaches to specifying the constraints involving $\hat{T}_X$, in particular how to set the variance V_X. In the simulations, we used the theoretical design-based variance, estimated via resampling from the true generative process, but this would need to be approximated in practice. Future research could examine the effectiveness of using the types of approximations described in Sect. 16.3. Additionally, one could investigate how different values of V_X affect the performance of the methodology. For example, using very small V_X could lead to a more efficient estimation of T_X; however, forcing the completed datasets to match very closely on T_X could affect the relationships among X and Y in the completed data in unpredictable ways. It could also lead to a less efficient MCMC sampler, since the set of imputations consistent with (16.13) would be smaller.

Third, future research could adapt this approach to other model specifications. For example, one could extend the approach to nonparametric models and semi-parametric models like those in Kim and Yu (2011), and Morikawa et al. (2017).

Fourth, one could extend the framework to handle imputation for unit nonresponse as well. In particular, we conjecture that analysts can follow the framework developed by Akande (2019, Chapter 4), who extends the AN model to unit nonresponse as well as item nonresponse in more than one variable in simple random samples. We expect that analysts can add the probabilistic constraint on the completed data totals on top of the models in Akande (2019, Chapter 4). We note that this requires survey weights for the unit nonrespondents, which often are not available.

Finally, we work with base weights instead of more complex "adjusted" weights, which are often inflated to adjust for nonresponse or poststratification. Since we take a model-based approach to handling survey nonresponse, there is no obvious justification for using adjusted weights that already account for the nonresponse. In fact, using such adjusted weights assumes that the weights are fixed, which is not often true as pointed out by Fienberg (2010). Since agencies often release those adjusted weights in practice, instead of the base weights, future work would explore the extension of our approach to adjusted weights as well.

Acknowledgments This research was supported by grants from the National Science Foundation (SES-1131897 and SES-1733835).

References

Akande, O. *Bayesian Models for Imputing Missing Data and Editing Erroneous Responses in Surveys* PhD dissertation (Duke University, 2019).

Bhattacharya, D. Inference in panel data models under attrition caused by unobservables. *Journal of Econometrics* **144**, 430–446 (2008).

Bishop, Y. & Fienberg, S. E. Incomplete two-dimensional contingency tables. *Biometrics* **25**, 119–128 (1969).

Chen, T. & Fienberg, S. E. The analysis of contingency tables with incompletely classified data. *Biometrics* **32**, 133–144 (1976).

Chen, T. & Fienberg, S. E. Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics* **30**, 629–642 (1974).

Das, M., Toepel, V. & van Soest, A. Nonparametric tests of panel conditioning and attrition bias in panel surveys. *Sociological Methods and Research* **40**, 32–56 (2013).

Deng, Y., Hillygus, D. S., Reiter, J. P., Si, Y. & Zheng, S. Handling attrition in longitudinal studies: the case for refreshment samples. *Statistical Science* **28**, 238–256 (2013).

Fienberg, S. E. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics* **41**, 907–917 (1970).

Fienberg, S. E. The analysis of incomplete multi-way contingency tables. *Biometrics* **28**, 177–202 (1972).

Fienberg, S. E. The relevance or irrelevance of weights for confidentiality and statistical analyses. *Journal of Privacy and Confidentiality* **1**, 183–195 (2010).

Fuller, W. A. in *Sampling Statistics* 1–93 (John Wiley & Sons, Ltd, 2009).

Gelman, A. Struggles with survey weighting and regression modeling. *Statistical Science* **22**, 153–164. https://doi.org/10.1214/088342306000000691 (2007).

Glynn, R. J., Laird, N. M., & Rubin, D. B. Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples* (ed Wainer, H.) 115–142 (Springer, New York, 1986).

Hausman, J. & Wise, D. Attrition bias in experimental and panel data: The Gary income maintenance experiment. *Econometrica* **47**, 455–473, (1979).

Hirano, K., Imbens, G., Ridder, G. & Rubin, D. Combining panel data sets with attrition and refreshment samples. *Technical Working Paper 230* (1998).

Hirano, K., Imbens, G., Ridder, G. & Rubin, D. Combining panel data sets with attrition and refreshment samples. *Econometrica* **69**, 1645–1659 (2001).

Horvitz, D. G. & Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685 (1952).

Kim, J. K. & Yu, C. L. A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association* **106**, 157–165 (2011).

Little, R. J. A. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134 (1993).

Little, R. J. A. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* **90**, 1112–1121 (1995).

Morikawa, K., Kim, J. K. & Kano, Y. Semiparametric maximum likelihood estimation with data missing not at random. *Canadian Journal of Statistics* **45**, 393–409 (2017).

Nevo, A. Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business and Economic Statistics* **21**, 43–52 (2003).

Pfeffermann, D. The role of sampling weights when modeling survey data. *International Statistical Review* **61**, 317–337. ISSN: 03067734, 17515823. http://www.jstor.org/stable/140363 (1993).

Reiter, J. P., Raghunathan, T. E. & Kinney, S. K. The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology* **32**, 143–150 (2006).

Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys* 258 (John Wiley & Sons, New York, 1987).

Sadinle, M. & Reiter, J. P. Sequentially additive nonignorable missing data modeling using auxiliary marginal information. *Biometrika* **106**, 889–911 (2019).

Schifeling, T. S., Cheng, C., Reiter, J. P. & Hillygus, D. S. Accounting for nonignorable unit nonresponse and attrition in panel studies with refreshment samples. *Journal of Survey Statistics and Methodology* **3**, 265–295 (2015).

Si, Y., Reiter, J. P. & Hillygus, D. S. Semi-parametric selection models for potentially non-ignorable attrition in panel studies with refreshment samples. *Political Analysis* **23**, 92–112 (2015).

Zhou, H., R., E. M. & Raghunathan, T. E. Synthetic multiple imputation procedure for multistage complex samples. *Journal of Official Statistics* **32**, 231–236 (2016).

Chapter 17
Insights into the Decision on Whether to Statistically Adjust the 2000 Census for Coverage Errors

John H. Thompson

17.1 Introduction

The decennial census is one of the cornerstones of our democracy. The Constitution requires that the Decennial Census be used for reapportioning the Congress of the United States and the Electoral College. The decennial census is also used for numerous other functions to support good policymaking and economic growth including: redrawing Congressional and local voting districts; allocating over $900 billion of federal funds annually; informing sound policy development; providing critical information for state, local, and tribal government planning; and supplying critical information to large and small businesses to generate growth and job creation. Inaccuracies or errors in the decennial census will have grave consequences for these uses not only in the year in which the census data are produced, but for the subsequent 10-year period until data from the next census are available.

Understandably, there are great concerns regarding the accuracy of the census by a wide array of stakeholders. No decennial census has been perfect, and historically, non-White and Hispanic population groups have been undercounted in the decennial census at a significantly higher rate than the White population. Renters have also been undercounted to a greater extent than owners. Statistically adjusting the decennial census to correct for such coverage errors, in particular the differential undercounts of non-White and Hispanic populations, has been a consideration for the 1980, 1990, and 2000 Censuses. In addition, for each of these censuses, lawsuits were filed over whether statistical adjustment should be employed. Statistical adjustment was not considered for the 2010 Census and was not included in the planning for the 2020 Census as explained below.

J. H. Thompson (✉)
Bend, OR, USA

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_17

One of the many interests of Stephen Fienberg was the conduct of the decennial census and how statistical methods could be used to improve the enumeration. One of his numerous publications included a book coauthored with Margo Anderson *Who Counts? The Politics of Census-Taking in Contemporary America,* Anderson and Fienberg (2001). That volume discusses the history of the decennial census and undercounts. The book includes a detailed discussion of the issues surrounding the 1990 Census regarding the use of statistical methods to correct for undercounts. It then describes the planning leading up to the 2000 Census that included the incorporation of statistical methods into the core 2000 Census processes to use sampling to reduce costs and to use statistical methods to adjust or correct the census results for coverage error. The book chronicles the controversy and litigation that emerged in opposition to this Census Bureau plan to use statistical methods and the final decision by the Supreme Court to rule out the use of statistical sampling to produce the census counts for apportionment. It concludes with the release of the 2000 Census data for apportionment and documents that in the next phase of the 2000 Census the Census Bureau would be considering the use of statistical adjustment for all non-apportionment uses of the 2000 Census including redistricting.

Subsequently, the Census Bureau determined not to use statistical adjustment to produce the redistricting data due to concerns with the accuracy of the coverage measurement methodology used in the 2000 Census. This chapter will build on the work of Anderson and Fienberg to describe the decision process used by the Census Bureau in reaching this conclusion and will provide insights into the deliberations of a statistical agency.

17.2 Background

In order to establish the proper context for the discussion of the Census Bureau deliberations on the matter of statistically adjusting the 2000 Census results for all non-apportionment purposes including redistricting, it will be necessary to provide a brief background on decennial census methodology and measurement of coverage errors.

The basic process that has been used to conduct the decennial census since 1970 is based on self-response and a subsequent follow-up for those households that do not self-respond. The methodology has changed little from the 1970 Census through the 2010 Census and is as follows: (1) The decennial census day is April 1 of the year in which the census is conducted. The decennial census counts every resident of the United States where they usually live or stay on April 1 of the decennial census year (Census Bureau 2018). Often the term “Census Day” is used as a reference to April 1. (2) The Census Bureau compiles an address list

of nearly all housing units in the United States.[1] The address list is validated and updated 1 or 2 years in advance of the census by address canvassers who walk virtually every street. The address canvassers add missing addresses and delete nonexistent listings. (3) Shortly before Census Day nearly every address is mailed an invitation to self-respond. Through the 2010 Census the invitation consisted of a paper questionnaire which respondents were asked to complete and mail back. The rate at which households provide a self-response has been about 65 percent since the 1990 Census (Census Bureau 2012). (4) Information on the completed questionnaires is electronically captured using a scanning technology.[2] (5) About a month after Census Day, the Census Bureau conducts a follow-up to collect information from those households that have not self-responded. This operation is referred to as "Nonresponse Follow-up" or NRFU. The NRFU is based on sending a census enumerator to visit each nonresponding housing unit a prescribed number of times to obtain an interview. NRFU is a massive undertaking requiring the Census Bureau to hire over 500,000 enumerators. Through the 2010 Census, NRFU was a paper-based operation. That is, no automated technology was employed to either assist or manage the enumerators. (6) And finally, starting in the 2000 Census, the Census Bureau has funded a large combined paid advertising and local partnership program to motivate self-response and to encourage full participation in the decennial Census.

The 2020 Census will differ from previous decennial censuses in four distinct areas (Census Bureau 2018): (1) The address canvassing operation will be conducted largely in an office environment using modern geospatial tools. (2) Respondents will be allowed to use both the Internet and telephone in addition to mail as means of self-response. (3) NRFU will be highly automated, taking advantage of mobile technology by equipping enumerators with smart phones and their supervisors with tablets. (4) And administrative records will be used for the first time to reduce the NRFU workload.

The Census Bureau is required by law, Title 13 US Code, Section 141 (b) and (c), to meet two deadlines for the release of certain data products. The first is the production of the 50 state population totals that are to be used to reapportion the Congress. Currently these data are required to be produced by December 31 of the year in which the decennial census is taken. The second legally mandated product is the data to support state redistricting. These data consist of census block level tabulations of Race and Hispanic ethnicity by voting age (McCully 2014). The Census Bureau must deliver these data to the states by March 31 of the year following the conduct of the decennial. These data are also used by the Department of Justice to enforce voting rights. It is important to note that the COVID-19 pandemic has forced the Census Bureau to delay a number of operations, and

[1] There are small portions of the United States where an address list is not prepared. In these areas, census enumerators will deliver a questionnaire or conduct a direct enumeration.

[2] The scanning technology has evolved from in-house mark recognition to optical scanning and intelligent character recognition.

the Census Bureau has requested that Congress extend these deadlines to April 30, 2001, and July 31, 2001, respectively (U.S. Census Bureau 2020). The actual release dates were April 26, 2021 for apportionment counts and August 12, 2021 for redistricting data.

Census accuracy in terms of undercounts and overcounts has been measured by two methods since at least 1980 – post enumeration survey with dual system estimation and demographic analysis. The statistical adjustment methodology that the Census Bureau has considered for correcting the decennial census for coverage errors has been based on using the results of the post enumeration survey and dual system estimation results. These methods are described very briefly as follows:

The post enumeration survey (PES) and dual system estimation (DSE) is described very well by Hogan (1993) and Mule (2012). These authors also discuss statistical adjustment methods. Essentially the methodology is based on conducting a PES that is independent of the decennial census and matching it to the decennial census. The following 2 × 2 contingency table as described in Hogan (1993) can then be constructed:

Census enumeration			
PES	In	Out	Total
In	N_{11}	N_{12}	N_{1+}
Out	N_{21}	N_{22}	N_{2+}
Total	N_{+1}	N_{+2}	N_{++}

Assuming that the PES is independent of the census, a dual system estimator (DSE) of the "true" population (N_{++}) can be constructed as:

$$\text{DSE} = (N_{+1})\ (N_{1+})\ /N_{11}$$

In actuality, the calculation of a DSE for the decennial census is much more complicated since the decennial census contains erroneous enumerations and includes enumerations which do not have enough information to match to the PES. This has required the Census Bureau to select an additional sample to measure erroneous enumerations so that the DSE can be modified to reflect erroneous census enumerations and the census enumerations that cannot be matched. Hogan (1993) also discusses this in detail. However, it will not be necessary for this chapter to go into that level of detail.

Undercount (U) is measured by:

$$\text{U} = (\text{DSE–Census})\ /\text{DSE}$$

It should be noted that negative values of U indicate an overcount.

The PES and DSE methodology allow for the estimation of coverage errors for a wide range of characteristics, including race, Hispanic ethnicity, age, sex, and tenure

(owner/renter). These estimates are also produced (using model-based techniques) for subnational geographic areas of the United States including states, large cities, and counties.

The PES and DSE methodology are subject to limitations as well. Most notably, the requirement for independence of the decennial census is not met when certain respondents who have a lower chance of being included in the decennial census also have a lower chance of being included in the PES. This phenomenon is often referred to as "heterogeneity" or "correlation" bias. The result of this bias is an underestimation of the actual undercount. In the case of the decennial census, this bias has often been observed for young- to middle-aged black males, and other groups that the Census Bureau considers to be hard-to-count. Therefore, the DSE is viewed as a conservative estimate of undercounts for these groups.

Table 17.1 summarizes the estimates of coverage error that resulted for the PES and DSE for the 1990, 2000, and 2010 censuses by race and Hispanic ethnicity:[3]

Demographic analysis is well described by Robinson et al. (1993) and Devine et al. (2012). Briefly, demographic analysis produces an estimate of the population of the United States at any point in time. Administrative records are used to account for births and deaths and legal immigration. Demographic analysis must make assumptions to estimate emigration and illegal immigration. Examples of administrative record sources are vital statistics on births and deaths and Medicare records for the population over age 65. As discussed in Robinson et al. (1993) and Devine et al. (2012), demographic analysis estimates of the total or "true" population are only produced at the national level and for age, sex, and Black and non-Black racial classifications.

Demographic analysis has provided estimates of decennial census coverage error for each census starting with the1940 Census (Robinson et al. 1993). In addition, even though demographic analysis does not have the granularity of the DSE, it is less subject to correlation bias. Therefore, estimates based on demographic analysis are valuable in assessing the quality of the DSE. This issue will be discussed in much more detail below.

Undercount (U) is estimated from demographic analysis (DA) as:

$$\mathrm{U} = (\mathrm{DA} - \mathrm{Census}) / \mathrm{DA}$$

Statistical adjustment of the decennial census is based on using the DSE as the starting point for a small area estimation approach to incorporate corrections for coverage error into the decennial census. The small area estimation methodology used for the 1990 and 2000 Censuses is referred to as "synthetic estimation." Hogan (1993) presents a comprehensive discussion of synthetic estimation. It can be briefly summarized as the following process:

[3] As will be discussed below the estimate of undercount for the 2000 Census in this table differs from the estimate initially produced in March of 2001.

- Census blocks are the basic tabulation unit for many important uses of the decennial census including the data that the Census Bureau is legally required to produce to support redistricting.[4] Therefore each of the census block counts must be statistically adjusted.
- The process starts by defining a number of post-strata that are designed to be homogeneous with respect to coverage error, are mutually exclusive, and include every component of the US population. For example, a stratum could be defined as Black renters in central cities of Standard Metropolitan Statistical Areas in the Northeast Region. For each post-stratum, a DSE is first calculated and then an adjustment factor (AF) is produced as:

$$\text{AF} = \text{DSE}/\text{Census}$$

- Then the count of each block component of the post-stratum is multiplied by AF. For the example of Black renters in central cities in the Northeast, suppose that AF = 1.05 (a 5 percent undercount). Then each block in the Northeast that had a count of Black renters in central cities would have the count multiplied by 1.05.
- At the end of this process, each census block would have a statistically adjusted count that would be used for all subsequent tabulation purposes.[5]

17.3 Census Bureau Consideration of Statistical Adjustment for the 2000 Census

As noted above, the initial plan for the 2000 Census included the use of sampling to reduce the NRFU costs and statistical adjustment to be included as an integral census process to produce the most accurate results possible. Anderson and Fienberg present a detailed discussion of the genesis of this plan; briefly it is described as follows.

The 1990 Census ended leaving many stakeholders (including many members of Congress) concerned with the outcome. The 1990 Census cost more than had been planned and undercounts of many population groups were higher than in the 1980 Census. Litigation had also been filed to seek adjustment of the 1990 Census to correct for undercounts (City of New York v. US Department of Commerce 1990). There was a widespread recognition that the basic methods for conducting the decennial census had to be reexamined in developing the plan for the 2000 Census.

[4]Technically the Census Bureau is not legally required to produce block level data; however, the Census Bureau produces this level of tabulation to meet the requests of state redistricting offices.

[5]This is an oversimplified description of the methodology. The actual process is more complicated and results in individual records being created on the base census data file that is used to produce all census tabulations.

In particular, Congressman Thomas Sawyer (D-OH), the chair of the Census Bureau's oversight subcommittee in the House, introduced a bill in late 1990 that required the Secretary of Commerce to engage the National Academy of Sciences (NAS) to examine ways for the Census Bureau to conduct the most accurate census possible in 2000 (House of Representatives 1991). In response, the NAS established a Panel on Census Requirements in the Year 2000 and Beyond National Research Council (1995).[6] The Census Bureau also asked the NAS to establish an additional Panel to Evaluate Alternative Census Methods National Research Council (1994).

Both Panels made a number of important recommendations regarding the conduct of the 2000 Census. Importantly, both panels recommended that the 2000 Census should use sampling to reduce the cost of NRFU and should integrate coverage measurement and statistical adjustment of the physical enumerations into the core census process to reduce, to the greatest extent possible, undercounts, including differential undercounts. This was frequently referred to as the "One Number Census."

The Census Bureau accepted these recommendations and in February 1996, then Census Bureau Director Marty Riche, announced the plan for the 2000 Census, (US Census Bureau 2009). The plan called for using statistical sampling techniques in the two principal ways recommended by the NAS Panels.

However, the composition of the Congress changed to a Republican majority in both the House of Representatives and the Senate. Under this new leadership, opposition to the plan to use sampling grew to the point at which the House of Representatives filed litigation to prevent the use of sampling for the purposes of apportionment, US Department of Commerce v. (US House of Representatives 1999). Ultimately the litigation went to the Supreme Court which ruled in January 1999 that the use of statistical sampling in the decennial census for purposes of determining congressional apportionment violated the Census Act.

The Census Bureau then revised the plan for Census 2000 to not use statistical sampling for the enumeration of the population and to return to the counting processes used in the 1990 Census. The revised plan did include a new use of paid advertising and local partnership to motivate response and cooperation to the 2000 Census.[7] The Census Bureau also announced that the revised plan would include a consideration of the use of statistical adjustment to correct the 2000 Census results for all non-apportionment purposes, including redistricting. It should also be noted that the Census Bureau began referring to the coverage measurement program based on a PES and DSE as the "Accuracy and Coverage Evaluation" or A.C.E. This terminology will be used extensively in the discussions that follow.

At this point, statistical adjustment was a highly charged political issue; the Democratic Party including the Clinton Administration supported the methodology, and the Republican Party opposed it. The Census Bureau understood the

[6] It should be noted that Stephen Fienberg was a member of this Panel.

[7] It is beyond the scope of this paper, but this may have been the most significant innovation adopted for the 2000 Census.

importance of maintaining its credibility as an independent and objective Statistical Agency in this environment. Therefore, it established the Executive Steering Committee on Accuracy and Coverage Evaluation Policy (ESCAP) to recommend whether statistical adjustment of 2000 Census results should be used for all non-apportionment uses including redistricting. The first task of the ESCAP was to develop a recommendation on whether the redistricting data (required to be released by March 31, 2001) should be statistically adjusted. The remainder of this chapter focuses on this decision process.

17.4 Developing a Recommendation on Statistical Adjustment of Redistricting Data

As noted above, the Census Bureau understood the importance of maintaining its credibility as a Statistical Agency, (National Academy of Sciences, Engineering and Medicine 2017). In order to be completely transparent on how it would determine whether to recommend that statistically adjusted 2000 Census data be used for all non-apportionment purposes, it issued a Federal Register Notice in June 2000 documenting how it would proceed (Federal Register 2000). The document set forth the rationale for the Census Bureau's preliminary determination that (1) statistically corrected census data can be produced within the time frame required by law and (2) that statistically corrected data will be more accurate. However, the Federal Register Notice made clear that the Census Bureau's final decision on what data to release would not be made until the Census Bureau conducted a review of both the census and the A.C.E. Importantly, the Federal Register Notice laid out criteria that would be used in this review and documented the establishment of the ESCAP.

The ESCAP consisted of 12 senior Census Bureau career staff. The ESCAP was established in November 1999 and was chaired by the Associate Director for the 2000 Census – the Career Executive with responsibility for managing all aspects of the 2000 Census. The other members of the ESCAP represented a body of senior career Census Bureau professionals, with advanced degrees in relevant technical fields and/or decades of experience in the Federal statistical system. All were exceedingly well qualified to evaluate the relative merits of both the 2000 Census and the A.C.E.

Three criteria were established that would guide the development of the ESCAP recommendation:

- Consideration of operational data to validate the successful conduct of the A.C.E. and 2000 Census
- Assessment of the consistency of the A.C.E. measures of undercount with historical patterns of undercount and independent demographic analysis benchmarks
- Review of quality measures for the A.C.E and the 2000 Census

The ESCAP reviewed extensive materials and analyses that were relevant for each of the above criteria, often the ESCAP required presentations from internal and external experts on various aspects of the 2000 Census and A.C.E. This extensive work is documented in the March 1, 2001, ESCAP recommendation report (ESCAP 2001a). To clearly demonstrate that the ESCAP deliberations were objective and independent of political interference, all briefing and meeting minutes were documented and made public with the release of the ESCAP report.

The results of the ESCAP review of the three criteria follows:

- Consideration of operational data to validate the successful conduct of the A.C.E. and 2000 Census

The ESCAP examined a number of operational measures for both the 2000 Census and the A.C.E. For the 2000 Census these measures included the mail response rates and metrics regarding the conduct of the address list development and NRFU. The measures examined for the A.C.E. included sampling error, missing data rates, and matching error. The ESCAP concluded that both the 2000 Census and the A.C.E. were efficient and effective operations that produced high-quality data. All major programs in the census were completed on schedule and within budget, and design improvements in both Census 2000 and the A.C.E. produced measurably better results than were observed in 1990 for the census and the coverage measurement operation equivalent to the A.C.E.

- Assessment of the consistency of the A.C.E. measures of undercount with historical patterns of undercount and independent demographic analysis benchmarks

This assessment was based on comparing the estimates of coverage error from demographic analysis and the A.C.E. The assessment identified a major inconsistency between the two methods of measuring census coverage. The A.C.E. estimated a net undercount of about 1.15 percent while demographic analysis estimated a net overcount of about −0.65 percent. This discrepancy was very troubling to the ESCAP, particularly since historically, demographic analysis produces higher estimates of undercounts than those resulting from a coverage measurement survey such as the A.C.E. (e.g., for the 1990 Census demographic analysis estimated a net undercount of 1.8 percent while the coverage measurement survey estimated a net undercount of 1.6 percent [Hogan (1993)]). The ESCAP investigated this inconsistency extensively, but in the time available could not adequately explain the result (subsequent analysis described below found that the A.C.E erroneously underestimated the level of duplicate enumerations, which explained much of the discrepancy with demographic analysis).

- Review of quality measures for the A.C.E and the 2000 Census

The review of quality measures was based on a total error model that had been developed for the 1990 Census coverage measurement process and DSE (Mulry and Spencer 1991). Under this approach the components of error that could occur in a coverage measurement survey were identified and combined in a model that could be used to describe the total error in the DSE. In 1990 a number of evaluation studies

were carried out to measure each error component and the results were incorporated into the total error model. The 1990 DSE could then be corrected for this total error, producing an estimate of the "true" population. The 1990 Census and the DSE produced by the coverage measurement could then be compared to the "true" population to determine which is more accurate.[8]

The ESCAP considered the 1990 total error model as a basis for evaluating the quality of the 2000 Census and the A.C.E. However, the evaluation studies designed to measure the components of the error were not scheduled to be completed in time for the ESCAP review for the purposes of making a recommendation on whether to release adjusted data for redistricting. The ESCAP was forced to use the results of the 1990 Census coverage measurement evaluations. Given the discrepancies between the A.C.E and demographic analysis noted above, the ESCAP could not rely on this analysis of quality, until more information from the evaluations of the actual components of error for the A.C.E. was available.

17.5 ESCAP Conclusion and Recommendation for the 2000 Census Redistricting Data

The ESCAP issued its first report on March 1, 2001, (ESCAP 2001a) and concluded

> There is a significant inconsistency between the A.C.E. estimates and demographic analysis estimates. Additionally, possible synthetic and balancing errors may affect the accuracy of the adjusted numbers. Until these concerns are more fully investigated and addressed, the ESCAP cannot recommend using adjustment. Accordingly, ESCAP has recommended that unadjusted census data be released as the Census Bureau's official redistricting data.

This recommendation was accepted by the Director of the Census Bureau and by the Secretary of Commerce and unadjusted redistricting data were released.

This decision was very difficult for the ESCAP, since it was clear that the 2000 Census data contained undercount that were differential between the White and non-white Hispanic populations. These undercounts would therefore be present in the 2000 redistricting data. However, the risk of introducing flawed A.C.E. results as a correction was viewed as too great as well as improper behavior for a Statistical Agency.

Following the release of the March 2001 report, the ESCAP began consideration of whether the A.C.E could be used to produce adjusted data for other non-apportionment purposes. The ESCAP was informed by the completion of the evaluation studies of the A.C.E during the summer of 2001. The results of this work are reported in the second report issued by the ESCAP on October 17, 2001 (ESCAP 2001b).

[8]This is an extremely oversimplified description of the process. The complete discussion is found in Mulry and Spencer (1991).

Table 17.1 Census net undercount estimates 1990, 2000, and 2010

Estimates of percent net undercount by Race/Origin domain						
	2010		2000		1990	
Race/Origin Domain	Estimate (%)	Standard error (%)	Estimate (%)	Standard error (%)	Estimate (%)	Standard error (%)
U.S. Total	−0.01	0.14	−.049*	0.20	1.61*	0.20
Non-Hispanic White	−0.84*	0.15	−1.13*	0.20	0.68*	0.22
Non-Hispanic Black	2.07	0.53	1.84*	0.43	4.57*	0.55
Non-Hispanic Asian	0.08	0.61	−0.75	0.68	2.36*	1.39
American Indian on Reservation	4.88*	2.37	−0.88	1.53	12.22*	5.29
American Indian off Reservation	−1.95	1.85	0.62	1.35	0.68*	0.22
Native Hawaiian or Pacific Islander	1.34	3.14	2.12	2.73	2.36*	1.39
Hispanic	1.54*	0.33	0.71	0.44	4.99*	0.82

Source: Mule, Thomas, *2010 Census Coverage measurement Estimation Report: Summary of Estimates of Coverage or person in the United States*, US Census Bureau, May 22, 2021
An asterisk (*) denotes a percent net undercount that is significantly different from zero
Negative values indicate a newt overcount

The ESCAP found that the A.C.E. had not correctly measured the level of duplicate enumerations in the 2000 Census. The result was that the A.C.E. was found to have overestimated the population of the United States by potentially three to four million people (ESCAP 2001a, b). Therefore, the initial estimate of a 1.15 percent net undercount from the A.C.E. could be in the range of an undercount of 0.01 to an overcount of −0.03 percent. The ESCAP concluded at this point "The significance of the error in the A.C.E. treatment of duplicates compels the recommendation that the current A.C.E. estimates cannot be used to adjust the Census 2000 data, (ESCAP 2001a, b)."

At this point the ESCAP was concluded and the Census Bureau went forward to produce revised estimates of the 2000 Census coverage error that appear in Table 22.1.

17.6 Conclusion

This chapter started with a discussion of the importance of the decennial census to our democracy and as a source of critical information to a wide array of stakeholders. Improving the accuracy and efficiency of the 2000 Census relative to the 1990 Census was the prime factor in the planning program that the Census

Bureau instituted. The Census Bureau followed sound scientific principles in developing the plan for the 2000 Census that included the use of statistical sampling and adjustment as means to reduce costs and increase accuracy. When this plan became the object of intense partisan political debate, the overarching concern at the Census Bureau was to maintain its status as an objective Statistical Agency free from any political interference. The deliberative processes described above have demonstrated how the Census Bureau achieved this goal. The Census Bureau would only recommend that statistical adjustment be used if there was clear and compelling evidence to support such an action. The decennial census cannot be the subject of arbitrary decision making. The decision to not statistically adjust the 2000 Census was supported by objective and careful research. Stephen Fienberg was intimately involved with the 2000 Census and was hopefully pleased with the final outcome.

One further point regarding the concerns for accuracy in the decennial census: The 1980, 1990, and 2000 Censuses resulted in litigation filed by stakeholders who felt that they would be disadvantaged by undercounts. For 1980 and 1990, litigation called for statistical adjustment, while the 2000 litigation was opposed to this methodology. The 2010 Census did not include any usage of statistical adjustment to correct for coverage errors, nor does the planning for the 2020 Census. However, as the past has shown, if the 2020 Census is found to contain serious undercounts for particular population groups, it would not be surprising to see litigation arise again seeking statistical adjustment as remedy.

References

2010 Census Planning Memorandum Series: 198 *2010 Census Mail Response/Return Rates Assessment Report.* Census Bureau, (2012)

2020 Census Program Memorandum Series: 2018.04 2020 *Census Residence Criteria and Residence Situations*. Census Bureau, (2018)

Anderson, Margo J. and Stephen E. Fienberg 2001 *Who Counts? The Politics of Census Taking in Contemporary America.* New York: Russel Sage Foundation

City of New York v. US Department of Commerce, 739 F. Supp. 761 (E.D.N.Y. 1990)

Devine, Jason, Renuka Bhaskar, Bethany DeSalvo, J. Gregory Robinson, Melissa Scopilliti, Kirsten K. West, *The Development and Sensitivity Analysis of the 2010 Demographic Analysis Estimates*. Population Division Working Paper No. 93. Census Bureau (2012)

Executive Steering Committee on A.C.E. Policy, *Recommendation Concerning the Methodology to be Used in Producing the Tabulations Reported to States and Localities Pursuant to 13 U.S.C. 141(c).* Census Bureau (March 1, 2001a)

Executive Steering Committee on A.C.E. Policy, *Adjustment for Non-Redistricting Purposes*. Census Bureau (October17, 2001b)

Federal Register Notice, Accuracy and Coverage Evaluation; Statement on the Feasibility of Using Statistical Methods to Improve the Accuracy of Census 2000, (June 20, 2000)

Hogan, Howard (1993) *The 1990 Post-Enumeration Survey: Operations and Results*, Journal of the American Statistical Association, 88:423, 1047-1060

House of Representatives (H.R.) 3280, Decennial Census Improvement Act of 1991. Enacted as P.L. 102-135, October 24, 1991

McCully, Catherine *Designing P.L. 94-171Redistricting Data for the Year 2020 Census*. U.S. Census Bureau, (2014)

Mule, Thomas (2010) *Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage in the United States*. US Census Bureau, (May 22, 2012)

Mulry, Mary H. and Bruce D. Spencer (1991) *Total Error in PES Estimates of Population*, Journal of the American Statistical Association, 86:416, 839-855

National Academy of Sciences, Engineering and Medicine 2017. *Principles and Practices for a Federal Statistical Agency: Sixth Edition*. Washington, DC: The National Academies Press

National Research Council 1995. *Modernizing the U.S. Census*. Washington, DC: The National Academies Press.

National Research Council 1994. *Counting People in the Information Age.* Washington, DC: The National Academies

Robinson, J. Gregory, Bashir Ahmed, Prithwis Das Gupta & Karen A. Woodrow (1993) *Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis*, Journal of the American Statistical Association, 88:423, 1061-1071

U.S. Census Bureau *2020 Census Operational Plan A New Design for the 21st Century, Version 4.0.* Census Bureau, (2018)

U.S. Census Bureau, *History: 2000 Census of Population and Housing (Volume 1).* Census Bureau (December 2009)

U.S. Census Bureau, *U.S. Department of Commerce Secretary Wilbur Ross and U.S. Census Bureau Director Steven Dillingham Statement on 2020 Census Operational Adjustments Due to COVID-19.* Census Bureau Press Release, (April 13, 2020)

U. S., Department of Commerce v. U. S. House of Representatives 525 U. S. 316, (1999)

Chapter 18
2030: A Sensible Census, in Reach

Kenneth Prewitt

Steve Fienberg, in 1998 as I prepared for the Census Directorship, told me that it is not whether we have a perfect method, but whether we have a sensible method. A few years earlier, asked to explain to a congressional committee how research is converted to a commercial product, Steve deftly pointed out that posing a narrow question about a complex transaction doesn't produce a sensible answer.

> The problem is that metrics used to assess any one aspect of the research system in isolation, without a strong understanding of the larger picture, may prove misleading. The benefits of research investments tend to arrive unpredictably, vary widely in eventual value, and require substantial additional investment (as well as investment in other fields of science) to realize their economic payoff through innovation. With few exceptions, approaches to measure the impacts and quality of research programs cannot depict the diffuse, interconnected and highly nonlinear pathways that lead from research to technologies and other innovations. (Fienberg 2014)

I link these two Fienbergian principles to underscore that every census travels on its uniquely diffuse, interconnected and highly nonlinear pathway and, every census, perfection not being an option, searches for the most sensible method for navigating its particular pathway.

The title of this chapter announces its target – a sensible census in 2030. Reproducing the current 2020 census design in 2030 is not an option. There are numerous reasons, the most important being the opportunity to lessen the reliance on labor-intensive field work by using data independently available – administrative data, links to other federal statistical data, and commercial/social media data. To the extent these data sources reduce the reliance on direct contact with millions of households, the decennial budget is reduced, and the budget for other survey-based data can benefit, notably the American Community Survey.

K. Prewitt (✉)
Columbia University, New York City, NY, USA
e-mail: kp2058@columbia.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_18

A less far-reaching, though possibly critical legacy of 2020, is the shadow of political interference; it is too early to know how much real damage is occurring. The census is still in the field.[1] But even the appearance of political interference is harmful, especially to the completeness and accuracy in the hard-to-count population. A related problem is, ironically, how a positive development in census-taking, starting with the 2000 census – below labeled a "census face-lift" – has inadvertently introduced practices inconsistent with census-taking. I sort this out below, using it and political interference to underscore the risks to today's census model and the rationale for adjustments in 2030.

18.1 Political Interference in the Census

Article 1, Section 2, of the Constitution directs that the actual census enumeration be taken every 10 years "in such Manner as [Congress] shall by Law direct." Because Congress is composed of the politically ambitious, it is no surprise that it has argued over the census. This was so from the first census, with its three-fifths clause advantaging slave-holding states. Since then censuses have varied in the degree to which they are shaped as much by politics as science – the 1840 question on whether a person was "insane or idiotic," which seemed to show (the data were eventually found to be useless) that free blacks were driven to insanity and the enslaved were spared that condition, giving Senator Calhoun ammunition to promote Texas joining the union as a slave-holding state. Or the unwelcomed surprise in 1920 that America had become more urban than rural, potentially sending a dozen or so congressional seats from rural regions to the urbanizing, politically threatening northeastern states (the Red Scare era). Congress, dominated by rural, conservative states, refused to reapportion – claiming census errors but in fact making a political decision.

These examples from census history (there are dozens more) remind us that census-taking is simultaneously political and scientific. What, then, constitutes political interference? Every census starts with its particular purposes, then becomes the production of statistics, and ends with its uses. Purposes and uses are political and belong to the politicians. Production belongs to the census professionals, guardians of the standards necessary for an accurate and fair census.

Consider an analogy. To defend the country, Congress funds a new fighter plane and then, when the plane becomes available, okays its use in a war zone. In between this purpose and this use is the production of the plane; this is assigned to scientists and engineers, insulated from congressional meddling. Political interference in the census is similar. It is the scientific production of statistics that advance the nation's

[1]This chapter goes to press on September 1, in the census year, 2020, with about 35 million households as yet uncounted and the non-response-follow-up, hard-to-count, and various critical data cleaning operations not completed. My primary use of the incomplete census is commenting on political interference with particular attention to how 2030 can be designed to minimize the politicization of the census, which in 2020 have interrupted the production function.

purposes and statistics that are then put to use. As discussed below, the boundaries between production, on the one hand, and purposes and uses, on the other hand, are blurry and have been especially so in the 2020 census.

This is the start of a definition authored a decade ago. I still believe it useful but for reasons made clear below, it doesn't fully deal with the blurry boundaries of today.

Political interference occurs when there is:

- "the politically motivated suppression of an agency's responsibility to offer its best judgment on how to most accurately and reliably measure a given phenomenon,
- the politically motivated decision to prevent an agency from using state-of-the-art science,
- the politically motivated insistence on preclearance of a major statistical product that is based on state-of-the-art science." (Prewitt 2010)

All three have played a role in 2020.

The 2020 Census After the 2020 census form was cleared by the Congress and was field tested – both standard practices – the Secretary of Commerce instructed the Bureau to add a question on citizenship status. The Bureau explained that it was too late for this; the census was in production. Further, the Bureau explained that adding the question would produce an intense legal battle that would attract extensive media coverage, largely negative, which could lessen census cooperation rates. The Bureau explained all of this to the Secretary, adding that it could produce a more accurate estimate of citizenship status using administrative records. The Bureau was unsuccessful in persuading the Secretary, who insisted that the data were needed for the Voting Rights Act.

As many readers will recall, the issue was litigated. There were multiple law suits, hundreds of media pieces, prolonged congressional hearings, and considerable distraction for the Bureau in the final weeks before the 2000 census was launched. The terminology was loudly, insistently political.

> What at first appeared to be a procedural and political fight turned into a made-for-TV movie in the wee hours before the Supreme Court ruled in response to the legal battle that this [instruction to add a citizenship question] triggered. The computer hard drives of a deceased Republican gerrymanderer ended up in the hands of a policy think tank as a gift from his estranged daughter. Tucked away in Powerpoints that acknowledged that said files never be public were a slew of materials that confirmed suspicions that the Trump Administration was not interested in upholding the Voting Rights Act, the cover it used in justifying the introduction of a citizenship question. Rather, as these explosive documents showed, a network of Republicans was invested in the citizenship question to strategically and intentionally disenfranchise non-citizens, the same Nativist agenda that helped undermine the [1920] census a century ago.
>
> In a surprise reversal from what all experts expected, the Chief Justice reprimanded the Trump Administration and told them to properly justify their need for this question in a legitimate fashion. The response was vicious, with career lawyers asking to be taken off the case because they couldn't properly defend their client, the Supreme Court telling them that they must stay on the case, and a game of musical chairs that has come to define the Trump Administration. Rather than justifying their ask, the Administration produced an Executive Order, demanding that the Census Bureau use administrative records to assign citizenship

> to every person in the census. In other words, they wanted their citizenship data, even if it wasn't going to come directly from a question on the census.
>
> The sensitivity of citizenship is profound. Civil rights groups, who have long struggled to get non-citizens and their family members to respond to the required census, encountered a wave of fear within their communities. No amount of promising people that Title 13 would prevent their data from being abused convinces people that participation is without risk. Surveys of attitudes about census privacy revealed stark differences between people based on race, ethnicity, and immigration status. With the 2020 count underway, we are already seeing significant race-based differences in participation. (Bouk and boyd 2020)

The Supreme Court ruling notwithstanding, the Administration was not finished with the citizenship issue. There was an unprecedented development a few months later, with the census underway. The Census Bureau has thousands of employees, only four of whom are political appointees and only one of these, the Director, is a presidential appointee. As is a long-standing practice the political appointees are carefully screened for experience and expertise in federal statistics matters. But not this time. The White House, without prior consultation with the Census Bureau Director, named two political appointees to senior positions at the Bureau, giving them titles that did not previously exist: Deputy Director for Policy and his Senior Advisor. The individuals chosen had extensive political experience, but thin to non-existent expertise in census-taking. The Commerce Department's Inspector General immediately requested documentation setting forth their suitability, which, months later, has not been produced. The American Statistical Association sounded an alarm.[2] The alarm was quickly followed by a large number of similar statements from professional associations, congressional inquiries, and extensive media coverage – all raising the specter of political interference. The New York Times (Michael Wines, June 23, 2020) quoted Terri Ann Lowenthal, a long-time and deeply informed consultant on census matters: "Their proximity to the director and lack of relevant expertise suggest a thinly veiled effort to interfere in the implementation and outcome of the 2020 census for the administration's benefit . . . It's hard to draw any other conclusion."

Representative Carolyn B. Maloney of New York, the Democratic chair of a House committee overseeing the bureau, called the appointees "political operatives"

[2] "The US Census Bureau produces statistics that are fundamental to our democracy, government, economy, and everyday life. The committed professional staff of the Census Bureau strives to ensure the bureau's products are accurate, reliable, and timely. The Census Bureau relies on the trust of its survey and census respondents, trust earned through objectivity, protection of personal information, transparency, and production of high-quality impartial data. The Census Bureau's addition of two political appointees to its top ranks undermines the work of the Census Bureau and federal statistical agencies because of the lack of transparency and justification, as well as the perception—if not reality—of improper political influence. We ask Director Dillingham to explain and provide rationale for the creation of the senior positions that includes their job duties in addition to the qualifications of the people appointed to the positions. In the absence of an explanation to US taxpayers for the need for these two senior political appointments, the ASA sees no justification for them." American Statistical Association Statement on the US Census Bureau Appointments, June 24, 2020

chosen by the Trump administration and accused officials of "using the census for political gain." I described the appointments as "a frightening development."

I had in mind the message being sent to the American people. Even if these appointments were, at the end of the day, politically harmless, the media storm was not harmless. Six decades ago, Japanese-Americans internment was facilitated by the Census Bureau. Every census since, the Bureau endlessly stresses that "you can trust us; your answers to the census cannot be used to harm you." And every census since, millions of Americans recall that that promise was violated in 1940. The violation cast a long shadow. The citizenship court case and now the hard-to-explain political appointments will cast a long shadow. As indicated above, the current census, by design, is highly visible. It matters if this visibleness has even a hint of political interference. The census *viewed* as a partisan tool has an uphill battle in securing public trust, particularly from population groups targeted – in this case, non-citizen immigrants and Hispanics. The census *used* as a partisan tool is a damaged census.

It turned out that the appointments were not harmless (and were not over, as later another political appointee was assigned to the Bureau, with even greater powers than his predecessors). With the census in full swing, and the mail out/mail back phase having achieved a self-response equal to that of earlier censuses (two-thirds of the population returning a form), the Bureau requested a schedule extension as it adjusted to the very disruptive Covid-19 pandemic. President Trump initially agreed to an extension of 4 months, commenting that this extension may be insufficient. The Bureau rearranged operations accordingly. A few weeks later the agreement evaporated and, in fact, the Bureau was instructed to complete its field work – then more than 56 million households yet to be reached – a month sooner. The Bureau had been told it was running in a marathon, only to find that it was actually a sprint.

The next census disruption was President Trump instructing the Census Bureau to separately count the non-citizens, in the process drawing a sharp distinction between the documented and the undocumented. His announced goal was to remove the undocumented from the apportionment numbers. This produced another media storm and another flood of legal challenges. This issue, by the way, has a several decade history and, thus far, has not received positive response from the Supreme Court,[3] though this does not prevent different legal strategies from trying again.[4] Irrespective of a Supreme Court ruling, experts consider Trump's instruction to be technically unfeasible. The Bureau cannot knock on doors and ask if any illegals live there. That leaves Administrative Records, which experts say is a bridge too far. *Science (*Mervis 2020a*)*, covering this story, cites the economist Amy O'Hara, who for years oversaw the agency's efforts to use administrative records: "I don't know what set of data sources the Bureau could identify for that purpose. And for the ones

[3] https://www.documentcloud.org/documents/7036772-State-of-New-York-Aug-14-2020-Amicus-Brief-of.html

[4] https://www.documentcloud.org/documents/7040829-State-of-New-York-Aug-19-2020-Memorandum-of-Law.html

they have, it's not clear how they would operationalize them." O'Hara pointed out that none of the existing government records was designed to do what Trump wants. Records with current citizenship status, for example, do not also indicate where the person is living. (The decennial census assigns each person to a specific address when it includes them in the overall count.)

Responding to these various interruptions and midway through the census, the Committee on Oversight and Reform of the House of Representatives called an emergency hearing to investigate whether the census itself was at risk from what was shaping up as an exceptionally politicized census. It featured four previous Census Bureau Directors, who unanimously urged that Congress extend the census schedule, as earlier agreed, into 2021. And we each pointed to the risks in what had become a rushed schedule, especially the likelihood of a sizeable undercount unevenly distributed across different demographic groups and geographic units.

My oral testimony included an unusual recommendation:

> How do I define a successful census? Easy – census numbers have specified purposes. The Bureau will know – as no other unit of the government can – if its numbers will accurately reapportion and fairly distribute federal funds for a decade. It knows that the census count is the denominator of every vital statistic we rely upon, whether it's the number of consumer prospects for a new business, the differential rate of infection across population subgroups in the pandemic, or disparities in the arrest rates across racial groups These numbers have a 10-year life.
>
> The Bureau not only knows these statistical facts, it knows the amount of damage that sub-standard numbers will inflict on society. It is not pretty – 10 years of homeless veterans because their hospitals are mislocated; 10 years of tropical storm disaster relief that is too little and too late because traffic congestion is underestimated; 10 years of poor planning by local school districts because they have flawed estimates of how many first graders are going to show up; and 10 years of misled Chambers of Commerce because predictions of population growth and characteristics were off base.
>
> The Bureau will not want to inflict that damage. It is too honorable, too scientific, too proud of its professional standards, too faithful to its constitutional duties. The Bureau will struggle with the enormous burden of whether to release substandard results.
>
> Recommendation: I urge the Congress to share that burden. Task a suitable independent institution – the National Academy of Sciences, or any apolitical and trusted institution of its choosing – to produce *predetermined quality metrics* that can assess if the final 2020 numbers reasonably match what the Bureau knows they should be. And if not, what steps the country should take.

This recommendation has not been acted on by the Congress and is not likely to be – though versions of it have been cited in the press, notably in *Science* (Mervis 2020b).

I note the recommendation here to emphasize the gravity of the various ways in which the Administration politicized the 2020 census, invited legal action, and,

more generally, complicated the non-response-follow-up phase of the census. It is this package that calls for such an unprecedented recommendation.

I return below to why political interference in 2020 should be taken into account when framing a sensible census for 2030.

18.2 From 1790 to 1990: Who/What Was the Face of the Census?

This is an odd question. Obviously, insofar as there is a face associated with the census, it is that of the census-enumerator. This was the experience of 100% of the American people for 18 censuses. For the 19th census (1970), the Bureau shifted to mail out/mail back as the primary data collection method, with enumerators used only for nonresponders.[5]

Soon, however, the mail out/mail back method faced an unexpected problem – a persistent declining response rate. In 1970, the self-response mail back captured 78% of the population. This dropped by 5% in 1980 and in 1990, another 10%. And over this period, the cost per household, in constant dollars, increased from $13 to $32. What was in store for 2000? It did not look good. Census Bureau statisticians initially arrived at an estimated 55% mail back (later adjusted to 61%, still a drop from 1990), and the likelihood of another sharp increase in costs. The Bureau was now dealing with a disgruntled Congress, telling the Bureau to figure out how to stop the response rate decline.

Which takes us to the "face-lift" or what became an unprecedented large-scale promotion of the census. An extensive paid advertising campaign ($165 million), first in census history, was used by the Census Bureau to urge mail back. It also established a partnership program, what was generally known as engaging trusted voices who volunteered to carry the census message to households across the country. The face of the census was less likely to be an employee of the Bureau, and more likely to be a school teacher, union leader, Mayor, employer, minister, priest, librarian, the chamber-of-commerce, leaders of race/ethnicity advocacy groups. If the measure of success is stopping the response rate from declining any further, bring out the champagne. Mail back in 2000 was 66%.[6]

[5] I skip many complexities here – long form/short form, update-leave for households lacking mail delivery, group quarters, military abroad, special arrangements for the hard-to-count, proxies and imputation as last resort.

[6] In fact, it was 78.4%, and, the final 1990 number was 75%. Why were the final numbers so much higher than what was initially announced? There are various factors. The most consequential is the length of time it takes for the Bureau to remove from the denominator the vacancies in the master address file. Not until this corrected denominator estimate is available is there an accurate response rate.

The strategy and its success were repeated in 2010, then again in the mail out/mail back phase of the 2020 census – despite the Covid-19 pandemic. (The final result for 2020 is not yet available.)

There was steady expansion of effort and costs across the three censuses. The 2010 paid advertising budget was, at $350 million, nearly double that of 2000. And, in 2020 what is called the "public education and outreach campaign" had a $500 million budget. The 2000 census had 140,000 partnering organizations; there were 255,000 in the next census and 380,000 in the present census. In all three censuses there were thousands of Complete Count Committees, working closely with local governments. The new face of the census had become a vast, volunteer labor force, loosely managed by the Census Bureau, and projecting itself into the census space not as enumerators but as neighbors, partners, and trustworthy messengers.

These numbers are impressive. Not impressive is that more advertising, more partners, more effort have only managed to hold the line. The self-response rate appears stuck in the two-thirds range.[7] It's an odd success we celebrate – from one decennial to the next, spend twice the dollars and double down on the effort but expect no improvement in self-response.

There was a parallel development, also new in census history. Historically all census operations were fully funded by the federal government. Of course, the majority of the funds in 2000 were still provided federally, but now states, cities, corporations, advocacy organizations, and complete count committees began to invest their own funds for census promotion. Private philanthropy gave support not to the census itself but to advocates for groups traditionally undercounted – children under five and indigenous people, for instance. Milwaukee invested $350,000 on a "I Will Count in Census 2000" campaign, one of several hundred cities that used their funds to build their census numbers. This occurred at the state level as well, led by California's $35 million investment. Such investments had two goals: to secure congressional seats and electoral college votes and to assure fair share of what was then $250 billion per year in federal funds distributed through formulas based on census counts. In 2000, I wanted to circulate a census poster picturing the IRS 1040 Form and the Census Form, noting that the former "Taketh Away" and the latter "Bringeth Back." The poster was rejected by wiser colleagues.

In any case, the Bringeth Back number by 2020 was *$1.5 trillion, annually*. New York City, for example, estimated one billion in federal funds every year for a decade based on its 2020 census count. In NYC, every low-income child missed in the census subtracts $2295 from the public-school budget; every lower-income worker missed subtracts $281 from job-training programs. With this federal funding at stake, we can hardly be surprised that NYC will dedicate its own funds to census promotion – $40 million, in fact. I noted above that California made an investment

[7] It is actually closer to three-quarters, but this cannot be determined until the denominator has been adjusted by removing vacant housing units. I believe that 2020 will be the last census afflicted with this misleading messaging. Third party sources have records on vacancies.

in 2000; two censuses later the investment had increased fivefold – nearly $200 million. Half of the States, on a per capita base, made similar investments.

This unplanned, unexpected consequence of the "face-lift" launched three censuses ago raises a troubling question. The Census Bureau is *totally* indifferent to the implication for red vs. blue states of any of its operations. It is only focused on providing a complete count of the complete country. Any operation or policy violating that standard results in an unfair distribution of federal funds, and unfairly reapportions the House of Representatives. That is not in the Bureau's game plan, and never will be.

However, the Bureau now finds itself linked to cities and states that are alert to the financial and political implications of how they promote, or don't, the census. A census in which half the states directly invest and half do not complicates the Bureau's basic fairness principle, and certainly risks sending signals that have no place in the census itself. They may, however, be useful for an advocacy group seeking a foundation grant or complete count committees in one state competing with committees in another.

At its most dangerous this messaging, not from the Census Bureau itself but still widely recognized as promoting the census, risks being misinterpreted. The public cannot distinguish between Census Bureau advertisements and similar looking promotional material from interested players. Has the Census Bureau, however inadvertently, created a condition in which individual states or cities making the biggest investments get the biggest awards – in congressional seats, electoral college votes, and federal funds? The federal funds are a fixed resource, distributed proportionally. If State A counts 100% of its population and State B counts 95%, funds that should have gone to State B will, instead, go to State A. Is the census for sale? I see more risks than benefits in the face-lift launched in 2000. I return shortly to the consequences of these two 2020 census features – political interference and a face-lift with some notable flaws – treating them as push factors in contrast with the next section, treated as a pull factor.

The face-lift had one other unexpected outcome. Among its partners are some very census-informed organizations – from academic circles (American Statistical Association, American Population Associations, Committee on National Statistics of the National Academies of Science, to name a few) and from advocacy organizations tightly focused on undercounts of demographic groups for which they take responsibility (Leadership Conference on Civil and Human Rights, Mexican American Legal Defense and Educational Fund, NALEO Education Fund, ACLU, and several dozens more). The leaders of this broad, impressive array of academic and advocacy organizations have four characteristics: when it comes to the census, they know their stuff; they completely understand that the best fix of the undercount is a census that counts everyone, which requires special efforts to erase the undercount; they are fierce defenders of the census and of the Bureau, and, when they have concerns, they are easily the strongest critics of the census and of the Bureau. The Bureau depends on them, but is also annoyed by them. This make for awkward partnerships.

18.3 New Data with an Eye on the 2030 Census

There are three data sources available for census use, each providing opportunities, though none likely to fully replace what the census offers.

(a) Government Administrative Record Data (AdRec). The Census Bureau has studied the possibility of using administrative records for three decades, coming to the conclusion that these data can, by 2030, be used in the basic count used for reapportionment and the distribution of federal funds. For these uses, all that is needed is an accurate count of how many people live in every household represented in the master address file. There is not yet consensus on how much to rely on these data, but that they can contribute to the basic count is unquestioned. This basic count does not include critical demographic statistics now collected on the decennial form, in particular, household composition and race/ethnicity. To be determined is whether these statistics can be saved with an enlarged and more frequent American Community Survey, using funds no longer required for the basic count.

(b) Data Linkage in Federal Statistics. A closely related asset is data linkage across the federal statistical system. "The public pays for and provides an incredible amount of data to governments and companies. Yet much of the value of this data is being wasted, remaining in silos rather than being shared to enhance the public good." (Groves and Neufeld 2017) There is strong pressure to link these data and good reason to expect major advances in this decade.

There are also frustrations with the barriers and difficulties. Evidently, however, difficulties can be conquered when the stakes are high, as demonstrated by the Covid-19 Household Pulse Survey, initiated by the Census Bureau in collaboration with five federal agencies producing data on the social and economic effects of the pandemic – employment status, spending patterns, food security, housing, physical and mental health, access to health care, and educational disruption. Statistics are provided at the state level and for the 15 largest Metropolitan Statistical Areas. The survey has a longitudinal feature that documented how household experiences changed during the pandemic and was expected to be in the field for 90 days, longer if the pandemic persists.

(c) Third-Party Sources. Nongovernmental sources include data collected by commercial transactions and social media, offering unprecedented frequency (daily credit card transactions) and granularity (location data from smart phones). The Dutch, for example, have smoothly moved their census entirely to government records and private sector data. In the United States, economic measurement experts are already producing studies linking administration records (taxes paid) and commercial data (bank account transactions) (Jarmin 2019).

Less advanced, but in active exploration are other sectors – agriculture, health, education, energy, and transport. In each sector there is attention to barriers and challenges: privacy, proprietary data, standardization, data security, broken trend lines, and trust in the government's numbers. In some sectors, progress will be quick

and impressive; in others, slow and frustrating. But the overall picture points to a future in which at least some non-survey data will add to what the census data now make available.

18.4 The 2030 Census, Early Steps

Every census in our history has been used as lessons learned for the next census – Federal Marshals on horseback, followed by continuous innovations that bring us today to the Internet. The Census Bureau and stakeholder interests, in partnership with academic statisticians and social scientists, are now from their respective vantage points in active discussion of a twenty-first-century information platform – with the multiple data sources summarized above as critical contributors.

The Census Bureau is designing a nationally integrated person-household-business-jobs frame. This integrated frame brings together the business register of establishments, the master address file of housing units, the Longitudinal Employer-Household Dynamics (LEHD) jobs frame, and a demographic frame of individuals. This universal frame can be augmented over the decade with multisource federal and private sector data sources. The Bureau's goal is to have this in place for the 2030 census. Independently, consulting with the Bureau at every step is an academic-led initiative, which includes representatives from commerce and advocacy groups. It is in the early stages of what is provisionally labeled: *Essential-Data Count of the Resident Population* coupled with a *21st Century Census Curated Information Platform*.[8]

Essential-data starts from the premise that a combination of the three data sources sketched above, on every April 1 that occurs in a year ending in zero, will enumerate nearly all of the persons resident in the United States, locating them at their normal place of residence. This count reapportions the House of Representatives and distributes federal funds. It will retain the key capacity of the census to allow all sample surveys – government, academic, media, and commercial – to use it as the gold standard. A number of possible designs are being considered. For example, based on lessons learned in 2020, for approximately 75% of the population in 2030, comfortable with the Internet, online responses might replace the mail out/mail back method with the remaining 25% being counted from a mixture of Administrative Records, other government surveys and commercial sources. Or, perhaps by 2030, Administrative Records will be positioned to provide more of the essential data, minimizing the need to contact every household. Any model is likely to need special attention to the hard-to-count population groups, which may require expert

[8]This initiative is housed in the Social and Decision Analytics Department, University of Virginia Biocomplexity Institute and is co-directed by Sallie Keller and the author, with initial funding provided by the Sloan Foundation. The labeling is provisional, and likely to be modified when it reaches the point of active consultation with numerous interested parties.

enumerators. We are confident that the essential data can, by 2030, be the foundation for a *sensible census* consistent with Fienbergian principles.

The second component, the Census Curated Information Platform will, first, help support the essential-data by connecting it where useful to other data sources. But the Platform has broader goals –demographic and geographic granularity, and temporality, especially for the American Community Survey. Much of what is brought together on this twenty-first-century information platform will be shaped and managed by a newly established Data Curator Department. This government unit – basically a redesigned Census Bureau – will assess the accuracy, coverage, privacy, and costs of commercially provided data products, making decisions about what can be successfully incorporated into the nation's statistical system. It will be responsible for data linkage across the federal statistical agencies and includes state-level statistical offices where appropriate.

These are sweeping statements of ambitious goals. They are being worked on by groups acutely aware of the various complexities and challenges – legal issues, privacy protections, preserving trend lines, creating new standards, cost effectiveness, and accurate data in the public's interest, dedicated to the public good.

18.5 The Push Factors, the Pull Factors

I noted above that visibility describes the three most recent censuses, especially their paid advertisement and partnership programs. Prior census were not invisible, but were experienced by each household independent of the next household, a quite different experience than the neighborhoods, towns, and cities that collectively mobilized around the census. This visibility is highly desirable, promoting the census as a shared civic experience, created by the founders and dedicated to the democratic principles outlined in the Constitution. Under optimum conditions, this conveys that the census belongs to the people, not to the government. The census produces numbers that Americans use to hold the government to account – numbers related to economic growth, public health conditions, social justice, educational achievement, and security. Characterizing the census this way was a goal of the face-lift and was successfully carried out in 2000 and 2010.

Through no fault of the Census Bureau, 2020 is a more mixed case – noting again that the census is still underway and faced with a number of uncertainties: court cases, possible schedule adjustments, more political interference, and the impact of Covid-19 on door-to-door follow-up. What is already apparent is the risk that the census is being experienced by a sizeable number of people as a tool of the political party in power. Whether true or not, some households feel that they are being targeted by the census, giving them reason to close the door to the census-taker. This partisan lens will be sharpened if the Supreme Court rules that undocumented residents are to be excluded from the apportionment count, though softened if the practice in place since the 1790 census is undisturbed.

It is likely though not certain that 2020 will leave another problematic message: perceived unfairness in how census counts indirectly[9] allocate federal funds geographically and demographically. When less well-counted states, counties, cities, and towns and different vulnerable population groups realize this unfairness, there will be press coverage and, again, legal action. Negative census publicity associated with political interference and what is interpreted as unfair fund allocation will cast a shadow over early planning for the 2030 census. This will push toward a census design that should minimize both political interference and unfair allocation. It will require laws and regulations much more robust than Title 13 (limited to issues of privacy and confidentiality).

New laws and regulations will no doubt draw on the Information Quality Act and the OMB issued *M-19-15* (April 24, 2019) on Improving Implementation of the Information Quality Act (IQA). The OMB memo summarizes the three core responsibilities of each federal agency under 2002 guidelines:

1. Agencies must embrace a basic standard of quality and consider quality in their information dissemination practices.
2. Agencies must develop information quality assurance procedures that are applied before disseminating information.
3. Agencies must develop an administrative mechanism for affected parties to request that agencies correct information of inadequate quality, with an appeal process and annual reports to OMB.

Further: The *Guidelines* explain that quality encompasses utility, integrity, and objectivity. Utility refers to the data's utility for its intended users and for its intended purpose. Integrity refers to the data's security. Objectivity refers to whether the disseminated information is accurate, reliable, and unbiased as a matter of presentation and substance.

The focus on the information's usefulness is critical; the *Guidelines* recognize that "information quality comes at a cost" and "that some government information may need to meet higher or more specific quality standards than those that would apply to other types of government information," depending on the information's expected use. The touchstone is "fitness for purpose"; information destined for a higher-impact purpose must be held to higher standards of quality. The census is held to the highest standards. It has "clear and substantial impact on important public policies or important private sector decisions" and publishes its own detailed Statistical Quality Standards.[10]

The challenge, of course, is to insulate these standards and guidelines that are specific to the production of census statistics from partisan motivated political

[9]Indirectly because funding formulae are based on Population Estimates and, to a lesser extent, the American Community Survey – both depend on the basic census count to correct for sampling error.

[10]For material on the Information Quality Act and related Guidelines, along with his helpful explanations that I have borrowed in this section, I am indebted to Andrew Reamer, Research Professor, George Washington Institute of Public Policy, George Washington University.

interference. Moreover, when political interference or non-controllable external conditions (a pandemic, national disasters) prevent the Bureau from meeting its quality standards in the time frame initially anticipated, schedule adjustment (not cutting corners) has to be the common-sense practice.

This challenge, too, is a push factor and is reason to build a census design for 2030 which incorporates IQA guidelines and standards, which has law clearly protective of the obligation of the census to meet well-defined quality standards, and which has predetermined metrics allowing nonpartisan review of whether those standards are realized.

The pull factors start from practices that date to 1790. Build the next census on lessons learned from the prior one. In 2030 this means adding to the census:

Data that are curated for accuracy, coverage, non-intrusiveness, privacy protection, and related positive features.

Data that can escape any negative judgments associated with 2020, or any prior censuses and that are, in fact, nonpartisan.

Data that have been examined from the perspective of the role of third parties, especially when they introduce independent resources for the benefit of demographic groups or geographic units that might be (mis)interpreted as rewarding some and ignoring others.

Add to this metrics that clearly define and measure must-meet quality standards – the IQA contribution.

Add a redesigned communication strategy that is fully transparent about the purposes, the production, and the uses of the census. That is persuasive on confidentiality and privacy. That matches the cost of the new design to the benefits provided, whether local, national, or anything between. That describes the new design as "the people's census" and means it. This last is critical. A census viewed as a tool of the "surveillance state" is a big step backward. The big step forward is a census viewed as indispensable to the health of America's society and economy and that strengthens its democracy – which means data designed to hold accountable whoever is in power.

To return to the Fienbergian principle – every census travels on its uniquely diffuse, interconnected and highly nonlinear pathway, and, every census, perfection not being an option, searches for the most sensible methods for navigating that pathway. This will again be the case in 2030.

Postscript – From Mid-September (2020) to Late April (2021)

The goal of this chapter was to outline how the 2030 Census should build on lessons learned from the previous census. I started the chapter on the assumption that by mid-September 2020, I could comment with confidence on my goal. However, by that date, the covid pandemic and the Trump administration had upended matters. Nonetheless, the publishing schedule required submission. Subsequent delays, then,

postponed publishing. And today, 35 weeks later, the editors received permission for this 500-word update. That opportunity allows comment on four relevant issues.

(a) Internet response rates. The Internet proved to be an efficient way to enumerate the population, easily on a level with the mailout/mailback method of recent censuses. Item nonresponse rates, however, were greater than expected. The Bureau is well positioned to study why and to strengthen an Internet-based census in 2030.
(b) Administrative records. There was also advance in using administrative records in the nonresponse follow-up phase of the census. I expect the Bureau to increase its use of administrative records in 2030, allowing it to reduce reliance on imputation and proxy response methods. This will result in higher quality statistics for harder to count population groups.
(c) Unprecedented quality control arrangements. The 2020 operations were disrupted by three features: the pandemic; political interference by the Trump administration; floods and fires at unusual levels, caused by climate change that are likely to be even more disruptive in 2030. The combination of these features led to serious concerns about data quality in 2020, even questioning whether the census would be fit for purpose. For a period, there was much anxiety among the census stakeholders, eventually replaced by unprecedented cooperation between the Bureau and informed stakeholders. In ways not previously imagined, with the census still in the field, the Bureau sought help from the Jasons, the National Academy of Sciences, and the American Statistical Association. This cooperation, I believe, will persist. Jointly designed quality control methods will be a feature of the 2030 census design, which will be made public. In time, the results of the quality checks – before, during, and after field operations – will be fully transparent. The public will come to see the application of quality controls as no less important than the already very visible enumeration census phase. The public will be invited to take notice of errors and their correction. Media coverage will be welcome. Trust in the census will follow.
(d) Differential privacy (DP) is a well-tested method that masks personal identifiers while still reporting the aggregated census statistics needed by federal, state, and local governments, by commercial institutions, by the technology industry, by academic researchers, and much more. DP is currently hostage to the legal system, and likely to reach the Supreme Court. If the Bureau wins in court, DP will be a model for privacy protection in 2020 and future censuses. If it loses, the redistricting data files will be delayed by months – prompting widespread anger toward the Bureau. It gets worse. There will be privacy breaches engineered by hackers equipped with computer power not previously available, seriously damaging the census counts. It would take at least a decade to climb out of that hole.

Final note, in the spirit of transparency. I was appointed Special Advisor to the Census Director on April 15, 2021.

References

Bouk, Dan and dana boyd, (2020) "Democracy's Data Infrastructure: The Entanglement of Politics and Science." To be presented at "Data and Democracy Symposium," Columbia University, October 15–16, 2020.

Fienberg, Stephen. The Federal Research Portfolio: Capitalizing on Investments in R&D. Statement to the Committee on Commerce, Science and Transportation U.S. Senate July 17, 2014 p. 3.

Groves, Robert M. & Adam Neufeld. *Accelerating the Sharing of Data Across Sectors to Advance the Common Good,* Beeck Center, Georgetown University. 2017

Jarmin, Ron S. Evolving Measurement for an Evolving Economy: Thoughts on 21st Century US Economic Statistics. *Journal of Economic Perspectives*. Volume 33, Number 1—Winter 2019—Pages 165–184.

Prewitt, Kenneth. "What is *Political Interference* in Federal Statistics?" in Prewitt (ed), *The Federal Statistical System: Its Vulnerabilities Matter More Than You Think*, Annals of the American Academy of Political & Social Sciences, September, 2010.

Mervis, Jeffery. "Can the Census Bureau actually meet Trump's demand to count noncitizens?" *Science* 07 Aug 2020a: Vol. 369, pp. 611.

Mervis, Jeffrey. "Will the Census be good enough, and how soon will we know? *Science* 24 Aug 2020b. Posted in Science and Policy and Scientific Community doi:https://doi.org/10.1126/science.abe4799

Chapter 19
Perturbed M-Estimation: A Further Investigation of Robust Statistics for Differential Privacy

Aleksandra Slavkovic and Roberto Molinari

19.1 Introduction

We live in a world of continuous data collection, storage, and sharing, with much of those data being sensitive, making data privacy a highly relevant societal topic.[1] Steve Fienberg has recognized the importance of data privacy and confidentiality and crucially the role that statistical science must play in this context. He had argued that the right methodology for collecting and sharing of sensitive data should rely on statistical principles of sampling, estimation and modeling, transparency of masking procedure, and the dualities of the data utility and the disclosure risk. Steve argued for these guiding principles in many congressional and government testimonies and followed them in numerous scholarly contributions on the topic of data privacy and confidentiality. In his first technical contribution in this area, he proposed a bootstrap-like approach for creating synthetic data, similar to the current synthetic data methodology that relies on multiple imputation (Fienberg (1994)). Here, we highlight a few additional representative publications of his—

[1] See, for example, https://www.nytimes.com/interactive/2019/opinion/internet-privacy-project.html.

This work was in part done when Molinari was a Lindsay Visiting Assistant Professor at Penn State University

A. Slavkovic (✉)
Department of Statistics, Penn State University, State College, PA, USA
e-mail: abs12@psu.edu

R. Molinari
Department of Mathematics and Statistics, Auburn University, Auburn, AL, USA

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_19

for example, see Fienberg and Steele (1998) on perturbation of categorical data, Duncan et al. (2001) for general disclosure principles and links to information loss, Trottini and Fienberg (2002) on Bayesian modeling of disclosure risk, Fienberg and Slavkovic (2005) on links between privacy-preserving data mining and contingency table releases, Fienberg et al. (2008) on distributed regression analysis and secure multi-party computation, Fienberg et al. (2010) on data privacy links to algebraic statistics and log-linear models, Hall et al. (2011) on how to perform distributed regression using homomorphic encryption, Wang et al. (2016) on KL-privacy and its links to differential privacy, and Lei et al. (2018)) on model selection under differential privacy; for a more comprehensive list, see Slavkovic and Vilhuber (2018).

Statistical data privacy, traditionally referred to as statistical disclosure limitation or control (SDL or SDC), is the branch of statistics concerned with limiting identifying information in released data and summaries while maintaining their utility for valid statistical inference. It has a rich history for both methodological developments and applications for "safe" release of altered (or masked) microdata and tabular data (see Dalenius (1977), Fienberg and Slavković (2011), Hundepool et al. (2012), Willenborg and De Waal (1996), and references therein). Besides traditional methods such as suppression and aggregation, many modern methods rely on sampling and modeling, such as synthetic data (e.g., Reiter (2005), Rubin (1993), and Snoke et al. (2018)), and aim to frame data privacy as a statistical problem that requires treating both the data utility and the disclosure risk as random variables. However, they often fall short of allowing for the transparency of masking procedures, which is important in order to achieve the right statistical inference, not the individual identification. Furthermore, the onslaught of big data has presented new challenges for traditional statistical data privacy methodology, and the so-called reconstruction theorem (e.g., see Dinur and Nissim (2003) and Garfinkel et al. (2018)) has identified a flaw in a probabilistic notion of disclosure as proposed by Dalenius (1977). Many practical examples have demonstrated increased privacy risk from the released data or summaries in presence of other "auxiliary" data that were previously either not considered or simply were not as readily accessible; see Dwork et al. (2017) for a survey of such attacks, and recent claims related to issues with the U.S. Census data (Abowd (2018)).

Differential privacy (DP) has emerged from theoretical computer science with a goal of designing transparent privacy mechanisms/methods with mathematically provable disclosure risk in the presence of adversaries with arbitrary priors, unlimited side information, and unbounded computational power, e.g., see Dwork et al. (2006) for the original proposal and Slavkovic (2013) and Slavkovic and Vilhuber (2018) for Steve's role in bringing computer scientists, statisticians and practitioners together to forge the new directions of formal privacy. Differential privacy guarantees that whether an individual is in a database or not, and the results of a DP method should be similar in terms of their probability distribution; this limits the ability of an adversary to infer about any particular individual (unit) in the database and at the same time allows the data analyst to carry out inference on a distribution not sensitive to outliers. DP quantifies the so-called privacy-loss budget,

ϵ, to how much the answer to a question or statistic is changed given the absence or presence of the most extreme possible person in the population.

Understanding the above risk-utility trade-offs under formal privacy constraints such as those imposed by DP and linking them to fundamental statistical concepts has been one of the key recent research threads in data privacy, as there are serious implications on how we carry valid statistical inferences if data are to be shared under the DP framework. Wasserman and Zhou (2010) were among the first to underline these links focusing on density estimation and offering a statistically flavored interpretation of DP. Over the past decade, numerous works have explored these links in different settings including parameter estimation (Duchi et al. (2013), Smith (2011)), hypothesis testing (Awan and Slavković (2018), Canonne et al. (2019), Gaboardi et al. (2016), Vu and Slavkovic (2009), Wang et al. (2015b)), confidence intervals (Karwa and Vadhan (2017)), model selection (Lei et al. (2018)), principal component analysis (Awan et al. (2019), Chaudhuri et al. (2013)), network data (Karwa and Slavković (2016)), and functional data analysis (Hall et al. (2013), Mirshani et al. (2019)), to name a few.

Dwork and Lei (2009) were the first to investigate links between differential privacy and robust statistics (e.g., see Huber (2011)). One of the fundamental concepts behind differential privacy is to define the maximum amount of change a query or statistic can undergo (*sensitivity*) when one row in the database is added or replaced by another arbitrary row. Once this sensitivity is defined, differentially private mechanisms add a proportional amount of noise in order to hide whether a change in output is due to a change in row or to the added noise; the amount of noise grows with the sensitivity of the query/statistic. Robust statistics aims at limiting the impact that an extreme observation can have on statistical estimation and inference. In this sense, using robust statistics can deliver statistics and/or analyses with bounded sensitivity. Based on this property, robust statistics can bound the (DP) sensitivity and therefore reduce the amount of noise required to ensure privacy and consequently improve utility of the private outputs. Dwork and Lei (2009) explore these links and make use of robust estimators (e.g., median and interquartile range) as a starting point for releasing differentially private estimators based on a Propose-Test-Release algorithm for interactive queries, while Lei (2011) proposes the use of (bounded) M-Estimators applied to differentially private perturbed histograms in order to enhance the utility of statistical estimations under DP. Chaudhuri and Hsu (2012) study convergence rates of differentially private approximations to statistical estimators and propose the use of (bounded) M-Estimation within the exponential mechanism. Most recently, Avella-Medina (2019) proposed a statistical inference framework where noise is added to the M-Estimators in order to ensure privacy.

In this chapter, we investigate the use of functions with bounded derivatives, such as those used for M-Estimation in robust statistics, within the objective perturbation mechanism (OPM), originally proposed in Chaudhuri et al. (2011) and modified by Kifer et al. (2012). We propose a new convex and bounded function called the robust hyperbolic tangent (RobHyt) function which can be used to produce a bounded M-Estimator with adequate statistical properties which itself can be easily integrated within the OPM framework. More specifically, we study the statistical

consistency of this bounded non-private M-Estimator. In the non-private setting, the choice of the bounding parameter (that is of the tuning constant) is usually made based on the asymptotic properties of the non-private M-Estimator. However, when integrating the M-Estimator in the OPM, the tuning constant can be used as a parameter regulating the trade-off between statistical efficiency and the amount of noise added for privacy. Thus, the non-private statistical properties of the proposed M-Estimator can provide a first rule to define this tuning constant when employed within the OPM. Based on the preliminary results, the resulting privacy mechanism, which we name the "Perturbed M-Estimation" mechanism, can greatly improve the utility of differentially private outputs while preserving the same level of privacy.

This chapter is organized as follows. In Sect. 19.2, we provide a summary overview of important definitions for differential privacy and then make links between these definitions and the framework of robust statistics. In Sect. 19.3, we briefly introduce M-Estimation theory and propose the RobHyt function to deliver a bounded M-Estimator. This estimator is then used to build Perturbed M-Estimation by integrating it within the OPM. In Sect. 19.4, we study the performance of the proposed method by using both the simulated and real-data examples, particularly focusing on gains in statistical utility in comparison to some existing methods. Finally, Sect. 19.5 concludes and provides possible future avenues of research in the proposed direction.

19.2 A Robust Parametric View of Differential Privacy

The basic idea behind differential privacy is to protect the privacy of an individual in the worst case scenario where an adversary is in possession of the data of all the other individuals in a database except for those of this particular individual. The release of differentially private data or analyses requires mechanisms (methods) to add noise, directly or via sampling, in such a way that an output of these mechanisms is (nearly) equally likely to occur whether or not an individual is included in a database. More formally, a mechanism $\mathcal{M}(\cdot)$ is defined to be (ϵ, δ)-differentially private if it respects the following condition:

$$\mathbb{P}[\mathcal{M}(D) \in S] \leq e^{\epsilon}\mathbb{P}[\mathcal{M}(D') \in S] + \delta,$$

where D and D' are two databases that differ in one row (i.e., neighboring databases) and S is a set of outputs belonging to the range of $\mathcal{M}(\cdot)$. This definition implies that, for the same output S, the probability of observing it given the database D' is within an "ϵ-range" of the probability of observing the same output given the database D plus an exception δ. This must hold for all measurable sets S and all pairs of databases D and D' that differ in one entry. The quantities ϵ and δ should be small. When $\delta = 0$, then we have the so-called pure differential privacy, while the presence of a small δ (e.g., decreasing polynomially with the sample size n) allows for the data of (some) individuals to be released entirely with low probability δ. The value

ϵ is the privacy parameter or the privacy-loss budget. Smaller values correspond to more privacy, but as it approaches infinity, there is no privacy guarantee.

DP mechanisms most often introduce some form of noise in the analysis (or data) or distort the problem definition underlying a query or estimation procedure in order to cover the variation due to the change in one individual's data. To determine the degree of "variation" and thus the amount of noise to be added, different notions of *sensitivity* have been proposed and discussed in the privacy literature. The *global sensitivity* is defined as

$$GS_f = \max_{D,D'} \| f(D) - f(D') \|,$$

where $f(\cdot)$ is any function (query, estimator, etc.) and this measure captures the maximum extent to which the function $f(\cdot)$ can vary between all possible combinations of neighboring databases. The *local sensitivity*,

$$LS_f = \max_{D'} \| f(D) - f(D') \|,$$

fixes the database of reference D and determines the maximum variation considering all other possible neighboring databases D'. Other notions of sensitivity exist and other norms to determine them are also considered (see, e.g., Dwork et al. (2014), Awan and Slavković (2020), and references therein). These quantities are important for improving the risk-utility trade-offs. The smaller the sensitivity, the smaller the amount of noise is required for privacy, which typically leads to better *utility* of the outputs and possibly better management of the privacy-loss budget. In Awan and Slavković (2020) and in this chapter, we show that for the same privacy cost, we gain better utility and more usefulness of data if we propose ways of adjusting the sensitivities of the outputs.

The above notions of sensitivity, which measure possible variations of estimating functions, is strongly related to the notions underlying the framework of robust statistics. The next section highlights the similarities between these notions and justifies the investigation of robust statistical tools for the purposes of achieving differential privacy since, by reducing the sensitivity of estimators (functions $f(\cdot)$), robust statistical approaches can require less noise in order to deliver more useful differentially private outputs.

Links with Robust Statistics

As highlighted above, the notion of differential privacy and the concepts based on which differentially private mechanisms are proposed are intrinsically linked with notions of function (query) sensitivity, centered on the space of neighboring datasets. Robust statistics, on the other hand, focuses on the sensitivity of the function with respect to the quantity it is meant to compute (estimate), which, in

general, corresponds to the output that would be observed if the function were applied to the entire population of reference. This is formalized within robust statistics by using a parametric framework where the population is described by an assumed parametric model $F_{\boldsymbol{\theta}}$, with $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ being the parameter vector defining the model. The goal in this setting is to estimate the parameter vector $\boldsymbol{\theta}$ (e.g., the regression coefficients and residual variance) through an estimator (function) with appropriate statistical properties.

However, the framework of robust statistics postulates that although we assume a model $F_{\boldsymbol{\theta}}$ for our data, this is at best an approximation to reality and what we actually observe is

$$F_\lambda = (1-\lambda)F_{\boldsymbol{\theta}} + \lambda G, \tag{19.1}$$

for small $\lambda > 0$ and with G being an unspecified "contamination" model (see e.g., Hampel et al. (1986), Huber (2011), and Maronna et al. (2019)). In this paradigm, the goal of an estimation and optimization problem would be to recover the value of $\boldsymbol{\theta}$ as best as possible by reducing the impact of the unknown model G. More specifically, let us define an estimator as a functional $T(F)$, where F is a general notation for a model (e.g., empirical or parametric). When we apply this functional to F_λ, we would want to obtain a good estimate that is output for the true value of $\boldsymbol{\theta}$, but this will depend on the properties of the functional. In order to determine these properties when observing F_λ, the notion of influence function (IF) was introduced (see Hampel (1974)) and is defined as follows:

$$\mathrm{IF}_T(z_0, F_{\boldsymbol{\theta}}) = \lim_{\lambda \downarrow 0} \frac{T((1-\lambda)F_{\boldsymbol{\theta}} + \lambda \Delta_{z_0}) - T(F_{\boldsymbol{\theta}})}{\lambda},$$

where Δ_{z_0} is a point-mass distribution in an arbitrary point z_0 which plays the role of the model G. In general terms, this quantity can be interpreted as the impact that an infinitesimal amount of contamination can have on a given functional T.

The IF is therefore an important notion in robust statistics since it can be used as a measure to understand the possible extent of asymptotic bias with respect to $\boldsymbol{\theta}$ introduced by the presence of G. An additional measure that is based on the IF is given by the gross error sensitivity (GES) defined as

$$\gamma(T, F_{\boldsymbol{\theta}}) = \sup_{z_0} \left| \mathrm{IF}_T(z_0, F_{\boldsymbol{\theta}}) \right|.$$

The GES measures the maximum impact that any point-mass distribution Δ_{z_0} can have on the estimator T. Then, an estimator T is defined as being (B-)robust if the GES is bounded, that is, if the IF is bounded—which is a sufficient condition.

Taking a deeper look at these definitions, one can see the similarities with the sensitivity definitions used for differential privacy. Let $\lambda = \frac{1}{n}$, and assume that the empirical distribution F_n (an estimator of $F_{\boldsymbol{\theta}}$) fully characterizes the database D.

Then, we could reformulate the contamination model from Eq. (19.1) as

$$F_\lambda = \frac{n-1}{n} F_n + \frac{1}{n} z_0,$$

which resembles another definition in robust statistics, i.e., *the sensitivity curve*. These types of robust measures resemble the definition of *local sensitivity* since they would measure the impact of one observation, z_0, on the database (model) of reference D (i.e., $F_{\boldsymbol{\theta}}$). The notion of *global sensitivity*, on the other hand, would require a contamination model where all possible versions of $F_{\boldsymbol{\theta}}$ are considered.

As mentioned earlier, other works have explored the similarities of differential privacy notions with those of robust statistics highlighted above. For example, after defining the above robustness measures, Chaudhuri and Hsu (2012) use the notion of GES to deliver convergence rates for differentially private statistical estimators, while Avella-Medina (2019) uses this measure to calibrate the additive noise to deliver differentially private M-Estimators. In the next sections, we explore another approach, suggested but not studied in Chaudhuri and Hsu (2012) and Avella-Medina (2019), where we investigate the use of *bounded* M-Estimation for differentially private estimation and prediction using the OPM. While empirical risk minimization, that objective perturbation is built on, can be classified as M-Estimation, it is not straightforward to integrate the standard bounded functions for M-Estimation as is. To address this problem, we propose a modified OPM that we call the Perturbed M-Estimation mechanism. More specifically, we propose the use of a new convex objective function, *RobHyt*, defining a bounded M-Estimator for which we first study its non-private statistical properties and convergence rates which then lead to its integration in a differentially private setting.

19.3 Perturbed M-Estimation

In this section, we present the Perturbed M-Estimation mechanism designed by integrating a new bounded function into the OPM of Kifer et al. (2012), thereby improving the overall utility of the differentially private output. Recall that the goal of robust statistics is to bound the impact of outlying observations on the output of an analysis. A popular class of estimators for this purpose is that of M-Estimators defined as

$$\bar{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho(\boldsymbol{\theta}; d_i), \tag{19.2}$$

where $\boldsymbol{\theta}$ is a parameter of interest we aim to release, $\rho(\cdot)$ is a convex loss function, and $d_i \in D \in \mathcal{D}^n$ is the ith row of a database with independent rows. In this form, the class of M-Estimators corresponds to the notion of empirical risk minimization. However, in order for the resulting estimator $\bar{\boldsymbol{\theta}}$ to be robust, we require the derivative

of the loss function to be bounded. The IF of an M-Estimator is given by

$$\mathrm{IF}_T(z_0, F_{\boldsymbol{\theta}}) = -\psi(z_0, T(F_{\boldsymbol{\theta}}))\, B(T(F_{\boldsymbol{\theta}}), \psi)^{-1},$$

where $z_0 \in \mathbb{R}$ is an arbitrary point, $\psi(z) = \partial/\partial z\, \rho(z)$ and $B(T(F_{\boldsymbol{\theta}}), \psi) = \partial/\partial \boldsymbol{\theta}\, \mathbb{E}[\psi(\boldsymbol{\theta}; z)]$ (see Hampel et al. (1986)). Thus, the IF of an M-Estimator is bounded if the ψ-function is bounded, which justifies why, in many cases within the robust literature, M-Estimators are also expressed directly with respect to their derivative as follows:

$$\bar{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argzero}} \frac{1}{n} \sum_{i=1}^{n} \psi(\boldsymbol{\theta}; d_i).$$

The class of maximum-likelihood estimators (MLEs) can be represented as M-Estimators, where $\rho(\cdot)$ would correspond to the negative log-likelihood and $\psi(\cdot)$ its derivative. But the MLE is not robust since, in general, the corresponding ψ-function is unbounded with respect to the data. Different functions have been proposed for $\rho(\cdot)$ in order to bound $\psi(\cdot)$, such as the Huber and Tukey Biweight functions (see, e.g., Hampel et al. (1981) and Maronna et al. (2019)). These functions, along with other bounded functions commonly used for robustness purposes, implicitly or explicitly assign weights to the residuals or score functions defined by the minimization problem thereby downweighing observations that lie far from the "center" of the assumed distribution of the residuals F. However, these functions typically have symmetric weights and can therefore be asymptotically biased with respect to the distribution of the residuals F (for example, if the latter is asymmetric). Hence, a correction factor is often added for Fisher consistency (e.g., see Huber (2011) and Cantoni and Ronchetti (2001)), which depends on the model F and the chosen bounded function $\psi(\cdot)$.

The definition of an estimator as an M-Estimator has additional advantages from a point of view of *parametric* statistical inference. Under a set of regularity conditions on the properties of the $\psi(\cdot)$ function and the parameter space, the asymptotic distribution of M-Estimators (see Hampel et al. (1986), Mises (1947)) is

$$\sqrt{n}(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\theta}_0$ represents the true parameter vector we aim to estimate, and

$$\boldsymbol{\Sigma} = \boldsymbol{M}_{\psi}(\boldsymbol{\theta}_0)\, \boldsymbol{Q}_{\psi}(\boldsymbol{\theta}_0) \boldsymbol{M}_{\psi}(\boldsymbol{\theta}_0)^T$$

is the asymptotic covariance matrix where

$$\boldsymbol{M}_{\psi}(\boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}[\psi(\boldsymbol{\theta}; d_i)]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0},$$

and

$$\boldsymbol{Q}_{\psi}(\boldsymbol{\theta}_0) = \mathbb{E}[\psi(\boldsymbol{\theta}_0; d_i)\psi(\boldsymbol{\theta}_0; d_i)^T].$$

Assuming one can define an appropriate M-Estimator for a given problem, it would be possible to use these properties to perform statistical inference thereby allowing for different parametric tests.

With respect to the use of M-Estimation for the purposes of differential privacy, as mentioned, in this work we aim to integrate the robust $\rho(\cdot)$ functions within the OPM. More specifically, the OPM requires computing bounds on the first and second derivatives of the objective (loss) function $\rho(\cdot)$ so that the adequate amount of noise can be added to this objective function to ensure privacy. In order to compute these bounds, we first propose a new specific function $\rho(\cdot)$ with bounded derivative $\psi(\cdot)$, delivering a robust M-Estimator that relies on a certain tuning constant. When this M-Estimator is used within the OPM, the tuning constant plays a role in determining the bounds of the above-mentioned derivatives and, consequently, plays a role in the amount of noise added for privacy. The following sections present the proposed bounded function, i.e., RobHyt, and the statistical properties of the resulting non-private M-Estimator (Sect. 19.3), and based on this, we then integrate this estimator within the OPM (Sect. 19.3) to obtain the proposed Perturbed M-Estimation mechanism.

The Robust Hyperbolic Tangent Function

The ρ-functions that are usually employed for robustness purposes are either non-convex (e.g., Tukey Biweight) or piecewise (and/or non-continuously differentiable) functions (e.g., Huber), which either make them unusable within the OPM or can make the computation of the required sensitivity bounds and/or asymptotic properties more complicated. There exist other smooth (and strongly convex) functions, such as the Pseudo–Huber loss function, but, given similar complexities in computing sensitivity bounds, we choose to address these issues by adapting the hyperbolic tangent (tanh) function (see Hampel et al. (1981)) to deliver a bounded function for M-Estimation. The tanh function has nice properties since it is (i) continuously differentiable, (ii) defined over the entire real line, and (iii) bounded between $[-1, 1]$ making it a good candidate for robustness purposes and for the derivation of the required sensitivity measures for the OPM. Given these properties, in Definition 1, we propose to modify this function by parametrizing it with a tuning constant $k \in \mathbb{R}^+$ that guarantees robustness when $k < \infty$ and converges toward the L_2-loss function when $k \to \infty$, similarly to the Huber loss function. To the best of our knowledge, although various modifications of the hyperbolic functions have already been proposed and used for robust optimization (e.g., see Chen et al. (2017) and Shen et al. (2019), to cite some recent work), we are not aware of a similar

parametrization of this function in either the statistical or the computer science literature to date.

Definition 1 The robust hyperbolic tangent (RobHyt) function is defined as follows:

$$\rho_k(z) := \frac{k^2}{2} \log\left(\cosh\left(\frac{2}{k}z\right)\right),$$

where $k \in \mathbb{R}^+$.

By definition, the proposed RobHyt function is convex with respect to its argument and has derivative given by

$$\psi_k(z) := k \tanh\left(\frac{2}{k}z\right),$$

which is bounded between $[-k, k]$. Hence, this function can be employed as a bounded function for robust M-Estimation since, as long as we choose $k < \infty$, we have that $\psi_k(z)$ is bounded and consequently so is the IF of the resulting M-Estimator.

Remark 1 The RobHyt function has the following important property:

$$\lim_{k\to\infty} \rho_k(z) = z^2.$$

Given the above definition and remark, this function can be seen as a smooth and differentiable-everywhere version of the Huber loss function (similarly to the Pseudo-Huber loss). Keeping this in mind, we next consider an M-Estimator based on the commonly used L_2-loss function, i.e.,

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} s(\boldsymbol{\theta}; d_i)^2, \tag{19.3}$$

where $s(\boldsymbol{\theta}; d_i)$ is a score function such that under the true model we have that $\mathbb{E}[s(\boldsymbol{\theta}; d_i)] = 0$. An example is given by

$$s(\boldsymbol{\theta}; d_i) := y_i - \eta(x_i^T \boldsymbol{\theta}),$$

which represents the non-scaled score function for a generalized linear model (GLM), where y_i represents the response variable, $x_i \in \mathbb{R}^p$ a vector of covariates, and $\eta(\cdot)$ a link function defined by the family characterizing the appropriate GLM model (see Nelder and Wedderburn (1972) and Cantoni and Ronchetti (2001)). If we plug this score function, or any MLE score function corresponding to the derivative of the log-likelihood function, into (19.3), then it is straightforward to see that the

estimator $\tilde{\boldsymbol{\theta}}$ corresponds to the MLE. This definition is particularly relevant since, based on Remark 1, it is also straightforward to see that the proposed M-Estimator

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \rho_k(s(\boldsymbol{\theta}; d_i)) \tag{19.4}$$

tends to the MLE as $k \to \infty$, in the same way as the Huber loss function.

In the robust statistical framework, one chooses a fixed tuning constant k based on the desired level of robustness and asymptotic efficiency with respect to the standard (non-robust) estimator. To do so, one usually requires an estimate of scale for the score function $s(\boldsymbol{\theta}; d_i)$, which could eventually be also obtained in a differentially private manner. If we let k diverge with n, thereby defining the sequence $k_n \in \mathbb{R}^+$, the estimator in (19.4) will inherit all the optimal properties of the MLE in terms of statistical accuracy based on the following assumptions (see Newey and McFadden (1994)):

(A1) The parameter space $\boldsymbol{\Theta}$ is compact.
(A2) $\mathbb{E}[s(\boldsymbol{\theta}; d_i)^2]$ is uniquely minimized in $\boldsymbol{\theta}_0$.
(A3) $\mathbb{E}[s(\boldsymbol{\theta}; d_i)^2]$ is continuous.
(A4) $1/n \sum_{i=1}^{n} \rho_k(s(\boldsymbol{\theta}; d_i))$ converges uniformly in probability to $\mathbb{E}[s(\boldsymbol{\theta}; d_i)^2]$.

While assumption (A1) is a standard regularity condition which can eventually be replaced by other (model-specific) constraints, assumptions (A2) and (A3) are generally verified when considering the MLE. We now state our key result on the statistical consistency of the proposed estimator in (19.4).

Proposition 1 *Under assumptions (A1)–(A3) and assuming* $s(\boldsymbol{\theta}; d_i) = \mathcal{O}_p(1)$*, for all* $k_n \in \mathbb{R}^+$ *such that* $k_n \to \infty$ *as* $n \to \infty$*, we have that*

$$\hat{\boldsymbol{\theta}} \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0.$$

This result, whose proof can be found in Appendix 19.5, implies that as long as k_n diverges at any given rate with n, the proposed estimator in (19.4) is statistically consistent and hence converges in probability toward the true parameter $\boldsymbol{\theta}_0$. If however we assume that the score function $s(\boldsymbol{\theta}; d_i)$ is *symmetrically* distributed, the following corollary delivers the convergence rate for a tuning constant $k_n \to 0$ with $n \to \infty$.

Corollary 1 *Let* $x_n \in \mathbb{R}^+$ *be a deterministic sequence such that* $x_n \to 0$ *and* $\sqrt{n}\, x_n \to \infty$ *as* $n \to \infty$*. Then, assuming* $s(\boldsymbol{\theta}; d_i)$ *has a symmetric distribution function and for any* $k_n \geq x_n$*, we have*

$$\hat{\boldsymbol{\theta}} \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0.$$

The proof of this corollary is in Appendix 19.5. These results are important since they allow us to define a region, which depends on the sample size n, within which we should define the tuning constant k_n in order for our estimator $\hat{\boldsymbol{\theta}}$ to be statistically consistent.

Remark 2 As stated earlier, for the purposes of robust statistical analysis, the "original" tuning constant k should be fixed and chosen, for example, with respect to the desired level of robustness and asymptotic efficiency of the resulting estimator compared to the non-robust alternative. However, for the purposes of privacy, we would require the constant to be chosen also with respect to the sample size and noise for privacy, in addition to the asymptotic efficiency. Therefore, given the above results, we want to define a tuning constant k_n that grows as slowly as possible since we want the statistical efficiency (low sampling variability) to dominate the noise added for differential privacy (which grows with k_n). A candidate could, for example, be $k_n := \log(\log(n))$ or any slowly increasing function in n. However, if we assume that the score function is (approximately) symmetrically distributed (e.g., linear regression with Gaussian residuals or logistic regression with probability $\pi \approx 0.5$), one could define, for example, $k_n := 1/log(n)$ for $n > 1$. At the same time, however, a k_n that is too small, despite allowing for consistency, can deliver an excessively inefficient estimator from a statistical point of view. Therefore, a rule for determining k_n based on the (asymptotic) efficiency under the constraint of consistency would be more appropriate and is left for future research.

The next section explores the use of the above proposed and studied M-Estimator within a differentially private mechanism in order to understand if the use of a robust M-Estimation framework can improve the utility of DP outputs for the same level of privacy. We also investigate the impact of the tuning constant k_n. For the purposes of notation, hereinafter we will simply denote the tuning constant as k and make its underlying dependence on n implicit whenever we let this constant diverge (or converge to zero).

Tuned Objective Perturbation

In this section, we propose the *Perturbed M-Estimation* mechanism that integrates the presented M-Estimator with the OPM framework; see Algorithm 1. The reason for considering the OPM as a good candidate for integration with the above described M-Estimation framework is that the OPM, being the result of an empirical risk minimization problem, produces an output that can indeed be classified as an M-Estimator as in (19.2). Following the definition in Kifer et al. (2012), the OPM for ϵ-differential privacy is defined as follows:

$$\bar{\boldsymbol{\theta}}_{DP} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\theta}; d_i) + \frac{\Delta}{2n}\|\boldsymbol{\theta}\|_2^2 + \frac{b^T\boldsymbol{\theta}}{n}, \tag{19.5}$$

where $l(\boldsymbol{\theta}; d_i)$ is a convex loss function, $\Delta \geq {}^{2\lambda}/_{\epsilon}$, λ is an upper bound on the eigenvalues of the Hessian $\nabla^2 l(\boldsymbol{\theta}; d_i)$, and $b \in \mathbb{R}^p$ is a random vector with density

$$f(b) \propto \exp^{-\epsilon \|b\|_2/2\xi},$$

where ξ is such that $\|\nabla l(\boldsymbol{\theta}; d_i)\|_2 \leq \xi$. Therefore, λ and ξ are two parameters that define the sensitivity measures of the loss function and consequently impact the amount of noise (perturbation) that is added to the loss function. Considering the definition in (19.5), we can now replace the loss function $l(\boldsymbol{\theta}; d_i)$ with the proposed loss function in (19.4) to deliver the new Perturbed M-Estimator.

Definition 2 The Perturbed M-Estimator is defined as follows:

$$\hat{\boldsymbol{\theta}}_{DP} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho_k(s(\boldsymbol{\theta}; d_i)) + \frac{\Delta_k}{2n} \|\boldsymbol{\theta}\|_2^2 + \frac{b_k^T \boldsymbol{\theta}}{n}. \tag{19.6}$$

From the above definition, we have that Δ_k and b_k (which depend on λ_k and ξ_k, respectively) are now quantities and variables that depend on the tuning constant k. Indeed, we have that ξ_k, and hence b_k, depends on the following quantity:

$$\nabla \rho_k(s(\boldsymbol{\theta}; d_i)) = \underbrace{\tanh\left(\frac{2}{k} s(\boldsymbol{\theta}; d_i)\right)}_{\in [-1,1]} k \, \nabla s(\boldsymbol{\theta}; d_i),$$

while λ_k, and hence Δ_k, depends on

$$\nabla^2 \rho_k(s(\boldsymbol{\theta}; d_i)) = \underbrace{2 \operatorname{sech}\left(\frac{2}{k} s(\boldsymbol{\theta}; d_i)\right)^2}_{\in [0,1]} (\nabla s(\boldsymbol{\theta}; d_i))^2 + \underbrace{\tanh\left(\frac{2}{k} s(\boldsymbol{\theta}; d_i)\right)}_{\in [-1,1]} k \, \nabla^2 s(\boldsymbol{\theta}; d_i).$$

From the above expressions, we observe that the tuning constant k can be directly related to a specific notion of DP-based sensitivity for $\nabla s(\boldsymbol{\theta}; d_i)$ and $\nabla^2 s(\boldsymbol{\theta}; d_i)$. Based on these expressions, for example, one could choose to define the tuning constant k as being inversely proportional to the sensitivity of these expressions according to the problem at hand.

Our proposed approach, highlighted in Algorithm 1, can therefore be seen as a form of "tuned" objective perturbation where we can calibrate the choice of k based on (i) sample size, (ii) required statistical efficiency, and (iii) known sensitivity bounds for the loss function. Indeed, we would generally want to choose a k that is "small" to achieve low sensitivity bounds (and add less noise for privacy), but in order to achieve statistical efficiency, we would ideally want k not to be *too* small. As stated earlier, the study of an optimal (private) choice of the tuning constant is left for future research.

19.4 Applications and Simulations

In this section, we investigate the potential utility of the suggested approach in Algorithm 1 in some applied and simulated settings. The examples are based on standard linear regression and logistic regression for small and large sample sizes and with a guarantee for pure differential privacy with $\epsilon = 0.1$. The parameter of interest $\boldsymbol{\theta}$ is represented by the regression coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^p$, and the utility of the estimators is measured via the L_2-norm (i) between the estimators and the reference value (non-private estimator or true value) or (ii) between the observed response and the predictions based on the different estimators (mean squared prediction error). The performance of the estimators is evaluated over $H = 100$ replications, and for each of them, different values of the tuning constant k are considered between $[0, 2]$. The latter range is considered since if $k < 1$, then the sensitivity measures for privacy are reduced while for values $k > 1$, the sensitivity is increased.

Remark 3 It must be noted that the OPM (and hence the proposed estimator) requires the optimization procedure to converge in order to guarantee differential privacy. In few examples, we did not have the convergence, but we still included them in the overall results to illustrate the potential gains in utility that this new approach could deliver. Hence, the results in this section should be considered as preliminary investigations rather than "conclusive" empirical results. Based on these observations, the goal would be to explore possibly more numerically stable privacy mechanisms for the considered approach using, for example, the stochastic gradient descent method (see, e.g., Rajkumar and Agarwal (2012), Song et al. (2013), Wang et al. (2015a), and Chen et al. (2019)) or the more recent KNG approach proposed by Reimherr and Awan (2019).

Algorithm 1 Perturbed M-Estimation—modified objective perturbation from Kifer et al. (2012)

INPUT: $D \in \mathcal{D}^n$, $\epsilon > 0$, a tuning parameter $k \in \mathbb{R}^+$, a convex set $\Theta \subset \mathbb{R}^p$, a convex loss $\hat{L}_k(\boldsymbol{\theta}; D) = \frac{1}{n}\sum_{i=1}^n \rho_k(s(\boldsymbol{\theta}; d_i))$ defined on Θ such that the Hessian $\nabla^2\rho_k(s(\boldsymbol{\theta}; d))$ is continuous in $\boldsymbol{\theta}$ and d, $\xi_k > 0$ such that $\|\nabla\rho_k(s(\boldsymbol{\theta}; d))\|_2 \leq \xi_k$ for all $\boldsymbol{\theta} \in \Theta$ and $d \in D$, and $\lambda_k > 0$ is an upper bound on the eigenvalues of $\nabla^2\rho_k(s(\boldsymbol{\theta}; d))$ for all $\boldsymbol{\theta} \in \Theta$ and $d \in D$.

1: Set $\Delta_k = \frac{2\lambda_k}{\epsilon}$

2: Draw $b_k \in \mathbb{R}^m$ from the density $f(b_k; \epsilon, \xi) \propto \exp(-\frac{\epsilon}{2\xi}\|b_k\|_2)$

3: Compute $\hat{\boldsymbol{\theta}}_{DP} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \rho_k(s(\boldsymbol{\theta}; d_i)) + \frac{\Delta_k}{2n}\|\boldsymbol{\theta}\|_2^2 + \frac{b_k^T\boldsymbol{\theta}}{n}$

OUTPUT: $\hat{\boldsymbol{\theta}}_{DP}$

Applications: Linear Regression

For the linear regression examples, let $\boldsymbol{y} \in \mathbb{R}^n$ be a vector of responses and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be a matrix of covariates, where the first column is a vector of ones for the intercept term. The score function is given by

$$s(\boldsymbol{\beta}; d_i) := y_i - x_i^T \boldsymbol{\beta},$$

where y_i is the response variable and $x_i \in \mathbb{R}^p$ is the vector of covariates for the ith row. We compare the following estimators:

- $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$: the least-squares non-private estimator that will be used as reference for the other estimators (i.e., considered as the true β we aim to estimate).
- $\bar{\boldsymbol{\beta}}$: the non-private robust estimator using $\rho_k(s(\boldsymbol{\beta}; d_i))$; we expect this to converge to $\hat{\boldsymbol{\beta}}$ as $k \to \infty$.
- K-norm sufficient statistics perturbation: this approach is proposed by Awan and Slavković (2020) and delivers differentially private estimators based on different norms considered for the sensitivity of the sufficient statistics $\boldsymbol{X}^T \boldsymbol{X}$ and $\boldsymbol{X}^T \boldsymbol{y}$ (see the functional mechanism of Zhang et al. (2012)). Based on these norms (including an optimal K-norm defined in Awan and Slavković (2020)), appropriate noise is added to the sufficient statistics to deliver differentially private estimators based on replacing an observation as opposed to removing an observation.
- $\tilde{\boldsymbol{\beta}}$: the proposed Perturbed M-Estimator in (19.6).

The first example we use is the "Attitude" dataset from Chatterjee and Hadi (2015), available in the `R` statistical software. This is a small dataset with only 30 observations and 7 variables capturing the percentages of favorable responses to a survey of clerical employees in a financial organization. A question of interest is how each variable contributes to the overall rating of the company ($\boldsymbol{y}$). The left plot in Fig. 19.1 reports the mean square prediction error for the different estimators over the different values of the tuning constant k for this data.

The second example, the "San Francisco housing" data, has been used for the evaluation of different statistical and differentially private methods; we use a dataset version from Awan and Slavković (2020). This dataset consists of 348, 189 observations on houses in the Bay area between 2003 and 2006. The main question of interest is in explaining the rent of the houses as a function of several other variables (e.g., square-footage, location, age of house, number of bedrooms, county). The right plot in Fig. 19.1 shows the mean squared error between all estimators and the parameter of reference $\hat{\boldsymbol{\beta}}$ (the non-private estimator).

In both cases, the data are preprocessed by taking the logarithm of some numerical variables and ensuring that all numerical variables lie between $[-1, 1]$. The latter bounding is not necessarily required for Perturbed M-Estimation since the tuning constant k can eventually compensate for a higher sensitivity due to

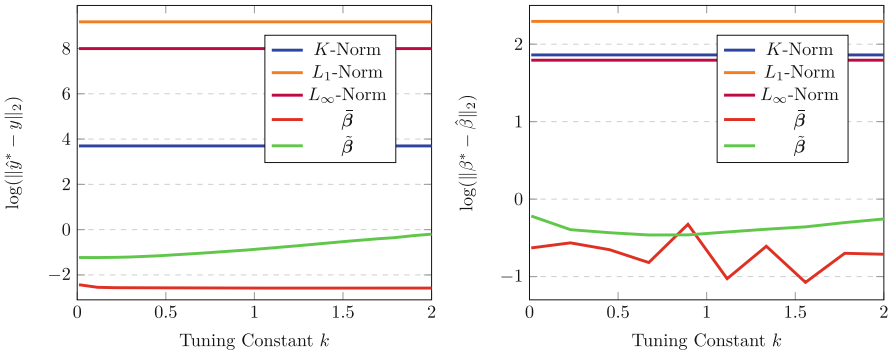


Fig. 19.1 Left: logarithm of the mean squared prediction error versus tuning constant k for the attitude dataset. Right: logarithm of the mean squared error versus tuning constant k with respect to the non-private estimator $\hat{\boldsymbol{\beta}}$ for the housing dataset

larger bounds on the variables; nevertheless, we perform this processing in order to compare it with the other estimators.

From Fig. 19.1, we can notice that the only estimators that depend on the tuning constant k (and whose lines therefore do not remain constant) are the robust non-private estimator $\bar{\boldsymbol{\beta}}$ and the proposed DP $\tilde{\boldsymbol{\beta}}$. For both datasets, as expected, it is clear that the robust non-private estimator $\bar{\boldsymbol{\beta}}$ (red line) has the best performance, and as the tuning constant k increases, this estimator improves its performance since it will converge to $\hat{\boldsymbol{\beta}}$. Our proposed DP M-Estimator (green line) appears to be the best alternative, and it significantly outperforms the other DP estimators in these settings. However, for both datasets, the performance of the Perturbed M-Estimator gets worse as the value of the tuning constant increases (although it still does better than the other DP estimators). This implies that the noise added for privacy starts to dominate over the statistical efficiency that is delivered through the increase of the tuning constant. Another effect that is more evident for the housing data (right plot) is that the performance of $\tilde{\boldsymbol{\beta}}$ is not optimal for the smallest values of k since it decreases and then starts to steadily increase around $k = 1$. This would indicate that for small values of k, the statistical inefficiency dominates the minimal noise added for privacy, while as k increases, this ratio starts to diminish as a result of the increasing statistical efficiency being overcome by the noise added for privacy.

Simulations: Logistic Regression

The simulation study in this section replicates the one in Awan and Slavković (2020) but with a smaller sample size of $n = 100$. We consider a logistic regression model where we generate uniformly distributed covariates $x_i \sim \mathcal{U}[-1, 1]$ and set the true parameter vector as $\boldsymbol{\beta} = (0, -1, -1/2, -1/4, 0, 3/4, 3/2)$. Based on these values, we

simulate uniform values $U_i \sim \mathcal{U}[0, 1]$ and define the simulated response values using the link function

$$\eta(x_i^T \boldsymbol{\beta}) = \frac{\exp(x_i^T \boldsymbol{\beta})}{1 + \exp(x_i^T \boldsymbol{\beta})},$$

as follows:

$$y_i = \begin{cases} 1 & U_i < \eta(x_i^T \boldsymbol{\beta}) \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the score function is defined as $s(\boldsymbol{\beta}; d_i) = y_i - \eta(x_i^T \boldsymbol{\beta})$. In order for the robust non-private estimator to be Fisher consistent, we would need to derive a correction term since the bounded function can introduce bias in the resulting estimator (see e.g., Cantoni and Ronchetti (2001)). For the purpose of this simulation, we do not apply this correction since, given our setting, the scores are approximately symmetrically distributed. Moreover, we assume that the performance of the proposed approach can only be improved if the correction term was introduced (and would be less relevant when $k \to \infty$).

We consider the following estimators for this simulation study:

- $\hat{\beta}$: MLE for logistic regression as a non-private reference.
- Objective perturbation estimators based on K-norms: private estimators based on the generalized OPM (adapted from Awan and Slavković (2020) and Kifer et al. (2012)) using different K-norms with change-DP, i.e., replace an individual. We consider the following norms, L_1, L_2, and L_∞ and another version of the L_∞-based OPM with an additional tuning constant to control the bias–variance trade-off set to $q = 0.85$ (instead of $q = 0.5$ for the other estimators, see Awan and Slavković (2020)).
- $\tilde{\boldsymbol{\beta}}$: the proposed Perturbed M-Estimator in (19.6).

The additional tuning constant q and the use of other norms could also be considered for our proposed approach in order to improve its performance. However, for this chapter, we keep it only depending on the tuning constant k. The mean squared errors with respect to the true parameter vector $\boldsymbol{\beta}$ are presented in Fig. 19.2.

The conclusions are similar to those of the previous section for the linear regression setting. Obviously, the MLE (red line) performs the best. However, our proposed DP M-Estimator (green line) is the best alternative, in some cases having substantially better performance than other DP estimators that are more commonly used with logistic regression. We also see that in this case the performance of $\tilde{\boldsymbol{\beta}}$ appears to worsen more rapidly as k increases but not much more than in the Attitude dataset; recall, both of these datasets are on a smaller scale with $n = 30$ and $n = 100$—settings where differentially private mechanisms, in general, have a harder time producing accurate statistics with small privacy-loss budgets, ϵ. In addition, we can see that the private estimator based on the tuned L_∞-norm (L_∞^*)

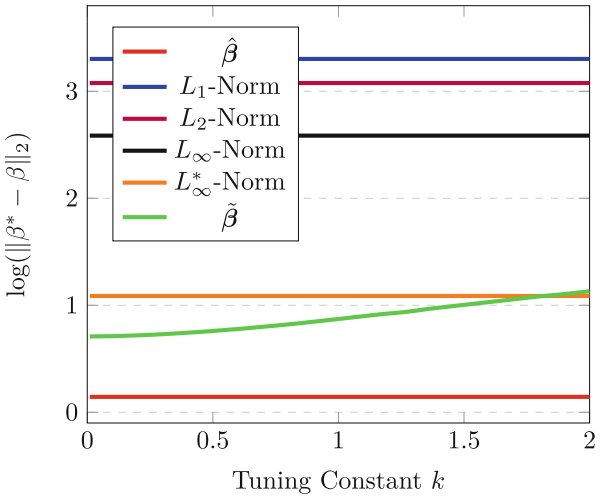


Fig. 19.2 Logarithm of the mean squared error versus tuning constant k with respect to the true parameter value $\boldsymbol{\beta}$ for the logistic regression simulation

also has a high utility, as argued in Awan and Slavković (2020), and is very close to our proposed approach. Nevertheless, as alluded earlier, it is possible that our approach could also benefit from consideration of other norms and the additional tuning constant q; we leave that to future work.

19.5 Conclusions and Outlook

In this work, we consider the use of methods from the field of robust statistics in order to improve the utility of differentially private mechanisms, that is, of their statistical outputs. More specifically, we propose a robust M-Estimator with well-defined properties, including consistency, and propose to employ it within the popular objective perturbation mechanism, thereby proposing a Perturbed M-Estimation mechanism. Our approach allows for calibration of noise needed to produce differentially private estimates, and it improves statistical utility of these outputs while removing the need to impose bounds on the parameter space and the response variables—this is a significant methodological and practical contribution as many current mechanisms require preprocessing of data such that it is bounded. There is still the need, however, to impose bounds on the covariates, for regression problems, for example, in order to determine sensitivity bounds. Our preliminary simulations and examples for linear and logistic regressions demonstrate significantly improved utility in estimation of parameter estimates under ϵ-DP in comparison to the currently used methods. It is also important to note that our proposed DP estimator works reasonably well for small sample sizes n. The setting

with small n is frequently problematic for DP since the noise needed to protect the privacy may overcome the sampling noise too much, making data unusable. While the choice of the tuning constant k for our robust estimator is more obvious in the non-private setting and is tied to n, in the private setting, the clear rules are yet to be determined and are part of future work.

Having investigated the possible use of robust statistical tools in the domain of differential privacy, it appears that it is worth to further explore this path and better understand properties and convergence rates of the proposed approach. Two improvements that can be considered jointly are the use of a Mallow's type estimator (see Huber (2011), Maronna et al. (2019)) and the redefinition of the expression for the OPM based on the properties of the proposed RobHyt function or of any other function with bounded derivative (and definable sensitivity bounds) and with similar properties of consistency. The Mallow's type estimator can automatically bound the covariates of a regression problem thereby possibly removing the need to impose any bounds on parameters and data. The redefining of the sensitivity bounds can be done, for example, by using the links between smooth sensitivity and the GES as highlighted in Chaudhuri and Hsu (2012) and Avella-Medina (2019)). Moreover, depending on the definition of the problem, rules to determine the value of the tuning constant k can be developed or appropriate methods to select an "optimal" k in a private manner can be studied. In the latter case, an intuitive approach would be to find the value of k based on the definition of the asymptotic variance for M-Estimators, which would possibly depend only on the model and the sample size, thereby allowing to determine it independently from the data (or find an approximation in a private manner). Another approach that will be worth investigating is the use of a private stochastic gradient descent mechanism (see Chen et al. (2019), Song et al. (2013), Wang et al. (2015a)), or methods such as the KNG mechanism in Reimherr and Awan (2019), in order to overcome possible non-convergence issues of the objective perturbation mechanism. Finally, once possible new sensitivity bounds are defined based on robust statistical measures, it would be possible to deliver the corresponding statistical inference framework that would allow to construct private confidence intervals and perform private parametric tests. And, nearly 50 years ago, after Andrews et al. (1972) provided an extensive survey of some 68 robust estimates of location, we can take a look back at those in order to move forward.

Acknowledgments The authors would like to thank Jordan Awan and Dan Kifer for the useful discussions and inputs as well as for sharing code to compare the results in this chapter with existing approaches. We would like to thank Stéphane Guerrier and Mucyo Karemera for their helpful suggestions and Marco Avella-Medina for sharing his working manuscript with us. This research was supported in part by NSF Grants SES-1534433 and SES-1853209 to Pennsylvania State University and by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1 TR002014. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF. Part of this work was done, while authors were visiting the Simons Institute for the Theory of Computing.

A Proofs

Proof of Proposition 1

Proof Provided the other assumptions hold (which is generally common or verified when considering the MLE), we need to prove that Assumption (A4) holds as well. The definition of uniform convergence for our setting is the following:

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| 1/n \sum_{i=1}^{n} \rho_k(s(\boldsymbol{\theta}, d_i)) - \mathbb{E}[s(\boldsymbol{\theta}, d_i)^2] \right| \to 0.$$

Let us denote $Q_n(s(\boldsymbol{\theta})) := 1/n \sum_{i=1}^{n} \rho_k(s(\boldsymbol{\theta}, d_i))$ and $Q_0(s(\boldsymbol{\theta})) := \mathbb{E}[s(\boldsymbol{\theta}, d_i)^2]$. We can therefore reexpress the above definition as

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| Q_n(s(\boldsymbol{\theta})) - Q_0(s(\boldsymbol{\theta})) \right| \to 0,$$

where, defining $\bar{Q}_n(s(\boldsymbol{\theta})) := 1/n \sum_{i=1}^{n} s(\boldsymbol{\theta}, d_i)^2$, by triangle inequality, we have

$$\left| Q_n(s(\boldsymbol{\theta})) - Q_0(s(\boldsymbol{\theta})) \right| \leq \left| Q_n(s(\boldsymbol{\theta})) - \bar{Q}_n(s(\boldsymbol{\theta})) \right| + \left| \bar{Q}_n(s(\boldsymbol{\theta})) - Q_0(s(\boldsymbol{\theta})) \right|. \quad (7)$$

Since we assume that $s(\boldsymbol{\theta}, d_i) = \mathcal{O}_p(1)$, we have that

$$\left| \bar{Q}_n(s(\boldsymbol{\theta})) - Q_0(s(\boldsymbol{\theta})) \right| = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right),$$

based on the weak law of large numbers and Markov's inequality. We therefore focus on the first term on the right side of the inequality in (7). For this reason, let us apply a second-order Taylor expansion of the two functions characterizing this term around the expected value of $s(\boldsymbol{\theta}, d_i)$ at the solution (i.e., zero):

$$Q_n(s(\boldsymbol{\theta})) = \frac{1}{n} \sum_{i=1}^{n} \left[\rho_k(0) + \frac{\partial}{\partial s_i(\boldsymbol{\theta})} \rho_k(s_i(\boldsymbol{\theta})) \Big|_{s_i(\boldsymbol{\theta})=0} s_i(\boldsymbol{\theta}) + \frac{\partial^2}{\partial^2 s_i(\boldsymbol{\theta})} \rho_k(s_i(\boldsymbol{\theta})) \Big|_{s_i(\boldsymbol{\theta})=0} s_i(\boldsymbol{\theta})^2 + R_{\rho_k} \right],$$

and

$$\bar{Q}_n(s(\boldsymbol{\theta})) = \frac{1}{n} \sum_{i=1}^{n} \left[0 + 2 \cdot 0 \cdot s_i(\boldsymbol{\theta}) + 2 s_i(\boldsymbol{\theta})^2 \right] = \frac{1}{n} \sum_{i=1}^{n} 2 s_i(\boldsymbol{\theta})^2,$$

since there is no remainder term for the expansion of $\bar{Q}_n(s(\boldsymbol{\theta}))$. As for the expansion of $Q_n(s(\boldsymbol{\theta}))$, by taking the required derivatives and evaluating them in zero, we end

up with

$$Q_n(s(\boldsymbol{\theta})) = \frac{1}{n}\sum_{i=1}^{n}\left[2\underbrace{\operatorname{sech}(0)^2}_{=1}s_i(\boldsymbol{\theta})^2 + R_{\rho_k}\right] = \frac{1}{n}\sum_{i=1}^{n}\left[2s_i(\boldsymbol{\theta})^2 + R_{\rho_k}\right].$$

Considering these expansions, we have that

$$Q_n(s(\boldsymbol{\theta})) - \bar{Q}_n(s(\boldsymbol{\theta})) = \frac{1}{n}\sum_{i=1}^{n} R_{\rho_k},$$

and therefore let us take a look at the remainder term that has the following structure:

$$R_{\rho_k} = \frac{\partial^3}{\partial^3 s(\boldsymbol{\theta})}\rho_k(s(\boldsymbol{\theta}))\frac{s(\boldsymbol{\theta})^3}{3!}.$$

Let us focus on the bound of the third derivative, and taking the absolute value, we have

$$\begin{aligned}\left|\frac{\partial^3}{\partial^3 s(\boldsymbol{\theta})}\rho_k(s(\boldsymbol{\theta}))\right| &= \left|\tfrac{6}{k}\tanh\left(\tfrac{2}{k}s(\boldsymbol{\theta})\right)\operatorname{sech}\left(\tfrac{2}{k}s(\boldsymbol{\theta})\right)\right| \\ &= \tfrac{6}{k}\underbrace{\left|\tanh\left(\frac{2}{k}s(\boldsymbol{\theta})\right)\right|}_{\leq 1}\underbrace{\left|\operatorname{sech}\left(\frac{2}{k}s(\boldsymbol{\theta})\right)\right|}_{\leq 1} \leq \tfrac{6}{k}.\end{aligned}$$

Hence, we have that

$$\left|R_{\rho_k}\right| \leq \frac{6}{k}\frac{s(\boldsymbol{\theta})^3}{3!} = \frac{s(\boldsymbol{\theta})^3}{k},$$

which, for $k \to \infty$ with n, implies that

$$\left|Q_n(s(\boldsymbol{\theta})) - \bar{Q}_n(s(\boldsymbol{\theta}))\right| = \mathcal{O}_p\left(\frac{1}{k}\right),$$

since $s(\boldsymbol{\theta}, d_i)$ is bounded in probability. Plugging this back in Eq. (7), we consequently have that

$$\left|Q_n(s(\boldsymbol{\theta})) - Q_0(s(\boldsymbol{\theta}))\right| = \mathcal{O}_p\left(\max\left(\frac{1}{k}, \frac{1}{\sqrt{n}}\right)\right),$$

which concludes the proof. □

Proof of Corollary 1

Proof This corollary is simply a consequence of the proof of Proposition 1. Indeed, if $s(\boldsymbol{\theta}, d_i)$ follows a symmetric distribution, we have that

$$\frac{1}{n}\sum_{i=1}^{n} s_i(\boldsymbol{\theta})^3 = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right),$$

and hence, following (7), we would have

$$\left|Q_n(s(\boldsymbol{\theta})) - Q_0(s(\boldsymbol{\theta}))\right| \leq \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right) + \mathcal{O}_p\left(\frac{1}{k\sqrt{n}}\right) = \mathcal{O}_p\left(\max\left(1, \frac{1}{k}\right)\frac{1}{\sqrt{n}}\right).$$

In order for this term to go to zero as $n \to \infty$, we need either of the following cases:

1. $k \to \infty$ (or in any case $k \geq 1$),
2. $k \to 0$ slower than $\sqrt{n}$.

□

References

John M Abowd. Staring-down the database reconstruction theorem. In *Joint Statistical Meetings, Vancouver, BC*, 2018.

D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. *Robust Estimates of Location: Survey and Advances*. Princeton University Press, 1972. URL http://www.jstor.org/stable/j.ctt13x12sw.

Marco Avella-Medina. Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association*, (just-accepted): 1–45, 2019.

Jordan Awan and Aleksandra Slavković. Differentially private uniformly most powerful tests for binomial data. In *Advances in Neural Information Processing Systems*, pages 4208–4218, 2018.

Jordan Awan and Aleksandra Slavković. Structure and sensitivity in differential privacy: Comparing k-norm mechanisms. *Journal of the American Statistical Association*, (just-accepted): 1–56, 2020.

Jordan Awan, Ana Kenney, Matthew Reimherr, and Aleksandra Slavković. Benefits and pitfalls of the exponential mechanism with applications to Hilbert spaces and functional PCA. In *International Conference on Machine Learning*, pages 374–384, 2019.

Clément L Canonne, Gautam Kamath, Audra McMillan, Adam Smith, and Jonathan Ullman. The structure of optimal private tests for simple hypotheses. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 310–321, 2019.

Eva Cantoni and Elvezio Ronchetti. Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96 (455): 1022–1030, 2001.

Samprit Chatterjee and Ali S Hadi. *Regression analysis by example*. John Wiley & Sons, 2015.

Kamalika Chaudhuri and Daniel Hsu. Convergence rates for differentially private statistical estimation. In *Proceedings of the International Conference on Machine Learning. International Conference on Machine Learning*, volume 2012, page 1327. NIH Public Access, 2012.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12 (Mar): 1069–1109, 2011.

Kamalika Chaudhuri, Anand D Sarwate, and Kaushik Sinha. A near-optimal algorithm for differentially-private principal components. *The Journal of Machine Learning Research*, 14 (1): 2905–2943, 2013.

Chen Chen, Jaewoo Lee, and Dan Kifer. Renyi differentially private ERM for smooth objectives. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2037–2046, 2019.

Yanbo Chen, Jin Ma, and Ji Zhang. Robust state estimator based on hyperbolic cosine function. In *2017 IEEE Power & Energy Society General Meeting*, pages 1–2. IEEE, 2017.

Tore Dalenius. Privacy transformations for statistical information systems. *Journal of Statistical Planning and Inference*, 1 (1): 73–86, 1977.

Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

George T Duncan, Stephen E Fienberg, Rammaya Krishnan, Rema Padman, Stephen F Roehrig, et al. Disclosure limitation methods and information loss for tabular data. *Confidentiality, disclosure and data access: theory and practical applications for statistical agencies*, pages 135–166, 2001.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *STOC*, volume 9, pages 371–380, 2009.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9 (3–4): 211–407, 2014.

Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4: 61–84, 2017.

Stephen E Fienberg. A radical proposal for the provision of micro-data samples and the preservation of confidentiality. *Department of statistics*, 1994.

Stephen E Fienberg and Aleksandra B Slavkovic. Preserving the confidentiality of categorical statistical data bases when releasing information for association rules. *Data Mining and Knowledge Discovery*, 11 (2): 155–180, 2005.

Stephen E. Fienberg and Aleksandra B. Slavković. *Data Privacy and Confidentiality*. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-04898-2. https://doi.org/10.1007/978-3-642-04898-2_202.

Stephen E Fienberg and Russell J Steele. Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, 14 (4): 485, 1998.

Stephen E Fienberg, Yuval Nardi, and Aleksandra B Slavković. Valid statistical analysis for logistic regression with multiple sources. In *Annual Workshop on Information Privacy and National Security*, pages 82–94. Springer, 2008.

Stephen E Fienberg, Alessandro Rinaldo, and Xiaolin Yang. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *International Conference on Privacy in Statistical Databases*, pages 187–199. Springer, 2010.

Marco Gaboardi, Hyun-Woo Lim, Ryan M Rogers, and Salil P Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. 2016.

Simson Garfinkel, John M. Abowd, and Christian Martindale. Understanding database reconstruction attacks on public data. *Queue*, 16 (5), October 2018. ISSN 1542-7730. https://doi.org/10.1145/3291276.3295691.

Rob Hall, Stephen E Fienberg, and Yuval Nardi. Secure multiple linear regression based on homomorphic encryption. *Journal of Official Statistics*, 27 (4): 669, 2011.

Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Differential privacy for functions and functional data. *Journal of Machine Learning Research*, 14 (Feb): 703–727, 2013.

Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the American statistical association*, 69 (346): 383–393, 1974.

Frank R Hampel, Peter J Rousseeuw, and Elvezio Ronchetti. The change-of-variance curve and optimal redescending M-estimators. *Journal of the American Statistical Association*, 76 (375): 643–648, 1981.

Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics*. Wiley Online Library, 1986.

Peter J Huber. *Robust statistics*. Springer, 2011.

Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. *Statistical disclosure control*. John Wiley & Sons, 2012.

Vishesh Karwa and Aleksandra Slavković. Inference using noisy degrees: Differentially private β-model and synthetic graphs. *The Annals of Statistics*, 44 (1): 87–112, 2016.

Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1, 2012.

Jing Lei. Differentially private m-estimators. In *Advances in Neural Information Processing Systems*, pages 361–369, 2011.

Jing Lei, Anne-Sophie Charest, Aleksandra Slavkovic, Adam Smith, and Stephen Fienberg. Differentially private model selection with penalized and constrained likelihood. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181 (3): 609–633, 2018.

Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

Ardalan Mirshani, Matthew Reimherr, and Aleksandra Slavković. Formal privacy for functional data with Gaussian perturbations. In *International Conference on Machine Learning*, pages 4595–4604, 2019.

R v Mises. On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics*, 18 (3): 309–348, 1947.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135 (3): 370–384, 1972.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4: 2111–2245, 1994.

Arun Rajkumar and Shivani Agarwal. A differentially private stochastic gradient descent algorithm for multiparty classification. In *Artificial Intelligence and Statistics*, pages 933–941, 2012.

Matthew Reimherr and Jordan Awan. Kng: The k-norm gradient mechanism. In *Advances in Neural Information Processing Systems*, pages 10208–10219, 2019.

Jerome P Reiter. Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21 (3): 441–462, 2005.

Donald B Rubin. Statistical disclosure limitation. *Journal of official Statistics*, 9 (2): 461–468, 1993.

Xingyu Shen, Xiang Zhang, Long Lan, Qing Liao, and Zhigang Luo. Another robust NMF: Rethinking the hyperbolic tangent function and locality constraint. *IEEE Access*, 7: 31089–31102, 2019.

Aleksandra Slavkovic. Steve the matchmaker: The marriage of statistics and computer science in the world of data privacy. *CHANCE*, 26 (4): 4–7, 2013. URL https://doi.org/10.1080/09332480.2013.868745.

Aleksandra Slavkovic and Lars Vilhuber. Remembering Stephen Fienberg. *Journal of Privacy and Confidentiality*, 8 (1), Dec. 2018. https://doi.org/10.29012/jpc.685. URL https://journalprivacyconfidentiality.org/index.php/jpc/article/view/685.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822. ACM, 2011.

Joshua Snoke, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181 (3): 663–688, 2018. https://doi.org/10.1111/rssa.12358. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12358.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248. IEEE, 2013.

Mario Trottini and Stephen E Fienberg. Modelling user uncertainty for disclosure risk and data utility. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10 (05): 511–527, 2002.

Duy Vu and Aleksandra Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, pages 138–143. IEEE, 2009.

Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient Monte carol. In *International Conference on Machine Learning*, pages 2493–2502, 2015a.

Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. On-average KL-privacy and its equivalence to generalization for max-entropy mechanisms. In *International Conference on Privacy in Statistical Databases*, pages 121–134. Springer, 2016.

Yue Wang, Jaewoo Lee, and Daniel Kifer. Revisiting differentially private hypothesis tests for categorical data. *arXiv preprint arXiv:1511.03376*, 2015b.

Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105 (489): 375–389, 2010.

Leon Willenborg and Ton De Waal. *Statistical disclosure control in practice*, volume 111. Springer Science & Business Media, 1996.

Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. Functional mechanism: regression analysis under differential privacy. *Proceedings of the VLDB Endowment*, 5 (11): 1364–1375, 2012.

Part VI
Introduction: Official Statistics

Judith M. Tanur

Steve cared about official statistics. A good deal of his interaction with official statistics came through his affiliation with the Committee on National Statistics of the National Academy of Sciences/National Research Council (henceforth CNSTAT). Steve's mentor, Fred Mosteller, was the vice chair of the President's Commission on Government Statistics, chaired by W. Allen Wallis. A major recommendation of that Commission was the establishment of such a committee. The first chair of CNSTAT, which was charged with evaluating statistical issues for the US government, including citizens' attitudes and behavior toward the census was Bill Kruskal, who had been a member of the Commission. Taking office in 1972, Kruskal got the committee off to a great start, raised funds, got agency buy in, and worked with Margaret Martin, who was the first executive director, to really accomplish administrative miracles at the National Academy of Sciences/National Research Council in creating an entity very different from the norm there. Bill was succeeded by Conrad Taeuber who held the Committee on a steady course during his term from 1978 to 1981. But Taeuber was born in 1906; Kruskal was born 1919. Steve, a baby boomer, had all the stereotypical characteristics of that generation, including energy and brashness. In an interview (reprinted later in this volume) for the series of conversations with distinguished statisticians and published in *Statistical Science* (2013), Steve described his feelings as he became a member of CNSTAT in 1978 and especially as he chaired it in 1981–1984 and then after a sabbatical year from 1985 to 1987.

> I got to join CNSTAT . . . This was . . . full of new ideas and problems to work on; I was like a kid in a candy shop! The committee didn't have a lot of projects then, but I just got to look around the Academy and the Federal government, and there were possibilities everywhere. I could only do so much, but I pushed the staff to do other things and got my friends on the committee to lead panels. By the mid-80s the committee was humming and there were all these neat activities on census methodology, on cognitive aspects of survey methodology, statistical assessments as evidence in the courts, sharing research data—there was just no end.—*Statistical Science* Interview, page 457

Under Steve's leadership, the Committee undertook informal reviews of the portfolios of the major statistical agencies. Steve kept getting the committee

involved in new projects and thus kept the staff badly overworked. Somehow he seemed to have more hours in the day than do the rest of us.

Let me talk in detail about one effort that took place during Steve's tenure as CNSTAT chair – chosen because I happen to know quite a bit about its genesis – the movement that investigates cognitive aspects of survey methodology and which is known by the acronym CASM. (We chose that name and acronym purposefully – it punned on the broad chasm between survey research and the cognitive scientists that we hoped the participants in the movement would bridge.) Although there were earlier efforts to bring together cognitive scientists and survey researchers, notably one headed by Al Biderman (see Biderman 1980) and focused on the National Crime Survey, and a major effort about the same time in Germany organized by ZUMA (see Hippler, Schwarz, and Sudman 1987), and although others got credit (in terms of authorship of publications and receipt of awards), one could make the case that Steve was, if not the progenitor of the CASM movement, at the least its presiding midwife. He talked me into chairing the 1983 St. Michael's seminar by bribery – he offered attractive surroundings, permission for my husband to visit along with other guests, the opportunity to choose the participants and thus include my friends and, most important, a great staff that included Tom Jabine, Miron Straf, and Roger Tourangeau. How could I refuse? At the seminar itself, Steve was an enthusiastic participant. Very soon thereafter he coauthored with Elizabeth Loftus, an eminent memory researcher, and me, a solicited article for the *Milbank Memorial Fund Quarterly* on cognitive aspects of surveys. It turned out the article as written was too long, so Steve suggested that we break it into three, and *Milbank* published them all. (That made quite a blip on my cv, but is hardly noticeable on Steve's or Beth's.) Then Steve got the Social Science Research Council (SSRC) (where he served on what amounted to the Board of Directors) to establish a Committee on Cognition and Survey Research, then talked me into co-chairing the committee (with Robert Abelson). Again, Steve was an active participant in the committee's meetings and workshops, especially enthusiastic in organizing a workshop on reporting pain and other symptoms. When the committee considered its work done (and its funding exhausted), Steve talked me into editing a volume of papers based on its work – and in 1991 *Questions about Questions* appeared.

The chapters in this section reflect Steve's interests in official statistics – but as usual his interests are too broad and interrelated to be easily pigeonholed. While Margo Anderson's chapter entitled "The History of 'Statistics in the Public Interest' in the United States" certainly belongs here because of its emphasis on how statistical issues arise from the very Constitution of the United States, its emphasis on the US Census could also have justified its placement in our section on Surveys and Censuses (where two former Census Directors, Kenneth Prewitt and John Thompson, discuss the interplay of statistics and politics on the Census). And the chapters that discuss issues of the US Census could as easily have appeared in this section.

More obviously germane to the topic of official statistics is the chapter by Connie Citro, Michael Cohen, and Porter Coggeshall, "Fostering Statistical Rigor for Evidence-Based Policy at the National Academies of Sciences, Engineering,

and Medicine," which chronicles Steve's enormous contributions to the Academies, starting even before his membership in the Committee on National Statistics and continuing until very shortly before his death.

Jan van den Brakel, a PhD advisee of Steve's, whose PhD work attempted to further extend the papers that Steve and I had been working on, by combining statistical methods from the fields of randomized experiments and sampling theory to improve the internal and external validity of experimental research. As Jan writes, that would have been a natural topic for his chapter in this volume, but an overview had been published recently. Jan is at the Research and Development Department of Statistics Netherlands and contributes a chapter touching on another of Steve's interest: "New data sources and inference methods for official statistics."

Kathy Wallman was Chief Statistician of the United States from 1992 to 2017. In that role at the US Office of Management and Budget, she provided coordination, guidance, and oversight for the Federal Statistical System of the United States. Her chapter explicates the organization of the US decentralized Statistical System and describes its mission.

References

Biderman, A. Report of a Workshop on Applying Cognitive Psychology to Recall Problems of the National Crime Survey. Washington, DC: Bureau of Social Science Research

Hippler, H.-J., Schwarz, N., and Sudman, S. (Eds.). (1987). Social Information Processing and Survey Methodology. Springer-Verlag Publishing.

Chapter 20
The History of "The Politics of Population" in the United States

Margo Anderson

20.1 Preliminaries

I met Steve Fienberg around 1989, when to the best of my hazy recollection, I got a phone call from him asking if I would be interested in team teaching a course at Carnegie Mellon on the census in spring 1990. We'd never met though we'd been in the same census universe, me as an historian, Steve as statistician. I had published the first edition of my history of the American population census the previous year (Anderson 1988). I lived in Milwaukee, WI, so that was an interesting logistical challenge, but I was also on leave so not teaching at the moment. After more conversations, I agreed and we tag teamed the class, with me flying to Pittsburgh for my sessions.

That initial collaboration led to more, including writing joint articles for *Chance* on the 1990 census lawsuits, and then a book level study (Anderson and Fienberg 1999a). We tried to integrate the long-term history of the American population census with the controversies about measuring and adjusting for the known differential undercount, particularly of minorities and the poor. The undercount had the effect of reducing the political power of affected communities and reducing the funding available through the formulas embedded in legislation. I traced the historical roots of the census's role in apportionment and funding allocations from its origins in the eighteenth century to the present. Steve did the statistical theory debate about adjustment methodology. We both collaborated on the ongoing policy (Anderson and Fienberg 1999b). We both served on the Committee on National Statistics panel which led to the report, *Modernizing the U.S. Census* (National Research Council 1995).

M. Anderson (✉)
Department of History, University of Wisconsin—Milwaukee, Milwaukee, WI, USA
e-mail: margo@uwm.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_20

After that burst of collaboration in the 1990s, we kept in touch, met at JSM or other events around the country. We shared our common interests in statistical confidentiality, and he published one of the papers I coauthored with William Seltzer in the first volume of the *Journal of Privacy and Confidentiality* (Anderson and Seltzer 2009).

So what might one suggest in honor of Steve about the question of "statistics in the public interest?" One might start by noting that, for an historian, the title of this volume is a bit of a redundancy. The origin of the term "Statistik" or "statistics" meant "of the state" and so "statists" who engaged in "statistics" gathered and analyzed quantitative or tabular information about politics, government, and public activity. In English, the term of art through the eighteenth century was "political arithmetic" (Porter 1986).

Today the discipline of "statistics" traces its roots to the seventeenth century both from the mathematicians who worked out the theory of probability and the measurement of uncertainty, and from the researchers who collected, classified, and tabulated the information we now call "numerical data" for analysis. These were somewhat separate developmental threads.

Early "data" work in the seventeenth and early eighteenth century was concerned with patterns of population growth, mortality, and the relative demographic strength of nations, and thus pioneered in the publication and analysis of the numerical patterns in existing data collections, such as Graunt and Petty's work on the London Bills of Mortality (Heyde 2001).

In the American colonies and the infant United States, as well, the issues of data collection were paramount, initially as the British Crown tried to understand what was going on in its far-flung colonies by asking for reports on the people, economy, and productivity of their "investments." Then, in the mid-eighteenth century, came the American Revolution, independence, and the need for the newly "united" colonies to establish a stable government and economic future that would survive what they realized would be further efforts by the British to regain control of their colonies (Cassedy 1969; Cohen 1982).

Meanwhile, back in Europe, as Stephen Stigler (1986) has shown, were the astronomers and mathematicians with a different problem. They were trying to understand the natural world, and wrestled with the problem of reconciling the differing measurements of allegedly the same thing, for example, a star's location. The measurements conflicted with one another, even after considering when or where the information was collected. Hence the notion of "error" was born not to mean a mistake, but an inevitable element of measurement to be quantified.

By the time the American Statistical Association was founded in 1839, "statistics" was no longer simply "political" arithmetic. ASA founders were wise enough to define their field widely in their constitution: "The objects of the Society shall be to collect, preserve and diffuse Statistical information in the different departments of human knowledge." But in the United States, well into the early twentieth century, data collection and the analysis of government policy issues and data continued to

be the backbone of the enterprise. As Fitzpatrick (1957: 15) noted, the American Statistical Association issued its first publication in the 1840s:

> The first part of this volume had appeared in 1843. It is entitled "Statistics of Towns in Massachusetts", and numbers 120 pages. The second part, was published in 1845, embracing 216 pages, and is called, "Statistics of Population in Massachusetts". The third part, appeared in 1847, consisting of about 375 pages, with the title "Statistics of Taxation in Massachusetts Including Valuation and Population". All three parts were "prepared by Joseph B. Felt", the Recording Secretary of the American Statistical Association from 1839 to 1859. [sic, punctuation in original]

JASA, which began publication under a variety of titles in the 1880s, was similarly concerned with government data.

20.2 Building a Politics of Population into the Structure of the American State

It turns out that complete count public data collections were fundamental to building the American state in the late eighteenth century. The 1787 federal constitution enshrined public data in two separate provisions, designed to define how to ground state authority in the "people," allocate the responsibilities and resources of the state among the citizenry, and provide mechanisms for stabilizing the state over time. This became the two-legged statistical system, grounded in demographic statistics on the population and economic statistics on the revenues and expenditures of the state.

This chapter, therefore, is a selective narrative of examples about how the data collectors in the discipline dominated into the first third of the twentieth century; how the mathematical breakthroughs in statistics, sampling, in particular, catapulted "statistics in the public interest" into the discipline we know today; and how the long historical experience in the problems of state making and measurement resonate even now in the big data world of the twenty-first century.

20.3 The First Leg of the System: Demographic Statistics

The US Constitution firmly grounds the authority of the state in the sovereignty of the "people:" "We the people of the United States. . . do ordain and establish this Constitution for the United States of America." The previous federal structure, the Articles of Confederation, in contrast, had located sovereignty in the states themselves, affirming that "Each state retains its sovereignty, freedom, and independence," in matters "not ....expressly delegated" to the United States. The 13 states of the United States were clearly identifiable entities in the 1770s and 1780s, with traditions of settlement and governance that stretched back a century or more. The concept of the "people" of the United States was a much murkier notion, a fine

rhetorical device, but filled with ambiguous connotations (Anderson 1988; Bailyn 1967; Morgan 1988; Wood 1969, 1992).[1]

One mechanism in the new governmental structure relied directly on the notion of the ultimate "sovereignty" of the "people." That was the institution of the decennial census or enumeration of the population required to allocate the seats among the states in the House of Representatives, the votes for each state in the Electoral College, and the apportionment of "direct taxes" among the states. Article 1, Section 2, paragraph 3 began: "Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers...." The full paragraph reads:

> Representatives and direct taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding the whole number of free persons, including those bound to service for a term of years, and excluding Indians not taxed, three-fifths of all other persons. The actual enumeration shall be made within three years after the first meeting of the Congress of the United States, and within every subsequent term of ten years, in such manner as they shall by law direct.

Just as an election count determined who would represent a local area in a legislative assembly, a population count would determine the strength of each state's House delegation. In the very grounding of the new state, the government would be required to count the population every 10 years, reallocate the seats in the House on the basis of that count, and reallocate the apportionment for direct taxes to be levied among the states. It is this provision that grounded the first leg of the statistical system: the periodic population count.

20.4 The Second Leg of the System: Economic Statistics

The second leg of the statistical system can also be found in the 1787 Constitution, which provided the basis for the collection and publication of economic data. Article 1 Section 9 mandated that federal appropriations required legal authorization and that "a regular Statement and Account of the Receipts and Expenditures of all public Money shall be published from time to time." The President was required by Article 2, Section 3 to "from time to time give to the Congress Information of the

[1] In the Constitutional Convention that ambiguity was evident as the framers struggled to "operationalize" their new government and provide concrete mechanisms for setting the thing in motion. The most obvious are the ordered election procedures, for Representatives, Senators, and the President. Even then though, the reliance on the procedures and traditions in the existing states indicated the framers' reticence with departing too greatly from the traditions within the states. Two senators were allocated to each state. The state legislatures chose the 2 senators for each state. The state legislatures chose the electors to the Electoral College who formally chose the President. Only in the case of members of the House of Representatives did the "people" vote directly for an individual who would serve in national office. Voter qualifications and the timing of elections for Senators and Representatives were also left to the state legislatures. .

State of the Union and recommend to their Consideration such Measures as he shall judge necessary and expedient." These two provisions guaranteed the administrative record keeping and publication of the fiscal condition of the federal government, which in turn permitted their compilation into a statistical record of administrative operations, both in the collection of taxes and in the expenditure of federal revenue.

20.5 Setting the Government in Motion: Implementing the Census Enumeration and Learning About Statistical Methods at the Same Time

Like the notion of the sovereignty of the people, counting the population is a deceptively simple idea, which on further examination is much more complex. Who actually are "the people"? How, practically, does one go about counting them? The first question of who "constituted" the people to be counted was the subject of intense debate in the Constitutional Convention because of the existence of slavery. Slaves were both "people" and a "species of property" expressly excluded from the possibility of political action. If sovereignty derives from the people, what about other "people" who exercise no political power and have no political authority: women, children, criminals, aliens, the poor? Practically speaking "voting," i.e., official participation in electing a representative was exercised by about 10% of the total population in the United States of the late eighteenth century. Nevertheless, the framers settled on a census which counted practically the entire population. The compromise in the Constitutional Convention was to use a very expansive rule. The enumeration would count everyone except "Indians not taxed" and distinguish the slave population from the free population so the slave population could be "discounted" to 60% for the apportionment of House seats and direct taxes. "Indians not taxed," that is, those American Indians who gave allegiance to their tribes rather than to the United States, were the only "people" within the geographic boundaries of the United States not required to be counted in the decennial census (cf. Anderson 2015; Lewis 1995; Zagarri 1987).

The second question, of the administration of the census, was left to the First Congress to resolve. The legislation enacted in early 1790 mandated a bare-bones administrative operation. It created no new administrative instruments. Rather it required the US marshals (at the time one per state) to appoint assistants in their districts to canvass each household and count the members. Six brief questions were mandated: name of the household head; a division of the free white males into cohorts of 16 and above and under 16; and free white females, "other free persons," and slaves. The assistants totaled the figures for their districts and sent them to the marshals who in turn totaled them for each state. The Secretary of State in turn tallied the national figures and sent them to the President. The first census was taken between August 1790 and March 1792 and recorded 3.9 million people. There were a variety of administrative headaches that developed, including worries

of under enumeration, and even some prosecutions for refusals to participate. On the whole, however, President George Washington recommended the results to the Congress as evidence of the success of the new state, and Congress undertook its first apportionment on the bases of the figures (Anderson and Fienberg 1999b).

Once the census results were available, a third problem emerged: how did one translate the population numbers into allocations of representatives among the states? How big should the House of Representatives be? Beyond requiring no more than one representative per 30,000 population, the Constitution was silent on this issue as well. As soon as they began to debate the issue, members of Congress articulated several different methods which led to allocations favoring one or another of the emerging factions. Once Thomas Jefferson and Alexander Hamilton, the leaders of the emerging Republican and Federalist factions in the new government, came to be identified with alternative apportionment bills, the issue was joined in a potentially dangerous way. Congress passed a bill to Hamilton's liking in the spring of 1792. Washington consulted Jefferson and vetoed the bill (using the Presidential veto for the first time). At this point, the new government could have faced a true crisis: i.e., the inability to "constitute" the House of Representatives. Congress relented and passed a "Jeffersonian" bill and everyone breathed a sigh of relief as the House was apportioned (Balinski and Young 1982).

This history of the creation of the census instrument in the Constitutional Convention and the first years of the republic reveals how high the stakes were in organizing this mechanism of the new government. The successful implementation of the population count and its use for legislative apportionment were absolutely essential to organizing the state itself.

I suspect the framers knew as much at the time, but I also suspect they recognized more deeply the importance of the census and apportionment instrument in future years. They did so because not only did the census provide a mechanism deemed fair to all parties for apportioning seats in the House among the states, it also provided a predetermined and automatic means to redistribute those seats each decade. Congress initially made only small incremental changes in the administrative structure for taking the census. The Secretary of State continued to oversee the count until 1850, when the temporary census office was transferred to the new Interior Department, and the temporary position of Census Superintendent was created to oversee the publication of an increasingly large volume of statistics. The US marshals and their assistants continued to serve as the field enumeration staff through the 1870 census. Occasionally Congress discussed the creation of a permanent statistical office to take the census and process other data, but the proposals did not materialize in legislative action.

The demographic results, on the other hand, were truly noteworthy. By the 1810s, Americans had taken three censuses and had allocated representatives to states of widely differing size. The allocations ranged from 1 to 23 members. But by the 1810s, with three data points and thus the capacity to calculate growth rates for two decades, they also recognized that the population was growing rapidly (to 7.2 million in 1810, or 30 to 35 percent a decade nationally). The populations of the states were growing at different rates. The House of Representatives grew from a

temporary constitutional apportionment of 65 members in 1789 to 186 members after the 1810 count. The western state of Kentucky, admitted to the union in 1792, with 2 representatives, had 10 after the 1810 apportionment, an allocation larger than 9 of the original 13 states.

By the time of the War of 1812, the shape of American population dynamics that would dominate nineteenth-century development and population politics was becoming clear. The population grew rapidly and differentially. The successful politician would be one who understood how to mobilize power of growing regions and local areas and in turn mobilize resources for such areas. Early national political leaders discovered that the decision to ground sovereignty in the people made real the "power of the people," to use a more modern phrase, in the structure of the state.

In short, the United States developed a politics of population even before the development of the mass party systems and the democratization of voting in the Jacksonian era of the 1830s. It is a lens through which to view a whole series of political issues of the early national era. Before the development of mass parties and the mass mobilization of voters for elections, political leaders had experience of a generation or more of learning that they had to learn to control (or manipulate) this popular base of the political system. The evidence of this learning process is clear in some of the standard political debates in the early national era.

At the most obvious, the fact of rapid and differential population growth required a delicate political debate about the size of the House of Representatives each decade. The House grew from the 186 members in 1810 to 243 members in 1860 on the eve of the Civil War. After the 1840 census Congress conducted a complex debate on apportionment methods and changed to a system developed by Daniel Webster. In 1850, it reformed the system again, using the Vinton method. For detail on the evolution of apportionment formulas and their political implications, see Balinski and Young (1982).

Congress also admitted new states to the union with a keen eye to the political implications of admitting the rapidly growing western territories to the union as states. The union grew in fits and starts. The story of the patterns of state admissions in the antebellum period is generally told as part of the history of the coming of the Civil War and the future of slavery, and so it is. But it is also a history of race, region, and the management of the relentless population politics built into the decennial reapportionment process. Only once in the 230-year history of the American republic did Congress fail to respond to the population changes of the census and reapportion. That was after the 1920 census and, though out of the scope of this paper, confirms the fundamental importance of the "politics of population" in American life (Anderson 2015).

The naming and refinement of the "gerrymander," or the conscious manipulation of the geographic boundaries of legislative districts to enhance the fortunes of one political group or faction, is another prominent "innovation" of the early national politics of population. Named for Elbridge Gerry, the Governor of Massachusetts, the term was coined to accompany a description of a cartoon of a legislative district in Essex County, Massachusetts, created after the 1810 census. Federalists who objected to the district ridiculed the "convex" shape of the district, and in later

years, the image and the name became common to criticize the designing motives of politicians who drew districts to their political advantage. The practice predates the early nineteenth century and currently is still litigated and debated in the United States in terms of "racial" and more recently "partisan" redistricting. What the practice does represent is the "discovery" of the capacity of an incumbent faction or party to use the reapportionment and redistricting process to its political advantage by literally drawing the rules of the game in the form of legislative districts (Griffith 1907; Grofman 1998; Grofman and Lijphart 1986; Kousser 1999).

A third aspect of the politics of population in the early national era is what would seem today a very strange political controversy about "slave representation." In 1800 Thomas Jefferson won the presidency and ended the domination by the Federalists of the national government. An intense period of factional conflict preceded the election, and the successful, and peaceful, election of Jefferson has often been dubbed "the revolution of 1800." The losing Federalists asked what went wrong and analyzed the votes in the Electoral College. Some pundits claimed that the margin of victory for Jefferson could be attributed to the votes of electors "representing" slaves in the South. Elaborate analyses of census results and the resulting House apportionments appeared, claiming a profound political impact for these "slave representatives." The writers proposed amending the Constitution to base representation on counts of the white population only. The sectional implication of doing so was quite obvious since northern states were in the process of legislating slave labor out of existence, and thus slavery was increasingly a labor system confined to the South. The rather crass proposal was to enhance northern strength in the national government and would continue to reverberate through later debates about sectionalism and the extension of slavery into new states. The writers did not follow the logic of their arguments about whether "female representation" or "child representation" or "alien representation" – all groups with no rights of suffrage – also had an impact on national politics (Anderson 2015; Dwight, Sereno Edwards (Boreas). 1812; Simpson 1941).

Finally, by the second and third decades of the republic, political leaders began to employ demographic arguments in their political rhetoric and respond to demands for service based on demographic need. The expansion of the postal network is a prime example, as Richard John (1995, 47, 49) has shown, of these demographic politics. In 1792, Congress granted "itself the authority to designate every postal route in the United States." John continues:

> The impetus for the expansion of the postal network owed a good deal to the structure of the House. Since congressional apportionment was based on population and constituents were constantly clamoring for new routes, there existed a built-in bias in favor of expanding the postal network on the basis of population rather than on the basis of the existing pattern of commercial demand.

Local citizens petitioned for service; every state generally received a seat on the House committee on the Post Office and Post Roads. The result was that the ratio of white population per post office declined from 43,000 in 1790 to around 1000 in 1840, and the number of post offices expanded from 75 to 13,468.

Congress also gave itself an expansive franking privilege and permitted the free mail exchange of newspapers among publishers. John describes the profound effect these combined policies had on American public life; they made possible, for example, the development of a "national community" well before the spread of new modes of transportation and communication (the telegraph and the railroad), which are often credited with the development of a truly national politics.

The emergence of a demographic political rhetoric is also evident in the spate of "statistical" publications by members of Congress – particularly after 1810 (Cohen 1982). The most elaborate of these was Adam Seybert's (1818) 803 page tome, *Statistical Annals*, with the rather long-winded subtitle, *Embracing Views of the Population, Commerce, Navigation, Fisheries, Public Lands, Post-Office Establishment, Revenues, Mint, Military & Naval Establishments, Expenditures, Public Debt and Sinking Fund of the United States of America* ((Seybert 1818), 1, 15, 17, 9). Seybert was a Pennsylvania Congressman who clearly understood the implications of the rapid population growth of the early republic. He began with some "preliminary observations":

> The state of civilized society and the resources of nations, are the tests by which we can ascertain the tendency of the government. It is to the condition of the people, in relation to their increase, their moral and physical circumstances, their happiness and comfort, their genius and industry, that we must look for the proofs of a mild and free, or of a cruel and despotic government.

Seybert opened his first chapter on population, noting, "That which most concerns every state is its population," yet most nations did not yet conduct regular population censuses. The success of the American census made "our practice . . . worthy of being followed by other nations." Seybert summarized his argument with a brief table of key indicators from the 1790, 1800, and 1810 censuses. He did not calculate growth rates, but the rapid population growth over the three census years was evident without it. Even more striking was the even more rapid growth in the revenue and expenditures of the federal government, which Seybert attributed to the "prosperity of the nation, because as far as our knowledge extends, *no individual* suffered from the *taxes* which were imposed, nor was he deprived of a single comfort, which he had before enjoyed" (emphasis in original). Reviewers of the volume outside the United States noted the hyperbolic rhetoric of Seybert's analysis, yet the dramatic patterns of growth did speak for themselves.

Finally Seybert compiled his tables from government documents available to Congress, noting that he had proposed a compendium because the "documents annually presented to Congress, are contained in more than one hundred and twenty volumes, and are too much diffused to be made the subjects of immediate reference... though they were intended to be used in debate." Thus Seybert (1818, vi) produced an "authentic book of reference" "to arrange the statements under a form that might be convenient and useful, and so that the details should be preserved."

Seybert's efforts point to growing Congressional awareness of the second leg of the federal statistical system, namely, the collection and publication of economic

statistics. These statistics would prompt the creation of permanent offices and officers to compile and publish the data.

The primary revenue base for the national government in the nineteenth century was the tariff. "Internal Revenue," derived from excise taxes or direct taxes, was a very minor portion of the federal revenue budget. Direct taxes, to be apportioned on the basis of population using the census, were levied in the late 1790s, during the War of 1812, and during the Civil War. An income tax, levied during the Civil War, was declared a violation of the Constitution in the 1890s and did not become part of the federal revenue system until the census clause was amended in 1913 to remove the requirement that such "direct taxes" had to be apportioned upon the basis of population (Anderson 2015). Thus the administrative record system of federal revenue collections was based primarily upon the records of annual imports and exports, and the tonnage of shipping.

The record keeping and reporting fell to the clerks in the Treasury Department and was done intermittently from 1790 to 1820, as Seybert noted. Legislative authorization of systematic annual reporting of such statistics came in 1820, as Congress provided that "the Register of the Treasury shall, under the direction of the Secretary of the Treasury, annually prepare statistical accounts of the commerce of the United States with foreign countries." The customs collectors were the field staff required to collect the information to submit to Washington. The Treasury Secretary was authorized to provide "such directions to the collectors and to prescribe such rules and forms as might appear proper." In 1844, Congress passed a joint resolution authorizing the Treasury Secretary to allocate "three or more clerks to be employed under his direction in collecting, arranging, and classifying such statistical information as may be procured, showing or tending to show each year the condition of the agriculture, manufactures, domestic trade, currency and banks of the several states and territories of the United States." In 1850, Congress authorized the regular publication and distribution of these annual reports and, in 1866, established a formal Bureau of Statistics in the Treasury Department with a bureau head in charge of statistical operations; monthly reporting of commercial statistics began. In 1878, the bureau inaugurated the *Statistical Abstract of the United States* ((Schmeckebier and Weber 1924), 9, 11, 21, and in later years became a staple of government publication for the Census Bureau, e.g., (U. S. Bureau of the Census 1999)). The 1878 edition was a modest affair, containing 150 tables in 157 pages, and focused primarily on the import and export statistics collected in the Treasury Department.

20.6 The Federal Statistical System Takes Shape

The inauguration of the annual publication of the *Statistical Abstract* provides a convenient marker to identify the emergence of what we now call the federal statistical system, including its characteristic decentralized structure, a combination of survey and administrative data collection, and at the time, commitment to

complete count data collection. The agencies which produced these demographic and economic statistics were diverse. A temporary Census Office in the Interior Department, created from scratch each decade, published voluminous information on the population, and oversaw decennial censuses of agriculture, manufacturing, and some other industries such as mining. The Bureau of Statistics in the Treasury Department published annual compilations of economic statistics, the *Statistical Abstract*, and monthly statistics on foreign commerce with the United States. The Bureau of Statistics relied on its administrative activities and personnel to provide the basic data for its statistical publications. The Census Office used a "survey-like" field staff of temporary employees who collected data for purely statistical purposes. Both agencies were beginning to develop more sophisticated technical expertise and define "statistical standards" for accuracy, consistency, efficiency, and precision.

The success of the demographic and economic statistics of the first two-thirds of the nineteenth century prompted Congress to propose further statistical efforts, using the same decentralized model. During and after the Civil War, for example, Congress put in place three ambitious new statistical initiatives, in agricultural statistics, labor statistics, and education statistics (Goldberg and Moye 1985; Leiby 1960; Smith 1923; Taylor and Taylor 1952; Warren 1974). In 1890, under pressure to tabulate the data quickly for apportionment and then produce dozens of volumes of statistical reports from more complex tabulations, the still "temporary" Census Office in the Department of Interior launched punch card machine tabulation of the census returns to mechanize the work.

By the turn of the century, there were a plethora of statistical agencies in the federal government, of varying competence, and with varying mandates and budgets. Different agencies published data on what might seem to be the same topic. Thus, the Census Office as well as the Agriculture Department produced agricultural statistics; the Treasury Department compiled statistics on agricultural commodities exported while the Agriculture Department published statistics on commodities produced. Immigration statistics were collected by the State Department and then by the Bureau of Immigration, part of the Treasury Department. Consular officials abroad, officials of the State Department, collected local wage data; the Labor Department collected wage data in the United States. Congress began to look to coordination and consolidation of the work as they considered proposals to create a new cabinet department charged with dealing with business, commerce, and manufacturing. With larger issues of government expansion on the agenda, statistical coordination and consolidation seemed politically possible.

The US population topped 76 million in 1900. In 1902 Congress made the Census Bureau a permanent statistical agency and removed the economic censuses from the decennial population census cycle. In 1903 Congress moved the Census Office from the Interior Department into the newly created Department of Commerce and Labor. Several other statistical agencies, including the Department of Labor (downgraded to a Bureau of Labor), the Bureau of Statistics of the Treasury Department, the Bureau of Foreign Commerce of the State Department, were relocated to the new Department. The organic legislation for the Department authorized the Secretary and the President to consolidate agencies and functions with an eye toward creating

a central statistical agency. These and later efforts at centralization failed (Anderson 2015; Bonnen 1984; Norwood 1995; Office of Federal Statistical Policy and Standards 1978; Parmelee 1910–11; President's Commission on Federal Statistics 1971).

The failure of centralization did not halt statistical innovation, as Americans' voracious appetite for statistical data did not abate and arguably prompted more informal efforts at statistical coordination and the reintegration of the mathematical side of the discipline into "statistics in the public interest."

Through the first third of the twentieth century, most federal statistics were still basically "complete count" tabulations of administrative data records, or true censuses. But policy demanded new kinds of data, surveys collected more frequently, to monitor conditions that changed rapidly, and for which lumbering complete count methods could simply not provide timely data. The crisis of the Great Depression forced the issue, particularly creating the need to measure unemployment, a rapidly changing phenomenon that required simple frequent data collection. Probability sampling became the solution, first in the experimental sample surveys that led to the "monthly report on the labor force," and after World War II to the Current Population Survey. Economic data were marshaled to build the National Income and Products Accounts, deployed successfully during World War II to guide the war effort. Computerization of statistical research soon followed, with wartime innovations in a digital computer, ENIAC, the Electronic Numerical Integrator and Computer, providing the inspiration for UNIVAC, Universal Automatic Computer, to tabulate the 1950 census (Duncan and Shelton 1978).

These mathematical and technical innovations in turn produced the rich and varied world of statistics in the public interest that we know today. Probability sampling methodology required the calculation of margins of error, just as the astronomers had been doing since the early nineteenth century. And sampling could be applied not only to conduct new types of surveys, public opinion polling par excellence, it also could be integrated into complete count data collections, to lessen the burden on respondents by moving questions to a "long form" sample for the census, for example, or by adding evaluation samples to measure bias and efficiency in survey administration. Anecdotal evidence of undercount could be replaced with demographic analysis of vital statistics, or post enumeration surveys to measure coverage. As computer technology advanced, government statistical agencies recognized that they could release microdata samples for external researchers, as well as the tabulations they traditionally published. The first of those came with the 1960 census; researchers could buy either the punch card or the tape version to use on their own computer platform.

This "revolution" in government statistics, as Duncan and Shelton call it, was also something of a fall from grace, as the greatly broadened statistical and policy community recognized not only the power of their data and techniques, but also their limitations. Census undercount, Steve Fienberg's and my original collaboration, became a particularly troublesome case in point.

Once the bureau was able to document scientifically, i.e., measure, the undercount, it didn't take long for politicians to recognize the implications for reap-

portionment, redistricting, and federal funds allocation. By the late 1960s, they pressed the bureau not only to measure the undercount but to "fix it" – improve the accuracy of the count. And when the Census Bureau admitted it could not do so, local jurisdictions and stakeholder groups sued the Commerce Department.

The Bureau faced suits to correct for the known differential undercount in the census cycles from 1970 to 2000. The most serious challenges were to the 1990 and 2000 censuses, where the litigation reached the Supreme Court for resolution in 1996 and 1999 (Anderson and Fienberg 1999a). The Census Bureau also improved its operational control over the census field enumeration. The innovations reduced but did not eliminate the differential between the white non-Hispanic count and minority undercounts. The procedure proposed for adjustment was a large post enumeration survey matched to the census returns and capture – recapture calculations (dual system estimation) to estimate the undercount of demographic groups and local areas. The Bureau built the adjustment methodology into the designs of the 1990 and 2000 census. In the 1990s, the courts ruled that the Commerce Secretary could authorize an adjustment if the results improved the accuracy of the "actual enumeration." In 1999, the Supreme Court clarified the legal status of adjustment further, ruling that the current statutory language of Title XIII, the Census Act, bars adjustment for apportionment numbers, but not for redistricting and other census uses.

In both decades, the Secretary decided not to adjust. The Census Director recommended adjustment for the 1990 census, and the courts upheld the Secretary's decision. The Bureau did not recommend adjustment after the 2000 census because of newly discovered problems with the adjustment methodology, particularly the discovery of a large number of duplicate enumerations which undermined the Bureau's confidence in the dual systems estimation results (National Research Council 2004). The 2010 census design included a post enumeration survey with dual systems estimation for evaluation, but not for adjustment.

20.7 New Challenges

The foregoing narrative highlights some key events in the development of "statistics in the public interest" in the United States. It is illustrative of my understanding of how to think about the long continuities in American statistical practice, and many excellent studies flesh out complementary details.[2] I'd like to conclude by

[2]The literature is voluminous, and a thorough catalog is beyond the scope of this essay. Nevertheless, in addition to the studies cited infra, see, for example, Fisher (1992); Citro and Michaels (1995) on poverty measurement; (Lopresti (2017) on statistics on women and gender; Hochschild and Powell (2008); Mezey (2003); Morning (2008); Nobles (2000); Perlmann (2018); Perlmann (2001); Perlmann and Waters, eds. (2002); Prewitt (2013); Rodriguez (2000); Samhan (1999); Schor (2017); and Williams (2006) on race, ethnic, and immigrant classification issues. On international comparisons and public policy, see, for example, Curtis (2001); Desrosières (1998);

contrasting some lessons from this narrative with the current challenge posed by the Trump administration's politicization of the 2020 census, starting with the effort to add a citizenship question to the 2020 decennial complete count census form.

On March 26, 2018, Secretary of Commerce Wilbur Ross informed Congress that "I have determined that reinstatement of a citizenship question on the 2020 decennial census is necessary to provide complete and accurate data in response to the *DOJ* request."[3] Within days, the controversy exploded publicly as civil rights organizations and state and local government officials announced their intentions to sue the Commerce Department to reverse the decision.

Ross publicly claimed that his decision to add the citizenship question to the form was not problematic, that the question had been asked for over a century on the census or the American Community Survey, and that it would not generate logistical or technical problems for the 2020 count. Yet word of Ross' intentions had been percolating among census stakeholders for several months and had generated a dramatic behind the scenes debate opposing adding the question. For example, in a January 2018 letter to Ross, six former census directors in both Republican and Democratic administrations wrote

> . . . we believe that adding a citizenship question to the 2020 Census will considerably increase the risks to the 2020 enumeration. Because we share your goal of a "full, fair, and accurate census," as the Constitution requires, we urge you to consider a prudent course of action in response to the Justice Department's untimely and potentially disruptive request.[4]

Critics charged that the Trump administration was intentionally trying to suppress census responses from immigrant communities. Ross' decision came too late for testing the efficacy of the question, and technical experts warned of known problems. The traditional "dress rehearsal" for the census in Rhode Island in spring

Patriarca (1996); Glass and Victor (1978); Higgs (1989); Deacon (1985); and Alonso and Starr, ed. (1987).

[3]Ross claimed that in December 2017, the Justice Department requested that he "reinstate a citizenship question on the decennial census to provide census block level citizenship voting age population ("CVAP") data that are not currently available from government survey data." The Justice Department claimed that "having these data at the census block level will permit more effective enforcement" of the Voting Rights Act. See Secretary Wilbur Ross to Karen Dunn Kelley, Under Secretary for Economic Affairs, March 26, 2018, available at https://www.commerce.gov/sites/default/files/2018-03-26_2.pdf

[4]The former directors continued: "It is highly risky to ask untested questions in the context of the complete 2020 Census design. There is a great deal of evidence that even small changes in survey question order, wording, and instructions can have significant, and often unexpected, consequences for the rate, quality, and truthfulness of response. The effect of adding a citizenship question to the 2020 Census on data quality and census accuracy, therefore, is completely unknown. Also of import, overcoming unexpected obstacles that arise as 2020 Census operations unfold would add to the cost, without assurances that such efforts would yield a more accurate outcome." For the text of the letter see Vincent P. Barabba (1973–1976; 1979–1981); Martha Farnsworth Riche (1994–1998); Kenneth Prewitt (1998–2001); Steven H. Murdock (2008–2009); Robert M. Groves (2009–2012); John Thompson (2013–2017) to Wilbur L. Ross, January 26, 2018 available at https://www.washingtonpost.com/r/2010-2019/WashingtonPost/2018/03/27/Editorial-Opinion/Graphics/DOJ_census_ques_request_Former_Directors_ltr_to_Ross.pdf.

of 2018 didn't have the question on the form. Critics also charged that the question had not been asked on the main census form since 1950 and that when it was asked on the main census form from 1890 to 1950, it was only asked of the foreign-born population. In 1990 and 2000, a question on citizenship was asked of sample households that received the detailed "long form" census form. Since 2000, it has been asked on the American Community Survey sample. In other words, the question has never appeared on the census to be asked of all persons.

Ross' action and the intense opposition it generated reignited the public debate about the functions of the census for reapportionment and redistricting, the dynamism and diversity of the US population, and the need for a fair census so that both the "winners" and the "losers" in the reapportionment and redistricting process will see the outcome as legitimate. The controversy reminded Americans that the census is an instrument of government, but it should not be a partisan or biased instrument with controversial questions that affect the quality of the results.

Since the spring of 2018, seven major federal lawsuits challenged Ross' decision to include a citizenship question on the 2020 census. The litigation process generated a detailed administrative record of the Trump administration decision making. Three trials were held, in New York City, San Francisco, California, and Maryland. All three federal courts ruled against the government on the grounds that Ross' action violated the Administrative Procedures Act. The judges in the Maryland and California cases also found that Ross' action violated the enumeration clause of the Constitution. The New York case was expedited to Supreme Court review, with oral arguments on April 23, 2019, so that the court could rule before its summer recess, and the Census Bureau could finalize the printing contract for census forms.

On June 27, 2019, Chief Justice Roberts affirmed the lower court decision and wrote the opinion for a slim 5–4 majority (Department of Commerce, et al. v. New York, et al. 2019). The court found that Ross' decision making had violated the Administrative Procedures Act. Printing was slated to start by July 1. For 2 weeks, the President pressed to find a way to get the question on the form, despite the fact that the Commerce Department authorized the printing contract to go into effect in the first week of July. Trump suggested an addendum, even postponing the census. On July 11, he relented and acknowledged that the 2020 census form would not have a citizenship question, but ordered the Census Bureau to produce small area citizenship data from the American Community Survey and administrative records. The potential constitutional crisis of a president defying a Supreme Court decision abated. But the threats and challenges to the 2020 count were not over. New questions immediately arose – about the impact of the controversy on people's willingness to respond to the census in April 2020, the cost of the census, on cybersecurity, on the readiness of the computer systems underpinning the new mass internet response option. And then in the spring of 2020, as the mail and Internet response options began, the coronavirus pandemic forced postponement of the labor intensive in person counting operations, nonresponse follow-up (NRFU), service-based enumerations, list-enumerate, among others. The administration requested relief on the statutory reporting deadlines in Title 13 in April 2020 and announced a

revised operational plan envisioning the delayed operations to resume in the summer of 2020.

Since, the Trump administration renewed its political meddling with census operations, appointing unqualified new political appointees in June 2020, reneging on the request for an extension of reporting, among other initiatives. See Prewitt discussion in this volume.

As of this writing, in mid-September 2020, the situation remains fluid. The administration has already lost two more legal challenges in federal court, and the dates for the completion of census field work and post processing are being litigated. At the moment, the ultimate resolution of the impact of the pandemic and the political manipulation on the quality of the 2020 census results is not clear, but the count is clearly under threat.

The statistical community almost universally objected to the introduction of the citizenship question on the 2020 census form, arguing that Ross' hasty and cavalier decision undermined the trust in the federal statistical system, by disregarding the technical and statistical expertise that have been carefully and systematically integrated into law and administrative practice over the past 200 years. It has resumed its expressions of alarm in summer 2020. As Teresa Sullivan (2019) argued in her President's Invited Address at the 2019 Joint Statistical Meetings, it is no mean accomplishment that American "statistics in the public interest," born in revolution and political decision making, nevertheless have come to be regarded worldwide as exemplars of scientific innovation and objectivity.[5] The Trump administration's actions since 2017 are an unprecedented attack on the statistical system, and the American statistical community is facing a truly existential moment in the years ahead.

Honoring Steve Fienberg's work and his legacy is a good way to begin to address that moment.

References

Alonso William and Paul Starr (1987) *The Politics of Numbers* (Population of the United States in the 1980s: A Census Monograph Series). New York: Russell Sage Foundation.

Anderson, Margo (2015). *The American Census: A Social History*, second ed. New Haven, Yale University Press.

Anderson, Margo (1988). *The American Census: A Social History*, New Haven, Yale University Press.

Margo Anderson and. Fienberg, Stephen E. (1999a). *Who Counts? The Politics of Census Taking in Contemporary America*. New York: Russell Sage Foundation.

Anderson, Margo and Fienberg, Stephen E. (1999b). "The History of the First American Census and the Constitutional Language on Censustaking: Report of a Workshop," Report to the Donner Foundation <http://lib.stat.cmu.edu/~fienberg/DonnerReports> July, 1999.

[5] A revised version of the address with commentary is available in the *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/

Margo Anderson and Seltzer, William, (2009) "Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues," *Journal of Privacy and Confidentiality* 1 (Spring 2009), 7–52; Comment on Article by Anderson and Seltzer, by C. L. Kincannon, 53–54; Rejoinder by M. Anderson and W. Seltzer, 55–58.

Bailyn, Bernard (1967). *The Origin of American Politics*. New York Vintage Books, 1967.

Balinski, Michel and Young, H. Peyton (1982). *Fair Representation: Meeting the Ideal of One Man, One Vote*. New Haven, Yale University Press.

Bonnen, James (1984). "Federal Statistical Coordination Today: A Disaster or a Disgrace," *Milbank Memorial Fund Quarterly/Health and Society*, 62 (1984), 1–41.

Cassedy, J. (1969). *Demography in Early America: Beginnings of the Statistical Mind, 1600–1800*. Cambridge: Harvard University Press.

Citro, Constance F. and Michaels, Robert T., eds. (1995). *Measuring Poverty: A New Approach*, Washington, D.C., National Academy Press.

Cohen, Patricia Cline (1982). *A Calculating People: The Spread of Numeracy in Early America*, Chicago: University of Chicago Press.

Curtis, Bruce (2001) *The Politics of Population: State Formation, Statistics and the Census of Canada, 1840-1875*. Toronto: University of Toronto Press.

Deacon, Desley (1985). "Political Arithmetic: The Nineteenth-Century Australian Census and the Construction of the Dependent Woman," *Signs*, vol. 11, no. 1 (1985), 27–47.

Department of Commerce, et al. v . New York, et al. (2019) https://www.supremecourt.gov/opinions/18pdf/18-966_bq7c.pdf

Desrosières, Alain. (1998) *The Politics of Large Numbers: A History of Statistical Reasoning*, Translated by Camille Naish. Cambridge: Harvard University Press.

Duncan, Joseph, and Shelton, William (1978) *Revolution in United States Government Statistics, 1926-1976*. Washington, D.C.: GPO.

Dwight, Sereno Edwards (Boreas). (1812). "Slave Representation", New Haven, No publisher, 1812, first appeared as two essays in the Connecticut *Journal*, New Haven.

Fisher, Gordon (1992). "The Development and History of the Poverty Thresholds." *Social Security Bulletin*. 55 (Winter 1992): 3–14.

Fitzpatrick, Paul J. (1957). "Statistical Societies in the United States in the Nineteenth Century." Author(s): *The American Statistician*, Vol. 11, No. 5 (Dec., 1957), 13–21. Stable URL: https://www.jstor.org/stable/2682565

David Victor Glass, David Victor (1978) *Numbering the People: the Eighteenth-Century Population Controversy and the Development of Census and Vital Statistics in Britain* (London: Gordon & Cremonesi.

Goldberg, Joseph P. and Moye, William T. (1985). *The First Hundred Years of the Bureau of Labor Statistics*. Washington, D.C., GPO.

Griffith, Elmer (1907). *The Rise and Development of the Gerrymander*. Chicago, Scott, Foresman and Company.

Grofman, Bernard, ed. (1998). *Race and Redistricting in the 1990s*. New York: Agathon Press.

Grofman, Bernard and Arend Lijphart, eds. (1986). *Electoral Laws and Their Political Consequences*. New York, Agathon Press.

Heyde, C.C. (2001). "John Graunt," in Heyde, C., & Seneta, E. *Statisticians of the Centuries*. New York: Springer, 14–16.

Higgs, Eddy (1989) *Making Sense of the Census: The Manuscript Returns for England and Wales, 1801-1901*. London : H.M.S.O.

Hochschild, Jennifer L. and Brenna Marea Powell (2008) "Racial Reorganization and the United States Census 1850-1930: Mulattoes, Half-Breeds, Mixed Parentage, Hindoos, and the Mexican," in *Studies in American Political Development,* 22 (Spring 2008), 59–96.

John, Richard (1995). *Spreading the News: The American Postal System from Franklin to Morse*. Cambridge, Harvard University Press.

Kousser, J. Morgan (1999). *Colorblind Injustice: Minority Voting Rights and the Undoing of the Second Reconstruction*. Chapel Hill, University of North Carolina Press.

Leiby, James (1960). *Carroll Wright and Labor Reform: The Origins of Labor Statistics*. Cambridge, Harvard University Press.
Lewis, Jan (1995). "'Of Every Age Sex & Condition': The Representation of Women in the Constitution," *Journal of the Early Republic*. 15 (Fall 1995): 359–87.
Lopresti, R. (2017). *When Women Didn't Count: The Chronic Mismeasure And Marginalization Of American Women In Federal Statistics*. Santa Barbara, California: Praeger.
Mezey, Naomi (2003) "Erasure and Recognition: The Census, Race and the National Imagination" *Northwestern University Law Review*. 97 (2003), 1701–1768.
Morgan, Edmund (1988). *Inventing the People: The Rise of Popular Sovereignty in England and America*, New York: Norton.
Morning, Ann (2008) "Ethnic Classification in Global Perspective: A Cross-National Survey of the 2000 Census Round," *Population Research and Policy Review*, 27, 2 (2008), 239–272
National Research Council. 2004. *The 2000 Census: Counting Under Adversity*. Washington, DC: The National Academies Press. https://doi.org/10.17226/10907.
National Research Council. 1995. *Modernizing the U.S. Census*. Washington, DC: The National Academies Press. https://doi.org/10.17226/4805.
Nobles, Melissa (2000) *Shades of Citizenship: Race and the Census in Modern Politics.* Stanford CA: Stanford University Press.
Norwood, Janet (1995). *Organizing to Count: Change in the Federal Statistical System*. Washington, D.C.: Urban Institute Press.
Office of Federal Statistical Policy and Standards (1978). *A Framework for Planning U.S. Federal Statistics in the 1980's*. Washington, D.C.: GPO.
Parmelee, Julius (1910–11). "The Statistical Work of the Federal Government." *Yale Review*, 19 (November 1910/February 1911): 289–308; 374–91.
Patriarca, Silvana. (1996) *Numbers and Nationhood: Writing Statistics in Nineteenth-century Italy.* New York: Cambridge University Press.
Perlmann, Joel (2018). *America Classifies The Immigrants : From Ellis Island To The 2020 Census*. Cambridge, Massachusetts: Harvard University Press.
Perlmann, Joel (2001) '*Race or People': Federal Race Classifications for Europeans in America, 1898–1913* (Annandale-on-Hudson, NY: Levy Economics Institute.
Perlmann Joel and Mary Waters, eds. (2002) *The New Race Question: How the Census Counts Multiracial Individuals.* New York: Russell Sage Foundation Press.
Porter, Theodore. (1986). *The Rise of Statistical Thinking, 1820-1900*. Princeton, N.J.: Princeton University Press.
Prewitt, Kenneth (2013) *What Is Your Race?: The Census And Our Flawed Efforts To Classify Americans*. Princeton, N.J.: Princeton University Press.
President's Commission on Federal Statistics (1971). *Report of the President's Commission* (2 vols.), Washington, DC: GPO.
Rodriguez, Clara (2000) *Changing Race: Latinos, the Census, and the History of Ethnicity in the United States*. New York: New York University Press.
Samhan, Helen (1999) "Not Quite White: Race Classification and the Arab-American Experience." in *Arabs in America: Building a New Future*. Michael W. Suleiman, ed. Philadelphia: Temple University Press
Schmeckebier, Laurence F. and Weber, Gustavus A. (1924). *The Bureau of Foreign and Domestic Commerce: Its History, Activities and Organization*, Baltimore, MD: The Johns Hopkins Press, 1924; reprinted 1974, AMS Press, NY.
Schor, Paul. (2017). *Counting Americans: How the US Census Classified The Nation*. New York, NY: Oxford University Press.
Seybert, Adam (1818). *Statistical Annals*. Philadelphia, Thomas Dobson & Son, 1818; reprinted by Augustus M. Kelley, 1970.
Simpson, Albert F. (1941). "The Political Significance of Slave Representation, 1787-1821," *Journal of Southern History*, 7 (1941): 315–42.
Smith, Darrell Hevenor (1923). *The Bureau of Education: Its History, Activities and Organization*, Baltimore, Johns Hopkins Press, 1923.

Stigler, Stephen. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Sullivan, Teresa. (2019). "Coming to Our Census: How Social Statistics Underpin Our Democracy (and Republic). Paper presented at the Joint Statistical Meetings, August 2019, Denver, DO. Available at https://ww2.amstat.org/meetings/jsm/2019/webcasts/.

Taylor, Henry C. and Taylor, Anne Dewees (1952). *The Story of Agricultural Economics in the United States. 1840-1932*, Ames, IA: Iowa State College Press.

U. S. Bureau of the Census (1999). *Statistical Abstract of the United States, 1998*. Washington, D.C.: GPO.

Warren, Donald (1974). *To Enforce Education: A History of Founding Years of the United States Office of Education,* Detroit, Wayne State University Press.

Williams, Kim (2006) *Mark One or More: Civil Rights in Multiracial America* (Ann Arbor, MI: University of Michigan Press, 2006).

Wood, Gordon (1969). *The Creation of the American Republic, 1776-1787*. Chapel Hill: University of North Carolina Press.

Wood, Gordon (1992). *The Radicalism of the American Revolution*. New York: A.A. Knopf.

Zagarri, Rosemarie (1987). *The Politics of Size: Representation in the United States, 1776-1850.* Ithaca, NY, Cornell University Press.

Chapter 21
Fostering Statistical Rigor for Evidence-Based Policy at the National Academies of Sciences, Engineering, and Medicine

Constance F. Citro, Michael L. Cohen, and Porter E. Coggeshall

21.1 Introduction

Stephen E. Fienberg ("Steve" to his legions of colleagues and friends) was deeply dedicated to the improvement of evidence for public policy and understanding by the application of rigorous statistical methods. His boundless energy, willingness to work on a wide range of applied statistics problems relevant to public policy, keen intellect, and—yes—commanding presence and impatience with incremental change enabled his stellar achievements.

An important arena in which he exercised his gifts to the full was the National Academies of Sciences, Engineering, and Medicine (see Box 21.1). Steve was elected to the honorific National Academy of Sciences (NAS), which has fewer than 2500 members, in 1999, but his pro bono service to the nation through the National Academies began over 20 years earlier when he first served on a National Research Council (NRC) expert consensus committee that produced the reports, *Rehabilitation of Criminal Offenders: Problems and Prospects* (1979) and *New Directions in Rehabilitation of Criminal Offenders* (1981). Steve served on consensus study committees, workshop planning groups, standing oversight boards and committees, and the all-important National Academies' Report Review Committee (RRC), which exercises life or death authority over NRC consensus study reports. He also served on the Governing Council of the NAS and many of

C. F. Citro (✉) · M. L. Cohen
National Academies of Sciences, Engineering, and Medicine, Washington, DC, USA
e-mail: ccitro@nas.edu

P. E. Coggeshall *(deceased)*
National Academy of Sciences, Vienna, VA, USA

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_21

its committees, including those involved in selection of NAS members. All in all, from 1977 until just a few months before his death in December 2016, Steve was continuously involved in National Academies' activities (NAS and NRC) over that 40-year period.

We three—Connie Citro, Mike Cohen, and Porter Coggeshall[1]—were honored to have known Steve and his work for the National Academies over many years. We discuss his National Academies' work by drawing on our personal knowledge and experience, supplemented by information from National Academies' internal databases. We address his service to the nation at the National Academies in three parts. First, Connie details Steve's extraordinary record of involvement in NAS and NRC activities. Then Mike discusses several consensus studies to which Steve contributed, their statistical aspects, and the centrality and impact of Steve's contributions for the public good. Finally, Porter describes Steve's integral role in assuring the quality of National Academies' consensus reports through his service as member and cochair of the RRC.

Box 21.1 The National Academies

On March 3, 1863, President Abraham Lincoln signed an act incorporating the National Academy of Sciences, a self-perpetuating, honorific society that "shall, whenever called upon by any department of the Government, investigate ... and report upon any subject of science or art." The NAS established the National Academy of Engineering in 1964 and the Institute of Medicine in 1970 (renamed the National Academy of Medicine in 2015). To tap the expertise in the broader scientific and technical community for its studies, the Academy created an operational and staff arm—the National Research Council—in 1916. In accordance with the original charter, experts serve pro bono on National Academies' study and oversight committees. Today the National Academies are an independent 501(c)(3) nonprofit organization; studies are conducted at the request of government agencies and foundations through contracts and grants.

The NRC has more than 50 standing committees and boards, which function to oversee and develop a portfolio of work related to their missions. Steve devoted much of his energies to the work of the Committee on National Statistics (CNSTAT), one of the oldest such standing committees, established in 1972 at the recommendation of the President's Commission on Federal Statistics, chaired by W. Allen Wallis.

[1]Connie Citro is a senior scholar with the Committee on National Statistics (she directed CNSTAT from 2004–2017); Mike Cohen is a senior program officer with CNSTAT; Porter Coggeshall directed the Report Review Committee from 1992 to 2017.

21.2 Service to the Nation Through the National Academies

Service to the National Research Council

At a very tender age, in terms of the usual age at which experts are appointed to National Academies' study committees, the then 35-year-old Steve Fienberg was appointed in 1977 to the study on research on rehabilitation techniques (cited above). In the course of this study, Steve often encountered Miron Straf, research director for CNSTAT, who had overlapped with Steve at the University of Chicago in the late 1960s. Steve also knew Bill Kruskal, the founding chair of CNSTAT, from his Chicago days. Those contacts likely led to the appointment of Steve to CNSTAT in 1978.

Steve served a 3-year term on CNSTAT and then became its third chair, serving from 1981 to 1987 (with a sabbatical year in 1984–1985 at the Center for Advanced Study in the Behavioral Sciences). During these years, he chaired a CNSTAT subcommittee that produced the landmark report, *Sharing Research Data* (1984), and was a member of the CNSTAT Advanced Research Seminar on Cognitive Aspects of Survey Methodology, which produced another landmark report, *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines* (1984). Indeed, the Seminar was his idea—he pushed hard for it to happen and then arranged for its work to continue under the auspices of the Social Science Research Council via a Committee on Cognition and Survey Research.

Connie well remembers a dinner in late 1983 at which Steve interviewed her for a position as director of a study that produced *The Bicentennial Census: New Methodology for 1990* (1985). Mike also worked on that study. We vividly recall Steve's enthusiasm for the study (he served as a member ex officio) and his out-of-the-blue proposal for a joint meeting with CNSTAT for part of the study committee's second meeting. As still very much newbies in the ways of the National Academies, we felt overwhelmed by Steve's blithe assumption that we could make this complicated event work, but as it turned out, he was right. Among the luminaries who participated in the meeting was John Tukey, who famously looked to be asleep for much of the time but offered penetrating remarks when called on.

Steve went on to serve on 15 more study committees (8 of which were overseen by CNSTAT) that produced the following reports:

1. *Priorities for the 1990 Census Research, Evaluation and Experimental (REX) Program* (1988) [link not available]
2. *Statistical Models and Analysis in Auditing: A Study of Statistical Models and Methods for Analyzing Nonstandard Mixtures of Distributions in Auditing* (1988) [https://doi.org/10.17226/1363]
3. *The Evolving Role of Statistical Assessments as Evidence in the Courts* (1989) (Steve served as cochair) [published by Springer]
4. *Providing Access to Epidemiological Data* (1990) [link not available]

5. *Assessing Evaluation Studies: The Case of Bilingual Education Strategies* (1992) (Steve served as chair) https://doi.org/10.17226/2014
6. *Modernizing the U.S. Census* (1995) [https://doi.org/10.17226/4805]
7. *The Polygraph and Lie Detection* (2003) (Steve served as chair) [https://doi.org/10.17226/10420]
8. *Measuring Racial Discrimination* (2004) [https://doi.org/10.17226/10887]
9. *Vaccine Safety Research, Data Access, and Public Trust* (2005) [https://doi.org/10.17226/11234]
10. *Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment* (2008) [https://doi.org/10.17226/12452]
11. *Secondhand Smoke Exposure and Cardiovascular Effects: Making Sense of the Evidence* (2010)[https://doi.org/10.17226/12649]
12. *The Potential Consequences of Public Release of Food Safety and Inspection Service Establishment-Specific Data* (2011) [https://doi.org/10.17226/13304]
13. *Using Science as Evidence in Public Policy* (2012) [https://doi.org/10.17226/13460]
14. *Options for Estimating Illegal Entries at the U.S.-Mexico Border* (2013) [https://doi.org/10.17226/13498]
15. *Furthering America's Research Enterprise* (2014) [https://doi.org/10.17226/18804]

To mention just one anecdote, Steve's service on the committee that produced *Measuring Racial Discrimination* involved him in friendly but hard-fought arguments about statistical rigor and causality with experts in other fields, including social psychology and econometrics. For the report, Steve came up with what we believe to be the unique idea of using an analogy based on *The Sneetches* by Dr. Seuss, to illustrate the issue of establishing a counterfactual for causal analysis (see Box 21.2).

In addition to serving on study committees, Steve served on oversight committees for the NRC. Having early on served as member and chair of CNSTAT, 20 years later he served two terms—from 1998 to 2004—on the advisory committee to the Division of Behavioral and Social Sciences and Education, which oversees the portfolios of a number of boards and standing committees, including CNSTAT. In 1999 Steve was elected to the NAS, and while he continued to serve on study committees, his oversight role shifted to that body.

Box 21.2 From *The Sneetches* to Causal Inference Theory

. . . Those who attempt to identify the presence or absence of discrimination typically observe an individual's race (e.g., black) and a particular outcome (e.g., earnings) and try to determine whether that outcome would have been different had the individual been of a different race (e.g., white). In other words, to measure discrimination researchers must answer the counterfactual

(continued)

question: What would have happened to a nonwhite individual if he or she had been white? ...

To illustrate the problem, we turn to a classic Dr. Seuss book, *The Sneetches* (published in 1961), which describes a society of two races distinguished by markings on their bellies. In the story, one race of Sneetches is afforded certain privileges for having stars on their bellies, and the other race, lacking these markings, is denied those same privileges. There are, however, Star-On and Star-Off machines that can alter the belly and therefore the race of both Plain-Belly and Star-Belly Sneetches. Thanks to these machines, an individual Sneetch's racial status and various outcomes could be observed more than once, both as a Plain-Belly and a Star-Belly Sneetch.

In *The Sneetches*, ... one could readily answer the counterfactual question, saying with certainty what would have happened to a Plain-Belly Sneetch had he or she been a Star-Belly Sneetch (or vice versa).... [Yet] it is nearly impossible in the real world to observe the difference in outcomes across race for a single person; one must instead draw causal inferences.

SOURCE: Chapter 5, Causal inference and the assessment of racial discrimination. National Research Council. 2004. *Measuring Racial Discrimination*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/10887.

Service to the NRC Report Review Committee

Deserving special mention is Steve's role as member and cochair of the RRC through which he profoundly affected the NRC study process in ways that elevated statistical rigor and evidence-based work NRC-wide. As Porter describes in the third section of this chapter, Steve contributed to the institutional review process at all three levels: (1) as reviewer providing insightful comments on draft reports, (2) as RRC monitor assessing whether or not the authoring committee had fully addressed all of the reviewers' comments, and (3) as RRC cochair recommending public release of the report after a comprehensive review process had been satisfactorily completed. In each role, Steve was well known for his high standards.

Not only Porter, but also Connie and Mike, saw Steve in action many times on the RRC because we were often consulted by project staff and the RRC regarding reports that raised particularly contentious issues regarding the quality of the evidence. We were also consulted by study staff from other units in the NRC as to how best to prepare for what they were sure would be thunderbolts from on high when their reports entered review. Indeed, Connie was often asked to accompany study staff when, at the very beginning of the study process, they had to seek approval of the statement of task and plan of work by the NRC Governing Board

Executive Committee (GBEC) and knew they would encounter Steve. He faithfully dialed into those monthly GBEC meetings and in his role as RRC cochair would raise questions about the scope and approach of a proposed study and invariably suggest that one or more statisticians be added to the committee slate. It is not too much to say that Steve was probably one of the most influential of the RRC's leaders over its long existence dating back to 1969.

Service to the National Academy of Sciences

Steve's participation in activities of the honorific NAS included service:

- On membership committees for most years from 2001 to 2009 and again in 2013–2014
- As chair of Section 32: Applied Mathematical Sciences, from 2006 to 2009
- On the Editorial Board of the *Proceedings of the National Academy of Sciences* in 2011–2012
- On the NAS Public Welfare Medal Selection Committee from 2014 to 2016
- On the NAS Council and its Executive Committee from 2013 to 2016
- On NAS Council committees on scientific programs, membership affairs, and international affairs from 2013 to 2016
- On the NAS Committee on Executive Compensation from 2013 to 2016

Through his NAS service, Steve was in a position to build the influence of statistics through promoting membership for outstanding individuals in the field. He also reached the highest levels of the NAS, serving on the NAS Council and its Executive Committee. All NAS members are giants in their fields, but few have contributed as much to the NAS and the NRC, and thereby to the nation, as Steve Fienberg.

21.3 Contributions to Public Policy Through Statistical Rigor

All 17 NRC consensus study committees that Steve served on (a few times as chair) had a methodological focus. Some studies involved meta-analysis, where a summary of the methodological findings was needed to provide readers with an understanding of what is known about the strength of the associations, especially the causal linkages, for some response of interest. This was generally accomplished through the application of one or more statistical models. Some studies described a variety of methodological approaches that were or could be used to answer policy questions of current interest, along with the pros and cons of adopting each of

these approaches, and whether any were preferable. Below we briefly describe five examples of these methodological studies that addressed important policy issues that are still highly relevant today: bilingual education, use of the polygraph, racial discrimination, terrorism, and illegal immigration.

It is important to point out that every NRC report reflects the consensus views of all committee members (unless a dissent is written). Therefore, while Steve was one of several voices on all these reports—albeit a very strong one—we do not know which of the various methods mentioned in each study were initially raised by him. What we do know is that he approved each committee's findings and recommendations. Furthermore, for two of the five studies we review—bilingual education and use of the polygraph—Steve chaired the committee and therefore played a leading role in arriving at committee consensus.

Assessing Evaluation Studies: The Case of Bilingual Education Strategies (1992)

This study, which Steve chaired, was tasked with evaluating the two then-most prominent studies on the benefits of bilingual education programs, one being a national longitudinal study of the benefits of various approaches to teaching limited-English-proficient students, and the other a longitudinal study comparing three different instructional strategies for this population of students: immersion where teachers understand Spanish but respond in English, early exit where students are placed in classes taught in English as soon as possible, and late exit where both languages are maintained over a longer period of time. The committee was asked to review the data collection and analysis carried out in these two studies, determine whether additional analyses would strengthen the findings, and suggest alternative ways to compare the different instructional strategies. The report made clear that statistical models used to support public policies needed to be carefully evaluated for their assumed associations and for their causal inferences:

> Users of statistical methods often wish to draw causal conclusions, for example, from programs to achievement outcomes. This is especially true in a policy setting. If one concludes that when a school follows approach *X* to bilingual education, the performance and achievement of the students will be *Y,* one is claiming, at least in a loose sense, that *X* "causes" *Y*. The notion of which [research] designs allow conclusions about the causal effects of treatments is critical to an appreciation of the evaluation of alternative bilingual education programs... There are no *general* sufficient conditions that can be used to declare and defend a claim that *X* "causes" *Y*. The evidence used to support such claims varies substantially with the subject matter under investigation and the technology available for measurement. Statistical methodology *alone* is of limited value in the process of inferring causation. Furthermore, consensus on causality criteria evolves over time among practitioners in different scientific lines of inquiry (p. 13).

Table 21.1 Expected results of a polygraph test procedure with an accuracy index of 0.90 in a hypothetical population of 10,000 examinees that includes 10 spies ... [if detection threshold is set to detect the great majority (80 percent) of spies]. (Box 21.3 defines "Accuracy Index")

Examinee's true condition			
Test result	Spy	Non-spy	Total
"Fail" test	8	1598	1606
"Pass" test	2	8392	8394
Total	10	9990	10,000

The Polygraph and Lie Detection (2003)[2]

The committee that issued this report was tasked with examining the scientific literature of 50 studies in a controlled laboratory environment to arrive at a summary estimate of the false-negative and false-positive rates given the application of polygraph technology to the screening of employees. The existing literature was substantially larger than 50 studies, but the panel winnowed this down to the final 50 studies by requiring the studies used to meet the following standards before being included in this meta-analysis: (1) documentation of each study needed to be sufficient to allow one to reproduce the analysis; (2) each study needed to use an independent assessment of the truth of each subject; (3) each study needed to include both innocent and guilty individuals; (4) each study needed to support an accuracy analysis given that all cases were classified as either deceptive or nondeceptive; (5) measuring and scoring had to be carried out by individuals who did not know which subjects were guilty or innocent; and (6) the study had to make use of an appropriate method for assigning experimental subjects to experimental groups.

A key result of this study is presented in Table 21.1 (see reference in Footnote #2, Table 2-1, p. 48) of what might result from using a polygraph screening tool on 10,000 employees where 10 are known to be bad actors: This table was instrumental in changing the view of Los Alamos concerning the use of polygraph methods to screen employees for spying and sabotage.

Box 21.3 Accuracy Index

Note: The accuracy index is a joint measure (described in the report) that combines the false-negative rate of a diagnostic's ability to identify spies for the subpopulation of spies and the false-positive rate of a diagnostic's ability to identify non-spies for the non-spy subpopulation. One way of getting an accuracy index of 0.90 is with a false-negative rate of identifying spies for

(continued)

[2]National Research Council. 2003. *The Polygraph and Lie Detection.* Washington, D.C.: The National Academies Press. https://doi.org/10.17226/10420

the spy subpopulation of 0.20 and a false-positive rate of 0.16 of identifying non-spies for the non-spy population. These rates are close to the average performance of polygraph instruments empirically tested in the literature.

Assume that there are 10,000 individuals in an organization, including 9990 nonspies and 10 spies. Then if everyone is administered the test, the number of individuals that will be incorrectly identified as spies in the non-spy population will be $0.16 \times 9990 = 1598$, presumably all of whom will have to be investigated. Consequently, the number of individuals correctly identified as non-spies on the non-spy subpopulation will be 8392. Further, the number of actual spies missed will be $0.20 \times 10 = 2$, and consequently the number of spies correctly identified as such will be $0.80 \times 10 = 8$. So two spies will be missed out of 10. If one wishes to keep the accuracy index at 0.90, and one wants the number of false positives to be reduced from 1598, the number of missed spies will grow (possibly much) larger than 2 out of 10.

Measuring Racial Discrimination (2004)

This study was tasked to "consider the definition of racial discrimination, assess current methodologies for measuring it, identify new approaches, and make recommendations about the best broad methodological approaches" (p. 1). In its comprehensive review of relevant methodologies, the committee noted the following:

> Research design is critical to the ability to draw causal inferences from data analysis. For purposes of causal inference, there is a hierarchy of approaches to data collection. As one moves from meticulously designed and executed laboratory experiments through the variety of studies based on observational data, increasingly strong assumptions are needed to support the claim that X "causes" Y. The more careful and rigorous the design and control, the stronger are the inferences that can be drawn, provided that the design and control are used to address the causal question of interest.... Alternatively, the available data may provide information on differential outcomes (e.g., wage rates) for racial groups together with other variables that the researcher may use to infer the possible role of race-based discrimination. In such passive observation, the researcher lacks control over the assignment of treatments to subjects and attempts to compensate for this lack by "statistically controlling" for possible confounding variables ... In such circumstances, causal inferences can be controversial (p. 83).

The key methodological contribution of this report can be found in Chapter 7, "Statistical Analysis of Observational Data." This is an extremely careful discussion of how far one can use regression-type models on observational data to assess the degree of discrimination. The chapter contains the arguments that are used to infer from a regression approach whether a difference should be attributed to discrimination. In addition, the caveats that one must be aware of are tied into the

modeling in a comprehensive discussion of this issue. The degree of care as to how far one can take this argument can be seen from the following:

> More generally, we will often be hampered in our ability to infer discriminatory behavior on the basis of regression decompositions because we can never be sure we have included all of the relevant controls in the model. We must be able to control for the relevant variables well enough to approximate closely the hypothetical counterfactual in which only race has been changed . . . .The use of statistical models, such as multiple regressions, to draw valid inferences about discriminatory behavior requires appropriate data and methods, coupled with a sufficient understanding of the process being studied to justify the necessary assumptions (p. 158).

Protecting Individual Privacy in the Struggle Against Terrorists (2008)

This study was an attempt to identify data mining methods that could be used to help identify individuals that were planning to help in or to commit acts of terrorism, in a way that protected the privacy of innocent individuals. It was pointed out that some machine learning methods could be very helpful in some tasks, such as identifying luggage that was used to transport weapons, etc. A major contribution of this study was a framework that could be used to evaluate the effectiveness of a program for use in identifying terrorists. The use of specific approaches that had a strong causal interpretation was greatly preferred to other approaches that were based on weak associations, partly because of the resulting large number of false positives that can result from the latter approach.

Options for Estimating Illegal Entries at the US-Mexico Border (2013)

The statement of task for this study was to assess the flow of illegal migrants at various portions of the border as a function of economic pressure, enforcement effort, and geographic factors. Clearly linking these variables necessitated the use of a statistical model. The committee discussed different approaches to accomplish this, including network sampling, various probability models (especially the geometric distribution for successive attempts at illegal entering), regression models, spatiotemporal processes, and agent-based modeling. The view that the committee expressed for how much to trust statistical models in this context was as follows:

> Even though the available surveys do not directly address all questions of interest to DHS . . . if a statistical model agrees with the findings of the surveys on those aspects of flow that the surveys do capture, then one can reasonably expect that the model has predictive power for estimating other relevant aspects of flows. Similarly, if a model produces results that are not supported by previous data, then one of three conclusions is possible: the model does not fit the data well, the migration process has changed significantly over time, or both of these

> conditions apply. The model must be flexible, and one should expect that it will be necessary to extend it when new factors come into play, leading to a new round of model retrofitting and validation. Beyond timeliness and the possibility of greater accuracy, modeling has additional advantages. A good model allows policy makers to explore "what if" scenarios by changing model inputs. In particular, DHS can explore the impact of different allocations of enforcement resource among border stations or the impact of new enforcement policies. More importantly, the process of building a good model can create a stronger understanding of the social process underlying immigration behavior (p. 97).

The committee was not sanguine about the predictive power of the older models that the Department of Homeland Security and the United States Border Patrol had employed previously:

> . . . the situation at the border has changed markedly since 1997, and the panel has no confidence that these older models, which antedate the drug corridors, modern enforcement technology, and other innovations, can provide good guidance for the current era. Since the older models are unlikely to have the correct form, it would probably be necessary to rebuild them rather than just refit them with new data. While the policy environment can be updated in a rebuilt model, another shortcoming of much survey-based regression type modeling is the endogeneity of many of the migration determinants. In the presence of endogenous covariates and dual causality, the ability to simulate counterfactuals is compromised (p. 100).

These various studies demonstrate Steve's interest in statisticians playing a central role in the various policy debates of our time, employing a variety of statistical models in a conscientious manner to better understand how different factors are related and at times to make causal inferential statements and recommendations. As the passages quoted previously make clear, these studies to which Steve contributed advocated the use of statistical models to answer questions about the impact of an intervention or the causal nature of a change in some factor that was hypothesized to be affecting some response of interest. It may be of interest to readers to contrast this view with that of another leading statistical figure, David Freedman, who was well known for his cautions against the use of statistical models in these circumstances unless the models were strongly validated (see Box 21.4).

Box 21.4 Steve Fienberg vs. David Freedman: The Use of Statistical Models in Support of Policy Analysis

David Freedman argued against the use of multiple regression, path analysis, and related methods in an effort to infer causation from association. His position can be summarized as arguing that unless the stochastic assumptions one relied upon were fully validated, it was preferable for statisticians to refrain from playing any role in policy debates. In contrast, Steve Fienberg was adamant that statisticians needed to play an active role in any policy question informed by data even when the statistical tools were somewhat flawed.

It can be argued that the outstanding statistical public policy question of recent times was whether the decennial census should be modified or

(continued)

"adjusted" to take into consideration information on groups subject to differential undercoverage. In recent decennial censuses, a post-enumeration survey and demographic analysis provided information on those missed and those duplicated or otherwise counted in error in the census, and as a result there was pressure to make use of this information to modify the census counts. This topic was the focus of several National Academies studies. Neither David Freedman nor Steve Fienberg served as a regular member on any of these studies, but they both attended National Academies workshops on this topic and both provided Congressional testimony. Steve Fienberg was a strong advocate of adjustment of the census counts while David Freedman was strongly opposed. To better understand their different views on the application of statistical methods to public policy problems (and applications in general), the reader may find Chapters 13 and 14 in W.M. Mason and S.E. Fienberg (eds), *Cohort Analysis in Social Research: Beyond the Identification Problem* (1985, Springer Verlag, NY), of interest.

21.4 Contributions to the Work of the Report Review Committee

Of his many contributions to National Academies reports, Steve Fienberg's broadest impact was through his services on the Report Review Committee (RRC), which is responsible for overseeing the reviews of all reports produced by Academies committees. Steve served on the RRC for 11 years (July 2005 to June 2016); and during the last 8 years, he cochaired this prestigious committee, composed of approximately 30 members of the National Academy of Sciences (NAS), National Academy of Engineering (NAE), and Institute of Medicine (IOM, now known as the National Academy of Medicine). As Steve succinctly described the role of the RRC, "Our mission is to save the Academies some embarrassment."

Contributions as Report Review Monitor

During his time on the RRC, Steve monitored the reviews of 35 reports.[3] The responsibilities of the RRC monitor are similar to those of a journal editor:[4]

[3] In addition to serving as RRC monitor and cochair, Steve reviewed a total of 15 National Academies reports. We thank RRC staff member Dalia Hedges, who provided the appended lists of the reports that Steve monitored and reviewed.

[4] One important difference is that the review comments on National Academies reports are submitted to the institution, which is represented by the RRC monitor and a review coordinator appointed by the Academies division overseeing the study. They together assess the adequacy of

(1) making certain the panel of reviewers reflects an appropriately diverse range of expertise and perspectives, (2) providing a summary of the most important issues raised by reviewers, and (3) evaluating the committee's responses to review comments and revisions to the draft report. As monitor, Steve paid particular attention to the composition of the review panel and frequently added statisticians and other experts, as warranted. Steve was also known for his rigor in evaluating the changes made to the report in response to review comments—frequently calling for further revisions. As he pointed out, one important distinction between the review of journal articles and Academies reports is that reviewers of Academies reports are specifically asked whether the draft report addresses the statement of task that had been approved by the institution and whether the committee's findings and recommendations are supported by evidence and arguments. Steve never hesitated to challenge committees when he believed that they had gone beyond their task statements or that their recommendations lacked convincing evidence.

During his 11-year tenure on the RRC, Steve monitored the reviews of more than three reports a year.[5] Moreover, he even handled reports well outside the field of statistics on such diverse topics as global security,[6] the FBI's investigation of the 2001 Anthrax letters,[7] K-12 STEM education,[8] family caregiving,[9] and even a classified report on unmanned undersea vehicles.[10] Steve told an RRC staff member, "Some people like to be narrow! I try to pick assignments to learn something new." Even when he lacked subject matter expertise for a report, he made a notable contribution to the review process. As Steve explained to his RRC colleagues, every report of the National Academies should be comprehensible to broad audiences of scientists and engineers, and furthermore many reports outside the field of statistics deal with methodological issues of interest to statisticians. [Lincoln Moses, a Stanford University statistician and former RRC member, went so far as to suggest that every Academies report should be reviewed by at least one statistician!]

---

the authoring committee's responses to all review comments. The reviewers themselves never see these responses. Nor do they see the revisions to the report until it is published.

[5]The review of a National Academies report typically takes 10–12 weeks to complete (i.e., from the time the draft is sent to reviewer to the date of RRC signoff).

[6]National Academy of Sciences. 2009. *Global Security Engagement: A New Model for Cooperative Threat Reduction*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12583

[7]National Research Council. 2011. *Review of the Scientific Approaches Used During the FBI's Investigation of the 2001 Anthrax Letters*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13098

[8]National Research Council. 2013. *Monitoring Progress Toward Successful K-12 STEM Education: A Nation Advancing?*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13509

[9]National Academies of Sciences, Engineering, and Medicine. 2016. *Families Caring for an Aging America*. Washington, DC: The National Academies Press. https://doi.org/10.17226/23606

[10]National Academies of Sciences, Engineering, and Medicine. 2016. *Mainstreaming Unmanned Undersea Vehicles into Future U.S. Naval Operations: Abbreviated Version of a Restricted Report*. Washington, DC: The National Academies Press. https://doi.org/10.17226/21862

Contributions as RRC Cochair

The total number and diversity of reports that Steve monitored is impressive. However, even more impressive was his leadership as RRC cochair, a role he shared for 8 years with NAE member Chris Whipple. The RRC cochairs are responsible for selecting the RRC monitors to oversee the review process as well as for "signing off on" (i.e., approving) the release of Academies reports, based on the advice of the RRC monitor and the review "coordinator" appointed by the Academies division overseeing the study. Chris and Steve split this responsibility, each signing off on nearly 400 reports during their 8-year tenure. While many of these signoffs were straightforward, for every report Steve made it a point to understand the committee's major recommendations, the most controversial issues raised by reviewers, and how these issues were resolved. Not infrequently Steve helped to adjudicate disagreements arising either at the end of the review process or even earlier when the RRC monitor expressed serious concerns about review criticisms. At the same time Steve understood the limitations of his authority to withhold RRC signoff—commenting once that "If I had two noses, I would hold them both while signing off" and similarly "Much like my father used to say, this [report] is now 'less worse.' I am ready to sign off."

As RRC cochair, Steve also served an ex officio member of the NRC Governing Board Executive Committee, which is responsible for approving all proposals for new National Academies projects. In this role Steve read every project prospectus and frequently commented on issues that he expected might come up later in report review (e.g., an ambiguous statement of task or flaws in the proposed study methodology). It is impossible to measure, of course, how many problems in report review were avoided because of Steve's comments on project prospectuses, but his impact was well recognized by Academies staff members whose project prospectuses he critiqued. Steve also was a strong advocate for involving NAS members on Academies committees and especially as report reviewers. In this regard, he made an effort to include at least one NAS member on every review panel and he frequently suggested candidates—even for reports outside the field of statistics. During his tenure as RRC cochair, the participation of NAS members in report review increased appreciably.

With responsibility for signing off on nearly 400 reports in 8 years, it is not surprising that Steve had e-mail and phone interactions with the RRC staff almost daily. Although at times serving as the RRC cochair seemed like a full-time job, he performed this role without any financial compensation.

Steve's leadership was very apparent at the RRC annual meetings, at which the RRC members, the presidents of the three Academies, and senior staff considered the lessons learned from 20 or so case studies, which described the review process for reports handled by the RRC during the past year. Steve and Chris used this meeting as a training session for the RRC members, who typically monitor one or more report reviews a year. The discussions focused on the substantive issues arising in review and how they were resolved. Steve was candid in his assessment of the

issues raised in the case studies and did not pull any punches. The RRC meeting was also an opportunity to suggest changes in institutional review practices and policies, one of which pertaining to original data collection and analyses is described below.

Examples: Forensic Science, Assessments of Doctoral Research Programs

While the enormity and breadth of Steve's involvement in the Academies report review process is staggering, his substantive contributions may best be understood by describing his direct involvement in three reports, based largely on material from case studies that were prepared by the RRC staff for the annual meetings.

Given his particular interest in the forensic sciences, it is not surprising that Steve eagerly volunteered to monitor the review of a report on assessing eyewitness identification.[11] He made certain that 3 of the 14 reviewers had statistical expertise. While the 14 reviewers raised several substantive issues, Steve focused on one particularly contentious concern that was mentioned by the three statisticians: the committee's overly enthusiastic endorsement of the receiver operating characteristic (ROC) curve analysis.[12] In their summary of key review issues, Steve and the review coordinator, a pathologist, urged the authoring committee to temper its enthusiasm for ROC:

> The committee needs to follow the suggestion of Reviewer H to 'dial back' its support of ROC analysis as the methodology to advance the applied research literature, as review comments indicated that the complexities have not been adequately considered. The discussion of the ROC approach appears to be simplistic and fails to recognize the need to set tools in the context of modern statistical research methodology.[13]

While the committee modified its detailed description of the ROC analysis, Steve was not fully satisfied and called for further changes: "We are not proposing that the committee should say that the ROC work should be dismissed. The report can still say that it opens a promising avenue of research, while at the same time acknowledging that it is far from a final word or definitive methodology."[14] To help resolve this disagreement Steve and the coordinator participated in a late-night conference call with 11 of the 14 committee members, who subsequently agreed to further modifications of the ROC description. It was later learned that a statistician

[11] National Research Council, 2014: *Identifying the Culprit: Assessing Eyewitness Identification.* Washington, DC: The National Academies Press. https://www.nap.edu/catalog/18891/identifying-the-culprit-assessing-eyewitness-identification

[12] The use of receiver operating receiver curves to measure the accuracy of eyewitness identification decisions has been widely debated. See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5256436/.

[13] From a case study that was prepared by the RRC staff for discussion at the 2015 RRC annual meeting.

[14] Ibid.

on the committee fully agreed with Steve's reservations, but was outnumbered by several members who were ROC enthusiasts. It took Steve's perseverance to effectuate this change in the committee's report.

The second and third examples of Steve's acumen as an RRC monitor involved two Academies reports on the assessment of doctoral programs: the first of which proposed a methodology to be used in the assessment[15] and the second presented the assessment results.[16] To understand the motivation for these reports, some historical information is helpful. The Academies had issued two earlier doctoral program assessments in 1995[17] and 1982,[18] both of which provided extensive data on doctoral programs in a wide range of fields in the physical sciences, life sciences, behavioral and social sciences, engineering, and humanities. Although both reports included objective measures of program size, faculty publication and citation rates, research funding, student outcomes and support, average GRE scores, and other program characteristics, most of the public attention focused on the reputational ratings derived from opinion surveys of faculty members in each field. The goal of this third assessment was to include only objective measures and avoid reputational rankings.

As might be expected, Steve was particularly interested in monitoring the review of the 2009 report (*Guide*) since it proposed novel methodologies. He actively collaborated with the division review coordinator, Lyle Jones, a former RRC member and psychometrician from the University of North Carolina at Chapel Hill, who had cochaired the committee that had authored the 1982 assessment report. The proposed methodology aggregated 20 objective measures, using two very different weighting schemes. One set of weights was determined by the relative importance of each measure, based on the opinions of a small sample of faculty members in the field. The second set of weights was derived from a stepwise multiple regression analysis explaining reputational program ratings that were based on the opinions of another small sample of faculty members in each field. The results of these two weighting schemes were then to be combined to provide an interquartile ranking for each doctoral program. Although reputational ratings for a subset of programs in

[15] National Research Council. 2009. *A Guide to the Methodology of the National Research Council Assessment of Doctorate Programs*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12676

[16] National Research Council. 2011. *A Data-Based Assessment of Research-Doctorate Programs in the United States (with CD)*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12994

[17] National Research Council. 1995. *Research-Doctorate Programs in the United States: Continuity and Change.* Washington, DC: The National Academies Press.

[18] Jones, Lyle V., Gardner Lindzey, and Porter E. Coggeshall. 1982. *An Assessment of Research-Doctorate Programs in the United States (five volumes)*. Washington, DC: The National Academies Press. This 1982 report was undertaken under the auspices of the Conference Board of Associated Research Councils in the United States, which included representatives of the American Council of Learned Societies, the American Council on Education, the Social Science Research Council, and the National Research Council. The 1995 report was undertaken by the National Research Council alone.

each field were compiled to calculate the second set of weights, they were not to be reported.

Given the widespread interest in this doctoral program assessment, Steve and Lyle thought that it was imperative to enlist a highly diverse review panel. Twenty reviewers were selected, including university presidents and provosts, graduate deans and faculty, higher education analysts, statisticians and survey methodologists, and even a few recent doctorate recipients.[19] Steve and Lyle identified two overarching concerns raised by these reviewers: (1) confusing and incomplete descriptions of the compiled program data and the proposed analytical approach and (2) apparent flaws in the proposed statistical methodologies.[20] To address the first concern, extensive revisions were made to the text and a new chapter was added with examples illustrating the calculation of the program rankings in two selected fields. With regard to the second concern, a reviewer, who expressed concern about the problem of collinearity in applying the stepwise linear regression to derive weights for the 20 measures, suggested that a principal components technique be used to transform the variables prior to the regression analysis. With Steve's encouragement, the committee adopted this reviewer's suggestion.

While Steve and Lyle were generally pleased with the extensive revisions to the draft report, they decided to take the unusual step of sending the revised draft to eight of the original reviewers for their overall reactions since so much new material had been added to the report. In this second round of review, additional technical issues were raised, based on a clearer understanding of the proposed methodologies. Of particular concern to Steve was the complex technique for estimating the interquartile range of program rankings, which the committee preferred to provide instead of a central ranking with a confidence interval. The estimation of this interquartile range for each doctoral program required calculating 500 ratings based on randomly generated program data and weights and comparing each rating with similarly generated ratings for other doctoral programs in the field. The 175th and 325th highest rankings for each program determined the endpoints of its interquartile ratings. After several exchanges with the committee statistician who had formulated this Monte Carlo approach, Steve agreed that the method could be used as long as it was clearly described in the report.[21]

The review process for the 2009 *Guide* involved two rounds of comments and took nearly 11 months to complete. From Steve's perspective the most helpful improvements were in clarifying and expanding the descriptions of the data and methodologies—especially the addition of a 20-page technical appendix that included a one-page schematic of the entire ranking process.

[19]The names of the 20 reviewers may be found in the Preface and Acknowledgments section of the 2009 report.

[20]From a case study that was prepared by the RRC staff for discussion at the 2009 RRC annual meeting

[21]The complex technique for estimating the interquartile range of program rankings is described in Appendix A of the 2009 report.

By the time that the RRC had signed off on the *Guide* in June 2009, the project was already 18 months behind schedule, and there was considerable pressure from the anxious academic community to issue the doctoral program rankings as soon as possible. The draft *Assessment* report[22] was ready for review 7 months later and sent to 16 reviewers, 7 of whom had submitted thoughtful comments on the *Guide.* Along with this draft report, reviewers were provided the final version of the 2009 *Guide* and access to Excel spreadsheets containing the data tables on the 20 measures and interquartile ranking for each of the 5000 (unidentified[23]) doctoral programs. In addition to the standard set of review questions, Steve and Lyle asked reviewers to consider three questions:[24]

1. Are the data, methodology, and statistical analyses described in a comprehensible fashion?
2. Are the data tables presented in a clear and usable fashion, given the amount of data included?
3. How can the User's Guide [i.e., the 2009 report] be improved?

While the reviewers offered numerous suggestions for improving the presentation of all the assessment material, none appeared to have tried to replicate the rankings using data from the spreadsheets. Steve felt strongly that the program rankings should be validated by an independent group of "auditors." He personally recruited three statisticians, who were given the electronic database and agreed to try to replicate the program rankings in 6 weeks. One of the auditors, using the same statistical software package that the committee had applied, was able to reproduce the committee's program rankings. However, the other two auditors, using different statistical software, found some appreciable differences for numerous program rankings. More importantly, all three auditors discovered that the ratings derived from the regression-based weights were very poorly correlated with the ratings using weights constructed from faculty-assigned importance of the 20 measures. Steve and Lyle summarized the auditors' findings:

> All three auditors have reached the same conclusion: *the committee's combined rankings should not be published.* Given the outstanding credentials of these three auditors and the compelling evidence provided in their detailed reports, we are deeply concerned. As Auditor #2 observed, "If the committee dares publish the rankings as is, there will be aggrieved departments complaining loudly on sensible grounds that, in their own careful and faithful

[22]The 2011 *assessment* report presents a brief description of the doctoral program data collected and how the rankings were calculated, including a detailed example of this calculation for a program in economics.

[23]The assessment data covered doctoral programs in 61 fields at 222 institutions. The names of the institutions and departments were not identified in the spreadsheet to prevent any reviewers, monitor, coordinator, and staff from leaking the assessment results before the report was publicly released.

[24]From a case study that was prepared by the RRC staff for discussion at the 2011 RRC annual meeting.

> implementation of the NRC Committee's methodology, they calculate a better ranking than they have been given."[25]

One of the auditors suggested that two separate sets of program rankings might be issued, based on the alternative weighting schemes. Steve strongly endorsed this suggestion since the two sets could be presented as *illustrations* of how the data compiled by the committee might be used and would emphasize that no single set of weights works well for all doctoral programs. The committee, after consulting with the Academies leadership, adopted this suggestion and also decided—given the uncertainties in the data—to expand the ranking ranges to the fifth and 95th percentiles (instead of the first and third quartiles). At Steve's urging, the following paragraph was inserted on the first page of the report's Summary:

> This report also includes illustrations of how the dataset can be used to produce rankings of doctoral programs, based on the importance of individual measures to various users. Two of the approaches provided in the report are intended to be illustrative of constructing data-based ranges of rankings that reflect values to assess program quality determined by the faculty who teach in these programs. Other ranges of rankings can also be produced reflecting the values of the users. The production of rankings from measures of quantitative data turned out to be more complicated and to have greater uncertainty than originally thought. *As a consequence, the illustrative rankings are neither endorsed nor recommended by the National Research Council (NRC) as an authoritative conclusion about the relative quality of doctoral programs* [emphasis added].[26]

While Steve and Lyle were pleased with this explanation, they expressed serious concerns that the body of the report was inconsistent with this caution—implying instead that the alternative program rankings were definitive and, in some fields, might even be combined. They asked for further changes to the discussions in the body of the report to emphasize that the program rankings presented in the assessment were *illustrative*.

Frustrated by the long delays in the review process (nearly 6 months), the committee chair informed the National Academies leadership that he would resign if the report did not clear review in 5 days.[27] Concerned that the chair's resignation at this late stage in the study would undermine the credibility of the report, the NAS and NAE presidents informed Steve that they would assume his and Lyle's responsibilities for the completion of the review process. The presidents promptly signed off on the report after some modest changes were made, although Steve's primary concern remained that the rankings would be interpreted by most readers as definitive program rankings by the National Academies committee.

While disappointed with the outcome, Steve understood that the RRC cochairs serve as representatives of the NRC chair (NAS president) and that the final decision about whether a National Academies report can be released is ultimately up to the

[25] Ibid.

[26] National Research Council. 2011. Op. cit., p. 1

[27] From a case study that was prepared by the RRC staff for discussion at the 2011 RRC annual meeting.

NRC chair. In the past there have been a few reports for which the RRC cochairs consulted the NRC chair about specific review issues before signing off. However, this *Assessment* report is the only instance where the RRC cochairs (and division coordinator) were overruled by the NRC chair.

As it turned out, Steve's concern proved to be well founded. The 2011 *Assessment* report received widespread criticisms from many in the academic community who did not perceive the program rankings to be "illustrative" and questioned their validity. Among the most vociferous critics was Stephen Stigler,[28] one of the statisticians who had reviewed both the 2009 and 2011 reports.

Steve's unsuccessful efforts to salvage this report were not entirely in vain; in fact, his efforts had an important and long-lasting impact on the report review process. After consultation with the National Academies leadership, the RRC adopted formal "Review Guidelines for Reports Involving Original Data Collection, Analysis, and Modeling Activities." These guidelines, which have been extensively used by the RRC,[29] raise three general questions for committees as well as reviewers of their reports to consider:

1. Is the analytical approach fully documented, and are the data publicly available so that the results can be replicated?
2. How have the data and analyses been checked for errors?
3. Is the analytical methodology appropriate and representative of best practices?

The two-page guidelines, which contain detailed advice for each question, are typically shared with committees and their project staff *before* they undertake any data collection and/or analyses activities so that subsequent difficulties in report review can be alleviated. Steve was instrumental in calling attention to this important issue and in formulating these guidelines, which may be his most important legacy to the RRC and to the National Academies deliberative process.

National Academies Reports Monitored by Steve Fienberg

[in chronological order]

National Research Council. 2005. *Measuring Literacy: Performance Levels for Adults*. Washington, DC: The National Academies Press. https://doi.org/10.17226/11267

Institute of Medicine. 2006a. *Disposition of the Air Force Health Study: Interim Letter Report*. Washington, DC: The National Academies Press. https://doi.org/10.17226/11483.

Institute of Medicine. 2006b. *Disposition of the Air Force Health Study*. Washington, DC: The National Academies Press. https://doi.org/10.17226/11590.

[28] *The Chronical of Higher Education*. September 30, 2010. "A Critic Sees Deep Problems in the Doctoral Rankings"; https://www.chronicle.com/article/A-Critic-Sees-Deep-Problems-in/124725

[29] Much to the surprise of many RRC members, it was discovered that as many as 5 percent of all National Academies studies involved some form of original data collection, analyses, or modeling.

National Research Council. 2006. *Once, Only Once, and in the Right Place: Residence Rules in the Decennial Census*. Washington, DC: The National Academies Press. https://doi.org/10.17226/11727.

National Research Council. 2007. *Engaging Privacy and Information Technology in a Digital Age*. Washington, DC: The National Academies Press. https://doi.org/10.17226/11896.

National Research Council. 2009a. *Responding to Federal Register Call for Comments: Letter Report*. Washington, DC: The National Academies Press. [link not available]

National Research Council. 2009b. *Review of the Methodology Proposed by the Food Safety and Inspection Service for Followup Surveillance of In-Commerce Businesses: A Letter Report*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12634.

National Research Council. 2009c. *A Guide to the Methodology of the National Research Council Assessment of Doctorate Programs*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12676.

National Academy of Sciences. 2009. *Global Security Engagement: A New Model for Cooperative Threat Reduction*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12583.

Transportation Research Board. 2009. *Review of U.S. Department of Transportation's Methodology for Assessing De Minimis Impacts of Changes to 4(f) Process*. Letter report. [link not available]

National Research Council. 2010. *Gender Differences at Critical Transitions in the Careers of Science, Engineering, and Mathematics Faculty*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12062.

Institute of Medicine. 2010a. *Research Priorities for Assessing Health Effects from the Gulf of Mexico Oil Spill: A Letter Report*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13036.

Institute of Medicine. 2010b. *Review of the Proposal for the Gulf Long-Term Follow-Up Study: Highlights from the September 2010 Workshop: Workshop Report*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13025.

National Research Council. 2011a. *Learning Science Through Computer Games and Simulations*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13078.

National Research Council. 2011b. *Review of the Scientific Approaches Used During the FBI's Investigation of the 2001 Anthrax Letters*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13098.

National Research Council. 2011c. *A Data-Based Assessment of Research-Doctorate Programs in the United States (with CD)*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12994.

National Research Council. 2011d. *Successful K-12 STEM Education: Identifying Effective Approaches in Science, Technology, Engineering, and Mathematics*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13158.

National Research Council. 2011e. *Research-Doctorate Programs in the Biomedical Sciences: Selected Findings from the NRC Assessment*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13213.

National Research Council. 2011f. *Increasing National Resilience to Hazards and Disasters: The Perspective from the Gulf Coast of Louisiana and Mississippi: Summary of a Workshop*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13178.

Institute of Medicine. 2012a. *Monitoring HIV Care in the United States: A Strategy for Generating National Estimates of HIV Care and Coverage*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13408.

Institute of Medicine. 2012b. *Monitoring HIV Care in the United States: Indicators and Data Systems*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13225.

Institute of Medicine. 2012c. *Ranking Vaccines: A Prioritization Framework: Phase I: Demonstration of Concept and a Software Blueprint*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13382.

National Research Council. 2012. *Research Universities and the Future of America: Ten Breakthrough Actions Vital to Our Nation's Prosperity and Security: Summary*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13299.
Institute of Medicine. 2013. *Ranking Vaccines: A Prioritization Software Tool: Phase II: Prototype of a Decision-Support System*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13531.
National Research Council. 2013a. *Measuring What We Spend: Toward a New Consumer Expenditure Survey*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13520.
National Research Council. 2013b. *Best Practices in State and Regional Innovation Initiatives: Competing in the 21st Century*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18364.
Institute of Medicine and National Research Council. 2014a. *Research Priorities to Inform Public Health and Medical Practice for Ebola Virus Disease: Workshop in Brief*. Washington, DC: The National Academies Press. https://doi.org/10.17226/19004.
Institute of Medicine. 2014a. *Health Standards for Long Duration and Exploration Spaceflight: Ethics Principles, Responsibilities, and Decision Framework*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18576.
Institute of Medicine. 2014b. *Discussion Framework for Clinical Trial Data Sharing: Guiding Principles, Elements, and Activities*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18610.
National Research Council. 2014a. *Identifying the Culprit: Assessing Eyewitness Identification*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18891.
National Research Council. 2014b. *Science and Technology Capabilities of the Department of State: Letter Report*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18761.
Institute of Medicine and National Research Council. 2015. *Potential Risks and Benefits of Gain-of-Function Research: Summary of a Workshop*. Washington, DC: The National Academies Press. https://doi.org/10.17226/21666.
National Academies of Sciences, Engineering, and Medicine. 2016a. *Assessing Health Outcomes Among Veterans of Project SHAD (Shipboard Hazard and Defense)*. Washington, DC: The National Academies Press. https://doi.org/10.17226/21846.
National Academies of Sciences, Engineering, and Medicine. 2016b. *Mainstreaming Unmanned Undersea Vehicles into Future U.S. Naval Operations: Abbreviated Version of a Restricted Report*. Washington, DC: The National Academies Press. https://doi.org/10.17226/21862.
National Academies of Sciences, Engineering, and Medicine. 2016c. *Families Caring for an Aging America*. Washington, DC: The National Academies Press. https://doi.org/10.17226/23606.

National Academies Reports Reviewed by Steve Fienberg

[in chronological order]

National Research Council. 1994. *Letter to Congressman Sawyer on Confidentiality of Health Care Data Letter Report*. Washington, DC: The National Academies Press. [link not available].
National Research Council. 1997a. *Letter Report Addressed to Dr. Martha Farnsworth Riche, Director, U.S. Bureau of the Census*. Washington, DC: The National Academies Press. https://doi.org/10.17226/9253.
National Research Council. 1997b. *Preparing For the 2000 Census: Interim Report II*. Washington, DC: The National Academies Press. https://doi.org/10.17226/5886.

National Research Council. 1999. *Measuring a Changing Nation: Modern Methods for the 2000 Census*. Washington, DC: The National Academies Press. https://doi.org/10.17226/6500.
National Research Council. 2000a. *Letter Report to the Office for Civil Rights, Department of Education*. Washington, DC: The National Academies Press. https://doi.org/10.17226/9880.
National Research Council. 2000b. *Panel on the 2000 Census: A Letter Report*. Washington, DC: The National Academies Press. https://doi.org/10.17226/9996.
National Research Council. 2001a. *Letter Report from the Panel to Review the 2000 Census*. Washington, DC: The National Academies Press. https://doi.org/10.17226/10234.
National Research Council. 2001b. *The 2000 Census: Interim Assessment*. Washington, DC: The National Academies Press. https://doi.org/10.17226/10210.
National Research Council. 2003. *Planning the 2010 Census: Second Interim Report*. Washington, DC: The National Academies Press. https://doi.org/10.17226/10776.
National Research Council. 2008. *Ballistic Imaging*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12162.
National Research Council. 2009d. *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12589.
National Academies of Sciences, Engineering, and Medicine. 2013. *Long-Term Stewardship of Safety Data from the Second Strategic Highway Research Program (SHRP 2) Letter Report: October 14, 2013*. Washington, DC: The National Academies Press. https://doi.org/10.17226/22484.
Institute of Medicine and National Research Council. 2014b. *The National Children's Study 2014: An Assessment*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18826.
National Academies of Sciences, Engineering, and Medicine. 2015. *Support for Forensic Science Research: Improving the Scientific Role of the National Institute of Justice*. Washington, DC: The National Academies Press. https://doi.org/10.17226/21772.
National Academies of Sciences, Engineering, and Medicine. 2016d. *Modernizing Crime Statistics: Report 1: Defining and Classifying Crime*. Washington, DC: The National Academies Press. https://doi.org/10.17226/23492.

Chapter 22
New Data Sources and Inference Methods for Official Statistics

Jan van den Brakel

Preamble and Acknowledgment

I met Steve Fienberg when I worked as a junior statistician at the Statistics Netherlands' Methodology Department doing research into design and analysis of experiments embedded in probability samples inspired by the papers of Fienberg and Tanur (1987, 1988, 1989). Steve visited our Department for 6 months as part of his sabbatical leave from Carnegie Mellon University in 1997 and got involved in this research project as my PhD advisor. My PhD work attempted to further extend the papers by Fienberg and Tanur by combining statistical methods from the fields of randomized experiments and sampling theory to improve the internal and external validity of experimental research. This would have been a natural topic for this chapter, but an overview has been published recently in another book (Van den Brakel, 2019).

The rise of new data sources typically not collected for statistical purposes, sometimes called big data, is an important topic in many fields and particularly in official statistics. The use of big data and official statistics are both topics that also interested Steve. Therefore this chapter is devoted to the rise of new data sources in the production of official statistics.

I am very grateful and honored that Steve Fienberg was my PhD advisor. He was an amazing, inspiring supervisor. His unreserved support, particularly during periods when submissions of the first papers were razed to the ground by some referees and the need to continue this research was questioned within Statistics Netherlands' Methodology Department, was crucial for a successful completion of

J. van den Brakel (✉)
Research and Development Department, Statistics Netherlands, The Hague and Heerlen, The Netherlands

Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands
e-mail: jbrl@cbs.nl

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_22

my PhD thesis. Steve's involvement at this very moment was decisive for the further course of my scientific career at Statistics Netherlands. Steve: "Thank you so much for everything you did for me."

The views expressed in this chapter are those of the author and do not necessarily reflect the policies of Statistics Netherlands. I am also grateful to Prof. J. Tanur, Prof. W. Eddy and Prof. A. Carriquiry and an anonymous reviewer for careful reading of an earlier draft of this manuscript and providing constructive feedback.

22.1 Introduction

National statistical institutes or agencies are responsible for the production of reliable statistical information about economic and social developments in a society. This information is often referred to as official statistics. The required data are obtained via registers or collected through surveys, usually on the basis of a probability sample. For decades, design-based and model-assisted inference methods have been the preferred methods for national statistical institutes to produce official statistics. The prevailing opinion at national statistical institutes is that official statistics should not be based on explicit statistical model assumptions that are hard to verify. On the other hand, there is increasing pressure for national statistical institutes to reduce administrative costs and response burden. At the same time, users of official statistics more and more expect that national statistical institutes provide this information in a more timely fashion, at a higher frequency and at a more detailed level. In addition, declining response rates compromise the quality of sample estimates. These developments stimulate the search for alternative sources of statistical information. Such sources could include administrative data like tax registers, non-probability samples or other large data sets – so called big data – that are generated as a by-product of processes not directly related to statistical production purposes. Examples of big data include time and location of network activity available from mobile phone companies, social media messages from Twitter and Facebook, sensor data, and internet search behavior from Google Trends. A common problem with this type of data sources is that the process that generates the data is unknown and likely selective with respect to the target population.

The question arises to what extent can national statistical institutes in the future afford to exclusively use traditional probability samples in combination with design-based or model-assisted inference procedures for the production of official statistics. The major drawback of this approach is that small sample sizes rapidly result in unacceptably large design variances. This hampers the production of timely statistics at a detailed level using design-based inference methods. Model-based methods, for example, known from the small area estimation literature, can be used to make more precise and timely predictions for detailed subpopulations. New data sources can potentially be used as covariates in these models, since they come at a high frequency and are therefore very timely and also cost effective. The advantage

of using big data sources as covariates in models for sample surveys is that problems with selectivity can be circumvented as detailed below. If, however, big data sources are directly used to produce statistical information, then the potential selection bias of these data sources must be accounted for. In this case statistical modelling also plays a vital role.

The purpose of this chapter is to discuss the potentials and risks for national statistical institutes of using these new data sources in combination with model-based inference procedures for the production of official statistics. The chapter is organized as follows. In Sect. 22.2, the traditional approach of probability sampling in combination with design-based inference methods is reviewed. In Sect. 22.3, the advantages of model-based inference procedures are described. In Sect. 22.4, the pros and cons of new data sources or big data sources are described. In Sect. 22.5, the potentials of using these new data sources as covariates in model-based inference procedures are discussed. In Sect. 22.6, different methods that account for selection bias of non-probability samples are reviewed. Section 22.7 concludes with a discussion of the challenges and issues of these new data sources and inference methods for national statistical institutes.

22.2 The Role of Probability Sampling for Official Statistics

National statistical institutes gather and publish reliable statistical information about finite populations, generally all people residing in a country or all enterprises registered in a country. This information is often defined as totals, means, or proportions. Consider a finite population U of size N. Let y_i, $i = 1, \ldots, N$, denote the values of a variable of interest of population unit i. Population totals are typically defined as $Y = \sum_{i=1}^{N} y_i$. Means are simply obtained as $\overline{Y} = Y/N$. This information is not only required at the national level but also for all kind of subpopulations, like municipalities, age classes, gender classes, etc. The population U can be divided in D subpopulations or domains U_d of size N_d. In this case domain totals are defined as $Y_d = \sum_{i=1}^{N} \delta_{i,d}\, y_i$, with $\delta_{i,d}$ an indicator taking a value equal to one if element i belongs to domain d, and zero otherwise.

The population values for these variables are generally unknown. Until the beginning of the twentieth century, this kind of information was obtained by a complete census of the target population. This is very laborious and expensive. At the beginning of the twentieth century, it gradually became clear that large data sets are not a sufficient condition for valid inference. Despite an impressive 2.3 million respondents, the 1936 Literary Digest poll completely failed to correctly predict the outcome of the US presidential elections, because both the sample and the response were selective and not appropriately dealt with (Squire, 1988). This and other polling failures provided a strong incentive to embrace the concept of random sampling, which had been developed, mainly on the basis of the work of Bowley (1926) and Neyman (1934), as a method of obtaining valid estimators for finite population parameters based on a relative modest but representative sample,

rather than on a complete census. Other important milestone papers are Hansen and Hurwitz (1943), Narain (1951), and Horvitz and Thompson (1952). Under this approach the probability sample s of size n is drawn from the target population U with $n \ll N$. Each element i in the population has a nonzero probability, say π_i, to be included in the sample. An estimator of the unknown population total is obtained as the sum over the observations in the sample, expanded with the so called design weights, i.e., $\hat{Y} = \sum_{i=1}^{n} d_i y_i$, with $d_i = 1/\pi_i$. This estimation procedure is called design-based since inference and is completely based on the randomization distribution induced by the sampling design. Statistical modelling of the observations obtained in the survey does not play any role so far.

National statistical institutes often have auxiliary information about the target population from external sources, e.g., censuses and registers. This information can be used to improve the precision of the sample estimates. One way is to improve the efficiency of the sampling design, e.g., stratified sampling with optimal allocation and sampling designs where selection probabilities are approximately proportional to the target variable. Another way is to use this auxiliary information in the estimation procedure via the so-called generalized regression estimator proposed by Särndal et al. (1992). The generalized regression estimator expands each observation in the sample with a regression weight such that the sum over the weighted observations is an approximately design unbiased estimator of the unknown population total. Let $\boldsymbol{x}_i$ denote a vector containing q auxiliary variables for which the population totals $\boldsymbol{X} = \sum_{i=1}^{N} \boldsymbol{x}_i$ are known from a register or census. The design weights d_i are adjusted such that the sum over the weighted auxiliary variables in the sample equates to the known population totals, i.e., $\sum_{i=1}^{n} w_i \boldsymbol{x}_i = \boldsymbol{X}$, where w_i are the regression weights. This results in a correction for groups that are underrepresented in the sample, for example, due to selective nonresponse. The regression estimator for the population total is now obtained as $\hat{Y}^R = \sum_{i=1}^{n} w_i y_i$. Generally the purpose of a survey is not limited to estimates at the national level but also to produce statistical information for subpopulations or domains. Direct estimates for domain totals are obtained by $\hat{Y}_d^R = \sum_{i=1}^{n} w_i \delta_{i,d} y_i$.

In the model-assisted approach developed by Särndal et al. (1992), this estimator is derived from a linear regression model that specifies the relationship between the values of a certain target variable and a set of auxiliary variables for which the totals in the finite target population are known, i.e., $y_i = \boldsymbol{\beta}^t \boldsymbol{x}_i + e_i$. Most estimators known from sampling theory can be derived as a special case of the generalized regression estimator. Examples are the ratio estimator and post-stratification. Generalized regression estimators are members of a larger class of calibration estimators (Deville and Särnal 1992).

The generalized regression estimator has two very attractive properties. Although this estimator is derived from a linear model, it is still approximately design-unbiased. If the underlying linear model explains the variation of the target parameter in the population reasonably well, then the use of this auxiliary information will result in a reduction of the design variance compared to the Horvitz-Thompson estimator, and it might also decrease the bias due to selective nonresponse (Särndal et al. 1992; Särndal and Swensson 1987; Bethlehem 1988; Särndal and Lundström

2005). Model misspecification might result in an increase of the design variance but the property that this estimator is approximately design-unbiased remains. From this point of view, the generalized regression estimator is robust against model misspecification. The linear model is only used to derive an estimator that uses auxiliary information, but the resulting estimator is still judged by its design-based properties, such as design expectation and design variance. This is the reason that this approach is called model assisted.

Design-based and model-assisted inference is a very powerful concept since it is based on a sound mathematical theory that shows how, under the right combination of a random sample design and estimator, valid statistical inference can be made about large finite populations based on relatively small samples. In addition, the amount of uncertainty arising from relying on small samples can be quantified under both approaches through the variance of the estimators. A strong advantage of probability sampling in combination with a design-based or model-assisted inference is that it has a built-in robustness against model misspecification. This is useful in a production process where there is not much time for extensive model evaluation. For these reasons, design-based and model-assisted inference is still used in modern statistical science and is the standard for most national statistical institutes for producing official statistics.

22.3 Towards Model-Based Inference in Official Statistics

Model-based inference refers to estimation procedures that rely on the probability structure of an explicitly assumed statistical model, with the probability structure of the sampling design playing a less pronounced role. This is the position taken by authors like Gosh and Meeden (1997), Valliant et al. (2000), and Rao and Molina (2015).

Results published by national statistical institutes must enjoy public confidence. For decades, this has resulted in the prevailing opinion that methods used to produce official statistics, particularly if they are used for planning and implementing policies, must be free from model assumptions and should therefore be based on the above-mentioned design-based and model-assisted approaches. The reason for this is that models depend on assumptions that are hard to verify, which raises concerns about the validity of the results. Design-based and model-assisted approaches, however, have some limitations. In the case of small sample sizes, the design variance of the sample estimates become unacceptably large, which makes the built-in robustness against model misspecification of less use, and these approaches do not handle measurement errors effectively. In such situations model-based estimation procedures can be used as an alternative. The rapid rise of large data sets – so called big data – that are generated as a by-product of processes not directly related to statistical production purposes is another incentive for national statistical institutes to move towards model-based inference procedures as will be detailed in Sects. 22.4, 22.5, and 22.6

Important quality aspects of official statistics are accuracy, relevance, timeliness, and comparability with preceding periods. Relevance of statistical information increases with the level of detail and the frequency of the information. For policy making, monthly figures at a low regional level are in general more relevant than annual figures at the national level. Figures for reference period t are more relevant if they become available in $t + 1$, instead of with a delay of multiple time lags. Together these desiderata result in a detailed breakdown of a target population in domains or subpopulations with respect to regions or socio-demographic classifications in combination with short reference periods. In such situations domain sample sizes rapidly become too small to produce sufficiently precise domain estimates with design-based or model-assisted procedures. As an alternative, model-based procedures, which explicitly use a statistical model, can be used to improve the effective sample size of a particular domain using information from other domains or preceding sampling periods. These methods are in the literature referred to as small area estimation, for example, Rao and Molina (2015) and Pfeffermann (2002, 2013).

Small area estimation is predominantly based on multilevel models. These methods can be classified as area level models (Fay and Herriot, 1979) and unit level models (Battese et al., 1988). These models are predominantly used to take advantage of cross-sectional sample information that is observed in other domains. In an area level model, the direct estimates of the domains are modelled in a multilevel model, while in a unit level model the sampling units are the input for a multilevel model. Both types of models consist of a regression component, where available auxiliary information is used to explain the variation in the survey data, and a random component, which describes the unexplained variation between the domains. Through the regression component, sample information from other domains is used to improve the precision of the estimates for each domain separately. To define an area level model, a measurement error model is assumed for the observed domain estimates; $\hat{Y}_d^R = Y_d + e_d$, for domain $d = 1, \ldots, D$, with e_d the sampling errors which are assumed to be normally and independently distributed; $e_d \sim N(0, \psi_d)$. Subsequently a linear model for the true population parameter is assumed; $Y_d = \boldsymbol{x}_d^t \boldsymbol{\beta} + v_d$, with $\boldsymbol{x}_d$ a vector of auxiliary information at the domain level, $\boldsymbol{\beta}$ a vector with regression coefficients, and v_d the random domain effects that are assumed to be normally and independently distributed; $v_d \sim N\left(0, \sigma_v^2\right)$. Assuming that the design variances ψ are known, estimates for $\boldsymbol{\beta}$, v_d, and σ_v^2 can be obtained with maximum likelihood methods or Bayesian methods. Finally model-based predictions for Y_d including approximations for its uncertainty can be derived. See Rao and Molina (2015) for details. With a unit level model, a similar multilevel model is defined but now at the level of the observations of the sampling units. Below we further focus on the area level model, since most auxiliary information from new data sources are fuzzy and difficult to match at the unit level but are often available at the domain level.

Most surveys conducted by national statistical institutes are conducted repeatedly over time. A natural approach for small area prediction is to extend the Fay-Herriot

model with related information from previous editions of the survey. Rao and Yu (1994) extended the area level model by modelling random domain effects with an AR(1) model. Other accounts of regional small area estimation of unemployment, where strength is borrowed over both time and space, include Tiller (1992), Datta et al. (1999), You (2008), and Pfeffermann and Tiller (2006).

Temporal information can be included in the area level by assuming a structural time series (STS) model for the unknown domain parameters. Similar to the area level model, a time series model for survey estimates observed with a periodic survey starts with a measurement error model, $\hat{Y}^R_{t,d} = Y_{t,d} + e_{t,d}$, where subscript t refers to the time periods of the survey, $t = 1, \ldots, T$. Subsequently a structural time series model is assumed for the domain parameters. For simplicity we assume a basic structural time series model, which assumes that a series can be decomposed in a stochastic trend model, say, $L_{t,d}$, for modelling the low frequency variation, a stochastic seasonal component, say, $S_{t,d}$, to model a cyclic pattern with a period of 1 year, and a white noise component, say, $v_{t,d}$, for the remaining unexplained variation. This leads to $Y_{t,d} = L_{t,d} + S_{t,d} + v_{t,d}$. This model can be extended with other cycles, regression components and AR or MA components. See Durbin and Koopman (2012) for an introduction in STS modelling. For the components stochastic models are assumed, which makes them time dependent. A frequently applied trend model is the local linear trend model, which is defined as

$$L_{t,d} = L_{t-1,d} + R_{t-1,d} + \xi_{t,d}, \quad \xi_{t,d} \sim N\left(0, \sigma^2_\xi\right),$$

$$R_{t,d} = R_{t-1,d} + \eta_{t,d}, \quad \eta_{t,d} \sim N\left(0, \sigma^2_\eta\right).$$

For the seasonal component, the dummy or trigonometric seasonal component can be used; see Durbin and Koopman (2012) for an expression. The white noise terms are independently normally distributed; $v_{t,d} \sim N\left(0, \sigma^2_v\right)$. Inserting the STS model into the measurement error model gives $\hat{Y}^R_{t,d} = L_{t,d} + S_{t,d} + \varphi_{t,d}$, with $\varphi_{t,d} = v_{t,d} + e_{t,d}$ and assuming that $\varphi_{t,d} \sim N\left(0, \psi_d \sigma^2_\varphi\right)$ with ψ_d assumed to be known. See Van den Brakel and Krieg (2015) for details. STS models can be fitted using the Kalman filter after writing them in state-space form; see Durbin and Koopman (2012) for details.

The univariate STS model can be seen as a form of small area estimation, where sample information from preceding periods are used to improve the effective sample size for the last period. This model can be extended in several ways. A first generalization is to combine the time series of all D domains in one multivariate STS model. In this case the D domain estimates for one period are stacked in one vector $\hat{\boldsymbol{Y}}^R_t = \left(\hat{Y}^R_{t,1}, \ldots, \hat{Y}^R_{t,D}\right)^t$. Each series has its own trend and seasonal component. By modelling the correlations between the level disturbances of the domains $\xi_{t,d}$, cross-sectional information from other domains can be used. This assumes a $D \times D$ full covariance matrix for the vector $\boldsymbol{\xi}_t = (\xi_{t,1}, \ldots, \xi_{t,D})^t$. In a similar way the correlation between the slope disturbances $\eta_{t,d}$ can be modelled as well as the

disturbance terms of the seasonal components. This results in a multivariate STS model that uses temporal and cross-sectional information to improve the effective sample size for the different domains. This approach is followed by Pfeffermann and Burck (1990), Pfeffermann and Bleuer (1993), Van den Brakel and Krieg (2016), and Boonstra and Van den Brakel (2019).

Another useful application of STS models is to account for non-sampling errors. As long as the survey design of a repeated cross-sectional survey is not changed, non-sampling errors like measurement bias and selection bias remain rather invisible. In some situations the effects of non-sampling errors become visible. The first example is rotating panel designs, which are frequently used by national statistical institutes for labor force surveys. In a rotating panel on each survey occasion, a new panel is added to the sample and followed for a number of periods according to a predetermined pattern, after which the panel is (normally) dropped and replaced by a new one. Generally there are systematic differences between the responses across the waves, a phenomenon which is referred to in the literature as rotation group bias (RGB, Bailar 1975). Pfeffermann (1991) proposed a multivariate STS model where time series of direct estimates of the different waves of the rotating panel serve as the input and the RGB is explicitly modelled. This model can be used as a form of small area estimation and also to account for RGB induced by the rotating panel design. Other occasions where non-sampling errors become visible are major redesigns of the survey process for a repeated survey. The necessary update of data collection and field methods generally causes a change in the series. Such systematic differences are distinct from the sampling error and are known as discontinuities. One way to avoid confounding real period-to-period change with discontinuities is to model the effect of a redesign with an STS model. In this case the above proposed model is extended with an intervention variable which changes from zero to one at the moment of implementing the new survey design. The corresponding regression coefficient can be interpreted as the discontinuity, see, e.g., Van den Brakel and Roels (2010).

Finally the STS model can be augmented with related auxiliary series. This can be done by extending the univariate STS model with a regression component or by defining a bivariate STS model where the input vector contains the survey estimate and the auxiliary series, say, $\left(\hat{Y}^R_{t,d}, x_{t,d}\right)^t$. Both series have their own trend and seasonal components. The correlation between level disturbance terms of both series can be modelled in a similar way as explained for the multivariate STS model for all domain estimates. Also the correlation between the disturbance terms of other model components can be modelled. In this way the additional information from related auxiliary series is used to improve the survey estimates, see, e.g., Harvey and Chung (2000) and Van den Brakel and Krieg (2016).

Improving precision of direct estimates is an argument for national statistical institutes to move towards model-based estimation procedures in the production of official statistics. Statistics Netherlands has made some steps in this direction. Boonstra et al. (2008) summarize the first research result in small area estimation at Statistics Netherlands. Based on this work, since 2010 Statistics Netherlands has

used a multivariate STS model in the production of monthly Labor Force figures to handle problems with small sample sizes, rotation group bias and discontinuities (Van den Brakel and Krieg (2015)). A similar model was implemented in 2017 for producing official figures for the Consumer Confidence Index. A Battese-Harter-Fuller unit level model has been in use since 2015 to produce annual municipal unemployment figures (Boonstra et al. (2011)). A multilevel time series modelling approach, based on an extension of the model proposed by Bollineni-Balabay et al. (2016), is implemented in 2019 to estimate official trend figures in time series of the Dutch National Travel Survey (Boonstra et al. 2019).

22.4 New Data Sources

The accuracy of statistics is measured by their variance and bias. The variance is inversely related to the sample size and will generally be a major uncertainty component for survey sample statistics, because sample surveys usually have limited sample sizes. A strong point of sample surveys is that a national statistical agency has control over the quality of the survey outcomes through the design of the sample survey. The precision of the sample estimates can be controlled in advance via variance and sample size calculations and the choice of an optimal sampling strategy, i.e., the combination of a sample design and estimator. In addition the national statistical institute is in control of the availability of the data source as well as its frequency. Repeated sample surveys are therefore a stable data source for measuring the evolution of social-economic phenomena over time.

Concerning bias, we distinguish between selection bias and measurement bias. The selection bias of sample survey statistics is approximately zero under complete response. In practice, however, selection bias arises due to selective nonresponse, undercoverage of the sample frame, and the extent that field work strategy misses a part of the target population. Nonresponse can be informative and result in biased estimates if not appropriately accounted for (Pfeffermann and Sverchkov, 2003, 2009). The measurement bias in sample statistics typically depends on the extent to which the conceptual variables to be measured are correctly implemented in the questionnaire, on the mode of data collection and on the quality and skills of the interviewers in the case of telephone and face-to-face surveys. Problems with measurement bias in surveys arise because measurements of the variables of interest are indirect in that respondents are asked to report about their behavior, introducing all kind of measurement errors.

Drawbacks of sample surveys are that data collection is costly, its quality is compromised by nonresponse and measurement bias, and generally surveys are not very timely. In addition survey samples induce response burden, which is particularly an issue in business surveys. For national statistical institutes, this is an argument to make more use of administrative data like tax registers, or other large data sets – so called big data – that are generated as a by-product of processes not directly related to statistical production purposes. Examples of

these include time and location of network activity available from mobile phone companies, social media messages from Twitter and Facebook, internet search behavior from Google Trends, information found on the internet, web scraping, scanner data and sensor data such as satellite images, aerial images, and road sensor data. A common problem with this type of data sources is that the process that generates the data is unknown and likely selective with respect to the intended target population. A challenging problem in this context is to use these data for the production of official statistics that are representative of the target population. There is no randomized sampling design that facilitates the generalization of conclusions and results obtained with the available data to an intended larger target population. Hence, extracting statistically relevant information from these sources is a challenging task.

A strong point of administrative data sources and some big data sources is that they contain direct measurements of people's behavior and are therefore unaffected by measurement bias induced by questionnaires. Examples include smart meters to measure electricity consumption, GPS trackers in mobile phones to measure mobility and travel of populations, search and purchase behavior on the internet. If similar information was collected via questionnaires, substantial measurement bias might occur. Big data sources that contain direct measurement of people's behavior without measurement bias, however, can be implemented in a practical manner in a limited set of cases only.

A problem with registers and big data sources is that a national statistical institute has no control over the quality, availability, and stability of these data sources. Major changes in the behavior of the public on social media and the internet have a disturbing effect on the comparability of series over time. Also the use of these media might fluctuate over time. For example, a Google-trend series on searches related to vacancies might track an official series on unemployment. It does not measure unemployment, however. Search behavior before the start of the financial crisis in 2009 might be completely different compared to the period directly after the financial crisis, invalidating the measurement of the intended concept. Another example is the frequency with which administrative data become available. For short-term business statistics, published on a monthly basis, Statistics Netherlands changed from survey data to administrative data of value added tax in a period during which businesses where required by law to declare value added tax on a monthly frequency. Later on this legislation changed and businesses were allowed to choose whether they declared tax on a monthly, quarterly, or even annual frequency. As a result, Statistics Netherlands had to change their publications to a quarterly frequency.

Particularly in the case of big data with immense volumes, the variance will often be a minor uncertainty component. The bias, however, might be substantial. The size of the selection bias, in particular, depends on the extent to which the non-probability data source represents or covers the intended target population. Currently, research is going on into the use of data obtained from smart meters, GPS trackers and internet behavior for the production of official statistics, because

they measure individual behavior very precisely in a cost-effective way. Here the question is how to account for selection bias.

The rise of the big data era is somewhat reminiscent of the development of probability sampling in the early twentieth century responding to problems with the use of large non-probability samples like the 1936 *Literary Digest* poll. The volume of big data might lure some into the same trap of narrowing accuracy to precision, ignoring selection bias. This paradox has been mathematically formalized by Meng (2018) who derived an expression for the error of estimates derived from non-probability samples. The error contains three components: (1) a data quality measure or data defect index which measures the level of departure from simple random sampling, (2) a data quantity measure which measures the fraction of the target population covered by the big data sample, and (3) a problem difficulty measure, which is the standard deviation of the target variable. This measure shows that selection bias in non-probability samples become an issue if the data defect index (1) becomes substantial even if the sample size is voluminous.

The non-probability nature of the data therefore requires dedicated methods of inference to produce statistics about the intended, finite target population. Broadly spoken, there are two ways to use non-probability data sources in the production of official statistics. The first approach is to use them as covariates in model-based prediction methods for survey data. The second approach is to use them directly as a data source for official statistics and correct for possible selection bias.

22.5 Big Data as Auxiliary Variables

Problems with selection bias of non-probability data sources can be circumvented, at least partially, if they are used as covariates in prediction models for sample survey data. One potential application is to small area estimation models. Most big data sources are fuzzy and volatile, and the records typically do not coincide with the units of an intended target population or the sampling units of a probability sample. Therefore linking units in big data sources with sampling units in a probability sample will often be a heroic task. These complications can be avoided, at least partially, by using area level models instead of unit level models for small area estimation. The area level model was briefly introduced in Sect. 22.4. Covariates traditionally used in small area prediction models are available from registers and censuses. The value of new data sources is multiple. First of all in developing countries and combat areas, the availability of registers frequently updated censuses and survey data is generally scarce. Satellite images and mobile phone data can have valuable information for making detailed regional predictions. Also in developed countries, new timely data sources offer valuable additional information, e.g., once a census, which is typically conducted with a frequency of 10 years, becomes outdated. The high frequency by which new data sources become available allow for more frequent updates of official statistics (Hand, 2018; Powell et al., 2017).

Parallel to the development of the small area estimation literature, several authors have proposed methods for combining survey data with non-probability data sources

available from, e.g., sensor data and mobile phone data with the purpose of making detailed regional predictions for well-being and poverty. Many applications apply machine learning algorithms to establish the relation between survey data and sensor or mobile phone data and use the latter data set in a second step to make detailed regional predictions. Noor et al. (2008) analyzed the correlation between nighttime light intensity from satellite images and survey sample data on household income in Africa. They report a high correlation and used this empirical finding as a motivation to use nighttime light intensity as an alternative measure of poverty. Although one can question whether nighttime light intensity is an efficient construct to measure poverty, their empirical findings illustrate the potential of using remote sensor information as covariates in small area prediction models. Engstrom et al. (2017) used daytime satellite images to predict well-being. In a first step they applied deep learning to extract features from satellite images that are potentially related to well-being, like number of cars, building type, roof type, etc. In a next step they applied a Lasso to construct a linear model that relates the relevant images features to survey data on well-being. This relation is used to predict well-being in fine regional detail in Sri Lanka. Blumenstock et al. (2015) applied machine learning methods to combine mobile phone data with survey data on poverty and used this to predict poverty and well-being on a small regional level in Rwanda. Steele et al. (2017) combine survey data and mobile phone and satellite data in a generalized linear model to predict poverty in Bangladesh. This literature illustrates the potential value of these new forms of data for official statistics.

Some caution is required, however, for making fine regional predictions with the use of machine learning algorithms for overreliance on a model is required. One step in this direction is made by Marchetti et al. (2015) who used mobility patterns of cars tracked with GPS as a covariate in a Fay- Herriot model for predicting poverty for small regions in Italy. This class of small area estimation predictions is specified as a composite estimator of a model-based prediction and a design-based estimate where the weights are based on their measure of uncertainty and provide mean squared error approximations for the uncertainty of the small domain predictions. Similarly Schmid et al. (2017) use mobile phone data as a covariate in a Fay-Herriot model to predict literacy in Senegal.

In Sect. 22.3, it was emphasized that STS models are particular appropriate as a form of small area estimation, since official statistics are based on repeated surveys. Multivariate STS models are therefore appropriate to borrow strength over both time and space. Multivariate STS models can be used in a similar way to combine time series obtained with repeated sample surveys with auxiliary series derived from registers or big data sources.

Such combining serves two purposes. Extending the time series model with an auxiliary series allows modelling the correlation between the unobserved components of the structural time series models, e.g., trend and seasonal components. If the model detects a strong correlation, then the accuracy of domain predictions will be further increased. Harvey and Chung (2000) propose a time series model for the Labor Force Survey in the UK extended with a series of registered people receiving unemployment benefits. Information derived from nontraditional data sources like

Google Trends or social media platforms is generally available at a higher frequency than series obtained with repeated surveys. This allows the use of this time series modelling approach to make predictions for the survey outcomes in real time at the moment that the outcomes for the big data series are available, but not yet the survey data. In this case the auxiliary series are used as a form of nowcasting, which refers to methods that make early estimates during the reference period of survey when no sample data or only a part of the sample data are available. Van den Brakel et al. (2017) applied a bivariate STS model to estimate the Consumer Confidence Index, based on a monthly cross-sectional sample, in real time using an auxiliary series derived from messages left on social media platforms. Google Trends in particular has been used in the economic forecasting literature for this purpose; see, e.g., Vosen and Schmidt (2011) and the references therein.

To exploit the timeliness of the auxiliary series obtained with big data sources, the multivariate STS model can be expressed at the high frequency of the auxiliary series. This requires a disaggregation of the unobserved time series components of the target series observed with a repeated survey at a low frequency to this higher frequency. After fitting the model, estimates for the survey parameters are obtained by aggregating the underlying components to a monthly frequency. Details of mixed frequency state-space models are described in Harvey (1989), Ch. 6.3, Durbin and Quenneville (1997), and Moauro and Savio (2005).

With data sources like Google Trends, a large number of potential auxiliary series are easily obtained. Combining them in a full multivariate STS model, as outlined before, limits the degrees of freedom for model fitting. Due to the so-called curse of dimensionality, prediction power of such models will be low. From this perspective, factor models are developed to formulate parsimonious models, despite the large number of auxiliary series considered. Factor models are developed and widely applied by central banks to nowcast GDP on quarterly frequency using a large number of related series observed on a monthly frequency (Boivin and Ng 2005; Stock and Watson 2002a, 2002b; Marcellino et al. 2003). More recently, Giannone et al. (2008) and Doz et al. (2011) proposed a state-space dynamic factor model. They propose a two-step estimator. In a first step a small number of common factors are extracted from a large set of series using principal component analysis. In a second step, the common factors are combined with the target series in a state-space model and are fitted using the Kalman filter. This approach is applied by Schiavoni et al. (2019) to estimate monthly unemployment figures in real time with a series of people receiving unemployment benefits and Google Trends series.

22.6 Big Data as Direct Data Sources for Official Statistics

If non-probability data sources are considered as a primary data source for compiling official statistics, then the question arises to which extent can results obtained with a non-probability data source be generalized to an intended, larger target population. Contrary to probability samples, the data generating process of

these data sources is generally unknown. As a result, statistical information derived from non-probability samples can suffer from a large selection bias if they are used for these purposes.

Different methods are proposed in the literature to account for selection bias in non-probability samples. Some authors apply standard weighting and calibration methods known from classical probability sampling to non-probability samples, which are referred to as pseudo-design-based inference methods (Baker et al., 2013). Several authors apply propensity scoring, proposed by Rosenbaum and Rubin (1983), to construct weights that correct for selection bias. Estimating response probabilities and using them in Horvitz-Thompson-type estimators to account for unequal selection probabilities is sometimes called pseudo-randomization. Valliant and Dever (2011) propose different models to estimate response probabilities in opt-in Web panels and discuss under which conditions they correct for selection bias. Deville (1991) proposed models for quota samples, which can be used to construct post-stratification estimators or linear weighting-type estimators (2008). There are many references in the literature where propensity scores are used to correct for selection bias in non-probability samples (see, e.g. Lee, 2006; Lee and Valliant, 2009; Schonlau et al., 2007, 2009). Buelens et al. (2018) compared pseudo-design-based, model-based, and algorithmic methods and conclude that auxiliary information typically available for weighting and calibration (e.g., demographic variables like age class, gender, regional classifications) do not sufficiently explain the data generating process of a non-probability sample to correct successfully for selection bias.

Another class of methods to correct for selection bias is to apply a statistical model to predict the units not in the sample (Royall, 1970; Valliant et al., 2000). This approach is based on the specification of an appropriate super-population model that captures the variation of the target variables instead of adjusting selection probabilities.

Some methods combine a non-probability sample that contains the target variable of interest and auxiliary variables with a reference sample that is based on a probability sample and only contains auxiliary variables. The reference sample is used to assess the selectivity of the non-probability sample. One approach, quasi randomization, is to construct propensity models to estimate selection probabilities for the non-probability sample (Elliot and Vailliant, 2017; Isaksson and Forsman, 2003; Valliant et al., 2013). Sample matching is also applied as an attempt to reduce selection bias in opt-in Web panels using covariates obtained in a small reference sample to construct propensity weights without collecting observations for the target variables (Rivers and Bailey, 2009; Vavreck and Rivers, 2008; Terhanian and Bremer, 2012). These ideas are related to approaches that are also used in microsimulation to match probability samples with population or census data (Tanton and Edwards, 2013). Kim and Wang (2018) proposed inverse sampling. In a first step, important weights are derived for the units in the non-probability sample, using the auxiliary variables in the reference sample and the non-probability sample. In a second step, a sample using unequal probability sampling proportional to the important weights is drawn from the non-probability sample, such that it can be

interpreted as a simple random sample from the target population. As an alternative, Kim and Wang (Kim and Wang, 2018) proposed data integration which implies that a parametric model is assumed to construct weights for the units in the non-probability sample, which are subsequently used in standard weighting methods. Rivers (2007) proposed imputation of the target variables observed in the non-probability sample in the reference sample using nearest neighbor imputation and subsequently apply standard weighting methods.

A consequence of combining a large non-probability sample with a high-quality smaller reference sample is that the precision of the large non-probability sample reduces to the standard error of the smaller reference sample. These methods nevertheless might improve the accuracy, in terms of mean squared error, of estimates derived from non-probability samples. The methods summarized above are based on strong ignorability assumptions and can lead to serious bias if these assumptions are not met.

In the case that the non-probability sample and the probability-based reference sample, both contain the target variable and some auxiliary variables. Kim and Tam (2018) propose a design-based inference method that can be regarded as a post-stratification estimator where one stratum is the subpopulation that is completely observed with the non-probability sample. Model-based approaches for informative sampling (Pfeffermann and Sverchkov, 2003, 2009), where the selection probabilities are related to the target variables, might potentially be fruitful to correct for selection bias in non-probability samples for situations where no reference sample is available.

Selection bias in non-probability samples is also a topic of concern in economic analysis. There is an extensive amount of econometric literature on correcting for selection bias in behavioral studies. Heckman (1976, 1979) has two early landmark papers on this topic. A more recent overview is provided by Heckman (2010). In this approach there is a random sample available for which two linear models are distinguished. The dependent variables in the first linear model suffer from selection bias due to self-selection, while the dependent variables in the second linear model do not suffer from selection bias and therefore can be considered as a random sample. Sample selection in the variables of the first linear model is considered as a specification error and correction for selection bias is achieved by considering the response on the variables in the first model as a censored or truncated sample using the second linear model.

This literature on selection bias in economic behavioral studies provides important methods that potentially correct for selection bias in non-probability samples in the context of official statistics and are related to the aforementioned methods that use a probability sample as a reference sample to assess selection bias in big data sources. In many situations, a non-probability sample might not contain variables for which it can be assumed that they are not affected by selection bias. This problem might be circumvented by matching a non-probability sample with a data source that is not affected by selection bias, e.g., a register. Further research into how to adapt these methods to correct for selection bias in new non-probability data sources in the context of official statistics is definitely worthwhile.

22.7 Discussion

National statistical institutes face multiple challenges. There is increasing pressure to reduce administration costs and response burden. Nonresponse is a gradually increasing problem which compromises the quality of traditional sample surveys. In order to remain relevant for data users, the level of detail, frequency, and timeliness of statistical information must increase. This raises the question whether national statistical institutes can continue to base official statistics solely on probability samples in combination with design-based or model-assisted inference methods. The advantage of this approach is its low-risk level. With sample surveys, a national statistical institute has full control over the availability of the data, as well as the quality and frequency of the statistical output. Model-assisted inference methods have a built-in robustness against model misspecification, which make these methods attractive for multipurpose surveys in the production of official statistics where there is usually very limited time for model building and evaluation. Repeated sample surveys therefore provide a safe method to produce consistent time series that measure period-to-period change in a reliable way.

In order to improve the level of detail, frequency, and timeliness of statistical information, without increasing sample sizes and thus data collection costs, model-based inference procedures known from the literature of small area estimation, time series analysis, and nowcasting can be considered. This, however, increase the risk level for a national statistical institute, since model misspecification can result in biased statistical information. The output, however, is primarily based on sample survey data, collected by the national statistical institute. This implies that the risks concerning availability, frequency, and quality of the data are still managed by the national statistical institute. In this context new data sources can provide useful additional information as covariates in small area prediction models, particular for countries without registers or timely census data. Many big data sources are available at a high frequency which makes them potentially useful to make more precise predictions of sample statistics in real time with nowcasting models.

Replacing sample surveys by registers or other type of non-probability data sources implies a substantially increased risk level, since in this situation a national statistical institute has no control over the availability, comparability, and quality of the data source over time. Another issue with using big data as a primary data source to compile statistical information is that of selectivity. Big data are used successfully in many different disciplines. The use of these data sources in the context of official statistics is, however, different. The problem, which is unique, e.g., official statistics, marketing research, and election polls, is the question to which extent statistical results can be generalized to larger intended target populations (Pfeffermann, 2019; Pfeffermann et al., 2015).

As highlighted in Sect. 22.6, there is a substantial amount of literature for correcting for selection bias in non-probability samples. There are nevertheless a lot of issues with the application of these methods in the daily practice of official statistics. One issue is that all methods are based on strong ignorability assumptions

conditionally on the available covariates, which are difficult to verify. A more practical issue is that all methods assume that the records in a big data source contain, besides the target variable, a set of auxiliary variables which correspond with the units in a target population or a reference sample. Unfortunately, these conditions are seldom met. Most big data sets are fuzzy, records do not correspond with units in the target population or a reference sample, and auxiliary information is generally not available since owners of the big data source are reluctant to provide them due to privacy issues. Mobile phone data, for example, are mostly a file of call detail records that contain time and location information generated by devices. Mobile phone companies generally do not provide the demographic information of the owners of the devices. As a result, methods summarized in Sect. 22.6 to correct for selection bias cannot be applied in a straightforward manner in these situations. Attempts to use these data to produce, for example, daytime population statistics are based on machine learning methods which attempt to derive demographic information from the observed mobility patterns of the devices, followed by rather naïve post-stratification corrections. It is not likely that this sufficiently corrects for selectivity.

At this moment it is not at all clear how big data can be used in the production of official statistics (Pfeffermann, 2019; Pfeffermann et al., 2015). National statistical institutes, nevertheless, have to investigate to what extent these new data sources, in combination with new inference methods, can be used to improve the level of detail, frequency, and timeliness of their publications, on the one hand, and to reduce data collection costs, on the other hand. The literature that uses satellite images and mobile phone data to make small area predictions for poverty and well-being on a fine regional level clearly illustrate the potentials of big data sources. Using these new data sources in the production of official statistics requires more research and insight into the quality of these data sources and an extension of the methodological tools to extract the right information from these new data sources. This is not only an extension from design-based to model-based inference but also to machine learning methods and artificial intelligence algorithms to extract information from satellite and aerial images or sensor data. An additional advantage of all these developments is that they make the life of an official statistician more exciting.

References

Bailar, B.A. (1975). The Effects of Rotation Group Bias on Estimates from Panel Surveys. *Journal of the American Statistical Association*, 70, pp. 23–30.

Baker, R., J.M. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile, and R. Tourangeau (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, pp. 90–143.

Battese, G.E., R.M. Harter and W.A. Fuller (1988). An error components model for prediction of county crop areas using satellite data. *Journal of the American Statistical Association*, 83, pp. 28–36.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, pp. 251–260.

Blumenstock, J., G. Cadamuro and R. On (2015). Predicting poverty and wealth from mobile phone metadata. *Science,* 350, pp. 1073–1076.

Boivin, J. and S. Ng (2005). Understanding and comparing factor-based forecasts. *International Journal of Central Banking*. 3, pp. 117–151.

Bollineni-Balabay, O. Brakel, J.A. van den and Palm, F. (2016). Multivariate state-space approach to variance reduction in series with level and variance breaks due to sampling redesigns. *Journal of the Royal Statistical Society, A Series*, vol 179, pp. 377–402.

Boonstra, H.J. and J.A. van den Brakel (2019). Estimation of level and change for unemployment using structural time series models. *Survey Methodology*, 45(3): 395–425.

Boonstra, H.J., van den Brakel, J.A., Das, S.: Multilevel time series modelling of mobility trends. Discussion paper October, 2019, Statistics Netherlands, Heerlen (2019).

Boonstra, H.J., Buelens, B., Leufkens, K., Smeets, M.: Small area estimates of labour status in Dutch municipalities. Technical Report 201102, https://www.cbs.nl/nl-nl/achtergrond/2011/02/small-area-estimates-of-labour-status-in-dutch-municipalities, Statistics Netherlands (2011).

Boonstra, H.J., J.A. van den Brakel, B. Buelens, S. Krieg and M. Smeets (2008). Towards small area estimation at Statistics Netherlands. *Metron International Journal of Statistics*, LXVI, pp. 21–50.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin de l' Institute International de Statistique* 22(Supplement to Book 1): 6–62.

Buelens, B., J. Burger and J.A. van den Brakel (2018). Comparing inference methods for non-probability samples. *International Statistical Review*, 86, pp. 322–343.

Datta, G., P. Lahiri, T. Maiti, and K. Lu (1999). Hierarchical Bayes estimation of unemployment rates for states of the US. *Journal of the American Statistical Association*, 94, pp. 1074–1082.

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

Doz, C., D. Giannone, and L. Reichlin (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, 164, pp. 188–205.

Dever, J. A., A. Rafferty, and R. Valliant (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods*, 2, pp. 47–60.

Deville, J.-C. (1991). A theory of quota surveys. *Survey Methodology*, 17, pp. 163–181.

Durbin, J. and S.J. Koopman (2012). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Durbin, J. and B. Quenneville (1997). Benchmarking by state space models. *International Statistical Review*, 65, pp. 23–48.

Elliot, M. R. and R. Vailliant (2017). Inference for non-probability samples. *Statistical Science*, 32, pp. 249–264.

Engstrom, R., Hersh, J., Newhouse, D.: Poverty from Space: Using high resolution satellite imagery for estimating economic well-being. Technical report (2017).

Fay, R.E. and R.A. Herriot (1979). Estimation of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Society*, 74, pp. 268–277.

Fienberg, S.E., and Tanur, J.M. (1987). Experimental and sampling structures: parallels diverging and meeting. *International Statistical Review*, 55, pp. 75–96.

Fienberg, S.E., and Tanur, J.M. (1988). From the inside out and the outside in: combining experimental and sampling structures. *Canadian Journal of Statistics*, 16, pp. 135–151.

Fienberg S.E., and Tanur, J.M. (1989). Combining cognitive and statistical approaches to survey design. *Science*, 243, pp. 1017–1022.

Giannone, D.L., L. Reichlin and D. Small (2008). Nowcasting: The real-time information content of macroeconomic data. *Journal of Monetary Economics*, 55, pp. 665–676.

Gosh, M., and G. Meeden (1997). *Bayesian Methods for Finite Population Sampling*. London: Chapman & Hall.

Hand, D.J. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society, A Series*, Vol. 181, pp. 555–605.

Hansen, M.H. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*. 14, pp. 333–362.

Harvey, A.C. (1989). Forecasting, structural time series models and the Kalman filter. Cambridge University Press, Cambridge.

Harvey, A.C. and C. Chung (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, A Series*, 163, pp. 303–339.

Heckman, J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *The Annals of Economic and Social Measurement*, 5, 475–492.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*. 47, pp. 153–161

Heckman, J. (2010). Selection bias and self-selection. In: *Microeconometrics*, 2. Ed. S. Durlauf and L.E. Blume, pp. 242–266, Palgrave Macmillan, London.

Horvitz, D.G., and D.J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, pp. 663–685.

Isaksson, A. & Forsman, G. (2003). A comparison between using the web and using the telephone to survey political opinions. In *Annual Meeting of the American Association for Public Opinion Research*, Nashville, TN, pp. 100–106.

Kim, K. and Z. Wang (2018). Sampling techniques for big data analysis in finite population inference. *International Statistical Review*, 87, pp. 177–191.

Kim, K., Tam, S.M.: Data integration by combining big data and survey sample data for finite population inference. Working paper (2018).

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22, pp. 329–349.

Lee, S. and R. Valliant (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods and Research*, 37, pp. 319–343.

Marcellino, M., J. Stock and M. Watson (2003). Macroeconomic forecasting in the euro area; country specific versus area wide information. *European Economic Review*, 47, pp. 1–18.

Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Perdreschi, S. Rinzivillo, L. Pappalardo and L. Gabrielli (2015). Small area model-based estimators using Big data sources. *Journal of Official Statistics*. 31, pp. 263–281.

Meng, X.L. (2018). Statistical paradises and paradoxes in big data. *The Annals of Applied Statistics*, 12, pp. 685–726.

Moauro, F. and G. Savio (2005).Temporal disaggregation using multivariate structural time series models. *Econometrics Journal*, 8, pp. 214–234.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, pp. 169–174.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, pp. 558–625.

Noor, A., V. Angela, P. Gething, A. Tatem, and R. Snow (2008). Using remotely sensed night-time light as a proxy for poverty in Africa. Population and Health Metrics, 6:5, doi https://doi.org/10.1186/1478-7954-6-5.

Pfeffermann, D.A.: Challenges in the production of official statistics with different methods of data collection. Paper presented at the Annual Workshop on Survey Methodology, Brazilian Network Information Centre (NIC.br). Sao Paulo, 20 May, 2019 (2019).

Pfeffermann, D.A. (2013), New Important Developments in Small Area Estimation. *Statistical Science*, 28, pp. 40–68.

Pfeffermann, D.A. (2002), Small Area Estimation – New Developments and Directions. *International Statistical Review*, 70, pp. 125–143.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economic Statistics*, 9, pp. 163–175.

Pfeffermann, D. and S.R. Bleuer (1993). Robust Joint Modelling of Labour Force Series of Small Areas. *Survey Methodology*, 19, pp. 149–163.

Pfeffermann, D. and L. Burck (1990). Robust Small Area Estimation Combining Time Series and Cross-Sectional Data. *Survey Methodology*, 16, pp. 217–237.

Pfeffermann, D., Eltinge, J. L. & Brown, L. D. (2015). Methodological issues and challenges in the production of official statistics. *Journal of Survey Statistics and Methodology*, 3, pp. 425–483.

Pfeffermann, D. and M.Y. Sverchkov (2003). Fitting generalized linear models under informative sampling. In *Analysis of Survey Data*, Eds. Chambers, R. L. & Skinner, C. J., pp. 175–195. Chichester: Wiley.

Pfeffermann, D. and M.Y. Sverchkov (2009). Inference under informative sampling. In *Handbook of Statistics*, Vol. 29, Ed. Rao, C., pp. 455–487. Amsterdam: Elsevier.

Pfeffermann, D. and R. Tiller (2006). Small Area Estimation with State Space Models Subject to Benchmark Constraints. *Journal of the American Statistical Association*, 101, pp. 1387–1397.

Powell, B., G. Nason, D. Elliot, M. Mayhew, J.J. Davies and J. Winton (2017). *Journal of the Royal Statistical Society A series*, 181, pp. 737–756.

Rao, J.N.K. and I. Molina (2015). *Small Area Estimation*, 2nd edition. New York: Wiley.

Rao, J.N.K. and M. Yu (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics*, 22, pp. 511–528.

Rivers, D.: Sampling for web surveys. *In 2007 JSM Proceedings, ASA Section on Survey Research Methods*, American Statistical Association (2007).

Rivers, D. and D. Bailey (2009). Inference from matched samples in the 2008 US national elections. In *Proceedings of the Joint Statistical Meetings*, Washington, DC, pp. 627–639.

Rosenbaum, P.R. and D.B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, pp. 41–55.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, pp. 377–387.

Schiavoni, C., Palm, F., Smeekes, S., van den Brakel, J.A.: A dynamic factor model approach to incorporate Big Data in state space models for official statistics. Discussion paper January, 2019, Statistics Netherlands, Heerlen (2019).

Schonlau, M., A. van Soest, and A. Kapteyn (2007). Are 'Webographic' or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? *Survey Research Methods*, 1, pp. 155–163.

Schonlau, M., A. van Soest, A. Kapteyn, and M. Couper (2009). Selection bias in web surveys and the use of propensity scores. *Sociological Methods and Research*, 37, pp. 291–318.

Schmid, T., F. Bruckschen, N. Salvati and T. Zbiranski (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. *Journal of the Royal Statistical Society, Series A*, 178, pp. 239–257.

Särndal, C.-E., and S. Lundström (2005). Estimation in Surveys with Nonresponse. New-York: Wiley.

Särndal, C.E., and B. Swensson (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *International Statistical Review*, 55, pp. 279–294.

Särndal, C.E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Squire, P. (1988). Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly*, 52, pp. 125–133.

Steele, J., P.R. Sundsøy, C. Pezzulo, V.A. Alegana, T.J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y.A. de Montjoye, A.M. Iqbal, K.N. Haddiuzzaman, X. Lu, E. Wetter, A.J. Tatum and L. Bengtsson (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Statistical Society Interface,* 14, 127.

Stock, J. and M. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Society,* 97, pp. 1167–1179.

Stock, J. and M. Watson (2002b). Macroeconomic forecasting using diffuse indexes. *Journal of Business and Economic Statistics*, 20, pp. 147–162.

Tanton, R. and K. Edwards (eds). (2013). *Spatial Microsimulation: A Reference Guide for Users*, Dordrecht: Springer.

Terhanian, G. and J. Bremer (2012). A smarter way to select respondents for surveys. *International Journal of Marketing Research*, 54, pp. 751–780.

Tiller, R.B. (1992). Time series modelling of sample survey data from the U.S. current population survey, *Journal of Official Statistics*, 8, pp. 149–166.

Valliant, R. and J.A. Dever (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods and Research*, 40, pp. 105–137.

Valliant, R., J.A. Dever, and F. Kreuter (2013). *Practical tools for designing and weighting survey samples*. New York: Springer Verlag.

Valliant, R., A.H. Dorfman, and R.M. Royall (2000). *Finite Population Sampling and Inference, A Prediction Approach.* New York: Wiley.

Van den Brakel, J.A. (2019). Design-based analysis of experiments embedded in probability samples. In: *Experimental methods in Survey Research: Techniques that combine random sampling with random assignment*, Eds. P.J. Lavrakas, M.W. Traugott, C. Kennedy, A.L. Holbrook, E.D. de Leeuw, and B.T. West, Wiley, New-York. Ch. 16, pp. 399–423.

Van den Brakel, J.A. and S. Krieg (2016). Small area estimation with state-space common factor models for rotating panels. *Journal of the Royal Statistical Society A Series*, 179, pp. 763–791

Van den Brakel, J.A. and S. Krieg, (2015). Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. *Survey Methodology*, 41, pp. 267–296.

Van den Brakel, J.A., E. Söhler, P. Daas and B. Buelens, (2017). Social media as a data source for official statistics; the Dutch Consumer Confidence Index. *Survey Methodology*, 43, pp. 183–210.

Van den Brakel, J.A. and J. Roels, (2010). Intervention analysis with state-space models to estimate discontinuities due to a survey redesign. *Annals of Applied Statistics*, 4, pp. 1105–1138.

Vavreck, L. and D. Rivers (2008). The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties*, 18, pp. 355–366.

Vosen, M. and T. Schmidt (2011). Forecasting private consumption: Survey-based indicators versus Google trends. *Journal of Forecasting*, 30, pp. 565–578.

You, Y. (2008). An integrated modelling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, 34, pp. 19–27.

Chapter 23
Understanding the United States' Official Statistics System

Katherine K. Wallman

Many if not most who learned and benefitted from Steve Fienberg's contributions to our Nation knew him as a singular contributor in statistical methods. But there was a special place in Steve's heart and work for the quality and vibrancy of our Nation's official statistics and the federal system that produces those statistics. For me, this somewhat rare interest among academic statisticians first manifested itself when we served together on the American Statistical Association's Committee on Law and Justice Statistics circa 1980. More importantly, in the early 1980s Steve took up the cause of reviving and strengthening the statistical policy function at the US Office of Management and Budget (Fienberg and Wallman 1983). At that time, we coauthored a few of Steve's less technical writings – and testified effectively to the US Congress.

Our democracy and economy demand that public and private leaders, governments at all levels, businesses, hospitals, academic institutions, and individual households – essentially every organization and person – have unbiased, relevant, accurate, timely, and accessible information on which to base their decisions. Statistics produced by the Federal Government, often referred to as "official statistics," inform decision makers in shaping policies, managing and monitoring programs, identifying problems and opportunities for improvement, tracking progress, and monitoring change. The programs of our statistical system furnish key information to guide responses to pressing challenges, including among others those associated with the economy, agriculture, crime, education, the environment, health, science, and transportation. In a very real sense, these statistics provide data users with a lens to focus the myriad activities of our society into a more coherent picture of the status, progress, and trends in our Nation.

K. K. Wallman (✉)
Statistical Policy, United States Office of Management and Budget, McLean, VA, USA

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_23

Our economy's complexity, growth, and rapid structural changes, for example, require data on real Gross Domestic Product (GDP), the Consumer Price Index (CPI), and the trade deficit to guide government spending, budget projections, and the allocation of public funds. These data are also essential inputs to monetary, fiscal, trade, and regulatory policy. Economic data, such as measures of price change, also have a significant influence on interest rates and cost-of-living adjustments that affect every American who runs a business, saves for retirement, or mortgages a home. Taken together, official statistics on demographic, economic, and social conditions and trends are essential to inform decisions that are made by virtually every organization and household.

Among the most influential statistics produced by our Federal Government is the Consumer Price Index (CPI) which measures the average change in prices over time for a fixed market basket of goods and services. As an economic indicator, the CPI is used by the Executive Branch, the Congress, and the Federal Reserve Board to determine and evaluate government economic policy. For example, reports of monthly changes in the CPI are a major input for Federal Reserve Board decisions in setting short-term interest rates and annual changes in the CPI affect components of the Federal income tax code.

As a consequence of statutory requirements, the CPI directly affects the incomes of millions of our citizens through Federal programs that deliver benefits to individuals. These include more than 100 million Social Security beneficiaries, food stamp recipients, and military and Federal Civil Service retirees and survivors. Changes in the CPI also affect children through adjustments to the School Lunch program and private sector workers whose wages are tied to the CPI under collective bargaining agreements.

Similarly, information from our Decennial Census of Population and Housing and its companion American Community Survey affects Americans every day. Data on the number and characteristics of the population are used by state and local governments to plan schools and highways, by the Federal Government to distribute hundreds of billions of dollars annually for health care and other programs, and by businesses in making their economic plans. And of course Census data have been used since 1790 initially to reapportion congressional (House of Representatives) seats among the States, and in addition since the passage of P.L. 94–171 in 1975, to draw legislative districts within states.

A handful of examples further illustrates the wide-ranging role of official statistics produced by the Federal Government in informing policies and decisions at every level and sector of our society:

- *Monetary Policy Decisions:* The Federal Reserve determines monetary policy based on monthly economic indicator data, such as economic growth and unemployment rates produced by the Bureau of Economic Analysis and the Bureau of Labor Statistics.
- *State Budget Decisions:* Nearly all states use Bureau of Economic Analysis state personal income statistics to project state budget revenues; 20 states have

established constitutional or statutory limits on state government revenue and spending that are tied to these statistics.

- *Federal Program Funding Allocations:* Federal programs, such as Medicaid, allocate over $400 billion annually to states and localities based on Bureau of Economic Analysis and Census Bureau statistics on income and population.
- *Federal Program Administration:* The Social Security Administration provides statistical analyses to help policymakers understand the potential distributional effects of alternative policy changes on social security beneficiaries and on the larger economy.
- *Private Sector Investment Decisions:* The private sector uses Federal statistics, such as agricultural production and workforce availability from the National Agricultural Statistics Service, to determine optimal locations for industries such as ethanol and slaughter plants.

So what is our Nation's "official statistics system" that was a beneficiary of Steve's attention? What is this system that brings to us a wealth of statistics on physical and mental health, educational attainment, employment, criminal victimization, transportation, energy use, the environment, and many other dimensions of our lives? In the United States, we have what is generally known as a "decentralized system" for the acquisition, compilation, analysis, and dissemination of official statistics. At the core of this decentralized system are 13 agencies that produce statistics as their principal mission; their work is complemented by activities carried out in more than 100 other agencies having primary missions to conduct research, administer Federal programs, or perform regulatory functions. Some – but not perhaps all – of the former group (the "principal statistical agencies") likely are known to readers (the Census Bureau, the Bureau of Labor Statistics) and agencies that focus on education, health, transportation, or other "topical" areas. Perhaps less often thought of in this framework are agencies such as Commerce's National Oceanic and Atmospheric Administration, Health and Human Services' Centers for Medicare and Medicaid Services, Homeland Security's Customs and Border Protection, Interior's Geological Survey, and a number of others that in fact have greater expenditures for statistics production than many of the "principal" statistical agencies. In total, the annual budget for the Federal Government's statistical work is in the neighborhood of $6 billion (exclusive of cyclical funding related to the Decennial Census). This total is roughly 0.04 percent of GDP. In Decennial Census years, this figure about doubles.

At the "hub" of this decentralized statistical system is the Statistical and Science Policy Office, headed by the US Chief Statistician, located in the Office of Management and Budget (the largest component of the Executive Office of the President). This office, whose statutory authority currently resides in the Paperwork Reduction Act (as updated in 1995), has been in existence for more than 80 years, providing oversight, coordination, and guidance to promote the quality and integrity of Federal statistical information. The primary functions of this office are to identify priorities for improving statistical programs, ensure the adequacy of funding proposed for statistical programs in the President's budget, establish

government-wide statistical and information quality policies and standards, evaluate statistical programs for compliance with OMB guidance, critique and approve statistical information collections promulgated by Federal agencies, and coordinate US participation in international statistical activities.

Advising the Federal statistical system and its component agencies are various councils and committees (such as the American Statistical Association committee on which Steve and I served) that assist the agencies that carry out data production and dissemination activities. Broadest in its scope is the Committee on National Statistics (CNSTAT) within the National Academies of Sciences, Engineering, and Medicine. Established in response to a request by the 1972 President's Commission on Federal Statistics, CNSTAT convenes panels of volunteer experts to advise on a variety of statistical programs. Steve chaired CNSTAT for two terms in the 1980s – and even more fundamentally contributed to the quality and integrity of federal statistical programs by serving on (and often chairing) CNSTAT study panels on Data Sharing, Statistical Assessments as Evidence in the Courts, Decennial Census Methodology, Review of Evaluation Studies of Bilingual Education, Census Requirements in the Year 2000 and Beyond, Review of the Scientific Evidence on the Polygraph, Methods for Assessing Discrimination, Committee on Technical and Privacy Dimensions of Information for Terrorism Prevention and Other National Goals, and Survey Options for Estimating the Illegal Alien Flow at the Southwest Border. Steve was also a member of the Advanced Research Seminar on Cognitive Aspects of Survey Methodology. The breadth of his contributions via these activities mirrors the scope of his interests and expertise.

As we look to the future for official statistics, it is clear that Steve would have had many more contributions to foster the viability of our Nation's statistical system and the quality of its products. Over the years, many challenges have confronted the US statistical system. While some of the more immediate have been overcome or addressed (e.g., devolution of government programs, deregulation), several – such as limited fiscal resources, declining survey response rates, and obtaining and retaining critical human capital – seem to be forever with us. Now, as we find ourselves at the center of growing attention to "Evidence-Based Policymaking" (statistics are, I believe, the heart of evidence), other challenges that had been emerging are coming to us with increasing acceleration, or as my former OMB colleague Don Arbuckle once noted, "we are not experiencing life in the fast lane, but life in the oncoming lane . . . ." Chief among these are calls for more finely grained demographic and economic data at more discrete levels of physical and political geography, desires for internationally comparable statistics, and demands for fuller and quicker access to data in more modern and creative forms and venues.

Essential to guiding the agencies as they address the challenges is adherence to the fundamental principles that underlie their work: in addition to meeting standards of relevance, objectivity, accuracy, timeliness, and accessibility, they must be viewed as credible, trustworthy, and free from political or other undue external influence.

- For federal statistical programs to effectively benefit their wide range of public and private sector users, the underlying data systems must be viewed as credible.

As the collectors and providers of these basic data, federal statistical agencies are data stewards – balancing public and private decision makers' needs for information with legal and ethical obligations to minimize reporting burden, respect respondents' privacy, and protect the confidentiality of the data provided to the government. To further ensure the credibility of our Nation's official statistics, the agencies responsible for these products disseminate their results under strict guidelines that reinforce both the reality and the perception of their freedom from policy perspectives.

- A growing challenge concerns the ability to garner the attention of the politically powerful, but avoid their interference. As former Census Bureau Director Kenneth Prewitt has offered, "political interference is the attempt to gain partisan or regional advantage by shaping the production of a statistical product against the judgment of a non-partisan and apolitical statistical agency." (Prewitt 2010) Prewitt elaborates with examples:

> (1) the politically motivated suppression of an agency's responsibility to offer its best judgment on how to most accurately and reliably measure a given phenomenon, (2) the politically motivated decision to prevent an agency from using the state-of-the-art science, or (3) the politically motivated insistence on preclearance [review and approval by political appointees] of a major statistical product that is based on state-of-the-art science.

  Political interference is not limited to the production of statistical products; for example, such interference can also occur when agency personnel are fired or reassigned because they produce politically inconvenient and thus unwelcome results. The fact that statistical series have become so important to the development and implementation of social and economic policies inevitably creates a temptation to manipulate the numbers for political gain. By refraining from participating in political discussion, and following careful, objective procedures for compiling our official statistics, federal agency statisticians are assuring the continued avoidance of political interference in our Nation's statistics.
- Last – but perhaps far from least – is the principle of trust. Though initially framed as a principle related to the statisticians' compact with data providers, it seems today that there is at least as great a challenge in gaining the trust of those who use the data.

Given the context in which official statisticians must operate, I want to address the importance of embracing, rather than resisting, our changing environment. The question is, how can we harness what is going on around us to meet information needs in new and perhaps better ways despite a host of challenges?

First, as any casual reader of the newspaper can tell you, the Federal budget process continues to be fraught with uncertainties. But one thing seems certain: the funding constraints affecting many statistical agencies over the past several years are likely just the beginning of what we should expect for the next few years. At the same time, with respondent cooperation at an all-time low, it costs more and more just to maintain historic respondent cooperation in household and establishment surveys.

Second, with talk of freezes on hiring and pay and of buyouts common parlance around Washington, and government employment perhaps not at its most respected status, it's harder than ever to hire and retain staff to help innovatively adapt our infrastructure and squeeze every last drop of usefulness out of the resources devoted to our Nation's statistical programs.

Third, the ever-increasing desires for access to data and for more flexibility in how and where the data can be accessed must be balanced with increasing public concerns about confidentiality and privacy. Government statistical offices derive their mandate for data collection and dissemination from a citizenry that demands at once both quality information to drive public policy and protection of the individual respondent from privacy invasion and administrative harm. Striking the appropriate balance between permitting access to accomplish compelling and legitimate research, and incurring the risk, however remote, of inadvertent revelation of individual information is a fundamental concern and challenge for official statisticians. As the same technologies that extend analytical capabilities also furnish the tools that threaten the confidentiality of data records, there is the potential to erode respondents' trust that the information they provide will be kept confidential.

Fourth, enabled largely by the presence of our arguably staid official statistics as benchmarks, entrepreneurs are releasing statistics that are available much faster and with more easily accessible means of presentation.

In the face of these challenges, our citizens continue to encounter statistics at every turn in their daily lives. Yet recent surveys suggest that many are distrustful of the providers of the numbers and most are unequipped with the statistical literacy required to evaluate the information presented to them. Twenty-eight years ago, when I was privileged to serve as President of the American Statistical Association, I defined "statistical literacy" as:

> The ability to understand and critically evaluate statistical results that permeate our daily lives – coupled with the ability to appreciate the contributions that statistical thinking can make in public and private, professional and personal decisions. (Journal of the American Statistical Association 1993, p. 1)

I take hope from signs that our profession's quantitative literacy initiatives are increasingly pervading the curriculum – but as Fred Mosteller challenged me many years ago, what do we do about the "over-30 crowd"?

Without money or staff even to maintain our current programs, and with a growing cadre of outsiders suggesting to the public that they are mimicking the official statisticians' work for less money, more quickly and in a flashier way, some providers of official statistics may be tempted to resist these pressures, owing to legitimate concerns for data quality and confidentiality, and perhaps less legitimate fear of competition. Rather than focusing on trying to point out the fragility of the methodology and the flaws of these pseudo-official statistics, maybe producers of official statistics need to take into consideration increasing demands for more detailed and timely data in flexible, accessible formats. The time has come to think

differently about the data and the federal statistical agencies' role. It is time to think openly and creatively about how we might collaborate, rather than withdraw.

Each challenge I have noted can be seen as an opportunity. And challenging times create a favorable climate for bold new actions. By embracing these challenges, the Federal statistical system can create the kinds of changes that allow it to remain relevant, playing a pivotal role in meeting the Nation's information needs in concert with a larger set of government and nongovernment actors.

So what are the specific opportunities?

- In the same way that personal computers opened up a new world to the public, Apps and APIs (application program interfaces) are giving the general public the ability to "play with" data. They want to use these data to help make personal decisions.
- Entrepreneurs want the underlying data, in a form that is timely, transparent, and flexible for creating innovative solutions to everyday problems.
- And constituents and stakeholders want to assess the data from their own perspective in order to inform their understanding of the role those data have in public policy.
- Rather than fighting to "control" the message, we who produce official statistics can take advantage of the insatiable appetite for information to extend our reach beyond researchers and policy wonks.
- We official [government] statisticians have always thought of our role as:
  - Developing the most rigorous statistical and scientific methods available, for use in censuses and surveys, with well-understood and documented data quality characteristics
  - Designing data access for trained professionals who offer their interpretations (often) through a peer-reviewed process and who are frequently the translators via the media to a broader public
- If, instead, data are thought of as a product from and for the general public to use to make personal, business, and public policy decisions, what would that mean for the Federal statistical system in practice?

I do want to emphasize that regardless of the financial, technological, and political pressures, the Federal Statistical System must never lose sight of its steadfast commitment to ensuring the relevance, practical utility, quality, and credibility of the information generated, as well as the widespread dissemination of those data and the maintenance of public trust. Official statisticians must take advantage of the opportunities that today's challenges bring, while remaining true to these core principles.

It is not our principles that need reexamination; it is the way in which we apply them. To me this means that we need to take a step back so that we are in a position to challenge long-held assumptions on how we do what we do. There are likely more configurations of methods, tools, and roles that allow us to streamline our operations while still being true to our underlying principles.

Information quality cannot be determined in a vacuum.

We may not actually need the most robust sampling design or the highest response rate to achieve a given goal – we need to challenge ourselves by asking when the gold standard is appropriate and when the balance of the trade-offs might actually be in favor of alternative methods. We need to ask, for example, "What frequency or what level of geographic detail is necessary in the data that are to be used for a given purpose?" As another example, pushing the boundaries of how we measure and adjust for nonresponse bias can help us make better trade-offs in our survey designs especially in ways that help us manage costs.

We need to be more open to leveraging administrative data in new ways – whether they are resident in a Federal statistical agency, housed in Federal or local program agencies, or collected by the private sector. If such data can help us generate new and useful statistics, we need to think about how we might best use them.

We come back to assessing data quality in the context of "fitness for use," as administrative data rarely meet the same standards for quality as one might design in a collection actually intended for analytic purposes rather than for administering programs or benefits. But done thoughtfully, incorporating administrative data can be consistent with our principles, including relevance, without tarnishing our products. Pushing forward on how to measure the quality of administrative records is another important frontier.

Transparency – a very popular theme these days – is not new to us. In fact, at times, we may appear to some to be "over-transparent" as we document, for example, the uncertainty of our estimates. But transparency enhances the credibility of official statisticians and the trust of those who use official statistics. We should make data sets more readily accessible to users, even knowing that many will use those data sets differently than we would – "mashing" them with sources that we'd never have considered due to our more traditional view of the world of data.

Rather than fearing that non-statisticians will use the data incorrectly, *we should empower them by ensuring that we provide sufficient metadata so that they can educate themselves regarding the fitness of the data for their use*. After all, there is much to be gained by getting many eyes on a dataset.

Wide dissemination arguably means something much more than it did in years past, given the increasing demand and capacity. It doesn't diminish our commitment to confidentiality or make the task of releasing detailed microdata any easier. In fact, now we are talking about not only our survey data, but also administrative data sets – potentially even those from the private sector. But these data can "sit behind" some of the Apps and smart disclosure efforts to repackage data in new ways that have immediate relevance to consumers.

Relevance, practical utility, quality, wide dissemination, credibility, and public trust remain top data principles. Determining how we best implement these principles in light of this broadened user base is our opportunity and challenge.

Official statisticians can take advantage of the energy and enthusiasm of this information age to reinvent how we collect, analyze, store, and disseminate data and, in the process, reach a new, broader audience. We must enhance the relevance of Federal statistics; ensure that the Federal statistical system evolves to reflect changes in society and the environment, providing the public and private sector with

access to the information necessary to inform their decisions; and adapt information generation approaches to take advantage of technological innovations and cultural expectations. We must consider, for example, thinking more broadly about data sources and using available data in creative ways to provide new types of data products and insights into interpreting existing data; ways for agencies to become more policy relevant without crossing the line into policy making; the nature of the relationship between data science and official statistics; and ways to enhance and broaden the skill sets of current and future agency employees to effectively deal with new technologies, data sources, types and volume of information. It will be the Nation's good fortune to benefit from Steve Fienberg's legacy and example as colleagues emulate his thoughtful and generous contributions for OMB and the statistical agencies to consider as they navigate this changing environment, where the relevance of traditional statistics is sometimes questioned and policy makers are demanding faster, more flexible approaches to meet their information needs.

References

Fienberg, Stephen E. and Katherine K. Wallman (1983) The Importance of Federal Statistical Policy and Coordination. *Amstat News* March, pp 4–5, American Statistical Association

Prewitt, Kenneth (2010)What Is Political Interference in Federal Statistics? *The Annals of the American Academy of Political and Social Science* 631 September 2010, p 228. American Academy of Political and Social Science

Wallman, Katherine K. (1993) Enhancing Statistical Literacy: Enriching our Society. Journal of the American Statistical Association, Vol. 88 No.pp 1–8, American Statistical Association

Part VII
Introduction: Steve and Joyce as We Knew Them

Judith M. Tanur

Steve was a man of many talents and many interests. We have tried to capture some of these in the preceding sections of this volume, but had we chosen to devote a section to each of his interests we would have been undertaking to produce a library, not just a volume. So we decided to allow this section to be considerably more eclectic than the earlier ones, providing some views of Steve as we knew him and of his varied interests. Joyce was also a many-faceted person, and so we include material reflecting both her personal life including her roles as Steve's wife and mother to Anthony and Howard as well as her professional life as a researcher in her own right. Gaea Leinhardt refers to her as a woman of grace and wisdom, a characterization with which we strongly concur.

We lead off with a short obituary written by Steve's long-time friend Steve Stigler, followed by a detailed reminisce of about Steve's time at Carnegie Mellon by another long-time friend and colleague, Bill Eddy. Then Nancy Reid writes about Steve's devotion to statistical service and how he served as a role model to her in that respect.

Moving to a more light-hearted mood, two substantive pieces speak to Steve's interests in graphic presentation and in fine dining. Howard Wainer and Michael Friendly, friends whose acquaintance with Steve originated at far different stages of his career, use an extended metaphor of a wedding between empiricism and visualization to tell the story of the origin of data visualization. They note that when Steve and Bill Eddy were starting *Chance* in the 1990s they asked Wainer to write a column on data visualization. He agreed to do so for a short time – but was still writing the column in 2020. Albert Madansky, building on his fame as the author of "The Gentleman Tasting Pastrami" which appeared in Volume 1, Number1 of *Chance*, chronicles Steve's secret career as "Belizaire," the restaurant critic for *Chance*. Al notes that although Steve tended to review rather high priced "fine dining" restaurants, he often included a moderately priced alternative, which he labeled as a "Sam Greenhouse" restaurant, commemorating Sam's usual choice of eating place in conformance with his government travel allowance. What Al doesn't note, because there is no documentary evidence for such a statement, but

which I can attest to, is that Sam informally labeled restaurants that he considered overpriced as "Steve Fienberg restaurants."

Moving back towards memoirs, we include a reminiscence I prepared for the celebration to mark Steve's transition to Emeritus status in October 2016, dealing mostly with the visits he and his family made to our summer home in Montauk over the years and our work on our planned volume on the parallels between experimental design and probability sampling. We were able to publish several papers that were destined to become chapters of that book (which we had decided to call *Reaching Conclusions: The Role of Randomized Experiments and Sample Surveys*), but we never got to finish the volume, perhaps because we were too busy having fun. After that we reprint, with permission, an interview that Miron Straf and I carried out with Steve and that was published in *Statistical Science* in 2013.

It is a tradition of Carnegie Mellon's Department of Statistics to acknowledge milestones for esteemed faculty with a collection of memories from those who worked closely with the person, frequently on a daily basis. The reminisces from colleagues, friends, and students appeared in two "memory books" in honor of Steve for, respectively, his 65th birthday in 2007, and his attainment of Emeritus status in 2016. We include a selection of those reminisces.

Gaea Leinhardt's piece about Joyce as a woman of grace and wisdom leads off our section on Joyce. She describes Joyce's activities over 30 years as an educational researcher. This is followed by two obituaries of Joyce which appeared in the *Tribune-Review*, written soon after her murder in the Tree of Life Synagogue shooting in October 2018. Then there is a series of tributes to Joyce from friends and from Steve's former students, also written shortly after the murder.

We conclude with a selection of photographs of Steve and Joyce with their family, with colleagues and with friends, though surely those categories overlap substantially, and with a list of Steve's PhD students.

Chapter 24
Stephen Elliot Fienberg, an Obituary

Stephen Stigler

Stephen Fienberg was the senior statesman of statistics in his era, both nationally in the United States and internationally. His advisory and editorial activities covered an amazingly broad set of areas and his engagement was always deep and effective. For the past few decades, no meeting or conference on pressing statistical issues such as census undercount or non-reproducibility or ethical experimentation would be complete without his lively and focused participation.

Steve was born November 27, 1942, in Toronto where he remained through his graduation from the University of Toronto in 1964. It was at that University that he first encountered the field of statistics, in a class taught by Don Fraser. The subject proved infectious; he went on to Harvard for his PhD, written under the supervision of Fred Mosteller. At Fred's suggestion, the work of his dissertation was considerably expanded in active partnership with two other Harvard researchers, Yvonne Bishop and Paul Holland, into the very influential book, *Discrete Multivariate Analysis*, finally published by MIT Press in 1975. That book, colloquially referred to as "Bishop, Fienberg, and Holland," did not invent loglinear models, but it played a crucial role in helping to develop them and inspired a major growth in research in the analysis of categorical data. At a workshop in his honor just two months before he died, Steve told a story that one of his sons was taking a statistics course in college; the instructor approached his son and asked, "Are you any relation to Bishop Fienberg of Holland?" He said the son replied, "I don't believe so, we are Jewish."

Reprinted with permission from *The Annals of Applied Statistics,* July, 2018

S. Stigler (✉)
University of Chicago, Chicago, IL, USA
e-mail: stigler@uchicago.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_24

Steve's first appointment after his PhD was in 1968 at the University of Chicago, where he was jointly in the Department of Statistics and the Department of Theoretical Biology. I first met Steve and his wife Joyce shortly after that, when I visited Chicago for a talk and they hosted a party for me at their apartment. We were never at the same institution, despite several determined attempts to accomplish that, but that first meeting was to be repeated regularly over nearly a half century. He left Chicago in 1972 to Chair a new Department of Applied Statistics at the University of Minnesota. In 1980 he moved to Carnegie Mellon's Department of Statistics, where he remained the rest of his life, save for a brief period as Provost at York University in Toronto.

Steve's research developed far beyond his thesis on contingency table models, into network analysis, methodology for confidentiality and for statistical privacy, algebraic statistics, and the application of statistics in science, particularly in social science. He even wrote on the history of statistics, including a paper that traced the history of the term "Bayesian" to R. A. Fisher in 1950, who used it in a pejorative sense.

Steve played a major role in what may be called the infrastructure of the profession of statistics. He followed two of his great role models in this, Fred Mosteller and Bill Kruskal. One part of this was editorial – early on in his career, he served as Coordinating and Applications Editor of *JASA* 1977–1979. Later he was Associate Editor and then Editor of *the Annals of Applied Statistics* 2006–2015. Steve was founding or cofounding editor of the *Journal of Privacy and Confidentiality, Chance Magazine*, and the *Annual Review of Statistics and Its Applicatio*ns. He wrote or co-wrote or edited about 30 books, including *Statistics and the Law* (with Kadane and DeGroot), *Intelligence, Genes, and Success* (with Devlin, Resnick, and Roeder), and *Who Counts?* (on the US Census, with Anderson). But beyond this, Steve played a major role as advisor and critic for many statistical agencies, including the US Census and, over many years, the committees of the National Academy of Sciences, where he played the major role in their 2003 report reviewing the *Scientific Evidence on the Polygraph*, and was on the National Academy's Report Review Committee, which he cochaired in 2008–2012.

Among Steve's many honors were the 1982 COPPS President's Award, and election to the US National Academy of Sciences, the Royal Society of Canada, and the American Academy of Arts and Sciences.

And he did not neglect other aspects of a full life, which for him included ice hockey as coach and player (well beyond the age some of us thought wise, yet with no noticeable loss of teeth). Dinners with Steve and his wife Joyce in great restaurants with fine wines and wide ranging discussion were always a treat. I particularly recall nights in Paris, in Strasbourg, and in Dublin, and one night in an Italian restaurant in Manhattan when after we were seated Jackie Kennedy Onassis came in with a friend and sat at the next table, and we realized that the check would likely set a new record cost for us (it did).

Despite his extensive international commitments, he was devoted to his 43 PhD students, and to judge their comments upon Steve's Emeritus attainment celebration in October 2016, this devotion was reciprocated. By that time his four-year struggle

with cancer was near the end, but no one without that knowledge would have guessed it in view of his vigorous presence throughout the celebration, continuing through the late party (including a small jazz combo) at his home. The energy he brought to his half-century career in statistics seemed undimmed. He died December 14, 2016, and left an indelible imprint on the literature of statistics, on statistical practice in many fields, and on countless colleagues and students.

Chapter 25
Reminiscences of Steve

William F. Eddy

Delivered on the occasion of Steve's Emeritus celebration in October, 2016.

I first met Steve in June, 1979, at the First International Bayesian Conference. At that time, Steve was the head of the Department of Applied Statistics at the University of Minnesota. We were both invited to the conference by Morrie DeGroot: Steve, to begin recruiting him to join us at CMU, and myself, to begin converting me to Bayesianism.

Our group of Steve, Morrie and I and wives Joyce, Marilyn, and Connie spent every night at a local nightclub until 3 a.m. Led by Morrie, we were all up bright and early each morning to attend the first session. Morrie's style apparently agreed with Steve, and later that year he agreed to come to CMU in 1980, and became head in 1981.

Steve had a variety of interests at that time – the National Crime Survey, cognitive methods in survey research, statistics and the law, and of course, the traditional topics he's interested in like loglinear models and computation. Those interests have only grown broader with time.

In 1984, Steve stepped down as head. A year later, he was named Maurice Falk Professor of Statistics and Social Science. In 1985, he came up with the idea for a statistics magazine. We developed the idea, and after some struggles, *Chance* was born in 1988. Steve and I had a lot of fun doing those early issues. We really enjoyed dreaming up covers and having a professional photographer create them for us. One of our favorites was connected to an article by Joel Greenhouse and his father, Sam. The photographer was Joel's brother, Richard. Our 30th anniversary is next year.

In 1987, Steve was appointed dean – no rest for the weary. In 1991, he returned to Canada as professor of statistics and the law and vice president for academic affairs at York University in Toronto, his childhood home. In 1993, then CMU Pres. Robert

W. F. Eddy (✉)
Statistics Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
e-mail: bill@stat.cmu.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_25

Mehrabian tried to recruit Steve back to the department. I recall Steve saying "there was no other place I'd rather be." Others may recall the department's tenure review of Steve, with an obvious outcome.

Then there was the time when I ran into Steve at Charles de Gaulle Airport near Paris. It's not surprising to run into a friend at PIT, but CDG? Steve was in transit returning from a meeting, and I was in transit going to a meeting. Curiously, despite our close friendship and regular interactions over all these years and despite our sharing an office wall, we only wrote a few papers together more than 30 years ago.

Speaking of that shared wall . . .

It wasn't long after Steve's return to CMU that a staff member went into his office to discover all his books on the shared wall in a heap on his floor. The shelves tore everything from the wall, nearly destroying it. The incident forced me to examine the shelves in my own office, which were teetering dangerously close to the same fate as his. The wall was rebuilt and reinforced on both sides, and there has been no problem since. I thank Steve for taking "the hit" for both of us.

As friends, we can discuss anything. Strong professional disagreements – while rare – are easily resolved. Really weighty matters – such as wine and single malt scotch preferences – are settled in taste tests at his home. Steve has been a critical element in the growth and development of the department. When he came there were about 10 faculty and 20 graduate students; during these years we've grown to over 30 faculty and over 75 graduate students.

Around 1995, Dick Cyert, former president of CMU, was trying to start a research center focused on the study of business/customer relationships. To generate corporate interest, he organized a "dog-and-pony" show for a group of VPs. About 15 core faculty gave presentations to about 15 VPs. The faculty included Steve, Tom Mitchell of CS, and me. We discovered our common interest in what was coming to be called data mining. Tom and Steve had continued discussions which eventually led to the creation of the Center for Automated Learning and Discovery in SCS with Tom as Head and Steve, Larry Wasserman, and me as inaugural members. It has since become the first Machine Learning Department in the world.

Despite his regular travels (those trips to Paris to see his five grandchildren there must really rack up the frequent flier miles!), we still talk regularly about the statistics department, about the profession, about our other departments, about other professions.

As many of you know, about 5 years ago, I became Emeritus. For those of you who don't know what that means, I gave up my tenured professorship, but I did not retire. My idea was I would, over several years, cut back on my activities, so by, say 2016, I would retire (because I had nothing to do). Instead Steve managed to get me involved in a census research project; in fact, of course, he has involved several other faculty and several graduate students. And, following that, he got me into another project, the Center for Statistics and Applications in Forensic Evidence (again, with other faculty and students). Several other people did similar things to me, and here I am working full time when I should be in Florida getting a suntan and growing old.

Steve has noticed that becoming Emeritus doesn't actually mean much except that you don't have to do things you don't want to do. Steve is now Emeritus also.

Steve, welcome to the club!

Chapter 26
Statistical Service

Nancy Reid

26.1 Introduction

The first time I met Steve, probably in 1981, he asked me why I had not applied for a larger grant from Canada's funding agency, the Natural Sciences and Engineering Research Council, as awards were limited to the amount requested. It was my first such application, and I suppose I asked for what I thought I might spend—I do not remember getting any particular advice from my department colleagues. I was surprised that he knew about the application, but (of course) he was a member of the grant selection committee. I was even more surprised that he told me he had argued at the committee meeting for a larger grant award for me. This was typical Steve, as I came to know him. Although living and working in the USA, he had agreed to serve on a Canadian grant panel, which would have involved a punishing week-long meeting in Ottawa, along with a great deal of preparation. In the midst of this, he found the energy to champion a first-time applicant, both within the committee and through personal contact. He probably broke some committee rules in doing this, but he had his eye on the big picture. And I learned a valuable lesson about grant applications, which I have passed on to my colleagues in turn.

In the many years since, I have done a good deal of statistical service myself, partly because I was asked and did not like to say "no," partly because it enabled me to engage with friends and colleagues around the world, but mainly because it opened my mind to different approaches to the advancement of the discipline. Many of these service activities intersected with Steve, because he was everywhere and doing everything. In the remainder of this essay, I will give an incomplete overview

N. Reid (✉)
Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada
e-mail: reid@utstat.utoronto.ca

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_26

of Steve's contributions, with emphasis on those in which I was directly or indirectly involved.

I have drawn heavily on Steve's *curriculum vitae* up to mid-2016—he sent it to me at my request so that our department could nominate him for an honorary doctorate from the University of Toronto. Unfortunately, he died before this honor could be awarded to him.

26.2 Editorial

In an article in the *Annals of Applied Statistics*, in the Fienberg memorial issue, I wrote [1]

> When I met Steve in recent years, at conferences or committee meetings, he always seemed to be busily tapping away on his iPad, and when he looked up he would say "Annals of Applied Statistics." (Although as it turned out he was editing several journals at the same time.)

During his career, Steve was editor, co-editor, associate or guest editor of 24 different journals and encyclopedias, from the *Annales de La Faculté des Sciences de Toulouse Mathématiques* to *The Statistician*. He was the founding co-editor, with Bill Eddy, of *CHANCE* magazine (1988), co-founder and later editor in chief of the *Journal of Privacy and Confidentiality* (2006), and founding editor of the *Annual Review of Statistics and Its Application* (2013). He was editor, senior editor, and editor in chief of the *Annals of Applied Statistics* (2006–2015), section co-editor of the *International Encyclopedia of the Social and Behavioral Sciences*, coordinating and applications editor of the *Journal of the American Statistical Association*, editor of *Statistical Science*, and a member of the editorial board of the *Proceedings of the National Academy of Science*. If there is a common thread, it is Steve's dedication to making the breadth and depth of statistical thinking available to many different audiences, from the general public to the advanced researcher.

An interview with Bill Eddy and Steve in Volume 20 of *CHANCE* gives a snapshot of the creativity, determination, and drive that led to the creation of the journal [2]. Although Volume 1 appeared in 1985, they had started the project three years earlier. While their original intent was to create a popular magazine, like *Scientific American* or *Psychology Today*, I think the impact has been more profound in the teaching of statistics. The first edition of Freedman's groundbreaking statistics text [3] was published in 1978, and like this book, *CHANCE* was a major contributor to a new approach to statistical education and to the creation of an active group of scholars focused on statistical education. My exposure to this was through the article describing a course developed by Laurie Snell and colleagues at Dartmouth [4]. The community established around this time continues to provide vibrant leadership in statistical education; see, for example, the data science roundtable and report by the National Academies' Committee on Applied and Theoretical Statistics [5].

I had a firsthand view of Steve's publishing vision in my role on the inaugural editorial board of the *Annual Review of Statistics and its Application*. Correspondence obtained from the president and editor-in-chief of *Annual Reviews* at the time, Sam Gubins, shows how quickly Steve moved to suggest a list of editorial board members, a list of potential invited papers, and a vision for the scope of the journal. Although the series of volumes published as *Annual Reviews* has an outsized role in many scientific fields, Steve was aware that it would take some effort to ensure that the new volume would have impact in the statistical sciences community and also emphasized the importance of ensuring that the journal could also speak to scientists and social scientists in a wide range of disciplines. He presented a clear and compelling vision to his editorial team and led the early meetings with authority and enthusiasm.

Annual Reviews, as the name suggests, focuses on overview articles summarizing the current state of the art on some topic. *Statistical Science* and, in some respects, *CHANCE* are also targeted on reviews, rather than original research. It is my impression that these types of articles are underappreciated by academic researchers and by promotion and tenure committees. I think this is a legacy of our roots in mathematics, which tends to view contributions to learning and teaching as distinct from "real research." At the same time, the articles are very much appreciated by students, researchers in other fields, and statistical researchers wishing to learn about a new topic. I would like to see our academic colleagues more willing to write, and to value, review articles.

26.3 Visiting Committees

One of the more interesting, though grueling, service activities for an academic is membership on an external review committee of a department or center. These run the gamut from an exercise in affirming the "world-class status" of the unit in question, to an exercise in emergency resuscitation, or pronouncement, on a unit that is on life-support. In all cases, this is, in my experience, therapeutic for the less powerful members of the department, a burden for the senior leadership, and a lottery in terms of the impact on the deans, vice-provosts, and others commissioning the review. The work of the external reviewers is mentally and emotionally exhausting—most of my colleagues who have contributed to this particular aspect of service have decided after fulfilling a handful of these efforts to gracefully decline further requests. In contrast, Steve reviewed 13 different Departments of Statistics, several of them twice, with the first in 1987 and the last in 2015. He also served on panels and review committees in public health, nutrition, sociology, education, and more.

We served on one such committee together; at the time, I was chair of our department and thought it would be a good learning experience to see another department close up. In addition to having a commanding view of the needs and opportunities of the department under review, and seemingly able to draft the report

in his sleep, Steve found time to give me advice about my new role. One of his more surprising questions was "Have you identified the next chair yet?"—I could barely focus on the current semester, let alone the end of my term. And yet, as experienced chairs know, this is indeed one of the most important files, for the health of the department and for one's own sanity.

26.4 Mentorship

In [6], we collected a series of vignettes from some of Steve's collaborators, as an incomplete attempt to illustrate the breadth of his scientific contributions. Many of these collaborators were former students, now colleagues, and nearly all of these added some comment about the importance to their career of Steve's role as a mentor. Similar text appears in the special section of the *Annals of Applied Statistics* [11] published in his honor in 2018 and an earlier special section of *CHANCE* [12].

What struck me as we compiled these contributions was the incredible range of people who expressed, in different words, more or less the same sentiment: gratitude toward Steve for his enthusiasm, encouragement, vision, engagement, collaboration, and leadership. Paul Holland wrote about the key role Steve played in writing and publishing what he called the "Green Monster" [7]. Bernd Sturmfels noted that "Steve had the vision to see this [the connections between algebra and statistics] before anyone else did." Philip Dawid described Steve's pivotal role in engaging him in a project that turned out to be "one of the most fascinating experiences of my professional life." Jana Asher wrote, in [11], "Steve was the only person that seemed to truly understand how my mind worked."

Steve was a member of the first Scientific Advisory Committee for the newly established Canadian Statistical Sciences Institute, which was launched by the Statistical Society of Canada in 2012 and received its first NSERC funding through the Canadian mathematical sciences institutes in 2014. He gave me a great deal of encouragement and advice when I became the Scientific Director in 2015. The Center for Statistics and Applications in Forensic Evidence was being established at the same time, under Steve's co-leadership, and with a budget many orders of magnitude larger than CANSSI's. But Steve took my questions and problems quite seriously and pushed the leadership of CANSSI and the Scientific Advisory Committee to think big, think ahead, and think broadly. His most memorable advice to me about funding was "Don't waste time fighting for a bigger share of the pie. Find a way to open your own bakery!"

26.5 The National Academies

Steve undertook an unusually wide range of service activities, but I think he may have been most proud of his work for the National Academies. He was elected to the National Academy of Science in 1999, chaired the Applied Mathematical

Sciences section from 2006 to 2009, and served on the NAS Council and several of the Council's committees from 2013 to 2016.

He also served on many committees of the National Research Council of the National Academies in three different divisions: Behavioral and Social Sciences and Education, Physical and Engineering Sciences, and the Institute of Medicine. In these capacities, he served on panels, organized and participated in workshops, and contributed to NAS reports. He was a member of the Committee on National Statistics for 10 years and Chairman for seven of those. In this capacity, he edited four reports published by the National Academies Press and contributed to dozens of other reports and workshops. One of the most significant of those was [8], the report of the Committee to Review the Scientific Evidence of the Polygraph, which Steve chaired. Earlier he had chaired the Panel on Statistical Assessments as Evidence in the Courts [9], which includes a case study on the use of hair samples in forensic analysis. He had a long-standing interest in forensic science, one important outcome being the current Center for the Statistical Analysis of Forensic Evidence, which at long last is bringing rigorous statistical thinking to forensic science.

To give a flavor of the breadth of his interests and remarkable range of activities, he was a member of the following: Committee on Law and Justice, the Standing Committee of the American Opportunity Study, the Committee on Technical and Privacy Dimensions of Information for Terrorism Prevention and Other National Goals, the Committee on Social Evidence for Use, the Committee to Assess the Value of Research in Advancing National Goals, the Committee on Applied and Theoretical Statistics, the Committee on Department of Energy Radiation Epidemiological Research Programs, the Computer Science and Telecommunications Board, the Committee on a Study of Food Safety and Other Consequences of Publishing Establishment Specific Data, the Committee on Review of the National Immunization Program's Research Procedures and Data Sharing Program, and the Committee on Secondhand Smoke Exposure and Acute Coronary Events.

In addition to all these committees, panels, and workshops, Steve served on the Report Review Committee from 2005 to 2016 and as co-chair of this committee from 2008 to 2016. In this position, he oversaw the review of 200 reports each year. The obituary published by Carnegie Mellon University [10] quotes Porter E. Coggeshall, the Executive Director of Report Review: "Steve's exceptional regard for integrity and quality has ensured that reports met the Academies' highest standards for objectivity and evidence."

What comes through in Steve's *vitae*, in comments from his colleagues, in the many obituaries, and so on is his remarkable dedication to advancing the careful use of statistical arguments in all areas of application, but especially in areas related to government, public policy, and the judicial system. Pleas to academic scientists to engage with the "outside world" are common, but few of us find this very easy to do, so we tend to avoid it. We are all of course very busy with our regular program within the academy; this engagement takes a very concerted effort, and any one effort may not have much impact. Steve took the charge very seriously and was unusually effective in this, but he is not here, and if many of us were to make a little extra effort, we could do his legacy proud.

26.6 Professional Societies

Not surprisingly, Steve was also very active in contributing to professional societies. He filled several different roles in the American Statistical Association, the Institute of Mathematical Statistics, the International Society for Bayesian Analysis, the International Statistical Institute, the Royal Statistical Society, the Statistical Society of Canada, and the Committee of Presidents of Statistical Societies. He served on the scientific advisory committees of the Banff International Research Station, the Centre de Recherches Mathématiques, MITACS, NISS, SAMSI, and the Canadian Institute of Statistical Sciences. He was active on the Council and in Section U of the American Association for the Advancement of Science and the Council of the Inter-university Consortium for Political and Social Research. He was on the organizing committee for the International Conference on Forensic Statistics for each of its triennial meetings from 1990 to 2011.

Among his many contributions to these and other organizations, he was influential in his role as President of the Institute of Mathematical Statistics, in finalizing the details of the arrangement between IMS and JStor, and between IMS and Project Euclid. The discussions with JStor began during my term as IMS President and moved as slowly as discussions between academics and lawyers tend to move, but they were completed in 1999. I think the leadership of IMS on this file led the way to other statistical societies signing on their publications with the foundation, an enduring legacy for our profession.

26.7 Closing

I have not even touched on many other professional activities that Steve undertook; his *vitae* lists 44 separate national and international organizations for which he consulted, including Statistics Canada (1975; 1992–1998), Statistics Netherlands (1997–1998), the Bureau of Labor Statistics (1985–1985), the UK's Engineering and Physical Sciences Research Council (2010) (which produced an influential report on their international review of the mathematical sciences), the Allegheny County Health Department (1988–1989), the Centers for Disease Control (1989–1990; 1994), and so on. He gave testimony to Congress and other government committees on 17 different occasions. He was the only statistician on the National Commission on Forensic Science (2014–2016) and, in 2016, served on the President's Council of Advisors on Science and Technology.

He combined all these efforts with an astonishing research output of more than 340 technical publications in statistics journals and in applied journals, six books authored or co-authored, and innumerable discussion pieces, short editorials, and more. While his enthusiasm for research was unbounded, his determination that this research should lead to better science, social science, health, and government

was an equally important part of his legacy. This involved countless hours of travel, meetings, phone calls, emails, presentations, and more meetings.

Robert Groves, executive vice-president and provost of Georgetown University and former director of the U.S. Census Bureau, was quoted in [10]: "Steve Fienberg's career has no analogue in my lifetime." While we can never fill his shoes, we can all try a little harder to contribute to statistical service.

Acknowledgments This work was partially supported by the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs Program.

References

1. Fraser, D.A.S., Reid, N., and Lin, W. When should modes of inference agree? Some simple but challenging examples. Annals of Applied Statistics **12**, 750–770 (2018)
2. Wainer, H.: Taking a CHANCE: An interview with William F. Eddy and Stephen E. Fienberg, CHANCE **20**, 33–39 (2007) https://doi.org/10.1080/09332480.2007.10722872
3. Freedman, D.A., Pisani, R. and Purves, R.: Statistics. (1st ed.) W.W. Norton & Co., New York (1978)
4. Snell, L.J., Finn, J. A course called *CHANCE*. CHANCE **5**, 12–16 (1992)
5. NAS: Roundtable on Data Science Postsecondary Education. (2020) https://sites.nationalacademies.org/DEPS/BMSA/DEPS_180066. Cited 30 Aug 2021
6. Carriquiry, A.L., Slavković, A.B. and Reid, N. Stephen Elliott Fienberg 1942–2016, Founding Editor of the Annual Review of Statistics and Its Application. Annual Review of Statistics and Its Application **6**, 1–18 (2019) https://doi.org/10.1146/annurev-statistics-030718-105334
7. Bishop, Y.M., Fienberg, S.E. and Holland, P. Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge (1975)
8. National Research Council. The Polygraph and Lie Detection. Committee to Review the Scientific Evidence on the Polygraph. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press 2003 https://doi.org/10.17226/10420
9. Fienberg, S.E. (ed.) The Evolving Role of Statistical Assessments as Evidence in Courts. Springer, New York (1989)
10. Rea, S. Stephen E. Fienberg, 1942-2016. Carnegie Mellon News (2016) https://www.cmu.edu/news/stories/archives/2016/december/obituary-fienberg.html. Accessed 23 Sept 2019.
11. Special section in memory of Stephen E. Fienberg (1942-2016) AOAS Editor-in-Chief 2013-2015. Ann. Appl. Stat. 12 (2018), no. 2, iii–x. https://doi.org/10.1214/17-AOAS122ED. https://projecteuclid.org/euclid.aoas/1532743469
12. Behseta, S. Editor's Letter – Vol. 26, No.4. CHANCE 26 (2013). https://chance.amstat.org/2013/11/editors-letter-26-4/. Accessed 29 Sept 2019.

Chapter 27
On the Origins of Data Visualization

Howard Wainer and Michael Friendly

27.1 Prelude

On December 24, 2018, we received an early Christmas gift from an old friend. Judy Tanur wrote to tell us that she had joined with two other friends to generate a memorial volume for Steve Fienberg, whose untimely passing in 2016 had saddened us all; this sadness exploded to tragedy with the murder of Steve's widow, Joyce, at her synagogue on October 27, 2018. We were honored to be included among the contributors to this volume and immediately set to work choosing an appropriate topic.

HW's history with Steve goes back to September of 1970 when we, as statisticians and new assistant professors at the University of Chicago, were both assigned to count votes in the faculty election. We became friends and continued to collaborate on various projects for almost 50 years. Two of those collaborations are relevant to today's topic, the use of data visualization both for data exploration and for the communication of quantitative phenomena.

1. In 1978, HW was in Washington, directing the NSF-sponsored Graphic Social Reporting Project, and as part of this project had convened a conference of interested scholars. Steve was one of the principal speakers. Al Biderman, HW's co-organizer of the conference, introduced Steve as "Holland's only Jewish Bishop." At that time, it was an insider's joke, but one that is not likely to be obscure to the audience of this volume. Steve's message in his talk, as nearly as we can remember it, was that a scientist's job was to speak truth – especially to

H. Wainer (✉)
Pennington, NJ, USA

M. Friendly
York University, Toronto, ON, Canada

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_27

those who were in positions to make policy – and that the best way to convey quantitative truths was using pictures, not numbers.

2. In 1988, Steve and Bill Eddy began the statistics magazine *Chance*, which they had imagined as readable by the general public, a parallel to *Scientific American* that dealt with the field of statistics, principally its applications. In 1990, Steve asked HW to write a column for *Chance* called "Visual Revelations" that would focus specifically on statistical graphics. He felt strongly that this should be a topic of primary concern to quantitative scientists generally and the readers of *Chance* specifically. The case he made was so convincing that HW agreed to take on the column for at least a couple of years; as of this moment, 31 years later, he is still writing it.

A decade later Steve took a leave of absence from Carnegie Mellon to return to his native Toronto and ascended to the position of Vice President for Academic Affairs at York University. It was here that he made the acquaintance of MF, who was then chair of York's Academic Planning Computing Committee. Together they immediately set about developing a plan to modernize the university's computing facilities. They planned to shift from a mainframe environment to the modern approach of distributed computing using the exploding developments of personal computers, augmented, for heavier tasks, with work stations. They were especially attracted to those developed by Steve Jobs (MACs and NeXT machines) whose graphical user interfaces were obviously (to Steve and MF) the future of computing. Steve's administrative leadership was as inspiring as it was rare and, within 18 months York's computing environment, had moved from the trailing edge to the forefront.

The subject of this essay is the origin of data visualization, and we begin with an extended metaphor of a wedding between empiricism and visualization. Among the details of the wedding that were only recently unearthed was that Steve Fienberg was the guest of honor at the ceremony. In his toast, he offered his wishes for a rosy future in which the offspring of the union would provide wisdom in choosing important questions and guidance in finding solutions. Among those joining Steve at the dais was a distinguished sextet of our graphical heroes, each with his own special gift to the couple. To his left were:

William Playfair, who presented the couple with a beautiful, hand-colored plot of England's national debt indicating how wars contributed to its skyrocketing growth.

Andre-Michel Guerry, whose gift of a shaded map of crime provided suggestions of potential causes and possible remediation, marking the birth of modern criminology and social science. The renowned Belgian statistician Aldolphe Quetelet had also figured largely in this but sent his regrets and best wishes to Steve.

John Snow, who gave an innovative dot-map of a cholera epidemic in London that marked the start of modern epidemiology.

And to Steve's right were:

Charles Joseph Minard, who had adapted his visual stories of commerce to provide a heart-rending tale of the horrors of war

Francis Galton, who had set aside his fascination with regression to construct a map of weather patterns decorated with multivariate glyphs that would allow the honeymooners to only go where the sun was shining brightly

Emile Cheysson, director of the statistical bureau of the French Ministry of Public Works who produced the *Albums de Statistique Graphique*, the most ambitious effort to make data of the state ("statistics") accessible to public inspection

27.2 One Wedding, No Funeral

What follows is the saga of a family. It begins, as do all such stories, with a marriage. The marriage is a good one, and we will learn how it came to be as well as how it evolved in both richness and depth. Our tale then branches to the issue from that union, taking time to include both the geneses of the offspring and their accomplishments.

Let us begin.

The marriage represents the joining of the epistemological approach of empiricism, as a window to understanding the world, to visualization as a way of connecting evidence to human experience. It was a match made in heaven, for it facilitated the easy connection of the seat of the intellect to the seat of the pants.

Empiricism begins, as do so many things, with Aristotle, but it was a tough sell, for it meant that any proposal, no matter how convenient or how beautiful, could be dismissed with a single reliable fact. Aristotle got away with such a rigid outlook only because he had Alexander the Great watching his back. Thus anyone who crossed Aristotle had to contend with Alex. But even Aristotle didn't buy in fully – he proposed that women had but 28 teeth. This was a rational conclusion; after all women were smaller and more delicate. What need they for the extra chomping power? Of course, had Aristotle truly digested his own epistemology he would have counted – he did have two wives. But apparently, this particular application of empiricism never occurred to him.

After the passing of the perfect storm of Aristotle, Alexander, and the Golden Age of Greece, empiricism faded. It briefly reappeared with Roger Bacon (1214–1292), who told us that,

> Reasoning draws a conclusion, but does not make the conclusion certain, unless the mind discovers it by the path of experience. (Bacon, 1897 in the section on the scientific method, *De Scientia Experimentali*, *page 244*)

But once again it slipped away only to gain a firmer foothold with the work of Francis Bacon (1561–1626), the second of the two fabulous Bacon boys. Following on quickly were the British empiricists John Locke (1632–1704), George Berkeley (1685–1753), and David Hume (1711–1776). And so by the end of the eighteenth century, one spouse was ready.

27.3 Picture that

The history of visualization is much longer. Perhaps the best known of very early examples is found in the Lascaux caves near the French village of Montignac. On the walls are remarkable drawings of animals, which carbon dating has estimated to be more than 17,000 years old. But although the Lascaux cave drawings are remarkable, and a bit stylized, they represent straightforwardly just what was seen (Fig. 27.1).

Jumping forward in time (to about 1400 BCE) and southward in direction, we arrive in ancient Egypt. The lives of most of Egypt's inhabitants revolved around the Nile. The regular floods of the river would wash away all but the most stubborn of property markers; thus, maps were prepared to indicate whose land was whose after the waters receded. Maps were a brilliant solution to a very practical problem, but, like the cave drawings in ancient France, they represented space in spatial terms. Aside from their miniaturization (and a rudimentary coordinate system of intersecting horizontal and vertical lines to enable a more precise placement of data points), this hardly represented a huge conceptual breakthrough.[1] For that we would have to wait more than two millennia.

But developments in cartography were to presage some critical future developments in the communication of evidence. For example, Descartes (1596–1650) is typically credited with the establishment of his eponymous coordinate system. But Hipparchus (ca. 140 BC) had a reasonably refined system for locating points in the

Fig. 27.1 Four sample paintings from the Lascaux caves, courtesy of the Bradshaw Foundation

[1]Maps were developed independently in the Far East. During the Warring States period in China (about 227 BCE), we find the first mention of a Chinese map being drawn. It showed a portion of Dukang that the Yan State was to cede to the King of Qin in exchange for peace.

heavens; its axes were called (translated into Latin) *longitudo* and *latitudo*. Roman surveyors used a coordinate grid to lay out their towns on a plane that was defined by two axes; the *decimani* running from east to west and the *cardi* that ran north to south. Music notation (as early as the ninth century) used a horizontal axis to represent time and the vertical axis for pitch; and the chessboard (with its associated notation to locate pieces) was developed in seventh-century India.

All of these visualizations were of something real and specific in the world. We might capitalize them as we do proper nouns – I saw Sam yesterday; my land and Gamal's before the spring flood; our chess game as we left it today. But there was another branch of visualization developing too – the common noun part – in which what was being depicted was theoretical. One well-known early example was published in Padua in 1486 on the first page of Oresme's *Tractus de latitufunus forarum* (Wainer, 2005, P. 10). The common noun structure of scientific visualizations is not a surprise once we consider that the reigning epistemology in natural science grew out of natural philosophy that favored a rational rather than empirical approach to scientific inquiry.

By the middle of the eighteenth century, the empirical seeds sown by Locke, Berkeley, and, especially, Hume (whose 1738 *Treatise on Human Nature* and his 1741 *Essays*, *Moral and Political* had a profound influence on Adam Smith, Jeremy Bentham, and Immanuel Kant) had started to bear fruit. The Scottish enlightenment, a magical period in the eighteenth century, gave rise to a torrent of practical innovations in mathematics, science, and medicine. James Watt revolutionized manufacturing; Adam Smith's *The Wealth of Nations* started modern economics; and the mathematician/geologist John Playfair's advocacy of Hutton's evidence-based theories yielded an estimate of the age of the Earth that was very much at odds with the 6000-year biblical estimate. But the star of our story is not the very worthy John Playfair (1748–1819), but rather his ne'er-do-well younger brother William (1759–1823).

Early on in his working life, William Playfair was a draftsman for James Watt. He later went on to become a pamphleteer typically focusing on political arguments based on economic data conveyed in vivid, original graphical forms. Thus was consummated the union between empiricism and visualization begun long ago during the Golden Age of Greece and completed in the Scottish Enlightenment. William Playfair's 1786 *Atlas* filled with spectacular and beautiful graphs of mostly economic data was not a natural outgrowth of what came before. In fact, we view Descartes' 1637 development of a coordinate system as an intellectual impediment that took a century and a half and Playfair's eclectic genius to overcome.

A wonderful example of Playfair's genius is his plot of England's national debt (Fig. 27.2) which is the first "skyrocketing debt" plot and exhibits many of Playfair's unique characteristics:

(a) It displays copious data, in this case England's national debt from 1688 until 1800.
(b) It uses a higher-than-wide aspect ratio to emphasize the skyrocketing.

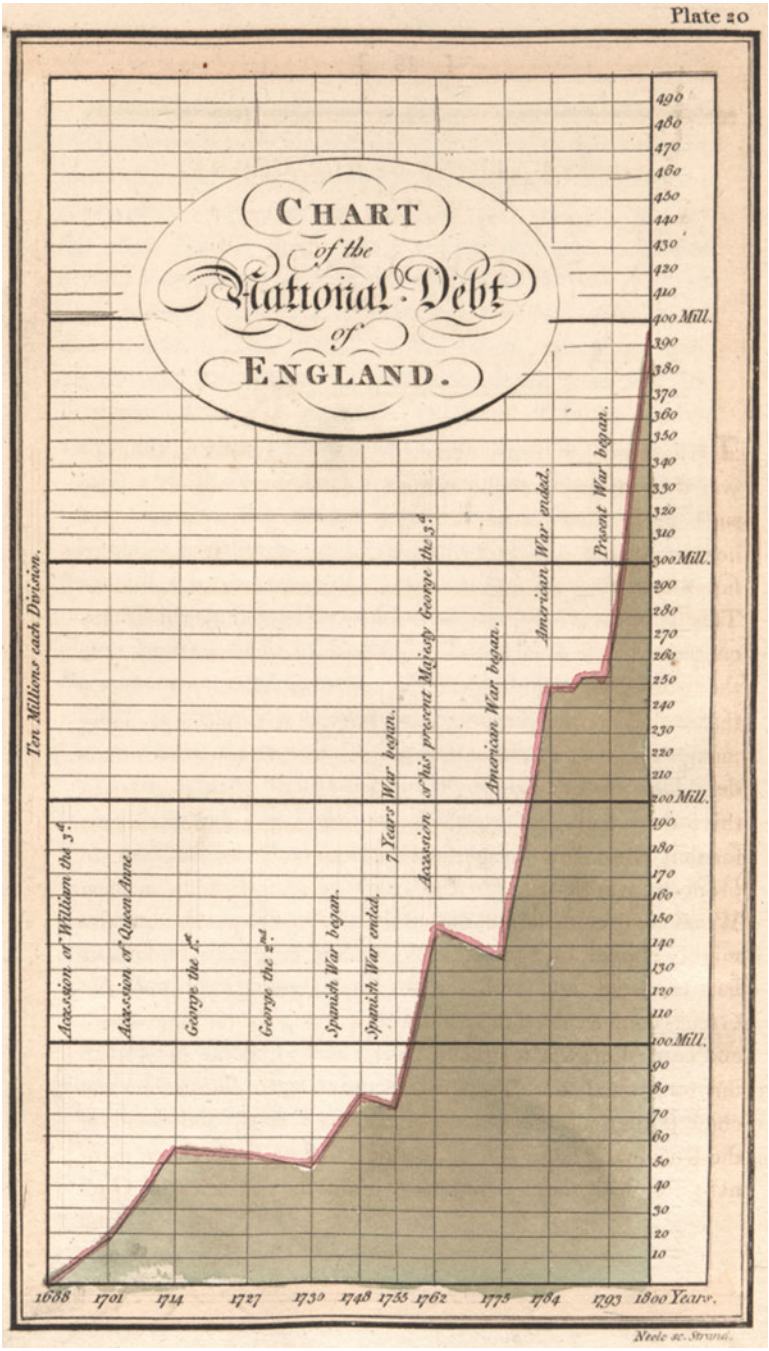


Fig. 27.2 Playfair's plot of England's national debt from 1688 to 1800 dramatically showing the adverse effect wars had on debt. Plate 20 (opposite page 85 in his 1801 Commercial Atlas)

(c) It has unevenly spaced years to be explicit about when events influential for the debt occurred.
(d) Those years are labeled with explanations (e.g., start of Seven Years' War).

27.4 Finding Unexpected Values

The plotting of real data had a remarkable, and largely unexpected, benefit. It forced the viewers to find what they hadn't expected. Thus was born the empirical modern approach to science that does not disdain the atheoretical plotting of data points with the goal of investigating suggestive patterns. Graphs that were in existence before Playfair (with some notable exceptions that we will discuss shortly) grew out of the same rationalist tradition that yielded Descartes' coordinate geometry – that is, the plotting of curves on the basis of an a priori mathematical expression (e.g., Orseme's "pipes" – discussed in Clagett, 1968).

Naked empiricism did not meet with universal approval. Luke Howard, a prolific grapher of data in the late eighteenth and early nineteenth century, as late as 1844, apologized for his methodology and referred to it as an "autograph of the curve . . . confessedly adapted rather to the use of the *dilettanti* in natural philosophy than that of regular students" (Howard, 1847, p. 38).

Now we can see the value of the grammatical metaphor that we introduced earlier, for it is accurate to think of early graphic displays as nouns, indeed common nouns that were used to depict some theoretical relationship. Thus we can conceive of the first major revolution in the use of graphic display in science as a shift from its use as a common noun (e.g., the theoretical relationship between supply and demand) to that of a proper noun (e.g., England's imports and exports from 1700 to 1800). This revolution seems to have begun in 1665 with the invention of the barometer, which inspired Robert Plot to record the barometric pressure in Oxford every day of 1684 and summarize his findings in a remarkably contemporary graph that he called a "History of the Weather" (Fig. 27.3) (Wainer, 2005, P. 14).

He sent a copy of this graph with a letter to Martin Lister in 1685 with a prophetic insight on the eventual use:

> For when once we have procured fit persons enough to make the same Observations in many foreign and remote parts, how the winds stood in each, at the same time, we shall then be enabled with some grounds to examine, not only the coastings, breadth, and bounds of the winds themselves, but of the weather they bring with them; and probably in time thereby learn, to be forewarned certainly, of divers emergencies (such as heats, colds, dearths, plague, and other epidemical distempers) which are not unaccountable to us; and by their causes be instructed for prevention, or remedies . . . we shall certainly obtain more real and useful knowledge in matters in a few years, then we have yet arrived to, in many centuries (Wainer, 2005, P. 15).[2]

[2]Plot's proposed method of crowd-sourcing weather data and his assessment of its potential value would later bear great fruit in Francis Galton's (1863) spectacular discovery of weather patterns in the northern hemisphere.

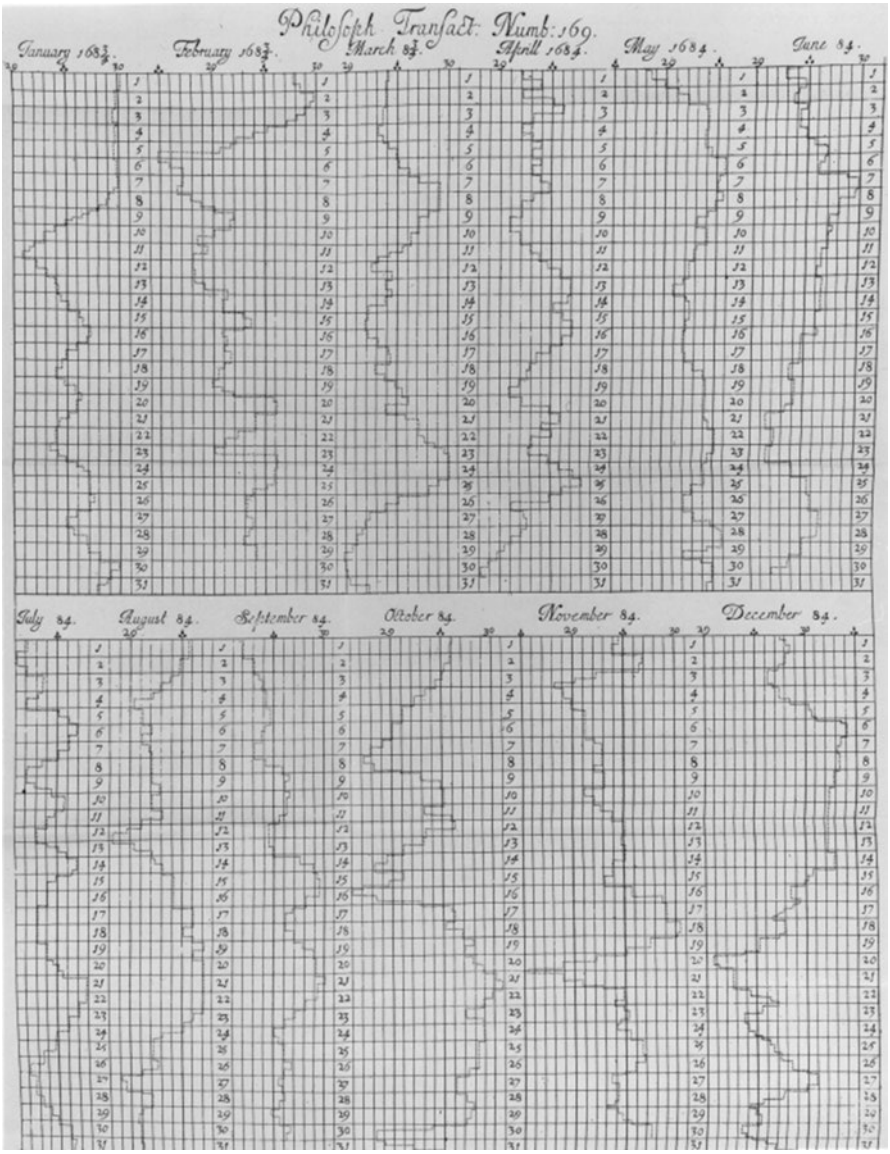


Fig. 27.3 Robert Plot's (1685) "History of the Weather" recording of the daily barometric pressure in Oxford for the year 1684

Plot and Lister's use of graphic display was scooped by the seventeenth century Dutch polymath Christiaan Huygens (1629–1693). On October 30, 1669, Christiaan's brother Lodewijk sent him a letter containing some interpolations of life expectancy data taken from John Graunt's 1662 book the *Natural and Political*

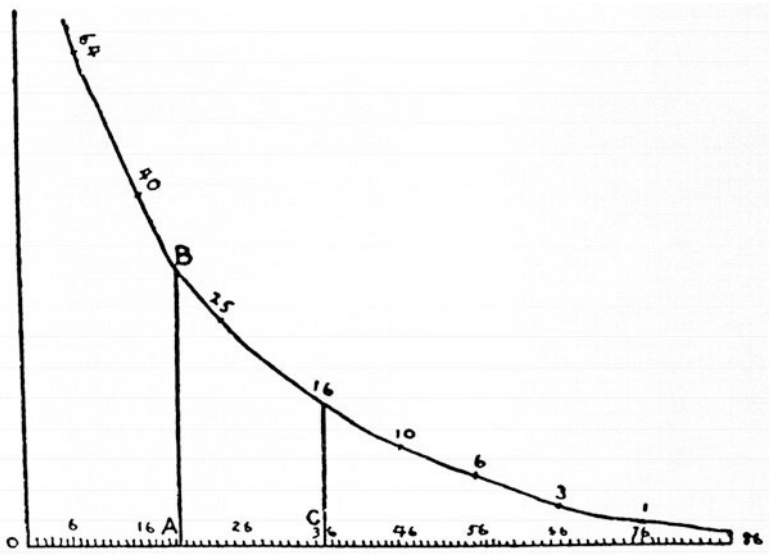


Fig. 27.4 Christian Huygens's 1669 curve showing how many people out of a 100 survive between the ages of infancy and 86. (The data are taken from John Graunt's *Natural and Political Observations on the Bills of Mortality, 1662*)

Observations on the London Bills of Mortality. Christiaan responded in letters dated November 21 and 28, 1669, with graphs of those interpolations. Figure 27.4 contains one of those graphs showing age on the horizontal axis and number of survivors of the original birth cohort on the vertical axis. The curve drawn was fitted to his brother's interpolations. The letters on the chart are related to an associated discussion on how to construct a life expectancy chart from this one – that is, analyzing a set of data to yield deeper insights into the subject. Christiaan constructed such a chart and indicated that it was more interesting from a scientific point of view; the alternative, he felt, was more helpful in wagering.

There were a smattering of other examples of empirically based graphs that appeared in the century between Huygen's letter and the 1786 publication of Playfair's *Commercial and Political Atlas*, for although some graphic forms were available before Playfair, they were rarely used to plot empirical information. In 1978, Albert Biderman argued that this was because there was an antipathy toward that as a scientific approach. This suggestion was supported by such statements as that made by Luke Howard. But at least sometimes, when data were available (e.g., Graunt's survival data, Plot's weather data, and several other admirable uses), they were plotted. Perhaps part of the exponential increase in the use of graphics since the beginning of the nineteenth century is merely concomitant to the exponential growth in the availability of data. Of course there might also be a symbiosis in that the availability of graphic devices for analyzing data encouraged data gathering. For whatever the reasons, Playfair was at the cusp of an explosion in data gathering, and his graphic efforts appear causal. He played an important role in that explosion.

The consensus of scholars, well phrased by Costigan-Eaves and Macdonald-Ross (in progress) (in their oft-cited, but as yet, unpublished manuscript), is that until Playfair "many of the graphic devices used were the result of a formal and highly deductive science . . . This world view was more comfortable with an arm-chair, rationalistic approach to problem-solving which usually culminated in elegant mathematical principles" often paired with elegant geometrical diagrams. The empirical approach to problem-solving, a critical driving force for data collection, was slow to get started. But the empirical approach began to demonstrate remarkable success in solving problems, and with improved communications, the news of these successes, and hence the popularity of the associated graphic tools, began to spread quickly.

We are accustomed to intellectual diffusion taking place from the natural and physical sciences into the social sciences; certainly that is the direction taken for both calculus and the scientific method. But statistical graphics in particular and statistics in general went the reverse route. Although, as we have seen, there were applications of databased graphics in the natural sciences, it was only after Playfair applied them within the social sciences that their popularity began to accelerate. Playfair should be credited with producing the first chart book of social statistics; indeed, publishing an *Atlas* that contained not a single map is one indication of his belief in the methodology (to say nothing of his chutzpah). Playfair's work was immediately admired, but emulation, at least in Britain, took a little longer (graphic use started up on the continent a bit sooner). Interestingly, one of Playfair's earliest emulators was the banker S. Tertius Galton (the father of Francis Galton, and hence the biological grandfather of modern statistics) who, in 1813, published a multiline time series chart of the money in circulation, rates of foreign exchange, and prices of bullion and of wheat.[3] The relatively slower diffusion of the graphical method back into the natural sciences provides additional support for the hypothesized bias against empiricism there. The newer social sciences, having no such tradition and faced with both problems to solve and relevant data, were quicker to see the potential of Playfair's methods.

The Prodigal Brother

Playfair's graphical inventions and adaptations look contemporary. He invented the line graph and the pie chart to suit particular purposes. He invented the statistical bar chart out of desperation, because he lacked the time series data required to draw a line showing the trade with Scotland and so used bars to symbolize the cross-sectional character of the data he did have. Playfair acknowledged Priestley's (1765,

[3]Ironically, had Galton paid close enough attention to his own graphs he would have been able to foresee the financial crisis of 1831 that created a ruinous run on his own bank.

1769) priority in this form, although Priestly used thin horizontal bars to symbolize the life spans of historical figures in a time line (Fig. 27.5).

Playfair's role was crucial for several reasons. It was not for his development of the graphic recording of data; others preceded him in that. Indeed, in 1805, he points out that as a child his brother John had him keep a graphic record of temperature readings. But Playfair was in a remarkable position. Because of his close relationship with his brother and his connections with Watt, he was on the periphery of applied science. He was close enough to know of the value of the graphical method, but sufficiently detached in his own interests to apply them in a very different arena – that of economics and finance. These areas, then as now, tend to attract a larger audience than matters of science, and Playfair was adept at self-promotion.[4]

In a review of his 1786 *Atlas* that appeared in *The Political Herald*, Dr. Gilbert Stuart wrote,

> The new method in which accounts are stated in this work, has attracted very general notice. The propriety and expediency of all men, who have any interest in the nation, being acquainted with the general outlines, and the great facts relating to our commerce are unquestionable; and this is the most commodious, as well as accurate mode of effecting this object, that has hitherto been thought of . . . To each of his charts the author has added observations (which . . . in general are just and shrewd; and sometimes profound . . . Very considerable applause is certainly due to his invention; as a new, distinct, and easy mode of conveying information to statesmen and merchants (Playfair 1801/2005, P. 31) . . .

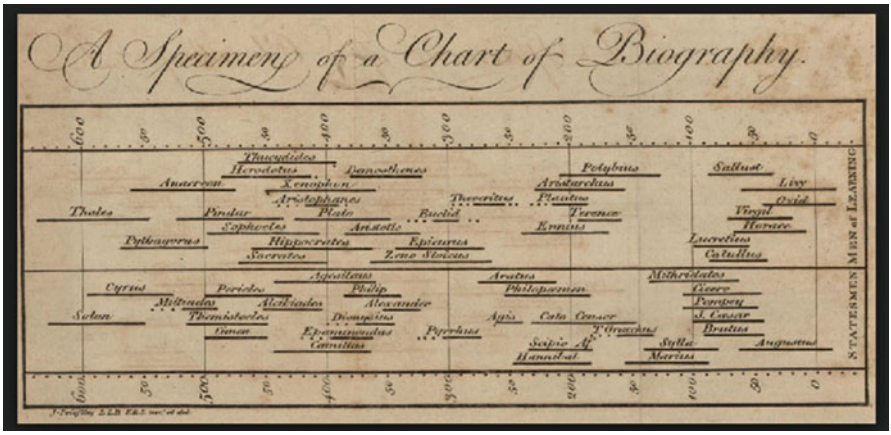


Fig. 27.5 Lifespans of 59 famous people in the 6 centuries before Christ (Wainer, 2005, P. 43)

[4]For more about the remarkable life and accomplishments of William Playfair (including the fascinating story of his attempted blackmail of Lord Archibald Douglas), the interested reader is referred to Spence and Wainer (1997, 2000), Wainer (1996), Wainer and Spence (1997), and, especially, Wainer and Spence's Introduction to Playfair (1801/2005).

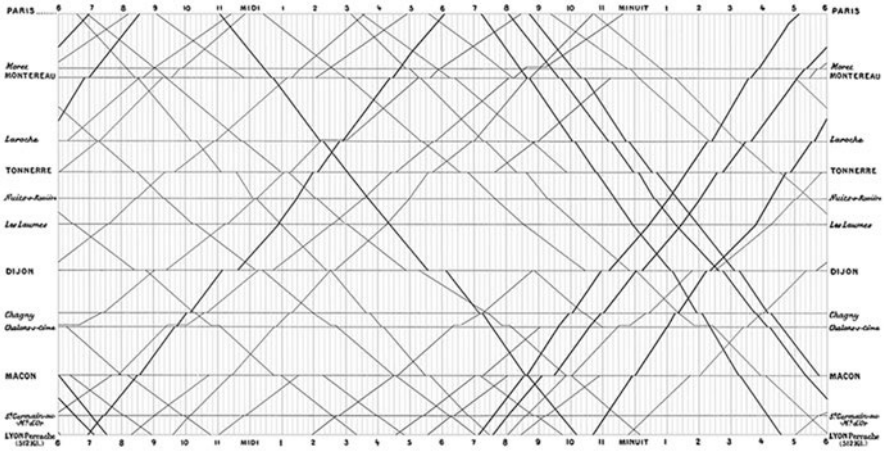


Fig. 27.6 Marey's graphical train schedule, showing all trains between Paris and Lyons each day (Wainer, 2005, P. 7)

Such wholehearted approval rarely greets any scientific development. Playfair's adaptation of graphic methods to matters of general interest provided an enormous boost to the popularity of statistical graphics.

The popularity of visualizations owes much to the almost religious fervor of scientists and epistemologists of the nineteenth century who sought to banish subjectivity from science. "The prophets, philosophers and preachers of objectivity drew on a number of techniques including inferential statistics, double-blind clinical trials, and self-registering instruments to hold subjectivity at bay."[5] But, as we have sketched, the oldest and most important of these was visualization.

By 1878, the French physiologist Etienne Marey, whose graphic schedule of all the trains between Paris and Lyons reproduced in Fig. 27.6 provides a powerful illustration of the breadth of value of this approach, expressed the feelings of most natural scientists of the value of graphical representation

> There is no doubt that graphical expression will soon replace all others whenever one has at hand a movement or change of state – in a word, any phenomenon. Born before science, language is often inappropriate to express exact measures or definite relations (Marey, 1878, p. iii).

Marey was also giving voice to the movement away from the sorts of subjectivity that had characterized prior science in support of the more modern drive toward objectivity. Although some cried out for the "insights of dialectic," "the power of arguments," and the "flowers of language" (All quotations on this page are drawn from Daston and Galison (1992), P. 86), their protestations were lost on Marey, who

[5]From page 17 in Lorraine J. Daston and Peter Galison's, marvelous 2007 book, *Objectivity* (Daston and Galison, 2007).

dreamed of a wordless science that spoke instead in high-speed photographs and mechanically generated curves – in images that were, as he put it, in the "language of the phenomena themselves."[6]

Historians have pointed out that "Let nature speak for itself" was the watchword of the new brand of scientific objectivity that emerged at the end of the nineteenth century. In their fascinating 1992 essay, Daston and Galison emphasize that "at issue was not only accuracy but morality as well: the all-too-human scientists must, as a matter of duty, restrain themselves from imposing their hopes, expectations, generalizations, aesthetics, and even their ordinary language on the image of nature" (p. 84). Mechanically produced graphic images would take over when human discipline failed. Marey and his contemporaries turned to mechanically produced images to eliminate human intervention between nature and representation. "They enlisted polygraphs, photographs, and a host of other devices in a near-fanatical effort to produce atlases – the bibles of the observational sciences" (p. 118) – documenting birds, fossils, human bodies, elementary particles, flowers, and economic and social trends that were certified free of human interference.

Daston and Galison conclude, "The problem for nineteenth-century atlas makers was not a mismatch between world and mind, as it had been for seventeenth-century epistemologists, but rather a struggle with inward temptation. The moral remedies sought were those of self-restraint: images mechanically reproduced and published, warts and all; texts so laconic that they threatened to disappear entirely. Seventeenth-century epistemology aspired to the viewpoint of angels; nineteenth-century objectivity aspired to the self-discipline of saints. The precise observations and measurements of nineteenth century science required taut concentration endlessly repeated. It was a vision of scientific work that glorifies the plodding reliability of the bourgeois rather than the moody brilliance of the genius[7]" (p. 118).

The graphic representation of scientific phenomena served two purposes. Their primary function was standardizing phenomena in visual form, but they also served the cause of publicity for the scientific community. They preserved what was ephemeral and distributed it to all who would purchase the volume, not just the lucky few who were in the right place at the right time with the right equipment. And, they served the cause of memory, for images are more vivid and indelible than words.

But the graphic display of natural phenomena was viewed as yet more. Marey, in an accompanying note to his design of a portable polygraph, which automatically registered a variety of measures, suggested that through the use of graphics scientists could reform the very essence of scientific research and scientific evidence. "The graphic method translates all these changes in activity of forces into an arresting

[6]Marey (1878, p. vi)

[7]Although with such contributors as Condorcet (1743–1794), von Humboldt (1769–1859), and Florence Nightingale (1820–1910), there was certainly room for genius in the eighteenth and nineteenth centuries. Indeed, Galton's weather maps, developed at the end of the nineteenth century, shows how plodding reliability when adjoined with moody brilliance can yield especially fruitful results, yet no one would doubt that Robert Plot was a plodding plotter.

form that one could call the language of the phenomena themselves, as it is superior to all other modes of expression" (Daston and Galison, 2007, p. iv). Such a language was, for Marey, universal in two senses. Graphical representation could cut across the artificial boundaries of natural languages to reveal nature to all people, and graphical representation could cut across disciplinary boundaries to capture phenomena as diverse as the pulse of a heart and the downturn of an economy. Pictures became more than merely helpful tools: they were the words of nature herself.[8]

The Wedding Banquet

Yet something even more remarkable occurred among the wedding planners in the latter part of the nineteenth century, as many forces combined to produce the perfect storm for data graphics we call the Golden Age. The table had been well set. Heaps of data on important societal issues (commerce, literacy, crime) had been ordered up; some statistical theory had been developed to allow the essential flavors to be extracted; technological advances in printing and reproduction now allowed serving a huge guest list. The guests were truly international, but they shared a common visual language and visual thinking.

Only one of the planners will be mentioned here: Charles Joseph Minard (1781—1870), a civil engineer in France and who later produced a now iconic[9] flow map depicting Napoleon's disastrous Russian campaign of 1812. Minard used the graphic method to design exquisitely beautiful thematic maps and diagrams showing all manners of topics of interest to the modern French state in the dawn of national concern for trade, commerce, and transportation: Where to build railroads? What happened to the production of cotton goods during the US Civil War (shown in Fig. 27.7)?

By the end of the nineteenth century, guests from the USA (Francis Walker in the Census Bureau), France (Émile Cheysson in the Ministry of Public Works), and others in Germany, Sweden, and elsewhere began to send their gifts to the happy couple – elaborate and detailed statistical albums tracing and celebrating their nation's achievements and aspirations – and decked out in the fancy colors and styles of what became the language of graphics.

We have gone beyond merely tracing the history of the bride and groom in the marriage of empiricism and visualization to also include some snapshots of the wedding, the honeymoon, and of a fair number of anniversaries attended by the

[8]This simple? Perhaps not. An alternative thesis to the one that characterizes science's task as capturing the glorious revelations by nature of her sublime design is one that sees humans imposing the order of their senses and their arts upon the unheavenly disorder they find themselves amidst.

[9]Marey (1878) first called attention to this work, saying it "defies the pen of the historian in its brutal eloquence." Tufte (1983) later bestowed the title of "the best statistical graphic ever drawn."

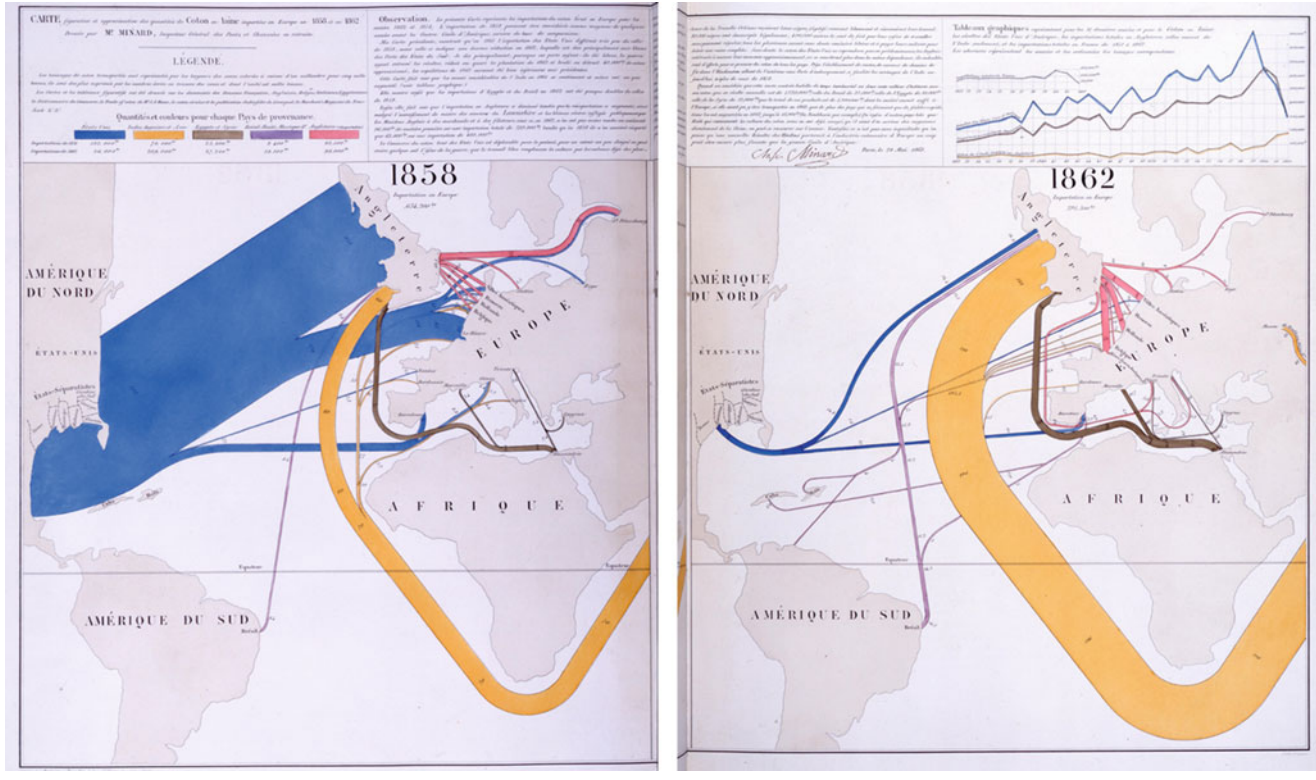


Fig. 27.7 Minard's before-after flow maps showing how the supply of cotton to English mills shifted dramatically from the southern states of the USA to India during the US Civil War, when the Union's naval blockades made reliable Confederate shipping problematic (Rendgen, 2019, P. 98–99)

many descendants of the initial pairing. Our goal was to provide a foreshadowing of the beauty and accomplishments that issued from this union.

For those who would value a fuller elaboration, we immodestly refer you to Friendly and Wainer (2021) from which this chapter has been abstracted.

References

Apel, W. (1944). The notation of polyphonic music. Cambridge, MA: The Mediaeval Academy of America.

Bacon, Roger (1897) Bridges, John Henry (ed.), The 'Opus Majus' of Roger Bacon: Edited, with Introduction and Analytical Table (in Latin and English), Vols. I & II, Oxford: Horace Hart for the Clarendon Press.

Beniger, J. R. & Robyn, D. L. (1978). Quantitative graphics in statistics: A brief history. *The American Statistician,* 32*,* 1–10.

Biderman, A.D. (1978) *Intellectual impediments to the development and diffusion of statistical graphics, 1637-1980*. Presented at 1st General Conference Social Graphics, Leesburg, VA.

Biderman, A. D. (1990). The Playfair enigma: Toward understanding the development of schematic representation of statistics from origins to the present day. *Information Design Journal, 6(*1*),* 3–25.

Clagett, M. (1968). Nicole Oresme and the Medieval Geometry of Qualities and Motions. Madison, WI: University of Wisconsin Press.

Costigan-Eaves, P., Macdonald-Ross, M. (in progress). The Method of Curves: a brief history to the early nineteenth century. Unpublished manuscript. Milton Keynes, England: The Open University.

Daston, L. & Galison, P. (1992). *Representations,* 40*,* 81–128.

Daston, L. & Galison, P. (2007). *Objectivity*. New York: Zone Books.

Friendly, M. (2002). Vision and re-visions of Charles Joseph Minard, *Journal of Educational and Behavioral Statistics*. 27 (1), 31–52.

Friendly, M. & Wainer, H. (2021). *A History of Data Visualization and Graphic Communication*, Cambridge: Harvard University Press.

Funkhouser, H. G. (1937). Historical development of the graphic representation of statistical data. *Osiris,* 3*,* 269–404.

Galton, F. (1863). *Meteorographica, or methods of mapping the weather*. London: Macmillan.

Graunt, J. (1662). *Natural and Political Observations on the London Bills of Mortality*, Martyn: London.

Gunther, R.T. (1968). *Early science in Oxford, Vol. XIII Dr. Plot and the correspondence of the Philosophical Society of Oxford*. London: Dawsons of Pall Mall.

Howard, L. (1847). *Barometrigraphia: twenty years' variation of the barometer in the climate of Britain, Exhibited in autographic curves, with the attendant winds and weather, and copious notes illustrative of the subject.* London: Richard & John E. Taylor.

Huygens, C. (1895). *Oeuvres completes, Tome Sixieme Correspondance*, 515–518, 526–539. Den Hague: Martinus Nijhoff.

Marey, E. J. (1878). *La Méthode graphique dans les sciences expérimentales et particuliérement en physiologie et en médecine.* Paris.

Marey, E. J. (1885), *La Méthode Graphique,* Paris: Boulevard Saint Germain et rue de l'Eperon.

McKie, D. (1972). "Scientific societies to the end of the eighteenth century." In A. Ferguson (Ed.) in *Natural philosophy through the 18th century and allied topics,* (p. 133–143). London: Taylor & Francis.

Playfair, W. (1786). *The Commercial and Political Atlas.* London: Corry.

Playfair, W.: *The Commercial and Political Atlas, Representing, by means of Stained Copper-Plate Charts, The Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the whole of the Eighteenth Century.* Edited and introduced by Howard Wainer and Ian Spence. New York: Cambridge University Press, 2005 (1801/2005).

Priestley, J. (1769). *A New Chart of History*. London. Reprinted: 1792, New Haven: Amos Doolittle.

Rendgen, S. (2019). *The Minard System*. New York: Princeton Architectural Press.

Spence, I. & Wainer, H. (1997). William Playfair: A Daring Worthless Fellow, *Chance,* 10*(*1*),* 31–34.

Spence, I. & Wainer, H. (2000). William Playfair (1759-1823): An inventor and ardent advocate of statistical graphics. In *Statisticians of the Centuries* (C.C. Heyde, ed.). International Statistical Institute: Voorburg, Netherlands.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Wainer, H. (1997). *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot.* New York: Copernicus Books.

Wainer, H. (1998). The graphical inventions of Dubourg and Ferguson: Two precursors to William Playfair. *Chance,* 11*(*4*)*, 39–41.

Wainer, H. (1996). Why Playfair? *Chance,* 9*(*2*),* 43–52.

Wainer, H. (2005). *Graphic Discovery*. Princeton: Princeton University Press.

Wainer, H. & Spence, I. (1997). Who was Playfair? *Chance,* 10(1), 35–37.

Chapter 28
"Belizaire": The *Chance* Restaurant Critic

Albert Madansky

The 1967 correspondence exchange between Richard Condon (of Manchurian Candidate fame), Martin Shubik (Seymour H. Knox Professor of Economics at Yale), Irv Roshwalb (Senior Vice President of Audits & Surveys, Inc.), and me about pastrami was a file folder (dubbed "The Pastrami Papers") which was privately circulated among our friends. One of its readers, Jack Gould (Economics Professor at the University of Chicago Graduate School of Business), suggested that a composite of this correspondence would make a great article for the *University of Chicago Magazine* and so introduced me to its editor, Don Morris, who came up with the format for the presentation of the material and who redacted the material, crafting what was to become a classic paper. It appeared in the Spring 1976 issue of the magazine, "Notes on an experimental evaluation of four leading delicatessens using as criteria their corned beef and pastrami sannawhitch, in accordance with the axiom: The deli is the sannawhitch." The centerpiece of the paper was a taste test of pastrami and corned beef sandwiches from four Manhattan delis.

In the 1980s, the GSB Statistics Group had periodic lunches in which someone would lead a discussion on a statistical topic. I decided to bring to one of these lunches pastrami and corned beef sandwiches from the two leading Chicago purveyors of delicatessen meats, Hungarian Kosher Foods and Romanian Kosher Sausage Company, and replicate the Shubik-Madansky experiment at that lunch. The result of that experiment was a privately circulated report, "Hungarian and Romanian: Whose Wurst is Best?"

When Steve Fienberg was planning the first issue of *Chance*, he called me and asked if I would fashion an article suitable for statisticians about these taste tests, and thus was born my paper, "The Gentleman Tasting Pastrami." I came up with that title as a homage to R. A. Fisher's famous "Lady Tasting Tea" of his 1935

A. Madansky (✉)
Booth School of Business, University of Chicago, Chicago, IL, USA
e-mail: albert.madansky@chicagobooth.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_28

Fig. 28.1 Cover of the first issue of *Chance* magazine

Design of Experiments book (not to be confused with David Salsburg's book, *The Lady Tasting Tea*, which was published in 2001). Concerned that the casual reader of *Chance* would not get the allusion, Steve included material from Fisher's book as a boxed insert in the paper.

Little did I know that this paper would define the cover art for *Chance*, as shown in Fig. 28.1.

Steve described the genesis of this picture in Howard Wainer's 2007 interview of him and Bill Eddy ("Taking a *Chance*: An Interview with William F. Eddy and Stephen E. Fienberg, *Chance*, 20:4, 33–39): "Unfortunately, we couldn't reconvene the panel of testers, since we had no budget, and the cover was simply a pastrami sandwich with a can of Dr. Brown's soda."

The paper's popularity may have contributed to Steve's notion that *Chance* needed a restaurant column. As Steve put it, "When we were starting *Chance*, we tried to have the third issue come out in time for distribution at the Joint Statistical Meetings, which turned out to be in New Orleans. So, I thought we could make *Chance* a hot item if it contained a statistician's guide to dining in New Orleans. Neither of us had enough knowledge to write this, so we turned to my brother, Lorne, who at the time was living in Jackson, Mississippi, and would regularly visit New Orleans with my sister-in-law. He agreed and wrote a terrific guide, but he

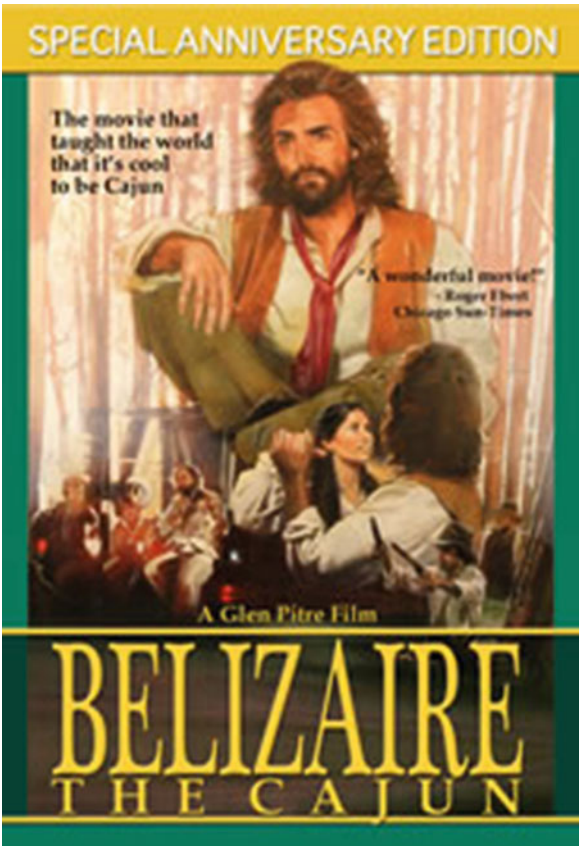


Fig. 28.2 *Belizaire the Cajun*

wasn't sure he wanted to be associated with statisticians, so he assumed the nom de plume Belizaire, after the title character in the memorable 1986 movie *Belizaire the Cajun*. The poster for the movie is shown in Fig. 28.2.

"The next year, when it came time for another dining column for JSM in Washington, DC, we asked my brother for a repeat performance, but he demurred. But, I knew Washington restaurants really well, so I decided to mimic my brother's style and write a new column, but this time with real statistical references and dining advice. And I used his nom de plume. Thus began my career as a dining critic. Bill collaborated on one of the columns and many friends offered advice, both on which restaurants to review and the content. Many years later, some of *Chance's* most faithful readers still didn't know that I was Belizaire and asked if he was going to revive his columns."

There were 12 Belizaire columns in *Chance*:

1. Way Down Yonder in New Orleans, 1988, 1:3 52–54
2. A Capital Place for Dining, 1989, 2:2 44–46
3. Gastronomie et Paris, 1990, 3:1 39–42

4. The Statistical Pleasures of California Cabernets, 1990, 3:2 48–50
5. Statistical Return to the Land of True Gastronomic Pleasures, 1991, 4:4 36–40
6. Toronto A World Serious City, 1993, 6: 1 38–40
7. A San Francisco Feast, 1993, 6:2 34–36
8. Temptations in Toronto Ethnic Variations, 1994, 7:2 41–44
9. Off to Orlando!, 1995, 8:2 26–27
10. Coming to Chicago, 1996, 9:2 27–33
11. Anaheim Appetite and Turkish Delight, 1997, 10:2 23–26
12. Virtual Dallas Dining, 1998, 11:2 31–33

One doesn't have to perform a Mosteller-Wallace-Federalist paper-like analysis of the Belizaire columns to determine which (one) was written by Lorne. Lorne's column on New Orleans lacked reference to statistics and statisticians and concentrated on 15 top restaurants in the area. Lorne did set the precedent for the closing of the Belizaire reviews—*bon appetit!* (Not being as enthusiastic as was Lorne, Steve ended all but one of his Belizaire columns with an exclamation-point-less *bon appetit*. That alone is enough to enable a discerning reader to distinguish between the two Belizaires.)

Steve's first Belizaire column covered Washington, D.C., and, though skewed to its top restaurants, did provide some tips for per diem constrained and budget-impaired statisticians. It was here that Steve introduced the Sam Greenhouse Award "for the best bargain meal" and included some "cheap eats" places in his list of 19 recommendations. (Those who do not know Sam Greenhouse can read about him in "Some Reflections on the Beginnings and Development of Statistics in 'Your Father's NIH'", *Statistical Science* 1997, 12:2, 82–87. As a government employee he was constrained at meetings by the government per diem reimbursement rules, and so was always looking for ways to not run afoul of those rules.)

To set a Bayesian spin on the Belizaire columns, the first two were "priors," giving gastronomical (and some oenological) guidance to attendees of the New Orleans and DC JSMs. The third column, though, was "posterior," namely, a trip report on Steve's meals while attending the Paris ISI meeting. Neither Steve nor any of the other statisticians he quoted engaged in a search for a Parisian restaurant worthy of a Sam Greenhouse Award. (To his credit, Steve eschewed Michelin three-star institutions and satisfied himself with a sampling of two-star bistros.)

Column 4 showed Steve at his schizophrenic best, where Belizaire, Steve, and a third party engage in a conversation about what to cover in preparation for the Anaheim JSM.

> Where are the meetings this year? I (Belizaire) asked somewhat gingerly. "Anaheim, California," was the answer. "You mean Disneyland and Knott's Berry Farm?" "Not quite what we had in mind," Steve Fienberg replied. "You're fairly creative; we're sure you'll find something interesting to write about!

This was merely a setup for Belizaire to report on a taste test in which he participated to ascertain whether one could distinguish between California and French cabernets. No doubt inspired by "The Gentleman Tasting Pastrami," this experiment exhibited the same flaws as those of its progenitor.

In the 11 October 1990 issue of "Nature" (Volume 347, pages 593–4), John A. Campbell of the Department of Computer Science, University College, London, had reviewed *Chance*, and in his review said, "But one can forgive much in a journal that carries a regular gastronomic column, provided that its columnist takes a more enterprising approach to selecting restaurants on his next visit to France." So Steve pulled a fast one in column 5; he chose not to go to the Cairo ISI meeting, but instead decided to go to France and publish another trip report on his meals there. Not knowing what Campbell meant by "more enterprising," I can't tell whether the restaurants Steve visited on this trip would have satisfied Campbell. (As a token gesture, he did, though, include two pictures of statisticians taking time off from the Cairo meeting to see the pyramids.)

Belizaire reported on two visits to Toronto. In one (column 6), he attended some World Series games and ran the gastronomic gamut from the Studio Café bistro in the Four Seasons hotel, through the informal Yves Bar and Grill, followed by three different hotdogs at the SkyDome stadium, to a Montreal "smoked meat" sandwich at the Pickle Barrel Restaurant. (Since I don't consider Montreal "smoked meat" as even mentionable in the same breath as pastrami, despite Belizaire's exhortation to me that I eat my heart out, I did not do so when I read this column.) In the other (column 8), he prepared the attendees of the Toronto JSM for the ethnic variety of eating possibilities to be found there. Belizaire recommended 19 restaurants, French, Italian, Greek, Chinese, and Jewish. The Sam Greenhouse Award went to a cafeteria, leaving me to wonder what criterion Belizaire uses to determine the award winner. Belizaire also promised to take me to the Centre Street Deli to taste their hand-sliced Roumanian-style smoked meat, but alas that excursion never took place. I don't remember seeing Steve at all at that meeting, and I ended up instead having lunch at three of the other delis in Toronto, all confirming the inferiority of "smoked meat" to pastrami.

A prelude to the San Francisco JSM was Belizaire's column 7. Thirteen restaurants were recommended, and one can see the bias in this review when one finds no Sam Greenhouse Award-winning restaurant. Similarly, column 9 was a prelude to the Orlando JSM, again Belizaire recommended 13 restaurants (6 of which are 35 minutes or more away from Orlando), and again one finds no Sam Greenhouse Award winner. One does, though, get a calibration of Belizaire's palate, in that he uses the Parisian restaurant Lucas Carton as his standard of comparison for the French restaurants he recommends. I note that Lucas Carton is a mere one-star Michelin restaurant, so Belizaire's was not holding French restaurants to the standard of the ten Parisian restaurants that achieved ❁ ❁ ❁.

It wasn't until column 10 that Belizaire gave a clue to his identity. "Belizaire is the *nom de plume* of a well-known statistician with a penchant for gourmet dining and an interest in sports statistics. He writes occasional articles for *Chance* on food tasting and dining, especially in connection with statistical meetings, with the input from friends and colleagues." Steve outdid himself, listing five Italian, five Greek, four Indian, six Thai, four Chinese, and two Vietnamese restaurants along with his usual collection of super-expensive places for fine dining (mostly French). He even came up with four restaurants as candidates for the Sam Greenhouse Award and

had a hard time coming up with a winner. In total Belizaire reviewed 45 Chicago restaurants for this column.

As this was a column about Chicago, Steve consulted me about deli choices (after having eaten a miserable sandwich at Mrs. Levy's Delicatessen, which deservedly closed in February 2006). I read him an excerpt from The Pastrami Papers describing the state of Chicago deli (which he duly quoted in his column), misquoted me on recommended cities (I only recommended New York; Steve, being a confused Canadian brought up on "smoked meat," added Montreal and Toronto to the list), and included my sole Chicago recommendation, Manny's. (At the last Chicago JSM, Michael Cohen, Ed George, and I went to Manny's to continue the tradition of hitting at least one deli during JSM.)

Two for the price of one was the theme of column 11. Belizaire covered cuisine in both Anaheim (JSM site, neglected in his column 4) and Istanbul (ISI site). The recommendation for Anaheim was to go to Santa Ana, Costa Mesa, Laguna Beach, or Newport Beach for a good meal (with six designated as $$$$). If, however, you wanted to eat at the Sam Greenhouse Award restaurant (Belizaire revealed his criterion for this award in this column—a good wholesome meal for under $10; see also column 12), you could do so by staying in Anaheim. As for Istanbul, Belizaire made some recommendations for the old city, the new city, and along the shores of the Bosporus; he only recommended Turkish restaurants, and only one was designated $$$$. (Steve and I met up in Istanbul at the ISI meeting and went to one of his recommended restaurants along the Bosporus shore for dinner to update his prior; our posteriors peaked that evening.)

The swan song for Belizaire was column 12, about Dallas. Here he recommended 19 restaurants, of which 4 were $$$$ ("second mortgage territory") and 2 were tied for the Sam Greenhouse Award (whose cap had increased to $15). Maybe this was prescience, but one can tell that it was to be Belizaire's last column; it ended with *bon appetit!*

In his column 11. Belizaire refers to himself as "*Chance's* restaurant *bon vivant*." The Cambridge Dictionary defines a *bon vivant* as a person who enjoys good food and wines and likes going to restaurants and parties. As can be seen from these columns, Belizaire's self-description was perfect. He is missed, and not just from the pages of *Chance*.

Chapter 29
Reminiscences of Steve

Judith M. Tanur

I met Steve in 1976 when he was teaching a short course at the JSM on Discrete Multivariate Analysis with Yvonne Bishop and Paul Holland – though Steve insists that we were together several years earlier – in the late 1960s at a planning meeting for what later became the book *Statistics: A Guide to the Unknown* (SAGTU). I have no recollection of meeting him there, but then I was so dazzled at being at that meeting at the invitation of Fred Mosteller that I probably would not have registered the appearance of anything short of a lightning strike.

In some sense, Steve and I shared a mentor – Fred was Steve's dissertation advisor and role model; for me he was the making of my career when he invited me to edit SAGTU. The later part of my long-term collaboration with Steve was the editing (with the collaboration of Bill Kruskal and Dave Hoaglin) of a 1990 Festshrift for Fred entitled *A Statistical Model: Frederick Mosteller's Contributions to Statistics, Science, and Public Policy*. And after Fred's death, we worked with Dave Hoaglin again to put Fred's unfinished autobiography into shape for publication, adding a chapter to bring it up to date because Fred had abandoned the effort in the 1980s.

Another mentor we shared was Bill Kruskal. Bill was in the habit of sending clippings or Xeroxes from his reading to a long list of correspondents. I was on that list because I had served as Bill's assistant in editing the statistics material in the *International Encyclopedia of the Social Sciences*. But Steve was not only on the list, he spent several years in the Statistics Department at Chicago with Bill and credits Bill with piquing his interest in surveys by plying him with clippings from the *Chicago Sun-Times* about its straw poll, leading to Steve's first sole-authored paper in JASA.

J. M. Tanur (✉)
Distinguished Teaching Professor Emerita, Department of Sociology, State University of New York Stony Brook, Montauk, NY, USA
e-mail: Judith.Tanur@stonybrook.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_29

So as I got to know Steve in the late 1970s, we had quite a bit in common. And when I was asked by the Social Science Research Council to write a piece on Advances in Large-Scale Surveys and Experiments for the NSF *5-year Outlook on Science and Technology* and learned that Steve (along with Bill Kruskal, Norman Bradburn, and Richard Berk) was on the advisory committee who would guide me through the production of that work, I was delighted to get to know him better. And it was from this project that our collaboration on the parallels between surveys and experiments and the embedding of experiments within surveys began.

Steve and his family started visiting every summer at our summer home in Montauk. Steve always brought a big stack of references that we would need and an ambitious schedule of work that we would accomplish. We did always accomplish a good deal, but never quite as much as we had expected or hoped. Things kept getting in the way, like a Jeep ride on the beach – or a ride in the boat or some water skiing or a lobster that had to be consumed or a tray of sashimi or a glass of good scotch.

Let me tell you a little more about those visits. Often Miron Straf and Carolee Bush would join us. That would mean six adults, and however many of Fienberg sons and Tanur daughters were part of that year's party and the Fienberg golden retriever, Princess. And any number of our Vizsla dogs – together with however many dogs, my daughters had parked with us, all in a three-room-one-bathroom house. There are lots of pictures of people asleep in the most unlikely places.

There are classic stories about those visits that Steve and I would repeat to whomever would listen. One day, as we often did, Steve and I went to the fish store to forage for dinner and brought home some soft shell crabs. Although both of us had enjoyed soft shell crabs in restaurants, neither of us – nor any other member of the house party – had ever cooked one. But my cookbook gave interesting recipes – including graphic and rather disgusting instructions about how to clean them. When dinner time approached, we each had a drink. And then another and still another. Finally we steeled ourselves for the messy task – we laid out knives, covered the kitchen counter with newspapers, took another drink, and unwrapped the crabs. What a relief – the fishmonger had already cleaned them for us.

Then there was the Labor Day weekend when we all took a ride on the beach in our Jeep. As we drove along, admiring the ocean and staying out of the way of the people playing volleyball, somehow one of the front wheels of the Jeep drove into a deep hole in the sand, so deep, in fact, that it was impossible to move the car either backward or forward. It was a holiday – no chance to get someone to tow us out of our predicament. It was Steve's idea to approach the young men playing volleyball to ask for help – they responded by descending en masse on the Jeep and simply lifting the front wheel out of the hole and placing it on solid sand. Luckily there was a keg of beer available to offer as a thank you.

In 2014, Steve and Joyce visited early in the spring, and since it was too cold to swim, we decided to take a Jeep ride along the Sound. As we were driving along the beach, we ventured on some rocks. We slipped sideways, right to the water line and found ourselves wedged between two enormous boulders. We could go neither backward nor forward and burned out the clutch (and probably the engine as well) trying to move. The tide was coming in and evening was coming on. The

adventure was becoming less fun every minute. Luckily, we had cell phones with us and sufficient reception to call a local service station. After what seemed like hours delay – while we all sat in the Jeep and wondered what would become of us – the tow truck arrived, having driven some five miles on the beach to get to us. The driver could not get in front of the Jeep or directly behind it and so had to leave the beach to return from the other end – and he was willing to bring the Fienbergs and me back to the house while my husband waited with the Jeep. Some hours later my husband joined us for dinner – by which time we were several drinks ahead of him. He had been towed for many miles – and it took two tow trucks to do it – and the Jeep was a total loss. The secret background of this adventure was that I had long wanted a four-door Jeep as I was getting too old to crawl into the back of the two-door model – and we were able to get a four-door with the insurance compensation we got for what our insurance company insisted on calling our "collision" with the rocks.

Unfortunately, that visit in Spring, 2014, was Steve's last to Montauk; he was making plans to visit again in mid-2016, but those plans never came to fruition before he died that winter. We finally prevailed upon Joyce to visit again in October, 2018 – she spent a long weekend with us from the 19th to the 22nd and really seemed to be enjoying herself, though clearly still missing Steve. On Saturday, October 27, she was killed in the Tree of Life Synagogue shooting.

Chapter 30
A Conversation with Stephen E. Fienberg

Miron L. Straf and Judith M. Tanur

MS: So, Steve, how is it that you came to become a statistician?

SF: It's actually a long story, because when I was in high school and entering university, I didn't even know that there was such a field. I was good at mathematics and I went to the University of Toronto, which was in my hometown—that's where the best students went if they could get in. I enrolled in a course called *Mathematics*, *Physics*, and *and Chemistry*. It was one of the elite courses at U of T, and during the first year, as I went through my chemistry labs, I never succeeded in getting the right result when I mixed the chemicals up in the beakers; I realized chemistry wasn't for me, and so the second year I did only math and physics. Then there were the physics labs, and I could never quite get the apparatus to work properly to get what I knew was the correct answer. I still got an *A* in the physics lab, because I could start with the result and work backward and figure out what the settings were and things

This is an electronic reprint of the original article published by the Institute of Mathematical Statistics in *Statistical Science*, 2013, Vol. 28, No. 3, 447–463. Reprinted with the permission of the Institute of Mathematical Statistics.

M. L. Straf *(emeritus)*
National Academies of Sciences, Engineering, and Medicine, Washington, DC, USA
e-mail: mstraf@nas.edu

J. M. Tanur
Department of Sociology, State University of New York Stony Brook, Montauk, NY, USA
e-mail: Judith.Tanur@stonybrook.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_30

Fig. 30.1 Miron Straf, Steve Fienberg and Judy Tanur at the University of Connecticut, October, 2009

like that; but it was clear to me that physics wasn't for me as a consequence. So that left me with mathematics, and it was in the second year that we had a course in probability. So I was being gently introduced to statistical ideas. Then in my third year there was a course in statistics that was taught by Don Fraser, and he was terrific. His course was a revelation, because I didn't know anything about statistics coming in. Don followed the material in his *Introduction to Statistics* book and he began with probability theory and he brought into play geometric thinking throughout. When he got to inference, it was like magic. Of course, in those days Don did what was called "fiducial inference"—he called it "invariance theory" and later "structural inference"—where you went suddenly from probability statements about potential observables given parameters to probability statements about the data. I recall the old cartoon by Sydney Harris that people like to reproduce of the two scientists pointing to a blackboard full of equations, and one of them points to an equal sign and says, "And a miracle suddenly occurs here." That's sort of what happened in Don's class. He was a great lecturer, he was friendly with the students, and it was very clear that statistics was a really neat thing to do. Thus, in my fourth year, I took three classes involving statistics and probability and then applied to graduate school in statistics. The rest, as they say, is history (Figs. 30.1, 30.2, 30.3, and 30.4).

MS: So it was mathematics by elimination and statistics by revelation. Let's go back a bit. When did you discover that you had an aptitude for mathematics and statistics? In elementary school? Or high school?

Fig. 30.2 Steve as a Toddler in 1940s in Toronto

Fig. 30.3 Steve at Camp Tamarack, near Bracebridge, Ontario in 1952

Fig. 30.4 Steve with Don Fraser and Nancy Reid at a conference on the occasion of Don's 75th birthday, June 2000

SF: Not at all. In those days statistics never showed its face in the K-12 curriculum—this was before *Continental Classroom*.[1] Actually, it was K-13 in Toronto where I was born and raised. They got rid of grade 13 only decades after I was in school. At any rate, although my mother thought I was genius—don't all mothers think that about their children—I don't have any memory of being anything other than just a good student. I was very good at what passed for mathematics, but even through high school I don't think I was truly exceptional, and, besides, we did pretty elementary stuff—algebra, Euclidean geometry, and then in grade 13 we had trigonometry. As I reflect on those days, I was good at mathematics, but certainly not precocious and I only took standard high school math and with a heavy component of rote and repetition. By the time I got to grade 13 I was at the top of my class, however, and in the province-wide exams at the end of the year I was No. 2 in my school. But I also played oboe in the orchestra and band, and drums in the marching band, as well as participating in several other extra-curricular activities. So math wasn't much of a preoccupation and I didn't know what statistics and probability were all about at all.

JT: So that explains your broad early work in math, physics, and chemistry as a kind of omnibus course rather than going directly into math or statistics. So after your undergraduate work at the University of Toronto, you applied to graduate school; where did you apply and where did you end up going?

SF: Well, at the University of Toronto there had actually been many people to go into Statistics from MP&C. Don Fraser was perhaps the first, but then there were Ralph Wormleighton, Art Dempster, and David Brillinger—they all went, by the

[1] *Continental Classroom* was a series of television "course" broadcast by NBC on a variety of college-level topics in the early 1960s. Fred Mosteller taught the course on Probability and Statistics during 1960–1961.

Fig. 30.5 Graduation portrait from the University of Toronto, 1964

way, to Princeton. The year before me there was John Chambers, and John had gone to Harvard. I knew John pretty well, and I asked him how it was at Harvard. He seemed pleased with what he was doing and I did apply to Harvard and was admitted. I also applied to Princeton, and in their wisdom they didn't think that I should carry on the tradition from the University of Toronto, and that made the decision easier for me (Figs. 30.5 and 30.6).

MS: Were you disappointed about not being admitted to Princeton?

SF: Clearly at the time I was. This was my first rejection, and it prepared me in a way for what was to come when I submitted papers for publication to major journals! But Sam Wilks, who was the key person at Princeton with whom I had hoped to work, died in the spring of 1964, before I would have arrived.

JT: By the time you went to Harvard you were already married, is that right?

SF: No, I had met my wife Joyce at the University of Toronto when we were both undergraduates. I was actually working in the fall of 1963 in the registrar's office, and on the first day the office opened to enroll people, Joyce came through. And one of the benefits about working in the registrar's office, besides earning some spending money, was meeting all these beautiful women students passing through. That first day I made a note to ask Joyce out on a date. The next day she came through again, this time bringing through another young woman who turned out to be the daughter of friends of her parents. And I thought this was a little suspicious, but auspicious in the sense that maybe I would succeed in getting a date when I asked her. And the next day, she came through again! This time with her cousin! Then I knew that this was really going to work out. And it did. We got engaged at the end of the summer

Fig. 30.6 Joyce and Steve in Portugal for a conference on privacy and confidentiality, 1998

of 1964 after I graduated, but we weren't married when I went away to graduate school. In fact, yesterday I was talking to one of the students at the University of Connecticut who was a little concerned about graduate school; it was wearing her down, and I told her I almost left after the first semester because I wasn't sure if I was going to make a go of it, in part because I was lonely. But I did survive, and Joyce came at the end of the first year; we got married right after classes ended, and we've been together ever since.

MS: And where were your children born?

SF: Ah, *conceived* in various places, born in others. We believe that Anthony, my older son, was actually conceived in Scotland, on the vacation we took just after I graduated from Harvard. He was born in Chicago, where I had my first academic appointment, and, indeed, as we traveled across the country, from Boston to Chicago, Joyce began experiencing morning sickness (all day long), which didn't make for such a great trip. Then Howard was born in Minnesota just after we had moved there and I had joined the University of Minnesota faculty.

JT: Tell us more about what happened when you first arrived at Harvard.

SF: Well, one of the reasons I went to Harvard is that they not only gave me a fellowship, but also a research assistantship to work with Fred Mosteller. The day after I arrived, I went into the department because I didn't quite know what a research assistant did, and I went to see Fred (at the time he was Professor Mosteller, of course—I didn't learn to call him Fred until later). Fred was busy, but his assistant, Cleo Youtz, said he would like to have lunch with me. So I came back for lunch, and we went to the Harvard Faculty Club. Fred was being very courteous, and he suggested I order the horse steak, a special item on the faculty club menu at the time. And the horse steak came—I'm not sure if you've had horse steak—it's not *quite* like the kinds of steaks we normally order, it's a *little* bit tougher. I cut my first piece of horse steak, I put it in my mouth and started to chew. And

Fig. 30.7 Steve dining with Fred Mosteller at ISI meetings in Paris, 1989

then Fred began to describe this problem to me. It was about assessing probability assessors. I didn't understand a thing, and he's talking away, and I'm chewing away. Then Fred asked me a question, and I'm chewing away. At this point, he pulled an envelope out of his pocket and on the back of it there were these scribbles. He handed it to me, and I'm *still* chewing because you really can't eat horse steak except in very small bites. It turned out that the scribbles were notes from John Tukey about this problem. In fact, this was a problem that John and Fred were working on for some larger project, and my job was to translate the chicken-scratches on the back of the envelope into something intelligible, when I didn't know anything about what was going on. I worked at it for a while, and then Fred slowly told me what John's jottings meant, and the key idea was that for assessing probability forecasts, you have to look not just at the equivalent of means, or the bias in them (known technically as calibration), but also at the equivalent of variability (how spread out the forecasts are). Actually, that was a very important lesson, although I didn't have any clue about it in my first months at Harvard.

Over the course of my first fall at Harvard, I discovered a paperback book called *The Scientist Speculates*: *An Anthology of Partially Baked Ideas*, edited by Jack Good, with whose work I later became very familiar. In it was a short essay by Bruno de Finetti on assessing probability assessors, and de Finetti's ideas went into the technical report I wrote up on the topic with Fred and John. Fifteen years later, at the Valencia I Bayesian meeting, Morrie DeGroot and I began to work on the problem and ultimately wrote three papers on the topic of calibration and refinement of probability forecasters, heavily influenced by that first research exercise with Fred (Fig. 30.7).

MS: I wanted you to talk about Fred. Fred has been a very influential person in your career, and not just during your thesis. Maybe you want to tell us a little bit more about how he influenced your life and also how you came to go from Harvard to Chicago.

SF: Well, during that first year I worked on several problems with Fred and I wrote up some memos, but they never quite moved into papers at the time. Fred was pretty busy, and I got interested in Bayesian inference and multivariate analysis. I had begun to take an interest in Bayesian methods, having participated as a first year student in a seminar across the river at the business school run by Howard Raiffa and Bob Schlaiffer. At the time, Art Dempster was the person who seemed to be most involved in these Bayesian things and multivariate analysis, so I began to meet with him. In the process of working with Art, I met George Tiao, who was visiting the Business School with George Box for the year. As a consequence, George and I wrote a paper together on Bayesian estimation of latent roots and vectors, but it just didn't look like it was going to be a thesis problem.

The next summer, Fred ran into me in the hall and said he had some problems that I might like to work on. Fred had become deeply involved in the National Halothane Study at the NRC and, unlike most NRC studies, he and others—Tukey, John Gilbert, Lincoln Moses, Yvonne Bishop, to name a few—were actually analyzing data and creating new methods as they went along. The data essentially formed a giant contingency table and Fred got me working on a few different problems that ultimately came together as the core of my thesis. In the process I collaborated on separate aspects of the work with John Gilbert, Yvonne Bishop, and Paul Holland. I did most of the work in 1967 and that was the summer of "The Impossible Dream," when the Boston Red Sox won the pennant. I would work into the wee hours and go to Fenway Park and sit in the bleachers for the afternoon games. Professional sports were cheap in those days. We also used to go to Boston Gardens for Bruins and Celtics games. Fred was also a Red Sox fan and he actually got tickets for some of the 1967 World Series games. I was envious, but when I returned to Boston in 1975 on sabbatical we both were able to get World Series tickets. I got tickets for game 6 and Fred got them for game 7!

Fred introduced me to lots of other statistical problems. I was also his TA 1 year, working with Fred and Kim Romney who was in the Social Relations department at the time. Then the time came to get a job, and Fred said to me, "Where would you like to go?" Things were different in those days, as you will recall from your days at Chicago. We went through the list of the best places in the field, at every one of which Fred had a friend. He called up John Tukey at Princeton, he called up Erich Lehmann at Berkeley, Lincoln Moses at Stanford, and Bill Kruskal at the University of Chicago. I either got offers without showing up for different kinds of jobs at these places or I got invited out for an interview. When I was invited to interview at the University of Chicago, it just seemed like a really neat place. All the faculty members were friendly. The temperature in January was really cold, but I liked everything about the university from the people to the architecture; it looked like a university. Leo Goodman was there on the faculty and he had done work that was directly tied to contingency table topics in my thesis. Chicago just seemed like a great place to go to, so I did.

JT: It was there that you first met Bill Kruskal and started being influenced by him?

SF: Bill Kruskal was the department chair at the time, and I barely got in the door before he began talking to me about a slew of different statistical problems...

JT: Without horse steak?

SF: Yes, without horse steak. Bill would just come and say, "What do you know about this?" And one of the first topics we actually discussed was political polls. This was the summer of 1968; there was a lot going on politically in the U.S.A., and the *Sun Times Straw Poll* was showing up in the newspaper regularly. Two of the key questions were: What was their real methodology? How accurate were their predictions? I began to save the data from the newspaper reports and work on the question of variability and accuracy. Then Bill got me to do a trio of television programs with Ken Prewitt and Norman Bradburn on a special series that aired at 6 o'clock in the morning when nobody ever watched. But right from the beginning, Bill and I interacted; he introduced me to Hans Zeisel in the law school, to people in the business school, in sociology. It was really hard to trail after Bill, because he was interested in everything in the university and outside, and almost everything we discussed seemed pretty neat. So, as I launched my professional career at Chicago, I tried to do something similar—not precisely the same as the way Bill did things—but similar.

MS: Bill was a real Renaissance man, and I presume you were a recipient of his many clippings from newspapers.

SF: Well, the clippings started when I was in my first year—he's the one that started to give me the *Sun Times Straw Poll* clippings. But it wasn't just clippings. Bill would leave library books for me in my box; he would go to the library, which was on the second floor of Eckhart Hall, the building we were in, and he would browse—people don't do that today—the stacks are closed. He would come back, armed with books, and he would share them with his colleagues and get Xeroxes of pages. And this continued up through the 1980s. I would always get packets of different materials from Bill, including copies of letters to somebody else that would say: "I hope you don't mind my sharing this with a few of my closest friends and colleagues." I had this image that he was making hundreds of Xeroxes to send around the world.

MS: And before that, carbon paper. So, tell us a bit about your life after Chicago.

SF: The University of Chicago really was a great place for me to work. I had a second appointment in theoretical biology, which was interesting because I had never taken a course in biology as a student. And actually it was a very formative experience, because it taught me that I could go into an area that I had never studied, never learned anything about, and learn enough for me to make a difference in the application of statistics. I wrote papers on neural modeling, and I wrote papers on ecology; I didn't do a lot of genetics, but I read genetics papers and books because I included that material in the course on stochastic processes that I taught. Unfortunately, Chicago wasn't the safest of places in those days, and Joyce made it pretty clear that she wanted to live in a place where our children could play in the backyard by themselves, not under adult supervision 100% of the time. So I began to be receptive to conversations with people from the outside, and soon I was approached by one of my former students, Kinley Larntz, who had just joined

Fig. 30.8 Judy Tanur, John Bailar, Steve, Henry Block, and Jim Press at a conference in Beijing, 1987

the University of Minnesota. They were looking for a chair for the newly created Department of Applied Statistics, as part of a School of Statistics. So after 4 years at Chicago, I became an administrator as well as researcher and teacher.

MS: Did you work with Seymour Geisser there?

SF: The School of Statistics was an interesting idea. Minnesota had had a statistics department, and it had run into some problems over the years. The university came up with this plan to reinvigorate statistics, and they created the School of Statistics. Seymour was the director, and the School was supposed to have three departments. There was the old statistics department, renamed as the Department of Theoretical Statistics, there was the new applied department that I was chairing, and there was the Biometry Department in the School of Public Health. But the biometry faculty didn't really seem to want any part in this, and so they resisted, and ultimately the school had two departments plus the Statistical Center—the consulting center that was associated with our department on the St. Paul part of the Twin Cities campus. Seymour and I interacted throughout my 8 years at Minnesota, but we never wrote a paper together (Fig. 30.8).

JT: I want to take you back a little more. You talked about these two giant figures who were colleagues and mentors—Fred Mosteller and Bill Kruskal. How do you see how they shaped your career, your interests—not only technical, but practical?

SF: One of the things I didn't know as a graduate student was how easy it would be to work on and contribute to new problems and new areas of application. The worst fear of a graduate student—well, the worst fear— is that they won't finish their thesis, the second fear is they won't have a new idea, and, in fact, 80% of students never publish anything other than their thesis. But Fred was going from area

Fig. 30.9 Steve and Seymour Geisser, attending a Bayesian Workshop in Rio de Janeiro, Brazil, summer 1990

to area: when I arrived at Harvard he had just published *The Federalist Papers* with David Wallace; while I was there, he was leading the effort on the Halothane report; I worked with him evaluating television rating surveys from Nielsen and other companies for a national network (that was a consulting problem). He just seemed to work around the clock on all sorts of different topics, and so I figured that's just what a statistician did. It's funny because, in some senses, clearly, everyone didn't behave like Fred, as we all know. But that was my model! So when I got to Chicago and Bill acted in the same way, and Paul Meier in addition, that seemed like a natural way for me to do work as a statistician. They seemed to work around the clock on statistics, so I did too (Fig. 30.9).

Now Fred liked art; in later years he actually took up reproducing art and it showed up in his office. When I was a graduate student I went into his office 1 day and there was a picture by Escher, the Dutch artist, called "The Waterfall" and I was very surprised because I had been introduced to Escher as an undergraduate. Escher's work showed up on the cover of a book called, *Introduction to Geometry*, written by Donald Coxeter—the great geometer at the University of Toronto. I had three courses on different aspects of geometry from Coxeter. This influenced some of my thesis research—and I still do some geometry—but I also learned about Escher from Coxeter! And there was this Escher print in Fred's office which I recognized immediately. Fred told me where he had purchased it, and shortly afterward I went off to the store. I still own two Escher prints as a consequence, ones that I couldn't afford to buy today, all because of Fred. Fred and I would occasionally go off to museums, and while we looked at the art we would talk about statistics, art and other topics.

Both Fred and Bill were Renaissance men and I didn't know how I would do things in the same way they did, but it became very clear to me that just doing

papers in the *Annals* and in *JASA* wasn't enough. While I had colleagues whose careers looked like that, I thought I should be doing something different with my career. I was easily seduced into all these other activities—and everything was so much fun. For example, Dudley Duncan, the sociologist, called me 1 day and asked me if I would join an advisory committee set up by the Social Science Research Council on social indicators in Washington. I hadn't been to Washington since I was 7 years old and I went off to this meeting and then spent 8 years interacting with giants in the field of sociology and survey methods! That experience just reinforced the way I was using my statistical knowledge in diverse applications.

And of course Bill and Fred would just sort of nudge me once in a while to get things done that they cared about deeply. In particular, Fred wanted to see the log-linear model work that his students had done for the Halothane study appear in a book. Fred was big on books. And as I left Harvard, he gathered together all the different students who had worked on different aspects of contingency table analysis—Yvonne Bishop, Dick Light, myself, and Paul Holland, who was a junior faculty member, for a meeting at his house. There were also a couple of other faculty members who sort of disappeared by the wayside in this enterprise, there were a few more graduate students—Gudmund Iversen who ended up at Swarthmore, for example—and Fred said, "We need to have a book on this."

But we didn't have Fred's grand picture in mind and the book didn't begin to take shape until long after I had joined the faculty at the University of Chicago. I taught a contingency table course in my first year there and it included the first three Ph.D. students I worked with—Tar (Tim) Chen, Shelby Haberman, and Kinley Larntz. Shelby extended Yvonne's code for multi-way tables and this inspired his thesis. I began to use iterative proportional fitting on new problems and this triggered a paper on multi-way incomplete tables and a draft of the first book chapter. But then everything progressed rather slowly, and the book took a full 6 years to produce. Fred kept pushing the book behind the scenes.

One of the things I learned is the time to produce a book goes up as the power of the number of authors. It would have taken less time if I had written the book myself instead of with Yvonne and Paul. But while we worked at the core of the enterprise, the three of us had different conceptions of some materials, and this slowed us down. Fred was a full partner, pushing us to "get the job done." He edited draft chapters over and over again, and Dick Light contributed big chunks to the chapter on measures of association, which Paul and I redid and integrated with the asymptotics chapter. If everyone who had come to Fred's house back in 1968 had become involved, we might still be working on the book today! Fred didn't want his name on the cover of the book. So we had this back-and-forth. The book ended up with five names on the title page; it's Yvonne Bishop, Stephen Fienberg, Paul Holland, with the collaboration of Frederick Mosteller and Dick Light; Dick had contributed to a chapter in the book and Fred had contributed to the whole enterprise.

JT: The book, which many have called the "Jolly Green Giant" because of its cover, really put you on the map. In fact, that's how we met, when I took the short course the three of you gave based on the book in 1976 at the Joint Statistical Meetings.

SF: We actually met earlier, when Fred organized a meeting in Cambridge to discuss the ASA-NCTM book projects that ultimately produced *Statistics by Example* and *Statistics*: *A Guide to the Unknown*, your first *magnum opus*. I was a bit intimidated since you seemed to be the organizer for *Statistics*: *A Guide to the Unknown*, and so we just didn't talk much.

MS: Steve and I met around the same time as well. I remember his coming to Chicago to interview and talking about the geometry of 2×2 tables. I asked him a question which he didn't really answer and then he wrote a paper about that problem several years later!

SF: But when I got to Chicago you were one of the few good students who didn't take my contingency table course. You were too busy campaigning for Hubert Humphrey and worrying about weak convergence!

MS: Well, one of the things that you have advanced in that book and elsewhere derives from the geometric structure that gave you so much insight into what's going on in these tables. Now, you mentioned taking geometry at Toronto, and we know R. A. Fisher was influenced by this, so how did that play out in the later research?

SF: It's come into play in an amazing sort of way. If you look at the cover of *Discrete Multivariate Analysis*, there is an artist's depiction of the surface of independence for a 2×2 table. You'd hardly know it was a hyperbolic paraboloid sitting inside a tetrahedron by the time the artist got done with it, and you see one dimension of rulings—a hyperbolic paraboloid has two dimensions of essentially orthogonal rulings—and those are things I actually learned from Coxeter in that course on the Introduction to Geometry. And so my first work actually drew upon that; I wrote a paper with John Gilbert on the geometry of 2×2 tables that appeared in *JASA* and published a generalization in the *Annals*, and I always thought about contingency tables and other statistical objects geometrically. Don Fraser thought geometrically, and so you're always up here "waving arms" in some abstract space, and he would always wave with his arms. And I think in high-dimensional space in some sense, although obviously we don't see in high-dimensional space. But a lot of statistics is projecting down into lower-dimensional spaces. I had left the geometry stuff behind, except for motivation, until I got into confidentiality research in the 1990s.

In the 1990s, there was a paper, unpublished for 5 years by Persi Diaconis and Bernd Sturmfels. Persi was at Cornell and Bernd had been at Cornell but moved to Berkeley. In the paper, they talked about the algebraic geometry structure associated with contingency tables. This turned out to be right at the heart of what I needed for my problem, and so I learned algebraic geometry, which I had not really studied carefully before. I learned at least enough to bring my problems to Bernd for help. And one of the things I realized is that figure on the cover of Bishop, Fienberg and Holland was being used by algebraic geometers in a different context; it's called a Segre Variety, named after Corrado Segre who was one of the fathers of algebraic geometry. That work is now reflected in the theses of a couple of my former Ph.D. students and lies at the heart of a lot of what I've been doing over the last several years, including recent work on algebraic statistics and network models (Fig. 30.10).

Fig. 30.10 Steve, Judy Tanur and Morrie DeGroot, Joint Statistics Meetings, 1978

JT: I think I derailed you sometime back where you were talking about the trajectory of your career. And we've left you at Minnesota. Can you tell us why you left?

SF: Minnesota was a giant bureaucracy. It was a big, big university, and one of the moments that convinced me of this was after I had presented a report, prepared with colleagues from around the university, to the president and the vice presidents on the teachings of statistics at the university, where I had pointed out that 40 different departments or units were teaching statistics or courses in which statistics represented a serious part of the activity. Virtually all of this was going on with little or no coordination with the School of Statistics. And then I met him [the president of the university] about a month later at a reception. Joyce and I were going through the reception line, and I shook his hand, and he asked what department I was from. I said applied statistics, and he said, "*Do we have a statistics department at the University of Minnesota?*" At that point I said to myself, "Oh my goodness!" and I understood where the School of Statistics and my department stood in the big picture of the university.

A year or two later, I was wooed by friends at another Big Ten university, but the right offer didn't quite come to pass. In the mid-1970s, I was working as an associate editor for the *Journal of the American Statistical Association*, initially with Brad Efron as theory and methods editor, and then with Morrie DeGroot. Later I became the Applications and Coordinating Editor of *JASA*, and so Morrie and I worked together. We had become friends a number of years earlier, drinking in a bar together at an IMS regional meeting. Morrie and Jay Kadane, who had joined the Department of Statistics at Carnegie Mellon in the early 1970s, and I would interact at the Bayesian meetings that Arnold Zellner organized twice a year. They both knew that I had flirted with the possibility of leaving the University of Minnesota, and they said, "You should just come to Carnegie Mellon; you could bring the rest of *JASA* over and we'd have the whole journal. Besides, it's a great place." So they

worked on the possibility of an appointment for me. When I came to interview, it wasn't just to meet with the Dean, and with Jay and Morrie and the people in the department that I knew. They took me to see the president of Carnegie Mellon (CMU), who at the time was Richard (Dick) Cyert. Dick was an economist but also a statistician! He took courses from Hotelling and Cochran at Columbia as a graduate student, and although his degree was in economics, he always thought that he was a statistician as well. In particular, he was a member and Fellow of ASA. Dick helped to fund the CMU Department of Statistics in the mid-1960s when he was the dean of the Graduate School of Industrial Administration. He was actually the acting chair at the outset until Morrie took over. So the staff ushered me into his office. I had never met Dick before, but that afternoon I spent 2 h with the president of Carnegie Mellon. And I told you about my interaction with the president of the University of Minnesota! Here I am sitting with the president of Carnegie Mellon, this great university, and he's telling me how important it is for me to come to Carnegie Mellon and what I'm going to do for the field of statistics. He said, "If you come here, everything you do will be called statistics. You will get to change the field." So I came. And I hope that I've changed parts of the field.

MS: Cyert was a visionary and really led the Graduate School of Industrial Administration to a high place among business schools and understood that he needed quantitative strength, and so he influenced you and supported you. I wanted to ask about one of your greatest honors, and that is your election into the National Academy of Sciences. Where were you and how did you get the word?

SF: Most people don't know what goes on at the National Academy—it's like a secret society—and its selection process is Byzantine, running over the course of one or more years. At the end, the NAS members meet in Washington at the annual meeting in a business meeting and they elect the new members. That happens between 8:30 and 9 in the morning; then they take a break in the meeting and everybody rushes out to find a telephone and they call their friends and the newly elected members to the section to congratulate them. This was in the spring of 1999, and I was teaching—actually that year I was teaching an introductory statistics class, so I had to be there relatively early—it was just at 9 o'clock, I was opening the door to my office, and the phone rang. I answered and it was several friends, mainly demographers—Jane Menken, Doug Massey, a couple of others—and there was a chorus on the phone saying "Congratulations, you've been elected to the National Academy!" I was floored, because I'm not quite sure whether they knew, a year or so earlier I wouldn't have been eligible, because I was born and raised in Canada, and I hadn't become an American citizen until January 1998. Thus being elected the next year was a special honor (Fig. 30.11).

JT: You have received many other awards and honors; that must be very exciting.

SF: Well I would be lying if I said that receiving honors and awards is not fun, and each is always very special. But I am reminded about something that Fred taught me. He said that awards and honors are really not for the people who get them, but they are for the field. Of course the person getting the honor benefits, but the field benefits more, for example, when statisticians get elected to the National Academy of Sciences. In that sense we don't have enough big awards.

Fig. 30.11 Richard Cyert, Dennis Gillings and Steve, at a National Institute of Statistical Sciences Board of Trustees Meeting, 1993

MS: There are some of our colleagues who are happy that there isn't a Nobel Prize in Statistics, and as a consequence statisticians cooperate more with one another than scientists in other fields. Do you agree?

SF: Well, I think if we follow Fred's reasoning we would all be better off with a Nobel Prize in Statistics because once a year all of the newspapers and media in the world would focus on our field and the accomplishments in it. What most statisticians don't know is that there almost was a Nobel Prize!

The story goes back several decades when Petter Jacob Bjerve, who was the director of Statistics Norway, began to raise funds for a Nobel Prize in Statistics. He was off to a good start when he ran into a political obstacle. Those in charge of the prize in Economic Sciences objected because they argued their prize encompassed a large amount of what was important in statistics. In the end Bjerve abandoned his quest, and the money he raised was left in a special account in Statistics Norway. Finally, the government auditors forced Statistics Norway to close this account and our colleagues there decided, among other things, to use the funds to host a special international seminar, to which they invited statisticians such as Fred Smith from the UK, Jon Rao from Canada, Wayne Fuller, me, and a few others. They paid for our spouses to come as well and we got the royal (small R) treatment, with relatively fancy hotel rooms and outstanding dinners. So in this sense you could say that I ate the Nobel Prize in Statistics, although there is no public record and it doesn't show up on my CV.

JT: You've been active in several committees and panels and so forth, including at the National Academies before and after your election as a member—what stands out particularly from those?

SF: Well, of course this is Bill Kruskal at work—most statisticians who are going to read this interview don't know the history—Bill Kruskal founded the Committee on National Statistics (CNSTAT) at the NAS. It was an outgrowth of

the 1971 Report of the President's Commission, chaired by Allen Wallis and co-chaired by Fred Mosteller, and Bill talked to the people at the National Academies, and the National Research Council (NRC, its operating wing), into creating a committee although there was no external funding, and the NAS really had to put up resources. Bill ultimately got some money from the Russell Sage Foundation to tide the committee over with a part-time staffer—Margaret Martin, who was and is absolutely fabulous and with whom the three of us have worked—and the committee slowly got going. Bill was succeeded by Con Taeuber. At that time I actually was on another committee, on the rehabilitation of criminal offenders, but Miron was working for CNSTAT and I would run into him on occasion. I got to join CNSTAT a year or so later, while I was still doing the work on criminal justice. Getting involved in CSTAT was like all these other activities I have been describing—I was exposed to lots of new ideas and problems to work on. I was like a kid in a candy shop! The committee didn't have a lot of projects then, but I just got to look around the Academy and the Federal Government, and there were possibilities everywhere. I could only do so much, but I pushed the staff to do other things and got my friends on the committee to lead panels. By the mid-1980s, the committee was humming and there were all these neat activities on census methodology, on cognitive aspects of survey methodology, statistical assessments as evidence in the courts, sharing research data—there was just no end.

MS: I wanted to ask about one of them in particular, which Judy chaired and which you were instrumental in creating, and that is Cognitive Aspects of Survey Methodology. When you were inducted into the American Academy of Political and Social Sciences, you referred to that in your speech as one of the most important activities that you had participated in. Why was this and how did it affect your work?

SF: Well, sample surveys is a very strange part of statistics. In my department, nobody else really does it, in the research sense. People think the theory is settled. But *doing* surveys is *really* hard. The measurement problems are enormous. Designing questionnaires is a big, big problem. In the 1970s, I got interested in the National Crime Survey on Victimization through the SSRC committee on social indicators in Washington on which I served. I learned about the difficulties in counting victimization events. In 1980, Al Biderman, who was involved in the re-design effort for the victimization survey, brought together a few people from the re-design project with cognitive psychologists to ask if we could learn something from cognitive science. I thought this was just terrific because I could see ways that I could take methodological statistical ideas and really intertwine them with the theoretical ideas that came out of cognitive psychology. As a consequence, I pushed for that CNSTAT activity even though others thought it made no sense. I was part of the CNSTAT workshop that you and Judy organized—Judy and Beth Loftus and I wrote a series of 4 papers on cognitive aspects of surveys afterward. I was also on the SSRC council, and we created a committee that followed up on those activities. It brought in new people to the enterprise, and it helped get these ideas embedded in the statistical agencies. Janet Norwood ran with the idea at BLS. It was part of the culture at NCHS at that time because Monroe Sirkin was at the CNSTAT workshop and a moving spirit in establishing a cognitive laboratory at NCHS. The Bureau

Fig. 30.12 Participants at 1983 CNSTAT Workshop on Cognitive Aspects of Survey Methodology watching a survey interview video, from left to right: Kent Marquis, Judy Tanur, Phil Converse, Lee Ross, Steve (in upholstered chair), Miron Straf

of the Census was actually the last of the big three agencies to create a separate laboratory facility—but they did—and the influence spread because the associated ideas changed research at the boundaries of survey methods and psychology in a variety of different ways. The reason I am especially proud of this activity is because you'd hardly know that there was any statistical theory or methodology lurking behind it, but there really was.

MS: It's really had a profound effect on the survey field, and now in many places it's commonplace—concepts of cognitive interviewing and all that (Fig. 30.12).

You've been especially close to your students, fostering them personally as well as professionally. Pictures of you attending weddings of your students appear frequently on websites in your honor. So could you tell us a little about your personal interactions with your students.

SF: Well, in the early years, the students were my contemporaries. In fact, I had a couple of students who were older than I was. Kinley Larntz was not only my Ph.D. student and collaborator, but we were good friends and remain so. Over the years I got a little older than my students, and when I moved to Carnegie Mellon I really had the opportunity to have a different kind of student, and with them different kinds of interactions. We were a small department in those days and I interacted with lots of students, not just those whose research I supervised. Each of the students I worked with then was interested in a somewhat different topic; they

Fig. 30.13 Steve with friends at the Objective Bayesian Analysis meeting in Rome, June, 2007. From left to right: Steve, Larry Wasserman, Jim Berger, Susie Bayarri, Robert Wolpert, Isa Verdinelli

went in different directions, and we remained close in most instances (Fig. 30.13, 30.14, and 30.15).

But then, something happened—first, I became a dean, and then 4 years later I left Carnegie Mellon, as you know. I had a second administrative career going on the side—actually, I had three careers, or four. There was also the committee work at the National Academy, which was a full-time job for awhile, there was the methodology I worked on in part with students in the Department of Statistics, and I was also an administrator—I was the Department Head for 3 years and then I was the Dean of the College of Humanities and Social Sciences. I was on an administrative track in the late 1980s and early 1990s, and my contact with graduate students actually tailed off toward the end of my time as Dean. I was also teaching, but there are only so many hours in the day and days in the week. In 1991, I left and went to the York University in Toronto as Academic Vice President (that's like a provost—they don't have that title at York) and so my regular ties with graduate students were severed. I resigned from Carnegie Mellon to go to York, although we didn't sell our Pittsburgh house, and I returned to Carnegie Mellon a few years later and re-joined the department.

I like to describe the move back to Carnegie Mellon as a promotion to the best position in the university—as a tenured professor with no administrative obligations.

Fig. 30.14 Steve with his wife Joyce and many of his former graduate students at a 65th birthday celebration at Carnegie Mellon, October, 2007. From left to right: Ellie Kaizer, Edo Airoldi, Elena Erosheva, Jason Connor, Sesa Slavković, Mike Meyer, Joyce, Steve, Alessandro Rinaldo, Justin Gross, Russ Steele, Adrian Dobra, Amelia Haviland, Elizabeth Stasny

I slowly began to work with graduate students again. Somewhere along the way I think I had learned something, which is you can't necessarily get graduate students to do what you want, and thus what you have to do is get them to do what they want to do in the best possible way. You have to get them to complete a thesis, but you have to be able to get them through and have them gain confidence in what they're doing so that they think they can make a difference. And I was lucky—I just had fabulous students; they were terrific people and all the rest of the stuff just sort of happened. I had the opportunity to give away in marriage one of my students, Stella Salvatierra, who was working in Spain, at a ceremony in the mayor's office in Bilbao, because her father had a heart attack and couldn't come to the wedding. And there have been several other weddings since! Because my students have been so great, the best thing I can do in some sense is to get them to do the things that they do best. That's in many ways a serious part of my legacy.

JT: I was going to ask you what advice you would have for graduate students in statistics, or undergraduates for that matter. Clearly, the best advice I could give would be for them to come to be your students, but since you can't spread yourself totally thin, failing that, what alternative advice would you offer?

Fig. 30.15 The longtime members of the Carnegie Mellon Department of Statistics in the DeGroot Library, 2011. Back row: Rob Kass, Mark Schervish, Steve, Joel Greenhouse; middle: Margie Smykla; bottom row: Jay Kadane, Bill Eddy, John Lehoczky

SF: Well, I really can't work with them all! It's really bad because now we've got this undergraduate program with upward of 150 majors. I can deal with one or two graduate students at a time. But my advice to budding statisticians is simple: statistics is an exciting field. There are all these neat problems. There are neat theories, neat methods, neat applications; we're in a new world. Big, big data sets. My joint appointments are now in the Machine Learning Department and in the Heinz College (of Public Policy and Management). I'm working with data sets that people couldn't conceive of dealing with a few years ago. And the students I'm working with have the ability to go and do things with those data sets that were unimaginable a decade ago. So my advice is simple. Work with data, take problems seriously, but you have to learn the mathematics and statistical theory if you want to do things right. And then you need to take seriously teaching people what you've done, not just doing the research. You need to get the descriptions of your work into a form that other people can understand—that's a really important part of what we do. That's what National Academy reports are all about. Academy reports don't have impact if they're badly written. Enormous effort goes into the executive summaries of reports, into the review process, and everything up the line. Learning how to do that as a student is time well spent. It's too late when you're a full professor and you still haven't learned how to write articles so that other people can understand what you've done (Figs. 30.16 and 30.17).

MS: So, of your vast experiences, what are you the most proud of?

Fig. 30.16 Steve and Bill Eddy celebrating the 20th anniversary of *Chance*, a magazine they co-founded in 1988, wearing their original Chance t-shirts

SF: I'm actually proud of a number of things. By the way, I didn't tell you what my fourth career was. I play ice hockey—I still play, that's number one, although the one for which I have the fewest skills or accomplishments.

MS: All right, let me interrupt you...

SF: Ha ha, no-no, as I left the locker room last Saturday night, one of the guys across the dressing room said to me, "So how many years have you been playing?" And I said, "62." He then said, "*62*?" and silence ensued. But maybe hockey is really number two; number one is my children and my grandchildren. They're really amazing. They're another part of my life. Joyce and I were really fortunate; I have two very smart sons, Anthony and Howard. They have independent careers, they have lovely wives... (Fig. 30.18)

MS: Where are they now?

SF: Anthony lives in Paris, and I have five grandchildren in Paris, four granddaughters, and a grandson. And Howard lives in the DC area and I have a lovely granddaughter in Vienna, Virginia. Howard actually has come very close to statistics, as government liaison for a consortium dealing with surveys and marketing. The grandchildren are terrific. I love being with them. We get to look after them every once in a while.

Then there are my students. They're really the people who are going to do the things that I can only imagine. As I look back over what I've done, I see a changed field of statistics. Fred Mosteller and Bill Kruskal were fabulous—and we've talked about how they shaped all three of *our* careers, not just my career. And they launched the Statistics Departments at their respective universities. I was part of both departments and their programs in retrospect look "traditional." They

Fig. 30.17 Steve (on the right) playing for the Division C national championship as a member of the Leiden Beaver Beer Team, in Eindhoven, March, 1997

emphasized mathematical statistics and probability. I like to think that when I left Chicago and went to Minnesota, I started to change what statistics did and how we thought about it. And applications today sit at the core of much of statistical theory and methods, and in my department at Carnegie Mellon our students come out having worked on multiple applied projects, and they're in demand, because that's the future of our field. People recognize that advances in statistical methods—and theory—are intertwined with real problems, major applications. I like to think that I contributed to the change that we've seen over the past 40 years.

MS: Very nice, Steve. What you talk about is a legacy, not the individual research that may wane in importance over the years…

SF: And it's not just my work, it's a collective…

MS: But it's the influence of your students, as well as your children. I wanted to interrupt, because I never thought you had four careers, I thought you had dozens of careers. You talked about these professors that, you know, worked 24/7, so that was your model. As long as I've known you, you're always multi-tasking, and you

Fig. 30.18 Steve with twin granddaughters, Tiffany and Selena, trying out their new bikes, Paris, 2006

were doing that before the word was even in vogue. You're fielding questions at a seminar or flying a hockey puck across the ice. Did any of that rub off on your sons, on your students?

SF: I don't think that either Anthony or Howard is quite as obsessed as I am with doing so many things simultaneously.

MS: How fortunate...

SF: That's right! But Anthony did play hockey in Paris for many years, and both Anthony and Howard have these terrific kids—since Anthony has five, they take up more of his time than mine did. Actually, Anthony has inherited some of this multitasking, at least at some level. He's created his own business in France—a subsidiary of a Dutch insurance company. His job went from finding the location to organizing the offices, to hiring the staff, to inventing the insurance policies and making sure that they were consistent with the ones of the parent company (Fig. 30.19).

My students also develop multiple facets of their careers and lives. I tell them when they come in and ask if they can work with me that there are a couple of things that are going to happen if the arrangement is going to succeed. One is they're going

Fig. 30.19 Steve, buried amidst files, in his CMU office, 2005

to live and breathe statistics. I see it everywhere. One of my favorite examples in my little contingency table book came out of the program from the symphony at the Minneapolis Orchestra one night when we were there in the 1970s. It didn't *quite* look like a contingency table, but I made it into one, in my book. Then I described why you shouldn't analyze it the way you would have otherwise, because the units of observation are not independent. At any rate, I tell the students that I expect them to live and breathe statistics. They'll get their ideas in the shower... they'll play hard too, but when all is said and done, if they're not into what they're doing, they should find another advisor, because other people have different attitudes about work and how to get your inspiration! Students of course have their own lives, and as I've said, you don't tell students what to do, they tell you what they want to do.

JT: What's next? For you?

SF: Wow. I'm too busy to stop at the moment to find out! I still have more than one job. I'm editing, with some others, the *Annals of Applied Statistics*, I have launched the *Journal of Privacy and Confidentiality*, I'm the co-chair of the Report Review Committee at the Academy.[2] I have a whole bunch of new Ph.D. students and post-docs. We've got some absolutely fantastic projects going on: research on confidentiality problems and on network modeling, which by the way, links to confidentiality. Judy and I also have a book on surveys and experiments to polish

[2]Steve took over as the editor-in-chief of the *Annals of Applied Statistics* on January 1, 2013, and is simultaneously serving as the founding editor of yet another publication, *The Annual Review of Statistics and its Application*, scheduled to launch within the year.

up for publication, as Fred Mosteller would say. I have six chapters that were, I had thought, pretty polished at one stage, but they are still in a drawer in my office. At least I know where the drawer is.

JT: And I know where my copies are…

SF: And so, I've got more books to write too—with good collaborators.

MS: Well, we're almost out of time, but I have one final question. How would you like to be remembered, Steve?

SF: Unfortunately not as a great hockey player. As long as my teammates just let me on the ice, I'm happy to be able to skate around and get off safely.

I guess I'd like to be remembered as somebody who produced really good students and who helped change the image of statistics in the sense that lots of people now work on serious applied problems and help solve them. And that's not just about statistics, that's real interdisciplinary scientific work, and that's the legacy I inherited from Fred and Bill Kruskal and Paul Meier, and all those other great people that I had a chance to work with, like Bill Cochran. I would just like for people to think of me in their kind of company, in some way or another. I suspect that a couple of decades from now, if anybody ever looks at the video we're making or reads this interview, they may not remember log-linear models for contingency tables and other forms of counted data because there will be new methodology, like the mixed membership and related models I now work with. What I know from students today is that, if it wasn't in the journals in the last 3 years, they're not sure it's worth their attention. So, if I am to have a legacy, it needs to be something larger. I have no theorems, well, I do have theorems, but none of them are named *Fienberg's Theorem*. And even if there were a Fienberg's Theorem, it probably wouldn't be important—what's important is the attitude, for what statistics is and how it's recognized by other people outside of our field.

MS: Well, you've changed statistics, and you've made it fun along the way. Thank you very much.

Chapter 31
Reminiscences from Steve's 65th Birthday Celebration

Margaret Smykla

It was a traditional practice in the Department of Statistics (and still is in today's Department of Statistics and Data Science; name changed in 2017) at Carnegie Mellon University that when a renowned faculty member reaches a milestone, that colleagues, students, staff, and alumni are given the opportunity to share their thoughts and experiences on working with that person.

Those "reminiscences" are then compiled in a so-called memory book to be shared with the honoree at an event in which he/she is so honored.

The reminiscences below were gathered for Steve Fienberg's 65th birthday celebration held on October 19, 2007, at Carnegie Mellon via a series of invited lectures and reception.

We share them here as written in 2007. Each reminiscence, in its own way, contributes to an all-round image of Steve as educator, mentor, scholar, innovator, colleague, advisor, friend, promotor of statistics as a force for good in science and government policy, and so much more.

Each reminiscence includes the author's name and current title/place of employment. For those authors who were students in the department at the time, the year in which they subsequently attained their PhD degree from the department is also listed. If it was obtained elsewhere, the name of that university is included.

The reminiscences are listed in alphabetical order by the author's last name.

(In the next section in this book are reminiscences about Steve from an October 14–15, 2016, event in honor of his attaining Emeritus status.)

*R*egarding academic matters, Steve is unstoppable. He once told me that the best way to show students what research is all about is by being an example. In point of fact, Steve works very hard. Throughout my graduate studies, for instance, and whenever I visit Carnegie Mellon to this day, it is typical of Steve to invite me to his office on Saturdays and Sundays,

M. Smykla (✉)
Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: mk74@stat.cmu.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_31

when the department is quiet and we can focus and be productive. His energy is indeed contagious. Talking to him about what is the big question that our new method or theory should be able to address, or trying to get around a subtle technical problem at the board, is stimulating and fun. On a more personal note, Steve is a wonderful man and a great mentor. He has always been there for me, even in situations where it would have been easier to simply let go. I enjoy his spirit and his strong opinions on everything (attending a seminar with Steve is never boring), and I look forward to collaborating with, and learning from, him in the years to come.

–Edoardo M. Airoldi (PhD 06 Comp. Sci.; Steve as advisor), Millard E. Gladfelter Professor of Statistics and Data Science, Director, Data Science Center, Fox School of Business, Temple University

*H*ow can I summarize the effect Prof. Stephen Fienberg has had on my professional and personal life in one story? He has been a wonderful advisor, mentor, and friend – and my stories about him range several years and three continents!

But my favorite memory of Steve is the role he played in my wedding. When I first became Steve's advisee, I was a single mom – but I met my husband in the course of my work (at the US Census Bureau). Steve and Joyce were naturally invited to the wedding, and they were delighted to attend.

There I was, pacing nervously in the Rabbi's office before the wedding, waiting for the signing of the Ketubah – the Jewish wedding contract. I had wanted Steve to sign the Ketubah as my witness, but hadn't had a chance to ask him yet! The minutes ticked closer to the start of the ceremony, but no Steve. As far as I was concerned, the wedding wasn't starting until Steve and Joyce were there.

Finally, Steve arrived at the door of the Rabbi's office, apologizing for being late due to travel delays! The Ketubah signing could proceed. By Jewish tradition, my husband and I were married then – even though the actual ceremony hadn't taken place. Only a small number of people witness the Ketubah signing – close family and friends – and the fact that Steve could be part of that moment will always be a wonderful memory for me.

–Jana Asher (PhD 16), Assistant Professor, Slippery Rock University

*W*hen I moved to Pittsburgh in Spring, 2000, it was to work at a startup on Craig St. called "Whizbang! Labs" (yes, with an exclamation point). I interacted with Steve briefly a few times during the next few years in his role as acting director of CALD, but I didn't really meet him until Spring, 2002, when I happened to run into him at Pittsburgh International Airport.

As I recall, it was a dark and stormy night, and we were both down on our luck. I was coming back from a discouraging trip to schmooze in D.C., I think looking for funding of some sort, and I was unemployed, Whizbang! Labs having recently gone the way of (most) Web 1.0 startups. Steve was returning from what was probably an even more discouraging trip to the Canary Islands – he'd gotten as far as the ticket counter and then discovered he'd left his passport at home. Now this was before 9/11, so he wasn't arrested, or detained for questioning, but he wasn't on his way to the Canary Islands either. He was waiting for the 28X to take him home.

To make the story not too long, we ended up discovering a shared technical interest in (passion for?) record linkage, and having a long and detailed and refreshing technical talk all the way back to Pittsburgh, which continued while we waited for a friend of mine to pick us up from the bus stop, and then again while all three of us cowered in the car for another 15 minutes waiting for one of Pittsburgh's "microburst" intense storms to blow over. This was a talk that also continued, on and off for the next several months, during which time Steve was instrumental in assembling the funding that eventually brought me to CMU, where I worked for several months on what I still think of as the "28X project."

So to summarize, I'm not sure if I'd be here at CMU if Steve had managed to remember his passport that day. And I'm not sure what the point of this story is, except maybe to say that sometimes we find the best of friends at unexpected times and places, and for unpredictable reasons.

–*William Cohen, Principal Scientist, Google*

I was Steve's PhD student from 2003–2006. I was fortunate to have an advisor generous with his time and expertise, but who gave me the freedom to investigate the statistical questions I found most interesting. From the beginning Steve hoped I would explore a particular model and I had a different idea in mind. While we had many conversations about the competing methodologies, I'm thankful he allowed me the freedom to explore what I wanted to explore and write the thesis that I wanted to write, all the while giving me his complete support.

Basically every time I discussed any statistical problem with Steve I learned something – even when it was a topic that I thought I understood! For that reason and many others, I always looked forward to spending time with him and always will.

–*Jason Connor (PhD 06), ConfluenceStat*

I first met Steve Fienberg 43 years ago, in the fall of 1964. I was a college sophomore taking Fred Mosteller's famous Statistics 122 (Statistics in the Social Sciences, or something like that), and Steve was a graduate student. As a math-oriented student in a sea of math-phobic psychology majors, I had come to the attention of our teaching assistant (actually universally known as "section man" in those days, regardless of gender – this particular "section man" was a fairly glamourous young woman) who invited me to a party for her fellow statistics graduate students. (They always needed more females and she thought I would fit in – she was right!) Steve was at the party. He was wearing lederhosen, for reasons then and now unknown to me. He was very friendly and we had quite a nice conversation going when I was yanked away by my "section man" who informed me that Steve was engaged and I'd better steer clear of him. I should note that two other grad students (one of them being Jonas) came in tuxedos. (I got yanked away from Jonas also, since it turned out my "section man" was dating him at the time.) All in all, a memorable party, definitely establishing that statisticians were interesting, good-looking and a lot of fun!

Jonas and I send our best wishes for a wonderful birthday!

–*Susan S. Ellenberg, Interim Chair, Department of Biostatistics, Epidemiology and Informatics, Professor of Biostatistics, Medical Ethics and Health Policy Perelman School of Medicine, University of Pennsylvania*

*W*hen I was looking for a dissertation topic and an advisor, I had a number of conversations with faculty members at CMU before coming to Steve's office. I remember the very first thing that he said was, "There is something you should know about me. I am very busy." Taking a mental note of this phrase, I did not expect much attention but I was clearly in for a pleasant surprise. Now, five years after obtaining my PhD, it seems that Steve has somehow always managed to give just the right amount of guidance, even if it had to be in a three-minute meeting at the end of which his only words would be "Go and write it up." The ever-increasing number of his students is the best testimony to his exceptional mentoring abilities.

–*Elena Erosheva (PhD 02), Professor of Statistics and Social Work, University of Washington*

*W*hen I first met Steve in December 1985, I was a graduating senior in statistics/operations research at the University of Cape Town. I did not have a plan, yet was fascinated by statistics and wanted to learn more. I had a vague idea that I might like to come to the US, but applying to different schools and taking care of the various application requirements was getting on top of me.

Steve was the keynote at the South African stats conference that year, and we met. That meeting completely changed my life. The next August I was on my way to Pittsburgh.

I arrived in Pittsburgh knowing no one, this naïve South African little girl. Steve and Joyce were there to welcome me and they hosted me and my mother as we navigated those overwhelming first weeks. Steve and Joyce became my family in Pittsburgh, and welcomed me for my first Thanksgiving and for Jewish holidays (thank you so much!).

Steve, I am so glad to have an opportunity to tell you how much of a difference you have made in my life. You opened the doors to me in this country; it all began with you. Since I came here I have found incredible career fulfillment and my life's partner.

I don't know that I can ever thank you enough.

–Ruth Etzioni (PhD 90), Professor, Fred Hutchinson Cancer Research Center

*I*t's not too much of an exaggeration to say that I wouldn't be at CMU if it weren't for Steve. When Steve called to invite me for a visit, it felt like getting a talking from an old friend. We discussed our common interests, the department, my plans. When I mentioned that my talk would not be Bayesian – Carnegie Mellon had a reputation, you know – I could almost hear him smile. He then laid out a vision of what the department was really about. Here is a person I could learn from, I thought, and a place I could thrive. Steve made the case even more compelling when I met him in person. We had been on opposite sides of the Census adjustment debate, and I remember sitting in Steve's office during my visit talking over the issue from top to bottom. Never once did I feel that we were on different sides but rather that together we were searching for the truth. It was invigorating.

For over a decade, Steve and I had neighboring offices. Before high-speed internet connections and wireless made it so easy to work at home, he and I spent a good deal of time working in the office on weekends, and we would talk about statistics and other things. Steve was always gracious, helpful, and interested. And he gives good advice. One regret I have is that he and I have not yet managed to write a paper together. But there's still time . . .

–Christopher R. Genovese, Head and Professor of Statistics, Dept. of Statistics and Data Science, Carnegie Mellon University

*W*hen I joined the Department in the mid-1980s, Steve was the head. An initial and enduring impression of the Department was the strong feeling of inclusiveness and collegiality. Not only that, but to my surprise the senior faculty not only expressed an interest in my work, but were forthcoming with their own ideas and opportunities for collaborations. In those days we didn't call it mentoring but that's what it was. I soon learned that this departmental climate was due in no small part to Steve's leadership and vision.

Steve's door has always been open. I have never seen him turn away a visitor, whether undergraduate, graduate student, or junior faculty. He is a demanding instructor and advisor, yet he is also realistic and understanding. He sets his expectations at a level that asks his students and colleagues to reach a little higher and achieve a little more. Steve is selfless about his work, and given his level of activity, there are plenty of good problems to share with students and junior colleagues, which he does.

Thanks, Steve, for being a model statistical scientist, outstanding colleague, and good friend. All the best for a very happy birthday. Looking forward to sharing many more simchas with you and Joyce.

–Joel B. Greenhouse, Professor of Statistics, Dept. of Statistics and Data Science, Carnegie Mellon University

*W*hen I was researching statistics departments for grad school, I solicited suggestions for faculty around the country with an interest in social science. I received a number of recommendations, but one name was on everyone's list: Stephen E. Fienberg. So I planned a visit to Pittsburgh and emailed Steve to ask if he would meet with me. A short time later, Steve had emailed me back to say that I was in luck – although he was on sabbatical and

about to leave for Paris, he would be in town the day of my visit and could make some time to chat with me about the program.

So I arrived for my appointment and was immediately won over. For someone so accomplished, Steve turned out to be surprisingly down-to-earth and exceptionally warm. Despite being on his way out of town, he wasn't the least bit impatient, and indeed engaged me in a long conversation about statistics, various departments I had been considering, and my possible research interests. When I hinted that I might be satisfied to just earn a master's degree, he brushed aside my implied lack of ambition and said, "Oh, you don't want to do that – we'll get you through the program."

That's not to say that Steve is averse to doling out a little tough love. One afternoon, I was sitting in the department lounge when Steve walked in. I had given a talk the day before and felt it had gone pretty well, but he took one look at me, tilted his head to the side, let out a slight groan and announced, "We really need to teach you how to give a presentation!" How much more gratifying it was, given his honesty, when a year or so later, I finished a departmental presentation and was greeted by Steve's broad smile and the words "Nicely done!"

Steve, you have touched the lives of so many colleagues, students, and friends. I am honored and blessed to be among them.

–Justin H. Gross (PhD 10), Associate Professor, Director, Division of Political Science, Department of Political Science, University of Massachusetts Amherst

I first saw Steve when I was a graduate student and he was receiving the Statistician of the Year award from the Chicago chapter of the ASA. Steve's predecessors included John Tukey, Fred Mosteller, L.J. Savage, G.E.P. Box, William Cochran, and Bill Kruskal. I recall Steve's saying, in accepting the award, that he wasn't in the same league as those guys. Well, every generation should stand in awe of its ancestors, and some of those earlier guys will be members of that rarified club who are remembered long past their own lifetime. But for those of us who are here now, Steve is his own kind of giant. Rough, but ready, like the hockey player he also is who pushes his way to the front of the net, waiting to turn trash into a score; seemingly tireless. By the way, does he ever sleep?

When I interviewed at Carnegie Mellon, there was a rather arcane, mathematical focus to my research, so it was not obvious to all on the faculty that I would be a good bet as a colleague. I am fortunate that Steve, as incoming department head, took a special interest in hiring me. And when he perceived some reticence on my part about coming to Pittsburgh, Steve showed me things I might appreciate about the town, while painting a convincing picture of the rosy future for the department. That was 26 years ago. I suppose the lingering memory may have influenced my reactions to Steve's ideas and pronouncements over the years. In any case, I have found Steve's instincts to be generally on target, his taste excellent, and his judgment very good. As I think about him, though, I realize that what I have appreciated most about Steve is his very strong desire to do the right thing, and to work hard to get the right thing done. It is his unusual combination of continual achievement and tireless commitment to improving the world that makes Steve an inspiration.

–Robert E. Kass, Maurice Falk Professor of Statistics and Computational Neuroscience, Department of Statistics and Data Science, Machine Learning Department, and Neuroscience Institute, Carnegie Mellon University

I was his student from the fall of 1986 until the fall of 1989. The 3 years were a period of voyage through a maze of misty sight. In those days, he had two offices: the Dean's office and his own office. Working in these two offices, he still found time to give thorough and detailed comments on my hand-written manuscripts. I remember his warm statement about my English: "'A' and 'the' seems to be difficult parts of English for foreigners like you." He then suggested inserting them when I was not sure. This comment called my attention to "a" and "the" whenever I read books in English. I might have tested his patience for a while with my "wild" English!

He is both a thesis advisor and a humanity mentor. As far as I am concerned, I would rather remember him for the latter role. He gave some ripe words concerning roles as a father and as a husband. Talking about kids with him, Steve once said, "Kids know better ways than you for their lives." These words helped enhance the level of my patience to a certain maturity when my wife and I were at heart-breaking moments with my kids. My experience with Steve as a thesis advisor and as a life-story consultant has led me to what I am now. Thank you for all of this, Steve.

–Sung-Ho Kim (PhD 89), Professor Emeritus of Statistics, Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology

I first met Steve around 1989–1990. I was in my compulsory military service in the Israeli army and thinking of going to graduate school, possibly in statistics. A friend of my parents told me that a friend of his, a statistics professor from the US, was visiting Israel, giving some talks, meeting with prospective graduate students, and he had agreed to take some time in his busy schedule to meet with me.

We met in the Statistics Dept. at Tel Aviv University. Steve sat with me for a long time – maybe more than an hour – and told me about the various programs in the US, their strengths and weaknesses, and encouraged me to apply to graduate schools in North America. It was the first time that I really had a sense that "maybe I can do this!"

Over the years, Steve grew to be a mentor and a friend. Now that I am in a position to mentor students at all levels, as well as junior faculty, I try to keep Steve's example in mind always: his generosity with both time and sharing of wisdom, his patience, and his humor. I can only hope that I have as favorable an impact on their lives as he has had on mine.

–Nicole A. Lazar (Dept. faculty 96-04), Professor, Department of Statistics, Pennsylvania State University

*I'*ve always been impressed by Steve's apparently unlimited energy. In January 2006, while I was in my first year here at CMU, Steve and I traveled to Colombo, Sri Lanka, to participate in a project with some human rights organizations there. The travel was exhausting and that, in addition to the time difference (11 hours!), just knocked me down to the ground for almost a week. Steve, however, just after a night of sleep was fresh as new and ready to work. We had lots of meetings and, on top of that, he gave a series of lectures at the University of Colombo, and even found time to meet with faculty and students there.

–Daniel Manrique-Vallier (PhD 10), Associate Professor, Department of Statistics, Indiana University Bloomington

I used to be the old man in the Mt. Lebanon Senior Hockey League. I retired last year and passed the mantle to Steve. This story applies to both of us. I had been urging my wife to attend a game and she finally showed up one day while the game was on. I was playing against Steve's team. Our jerseys have no numbers, and we wear face masks which hide our faces. So my wife couldn't immediately make out which of the skaters either I or Steve was. But she said she quickly figured it out. "How?" I asked. "Simple," she said. "I looked for the slowest skater on each team!"

Steve, don't let this bother you. It's amazing that you can still skate at this age, and within a year you will have broken my record as oldest player. Best wishes,

–Ralph Roskies, Vice Chancellor for Research Computing, University of Pittsburgh

*S*teve, under the least likely circumstances, you drew me into the world of statistics, and bestowed some of your magic upon me. You helped me unravel statistical incantations, but rarely without a few extra tips on good wines, great books, and special travel spots. You are not just my mentor, you and Joyce are my true friends who are there in times of joy and sadness. Maybe one day you will share the secret of your relentless enthusiasm and support for those around you and beyond.

Cheers to you, Steve, and your legacy. Let the magic continue rolling!

–Aleksandra (Sesa) Slavkovic (PhD 04), Professor, Departments of Statistics and Public Health Sciences, Associate Dean for Graduate Education, Eberly College of Science, Pennsylvania State University

*S*teve was my dissertation advisor. I obtained my degree in December 1983, but Steve has continued to encourage and aid me in my career even twenty years later.

When Steve asked me if I would like to work with him on a grant from Statistics Canada in my second year of PhD studies, I didn't realize how fortunate I was. He pushed me to take a very active role, not just in research for the grant, but also in other aspects of the work for the grant. For example, with guidance from Steve, I traveled to Statistics Canada to meet with the research sponsors; I worked with them over the phone to determine the form of the data we would receive; I wrote up a research proposal for the grant; I drafted a paper based on our preliminary research; and I presented the results at conferences. Although all of this made me very nervous, there was no question that I would do it – Prof. Fienberg expected me to be able to do it. It was, of course, invaluable experience when I started working on my own grants as a new assistant professor.

Also, while I was in graduate school, Steve introduced me to a number of important people in my area of research, particularly women (Barbara Bailar, Janet Norwood, and Judy Tanur), who have been vital contacts and role models for me over the years. When I was nearing graduation, Steve encouraged me to apply to good schools for the academic position I desired.

I have noticed that at national meetings, Steve makes it a point to introduce other women to me, often those just starting their careers. Then he leaves to give us a chance to talk, share experiences, and trade advice. I know that several of my friends from graduate school, and even some who did not attend CMU, have been similarly aided and encouraged in their careers by Steve over the years. I had no doubt that I owe a good proportion of my success in my career to his behind-the-scenes encouragement to get involved. For that, I am very grateful.

–Elizabeth Stasny (PhD 83), Professor Emeritus of Statistics, The Ohio State University

*A*s an undergraduate student at Carnegie Mellon, I was fortunate enough to work with Steve on disclosure avoidance and contingency tables for my undergraduate honours thesis. As my research supervisor after my sophomore year, Steve gave me bits and pieces of the problem to work on, step by step, until I unknowingly had programmed an algorithm using computational algebra to generate proposals for a Metropolis Hastings algorithm under a model of independence for multi-way contingency tables. It would be four more years and almost three degrees later that I would understand completely what I had accomplished. It's unthinkable, in retrospect, what I was able to accomplish and I know that it is due mainly to Steve's support, attention, and the way that he never let me feel like what we were trying to do was impossible. It's with a lot of pride that I can say that I've supervised 14 undergraduate research projects and courses in my five-plus years at McGill University. . . . I learned from Steve that it is never too early in their academic life for someone to begin doing research, as long as they know that they're not doing something impossible and that they're not doing it alone.

–Russell J. Steele (PhD 02, University of Washington), Associate Professor, Department of Mathematics and Statistics and Associate Dean, Graduate and Postdoctoral Studies, McGill University

*I*t is my great pleasure and privilege to offer a tribute to Steve on the occasion of his 65^{th} birthday. John Lehoczky once conjectured to me that, if n is the number of projects an ordinary human being can handle at one time, then Steve will take on n+k, where n and k are positive integers. For me, Steve's accomplishments come most clearly into focus when one thinks of the various roles a university professor can assume: teacher, researcher, administrator, advisor. What I find truly remarkable about Steve is that, by my accounting, he has filled all these positions many times in his distinguished career and,

invariably, has succeeded brilliantly. As a student in his classes I enjoyed Steve's well-organized and provocative lectures. As a teaching assistant, I admired his insistence that students in the introductory statistics course should engage the world at large with their newfound knowledge and skills. As a scholar, I respected the deep contributions Steve has made to our field, both in methodological research and in an incredibly broad array of public policy studies. As an alumnus, I valued his contributions as a leader of both the college and the university. And to this impressive list, one can add Steve's legendary talents as a hockey player.

On a personal level, I am most grateful for Steve's generous and wise counsel. At critical points in my own career, Steve has listened patiently, brought opportunities to my attention, and offered his sage advice and perspective. I consider myself truly fortunate to count Steve as an influential mentor, colleague, and friend.

Congratulations, Steve, and many happy returns! Cheers,

–Duane Steffey (PhD 88), Principal Scientist and Director, Statistical and Data Sciences, Exponent

I was sitting in my apartment in Toronto in 1987 working on my thesis when I got a call. "This is Steve Fienberg. Mind if I come over for a cup of coffee?"

A phone call from a famous statistician was the last thing I expected. What could I say but, "Sure. Come on over."

A few hours later, my (soon-to-be ex) wife announced: "Some guy who looks like Rodney Dangerfield is at the door."

I let Steve in and he immediately put me at ease. We had a nice chat about CMU, Pittsburgh, and probably a few other things that I now forget. Mostly what I remember was how welcome Steve made me feel. I had a strong sense that the CMU Statistics Dept. would be a nice place to visit.

Well, my 2-year visit turned into a lifetime career. If Steve had not made this unexpected visit, who knows where I would have ended up.

To this day, there is no one in the department who goes out of his way as much as Steve does to recruit students and faculty for the department. Every year, when I do admissions, I can always count on getting one or two students thanks to Steve's proactive approach.

This is a small example of Steve's selfless devotion to the department and we are all in his debt because of this.

–Larry Wasserman, UPMC Professor of Statistics, Dept. of Statistics and Data Science, Carnegie Mellon University

Chapter 32
Reminiscences of Steve Offered on the Occasion of his Emeritus Celebration

Margaret Smykla

The reminiscences below were gathered for Steve Fienberg's Emeritus celebration held on October 14–15, 2016, at Carnegie Mellon consisting of a series of invited lectures, short informal presentations of reminiscences, and a reception.

We share them here as written in 2016. Each reminiscence, in its own way, contributes to an all-round image of Steve as educator, mentor, scholar, innovator, colleague, advisor, friend, promotor of statistics as a force for good in science and government policy, and so much more.

Each reminiscence includes the author's name and current title/place of employment. For those authors who were students in the department at the time, the year in which they subsequently attained their MS/PhD degree from the department is also listed. If it was obtained elsewhere, the name of that university is included.

The reminiscences are listed in alphabetical order by the author's last name.

(In the prior section in this book are reminiscences about Steve from an October 19, 2007, celebration in honor of his 65th birthday.)

> *B*efore I started my thesis at CMU, I heard: "Steve can be really hard on you when you are going through your thesis work, but he stands out for being extraordinarily supportive, and for his confidence in his students." What I will never forget was: "He will never let you down." These words will always remain in my mind, as it was what I found in you as my advisor. There were many difficult times during those years working on my thesis. You were always there to support me, even when my work was stuck, and helped me to get through.
>
> I am extremely grateful to you, as well as to Joyce, who played a fundamental role for her support and love, while I was living in Pittsburgh.
>
> *–Anita Araneda (PhD 04), Associate Professor, Catholic University of Chile*

> *I*f we are to reduce the amount of human violence in the world, the first task is to determine how much of it there is. And no one in the human rights field really understood how to do that until we re-read – very closely – Chapter 6 of Bishop, Fienberg, and Holland.

M. Smykla (✉)
Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: mk74@stat.cmu.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_32

Of course, people knew how to estimate hard-to-reach populations before then. Nonetheless, it is the principled approach explained in Chapter 6 that gave us the tools to make the estimates in a truly rigorous way.

In the international human rights movement, we remember the dead. And by remembering the known dead, methods that Steve put in our hands allow us to remember the unknown dead. From Guatemala to Colombia to East Timor to Perú to Kosovo, work based on Steve's thinking has helped move the world a little closer to justice. Thank you.

–Patrick Ball and Megan Price, Director of Research and Executive Director, respectively, Human Rights Data Analysis Group

I vividly remember the day that I interviewed for my position in the Statistics Department. I was ushered in to meet with Steve in his office as the department was looking to hire a new assistant for him. I remember being struck by the stacks and stacks and stacks of books and journals and papers. In my memory they were stacked absolutely from floor to ceiling, teetering towers lined up side-by-side on the desks and all about the room. We chatted about the amazing shared musical company that we had in common: he was a loyal patron of the Pittsburgh Symphony and played hockey with my clarinet teacher, Michael Rusinek. Not only that, Michael's longtime mentor Avrahm Galper (famous clarinet pedagogue and former Principal Clarinet of the Toronto Symphony) had a connection to Joyce!

I was a very green new-to-the-workforce musician right out of graduate school whom Steve was convinced could learn LaTeX quickly enough to complete a book project that he was aiming to publish – the autobiography of his mentor, Statistics giant Frederick Mosteller. I jumped in headfirst, honestly not even knowing how to open LaTeX on my computer, let alone how to compile an actual book with it, but Steve trusted I could learn. Eventually we completed the book, launched the online Journal of Privacy and Confidentiality, and even published a second book: The Handbook of Mixed Membership Models. Each project a mammoth undertaking that Steve took in stride (as they ran in parallel conjunction with all the other projects that occupied his time, including at this point in his history the inception of LARC and the NSF Census Node Project). All the while Steve has supported my musical ambitions, always asking about my latest projects and upcoming auditions. It's a tough business out there but my work in this department has grounded me and I'm extremely grateful for it. Not unlike my musical peers, Steve demands greatness and ceaseless effort not only from himself but from those around him. It serves to elevate everyone's work. He knows no other way – it's simply the way Steve gets things done. He cares tremendously about his field and it shows. It has been an immense pleasure and privilege to work alongside him.

–Kira Bokalders, Assistant Business Manager of Finance/Sponsored Research Manager, Dept. of Statistics and Data Science, Carnegie Mellon University

I am still working with Steve, so I create new memories every time we meet! But if I had to describe what makes him stand out as a professor it is that he treats me with respect. From the first time we spoke by email when I was applying to the PhD program, to our last few meetings, Steve has made it clear that he thinks my ideas are worth discussing, even when I had no credentials under my belt. He has argued with me, pointed me in the right direction many times, and most importantly, he has never told me what to do. The research I have done under his guidance has been mine, and I can never repay him for that.

After working with Steve I now know how to write better ("PGP! Particular-general-particular!"), I know that statistics is defined very broadly as the set of tools you need to answer interesting questions ("Statistics is what I do!"), and I know that I should not let bureaucratic rules guide my research, no matter how strict they might seem ("Don't think about the requirement for now. Think about the research."). I know it is ok to follow my heart in my research, and I have a better idea of how to do this.

If I am ever in the position of mentoring students, I hope I can be as respectful, available, and supportive of them as Steve has been of me.

–Maria Cuellar (PhD 17), Assistant Professor, Dept. of Criminology, University of Pennsylvania

*W*e have had so many great adventures together, over so many decades, that the beginnings of our friendship fade into the mists of my memory. Surely I knew you even before I first visited CMU in 1977? Certainly I already knew of you, through my immense admiration of your ground-breaking book "Discrete Multivariate Analysis" with Bishop and Holland. At any rate, from then on, in various visits I made to CMU (during which you showed me many great kindnesses), or in your visits to Europe, as well as at very many international meetings—especially Bayesian and Forensic—we were able to spend increasing amounts of time together, and interact, both professionally and personally, in most delightful ways. I particularly recall the 1991 Valencia Bayesian Meeting in Peñiscola, where we sallied forth every evening in search of a dinner rather better than available on site—and though it wasn't easy to meet your invariably high standards, we did more than well enough for mine.

Quite as much as I have always valued our friendship, I have had enormous respect for both the depth and the breadth of your fundamental contributions to Statistics. I honestly can't think of anyone who can match you on this. As just one indicator of your unparalleled versatility: I was recently at a workshop at the Isaac Newton Institute in Cambridge, which was running 3 parallel statistical programmes, on "Data linkage and anonymisation," on "Theoretical foundations for statistical network analysis," and on "Probability and statistics in forensic science." I saw there that you were listed as "Simon's Foundation Visiting Fellow" for every one of these three very different programmes—whereas no one else was involved in more than one. It is a great pity that you were not, in the end, able to come to any of these: I know just how much many others share my regret at missing the insight and clarity that you always bring to everything you touch.

Recently we have been collaborating fruitfully on topics in statistical causality, with a focus on legal issues. And I owe it to you that, through this, I was invited to serve on a MacArthur Foundation interdisciplinary research committee, looking into the tricky issue of making group data relevant to individual cases—an utterly fascinating experience.

Steve, it is indeed gratifying that your statistical acumen, prowess, originality and leadership have been very widely appreciated and frequently called upon. And in applying them to important matters of public policy you have made vital contributions to both national and international affairs. We all owe you so much, in so many ways.

I would love to have been able to join personally in the celebrations under way for you, but alas! can only participate telepathically. I will be thinking of you, and raising a glass to your illustrious career—and another, even larger, to the warmth of our friendship.

Your good friend and great admirer,

–Philip Dawid, Emeritus Professor of Statistics, University of Cambridge

I was Steve's PhD student from 1987 to 1990. Steve encouraged students to think big. When a famous statistician came to give a department seminar, he challenged me to "ask some tough questions." When I raised concern about an aspect of thesis research being ambitious, he said "I want you to be as ambitious as possible."

My interactions with Steve and other faculty, as well as the outstanding graduate training at CMU, have greatly benefited my career development. From my thesis work with Steve on modeling matching errors in capture-recapture census, I knew I could enjoy developing novel methods for real application problems, whereas pure theoretical work is not for me. Although still not a big risk taker, I learned to pick highly significant problems to work on in my field of RNA computational biology. My research program has been continuously funded by NSF and NIH since 2002, with total accumulative funding of nearly $8 million,

including a recent 4-year $2.2 million grant award from NIH. I congratulate Steve on reaching Emeritus status, and wish him the very best.

–Ye Ding (PhD 90), Research Scientist, New York State Dept. of Health

I started working with Steve a few months after becoming a graduate student in the department. Even after graduation, Steve continued to encourage and help me navigate the sometimes uncharted waters of academia for many years. I appreciate Steve not only as a mentor, but also as a person I could fully trust, and as an example that I constantly strive to follow. I am sure there are many PhD advisors that know how to train and guide their students well towards great careers. Among them, Steve stands out for the following key reason: throughout the years, he constantly kept in touch with me and with his other PhD students. He knew how to create a community of students whose shared research interests stimulated growth and professional development. I owe a lot to this man on a professional and personal level. He never hesitated to sacrifice his time to help me out when I needed it most. And I know he always did the same for his other PhD students.

–Adrian Dobra (PhD 02), Professor, Dept. of Statistics, University of Washington

I first met Steve and Joyce in 1979 at Las Fuentes in Spain. We've been good friends ever since. Steve immediately impressed me and continues to impress me with his tenacity. In those early years, the four of us had some fun trips together. I remember one night in Buenos Aires, Bill had the rigors (Bill had some kind of serious ongoing intestinal infection). I called Steve (across the hall) and he came over and sat on Bill's legs for quite a while to try to control those shakes. Most recently, because of their cancers and mine, I've been thinking a lot of Steve and my sister, Cathy. Now that I'm fighting cancer again, they both have inspired me as they have fought with great strength and tenacity. A few months ago my sister finally lost the battle after six-and-a-half years. I hope Steve can do even better.

Steve, it's wonderful to have you in Pittsburgh as a good friend. My love to you and Joyce.

–Connie Eddy, late wife of William F. Eddy, and close personal friend of Steve and Joyce Fienberg

*W*hen I visited graduate schools as an undergrad, I talked to many people. A large number of them told me to say "hello" to Steve! I realized then how many lives Steve had touched, including my own.

–Shannon Gallagher (PhD 19), Postdoctoral Fellow, National Institute of Allergy and Infectious Diseases

*W*hen I first met you in 1996 I was trying to figure out what I wanted to study in graduate school. I took your class on survey sampling. You took me on a trip in your sporty Mazda to Washington DC, and you introduced me to the protest songs of Pete Seeger. (I remember being surprised that there was a song with the word "garbage" repeated several times.) I'm quite sure I hadn't done anything to deserve such generosity but you took me (and we met up with another undergraduate, Larissa?) to DC to get a taste of what statisticians did for a living. There I was, a wide-eyed 21-year-old, being introduced to the many important people you knew in various agencies in DC. Between that and your encouragement to pursue statistics, I ended up doing a PhD in it even though I was a computer science undergrad. Turns out I really like the subject and couldn't have asked for a better profession. Thank you!! Now when I take time out of my schedule to work with and encourage undergraduates, I am reminded of you and Bill (Eddy). So the next generation of undergrads benefit from your generosity as well.

My gratitude and very, very best wishes to you today and always.

–Murali Haran (PhD 03, Univ. of Minnesota), Professor and Head, Dept. of Statistics, Pennsylvania State University

I will always remember the great parties Steve and Joyce hosted for our department. Many thanks to Steve and Joyce for your warm hospitality over those years. It was wonderful to all be together with faculty, staff, students and the kids. Fun times and great memories and, of course, I did take a lot of pictures.

Wishing you all the very best and in deep appreciation and gratitude.

–Cristina Ilangakoon (MS 96), independent consultant

*S*teve—not only are you a brilliant, nationally renowned statistician with an influential record of scholarly contributions, you are also a remarkable Carnegie Mellon citizen, mentor, teacher, and of course, a valued friend and colleague. Thank you for your good company and wise counsel over the last few years. May this next chapter hold many opportunities for relaxation, research and continued engagement with our campus community.

–Farnam Jahanian, President, Henry L. Hillman President's Chair, Carnegie Mellon University

*O*ne of the best moves I made as department head in the 1970s was to hire Steve. The final negotiating item was to find opportunities for Steve and his son, Anthony, to play amateur hockey. Although I knew nothing about hockey, I did manage to find such a hockey league, so Steve accepted.

Hockey was a personality theme for Steve in those days. He brought the same focus and attention to being our Department Head and, later, our Dean. He always worked hard, whether digging the puck out of the corner, or finding opportunities for his many graduate students. He made important contributions toward building the department to the hive of activity it is today.

Steve has also made very important contributions nationally, particularly in his work for the National Research Council, and as an editor.

He no longer skates as hard as he did, but he still has the fire in his eye to move forward on his many projects.

–Joseph B. Kadane, Leonard J. Savage University Professor of Statistics and Social Sciences, Dept. of Statistics and Data Science, Carnegie Mellon University

*T*hank you so much for contributing to the department that I came to regard as a second family. Your untiring efforts are what helped make it a great place to work. I was always so impressed by your many projects and, most especially, the amount of time you spent with your students, both grad and undergrad.

I did not do much work for you directly, but when I did you were generous with your thanks. When I worked for you on the Encyclopedia, you appreciated my efforts and made several very kind comments. You gave me a beautiful Cross gift set at the end of the project.

I will always be grateful for your understanding of my missteps.

I so enjoyed meeting your granddaughters and the sharing of our twin stories. Twins are a never ending source of joy and craziness. Their teen years will be so interesting!

Steve, I wish you all the best. Enjoy your special Event. You certainly deserve it.

–Rose Krakovsky, Retired Receptionist, Dept. of Statistics and Data Science, Carnegie Mellon University

*I*n my first semester, Steve thoroughly lost me with a lofty discourse on multivariate analysis. When I showed up at his office hours, he patiently explained the difference between the log and the logit. His response to my request to drop his course surprised me: "Just do your best and don't worry about it." The loglinear models he taught in that course became a central building block of a thesis draft I submitted to the committee less than four years later. Upon reviewing this draft, Steve was the loudest to object: "You've got more work to do." Kudos to Steve for guiding me to an ever-higher bar!

–Zachary T. Kurtz (PhD 14), Data Scientist, Argo AI

I first met Steve when I was working at Google. During our thirty-minute chat in a Google cafeteria, Steve told me about his vision of research, and how I could do research and interact with people at CMU. One week after that, I was no longer lost about my career plan, and I decided to leave my full-time job to join the CMU Statistics Department as Steve's postdoc.

Steve is a great mentor and collaborator. He has many amazing abilities, such as putting the right people working together, and finding deep insights beyond technical details.

–Jing Lei, Associate Professor, Dept. of Statistics and Data Science, Carnegie Mellon University

*F*einberg, Fineberg, oh it's Fienberg — got it. However, Steve is most definitely a fineburg, the finest in Pittsburg and well beyond. I don't remember where or when we first met in person, but am quite sure that my first contact with him was upon publication of BFH/DMA. Happily, both social and professional in-person contacts quickly followed and have enriched my professional and personal lives. Dinner in Leiden; floating in (almost on) the Mediterranean in Crete, working to make the streets safe for hospital comparisons via a COPSS white paper, collaborating on the Annual Review of Statistics and its Application; these are but a few highlights. Steve, enjoy your emeritus status (though your CMU colleagues report no detectable difference), and continue your stewardship of almost all things statistical. I treasure our relationship; our profession and country treasure your contributions.

–Thomas A. Louis, Professor Emeritus, Johns Hopkins Bloomberg School of Public Health

I met Steve just over 10 years ago, when he was chairing a meeting on problems in the forensic sciences at the National Academies. At the time, I was a journalist who'd just finished a documentary on that subject. Steve kept in touch with me over the years, answering my questions about statistics when I was reporting, even inviting me to my first ever JSM, in Salt Lake City. Eventually, I turned to him for advice when I decided to go back to school – and throughout my time as a graduate student. Like so many people, I was continually amazed by his ability to be 3 places at once. When I decided to come to CMU to this fall, my main fear was that Steve would realize I can't keep up with him. Fortunately, he's remained as gracious as ever. I count myself incredibly lucky to have the chance to work with him.

–Robin Mejia, Statistics and Human Rights Program Director, Center for Human Rights Science, Carnegie Mellon University

*I*t has been an honor, and an education, to work with Steve as we jointly created CMU's Center for Automated Learning and Discovery in 1997, and evolved it into today's Machine Learning Department. Throughout, I've been inspired by Steve's broad and deep understanding of science, and by his enthusiasm and energy for really making things happen. Congratulations Steve on this next step forward – I plan to still call on you for advice!

–Thomas Mitchell, University Professor, Carnegie Mellon University

I first became aware of Steve when I was an undergraduate at CMU, when he was Dean of H&SS, and then later left for Canada to be a super dean or something. My general understanding was that academics become deans when they are sick of research, and then they move off up the administration food chain. Thus I was surprised when I later became a graduate student, finding Steve back at CMU, and slowly discovering that he not only had *not* left research, but that he was also was a research powerhouse. Aside from being at the front edge of the discipline, whether in data disclosure or network analysis, he was a journal editor, a book series editor, a National Academy work horse, and all the while being a nurturing mentor. (I still remember two takeaways from notes he distributed on dissertation writing: "Write! You need something to edit," and on footnotes "Avoid them.

Footnotes are rat's nests for pedants.") How a person could carry out so much work with such vigor, and so much rigour, has always left me speechless.

–Thomas Nichols (PhD 01), Professor of Neuroimaging Statistics, University of Oxford

*S*teve, you have always been present for me: as an advisor, at my wedding(s!), and as a mentor and colleague. You have been both a professional and a personal role model. I and my family owe you a big debt of gratitude.

–Alessandro Rinaldo, Professor, Dept. of Statistics and Data Science, Carnegie Mellon University

*M*y most vivid memory of Steve occurred 11 years ago when we were working together to edit a book and had weekly meetings every Monday from 3 pm to 4 pm. When I committed to this joint project I hadn't fully recognized the meaning Steve gives to "every Monday." When there is work to be done, Steve is working, and so is everyone else in the vicinity. At the time we commenced the endeavor I was pregnant, and eventually I was very pregnant. Steve had even hosted a baby shower for me! We were almost done with the book when we met August 19, 1996. In fact we declared it temporarily our last meeting, since we were ready to send material to the publisher—one last meeting to tie up loose ends. That morning I was feeling kind of odd. By noon it was obvious I was in labor. But with Steve would you cancel a meeting for such a minor inconvenience as an impending birth? I thought no. The meeting proceeded as usual. By 9:45 pm, my daughter was born.

Thus I wanted to keep up with the high standards Steve has always set for hard work and dedication to purpose. An excellent example of Steve living up to these standards occurred when he was chosen to give the Fisher Lecture. I frankly told him, "Steve, please don't focus on Fisher in your lecture. It's been done to death and it's always boring." Not to be done in by my lack of enthusiasm, Steve worked for weeks to construct a brilliant story that meshed with all of his research initiatives and yet reflected on Fisher's principles. It was beautiful.

–Kathryn Roeder, University Professor and the UPMC Professor of Statistics and Life Sciences, Dept. of Statistics and Data Science, Carnegie Mellon University

I want to share the story of how I started interacting with Steve, because I think it shows how open and generous he is. Back in Colombia when I was an undergraduate student, I was involved in a project that had the goal of combining different sources of information to estimate the number of people who had been displaced due to the Colombian armed conflict. This is how I came across record linkage and capture-recapture methods. After learning the basics of these methodologies, I found that the literature didn't deal with some of the problems I was facing with my data. I realized that there was an author that kept coming up in the literature of these two areas: Stephen E. Fienberg. I decided to email Steve with my very rudimentary English of the time: "Dear Professor Fienberg, . . . " (Of course at the time I didn't know how famous and busy he was). I explained the problem I was working on and the difficulties that I was finding. To my surprise Steve was very prompt to reply, provided very helpful literature and references, and his email was quite extensive. Nevertheless, I realized that many of the issues that I was facing had not been fully addressed in the literature, so it seemed like this was an area that needed more research. Months later I emailed Steve saying that I wanted to do research on this, and that I wanted to "study a masters in capture-recapture estimation," to which he nicely replied that there wasn't such a thing and that I'd have to enroll in a PhD program. At the time probably not many of the things that I was writing made much sense, so I now realize how nice he was by taking me seriously and providing me guidance! Our email exchanges continued for around two years, until I finally enrolled in the PhD program at CMU, where I had the fortune of having him as my advisor. I will always be grateful for the trust that Steve has put in me and for all the support and mentoring that he has given me.

–Mauricio Sadinle (PhD 15), Assistant Professor, Department of Biostatistics, University of Washington

I first met Steve when he interviewed at CMU in 1980. Or was it Steve who interviewed CMU? Whichever it was, both were ready to join forces. And what a force Steve has been for our department. He led both us and the whole profession into the age of statisticians as scientific collaborators rather than consultants. When he left for a few years to become an administrator at another university, I felt that both the department and the profession had lost a valuable asset. Fortunately, he came back and led the drive to embrace the contributions of computer scientists to statistical methodology without losing our own identity. I feel honored to have been able to participate in two revolutions in our field, and particularly honored to have been a colleague of Steve's.

When Steve was department head and I was a junior faculty member trying to balance my roles as teacher and researcher, I remember Steve being a mentor in both endeavors. When students were not happy (euphemism) with my teaching, we went over my technique and the student comments and found ways for me to be a more effective teacher (more effective than before, that is.) He also joined with Morrie DeGroot to offer me opportunities to present my research at international meetings.

A distinct memory is of Valencia 7 in June, 2002. I was trying to sleep late one night in my hotel room but there seemed to be a loud party going on nearby. Was nobody planning to attend the early sessions the next morning? I began to think that I recognized one or more of the voices. So, I walked out on the balcony and there was Stephen and several former students in the next room behaving like Spaniards. Fearing that some trouble might arise, I decided that the best course was to take a chair next door and keep an eye on them. Did I mention that Stephen always seems to be able to locate the best wines no matter where he travels?

I owe a lot to Steve, and I want to thank him for everything that he did for me and my career. I also want to thank him for everything he did (and continues to do) for our department.

–Mark J. Schervish, Professor of Statistics Emeritus, Dept. of Statistics and Data Science, Carnegie Mellon University

*S*teve, please make sure to provide your cloning recipe for those of us forever learning to manage our time properly. Somehow you've managed to perfect the art of multitasking and time efficiency! I marveled at this skill all of those years ago when you first started serving as a faculty mentor for me, and I continue to marvel at your abilities to this day! Something tells me that attaining Emeritus status will not slow you down in the least, so feel free to continue mentoring me (. . . I'm just saying . . . I'm forever accepting tips; smile).

–Kimberly F. Sellers, Associate Professor of Statistics, Department of Mathematics and Statistics, Georgetown University

*S*teve, thank you for being my inspiration and role model in becoming a better researcher, advisor, colleague, friend and parent. You have motivated and inspired many careers but also touched many hearts. I would love to know your secret about how you manage to be a worldwide renowned statistician while still having time to take us all out for nice dinners and having time to put up with your 5+ grandchildren! In my travels, fun or work, you have been my super-powered connection to the world, from Buenos Aires to Tokyo! The name Steve Fienberg travels with the speed of light across the world. I am grateful to have had the honor to have you as part of my life.

–Nicoleta Serban (PhD 05), Professor, Georgia Institute of Technology

*I*n Spring 2001 I was lucky to co-teach with Steve the course "Sampling, Surveys and Society." Until then, I only taught courses on intro and industrial statistics to engineers. Co-teaching with Steve was a real eye opener about what statistics means when it's about humans. About us. Steve's teachings about IRB, ethics, confidentiality, and controversial use of data was fundamental—he was clairvoyant in seeing today's world of behavioral big data and the convergence of engineering with human and social data. I am so grateful to him for opening my eyes to the meaning of statistics beyond Greek letters and numbers. It

is this thoughtfulness and awareness that have led me to where I am today, a statistician in the world of behavioral big data.

– Galit Shmueli, Tsing Hua Distinguished Professor, Director, Center for Service Innovation and Analytics, College of Technology Management, National Tsing Hua University, Taiwan

*B*rother Steve—As I tally up, I think we've known each other for something over 40 years. With direct intersections via CNSTAT, SSRC, a nice paper on large scale social experiments, Elena Erosheva's thesis defense, and your kind appearance at a little Princeton celebration a few years ago, I regret that there wasn't much more face-to-face contact over the years. At this Emeritus juncture, I feel it a pleasure and an honor to be with you. While often out of sight, you have never been out of mind. On many dimensions, you have been a wonderful role model.

–Burton Singer, Emerging Pathogens Institute, University of Florida, Gainesville FL

*F*or three years I served as Steve's Associate Dean in the College of Humanities and Social Sciences. I had originally agreed to take on this post reluctantly for one year, since I was anxious to get back to my research after a stint as Acting Dean of the School of Urban and Public Affairs. The one year turned into three years, as I had the pleasure of working closely with Steve, observing his keen intelligence, his capable leadership as Dean, and his devotion to the college and to CMU. Steve could be tough, but he could also be compassionate and a wise decision-maker. It was a privilege to work closely with him.

–Joel A. Tarr, Richard S. Caliguiri University Professor of History and Policy, Carnegie Mellon University

*F*rom the day I met him as a grad student, I knew Steve had every skill I ever wanted to learn: incisive insights into every problem he found time to study; the ability to find more time to do things than anyone I ever met; the exact degree of patience required of someone who needed to grow or of someone who needed to be disabused of their wrongheaded thinking—all of which were on full display to those who attended that conference with us. I felt like I was meeting the first combination of true intellectual, decent human being and prizefighter I'd ever encountered, and to date he's still the only one I'd classify as such. I am continually in awe of his networking, his stamina, and his humour, and my life is far richer for having met and worked with him. I still hope I get the chance to skate with (or against) you in the near future.

–Andrew C. Thomas, Director, Data Science, SportsMEDIA Technologies (SMT)

I've never worked directly with Steve on a project, but I've definitely felt his presence and influence as my senior colleague. One of the things I admire most about Steve: he's always working hard to do good things for all those around him, i.e., his students, his post docs, and his colleagues at all levels. We are all beneficiaries of his great generosity in this regard, and I think we all owe him—in one way or another—a big "thank you" for his tremendous support over the years. So, Steve: thank you not only for your personal guidance from time to time, but for all your efforts that have brought so much success and happiness to CMU!

–Ryan Tibshirani, Associate Professor, Dept. of Statistics and Data Science, Carnegie Mellon University

*Y*ou are the most efficient and caring person we have ever known. Your positive attitude and forward looking spirit always served as excellent guidelines for us. Teresa and I were so lucky to start our career at Carnegie Mellon when you served as the department head. We learned so much from you, and always appreciate your help and encouragement. I still remember clearly that you drove me to a Bayesian conference held at the Ohio State University all the way from Pittsburgh in your RX7. We wish you the best and look forward to riding with you again.

–Ruey S. Tsay, H.G.B. Alexander Professor of Econometrics and Statistics, Booth School of Business, University of Chicago

I remember the first time I gave a talk about my research with Steve in the room. My nervousness subsided a few slides into my talk when I glanced at Steve, and he appeared to not be paying attention. "Whew, good thing I don't have to worry about answering a difficult question from him," I thought. You can imagine my surprise when, at the end of the talk, Steve immediately asked an extremely difficult and insightful question about my work. Somehow, he knew more than I did about my own work, and he had barely even glanced at the slides! The kicker was when I checked my email afterwards: I had two emails in my inbox from Steve, both of which were sent during my presentation.

–Sam Ventura (PhD 15), Director of Hockey Research, Pittsburgh Penguins

Steve is one impressive dynamo. Family, academics, travel, hockey, and what not. And then he and Joyce still find time to attend every performance of our fantastic Pittsburgh Symphony; orchestra row P. For others who might consider attending these concerts (and you should), the Pittsburgh crowd is a little noisy compared to what we expect in Europe: talking, coughing, snoring, and the dreaded candy wrapper are regular staples. And then you will hear pages being flipped, and you can blame that on one of two people: Mark Kanny, the art critic of the Trib, following the score, and our own Steve, editing students' papers and theses to beautiful tunes! (Apparently, he can also see in the dark.) I hope to hear this particular noise for years to come!

–Valérie Ventura, Professor, Dept. of Statistics and Data Science, Carnegie Mellon University

Chapter 33
Joyce Fienberg: A Woman of Grace and Wisdom

Gaea Leinhardt

Joyce Fienberg and I first met in 1968 in Cambridge Mass. Our respective spouses (Steve and Sam) were working at Harvard, and they knew each other. Joyce and I were friendly acquaintances through them. That was 52 years ago. After Joyce and her family moved to Pittsburgh in the early 1980s, we reconnected. Joyce had been volunteering at a local hospital. I managed to convince her that working for pay might be just as interesting, so Joyce started working with me on educational research projects at the Learning Research and Development Center at the University of Pittsburgh. We saw each other and talked at length five days a week for nearly 30 years. We talked about children, grandchildren, the details of research, the plans for the next set of studies, and a million other things. We were very close friends and colleagues. Perhaps the clearest signal of our relationship is the fact that Joyce and Steve agreed to be the guardian for our daughter Zoe should the need arise.

Over the 30 years we worked together, the research projects focused on the nature of learning and teaching in a variety of subject areas (mathematics, history, geography, and chemistry). Joyce played a critical part in all of the work that we did. Her unique and gracious personality was the glue that kept things on track and helped to maintain a steady flow of communication. The research projects usually involved going into classrooms to observe and then interview. In mathematics classes, for example, we would interview the teachers before a particular class, videotape the class, and then interview the teacher as they watched the tape of the class afterward. This procedure would go on for at least a month of teaching. We would also interview the students at various times during the overall study. Sometimes we would sit next to a student and interview them during class to understand how well they were following the class in real time. It took a careful

G. Leinhardt (✉)
Professor Emeritus, University of Pittsburgh, Ames, IA, USA
e-mail: alicia@iastate.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_33

balance to be able to do those in class interviews without disrupting the class as a whole. Joyce was always very sensitive to the needs and concerns of the teacher and the students. That sensitivity translated into good decisions about how to treat the situation. Joyce carried that same level of sensitivity through to the work and friendship with doctoral students and post docs, both those who worked with Steve and those who worked with me.

Starting in 1998, our research migrated to museums where we investigated the nature of learning in informal settings. Joyce would observe and interview small groups of visitors. Joyce was wonderful at these tasks because she was so honest, sincere, and caring. This sense of real engagement with whomever Joyce was talking to was deeply appreciated by visitors, teachers, and students. It led to important understandings about the nature of the learning that was going on regardless of setting. But Joyce was not only a warm and gentle person, she had a vibrant sense of humor and was surprisingly forceful when she felt it was necessary. For example, another aspect of her work with me was to act as gatekeeper in chief on research papers leaving the project – almost nothing went out for publication from the project without her demanding edits. There was nothing soft and fuzzy about those edits! Everyone on the project would hand her their papers with some trepidation. No run on sentences slid by, no unintelligible metaphors were allowed, and no flowery titles that didn't do their job stayed. Joyce knew the rules but she also sensed the nuance and recognized the purposes of writing; she herself was a lovely, clear, and elegant writer; it was always a joy to write with her. It was my good fortune to have had such a friend and colleague for more than 30 years.

Chapter 34
Obituaries

Margaret Smykla

On Saturday, October 27, 2018, Joyce Fienberg, a cherished member of the Carnegie Mellon Department of Statistics & Data Science community, was killed inside the Tree of Life Synagogue after a gunman opened fire.

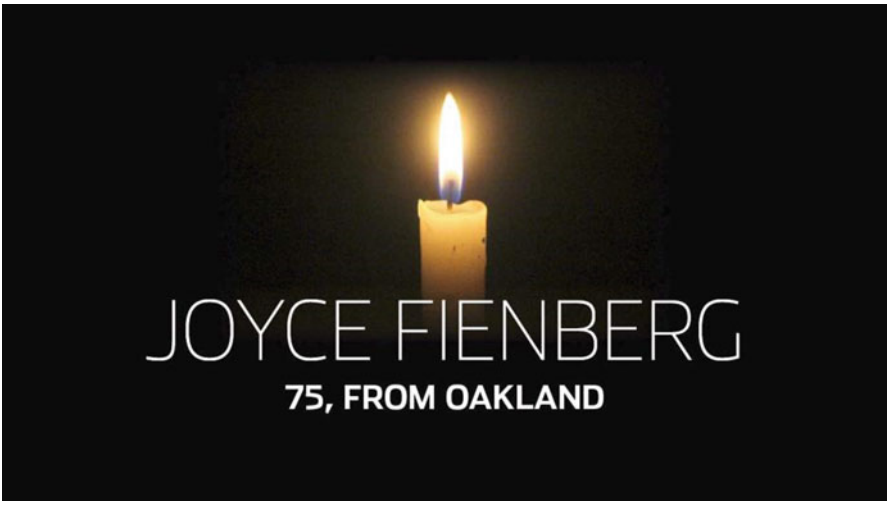
Besides being the widow of Stephen E. Fienberg, a renowned university professor of statistics and social science, Joyce was a beloved mother to two sons and a grandmother to their six children; a retired research specialist; a dedicated congregation member and volunteer; and a treasured friend to countless department members/visitors and their families.

This obituary by Jamie Martines appeared in the *Tribune-Review* on October 28, 2018. The piece that follows, "Tree of Life shooting victim Joyce Fienberg 'will continue to watch over us,'" is by Nathan Duke and appeared in the *Tribune-Review*

M. Smykla (✉)
Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: mk74@stat.cmu.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_34

on October 31, 2018. It was part of a series of profiles of people killed in the Tree of Life attack.



Synagogue shooting victim Joyce Fienberg was a "driving force" of good

Personalized, handwritten holiday cards, beautiful dinners, and thoughtful advice: This is what Joyce Fienberg shared with the graduate students she welcomed into her home, said Aleksandra Slavkovic, now a professor in the statistics and public health sciences departments and associate dean for graduate education at Eberly College of Science at Penn State University.

Slavkovic described Fienberg as an esteemed social scientist, as well as an elegant, magnificent person who was "kind beyond belief."

Fienberg, 75, of Oakland, was one of the 11 people killed Saturday during a shooting at Tree of Life Congregation in Squirrel Hill. She was married to the late Stephen E. Fienberg, a professor of statistics and social science at Carnegie Mellon University. They had two sons, Anthony and Howard.

Slavkovic met Fienberg through her husband, who served as Slavkovic's adviser while she completed her doctoral studies at Carnegie Mellon. Their home was open to all of the students he mentored, Slavkovic said.

"That's just goodness, and no limit to what they were willing to offer and show kindness to people," Slavkovic said of the Fienbergs. "I felt Joyce was really a driving force to that."

Their home was warm and inviting – a feeling Slavkovic attributes to Fienberg.

"It didn't matter what race, religion, ethnicity you were," she said. "Their home was open to you. ... A big part of this was because of Joyce."

Fienberg earned her degree in psychology at the University of Toronto, where she was a student research assistant in social psychology, according to a post on Facebook from the Learning Research and Development Center. She later worked with children with emotional and behavior needs at a residential treatment center.

Fienberg went on to work as a research specialist at the Learning Research and Development Center from 1983 until she retired in 2008.

"We worked full work weeks, and she managed to be a fabulous mother," said Gaea Leinhardt, a professor at the University of Pittsburgh's Learning Research and Development Center, where Fienberg worked as a researcher.

She and Leinhardt worked together on several classroom-based research projects.

"Joyce was a magnificent, generous, caring and profoundly thoughtful human being," said Leinhardt, who regularly saw and spoke with Fienberg since her retirement.

Leinhardt met Fienberg in 1968 in Cambridge, Massachusetts. They worked together on many projects related to teaching and learning, as well as projects that were part of the Museum Learning Collaborative, studying how people learn from visiting museums. Leinhardt described Fienberg as an "unbelievably warm person" who was an outstanding mother and adoring grandmother.

After retirement, Fienberg spent time volunteering at the Allegheny County Courthouse in Downtown Pittsburgh, Leinhardt said.

Tree of Life shooting victim Joyce Fienberg "will continue to watch over us"

Joyce Fienberg traveled the world during her lifetime, but she couldn't dream of living anywhere other than her beloved Pittsburgh.

The Tree of Life Congregation, "became a refuge for her," her son told mourners Wednesday during a funeral.

"Mom spent an inordinate amount of time worrying about other people's needs, not hers," Howard Fienberg said.

Hundreds of people turned out Wednesday morning for Joyce Fienberg's funeral at Beth Shalom Congregation on Beacon Street. She was one of the 11 victims in Saturday's shooting at the nearby Tree of Life synagogue in Squirrel Hill.

Outside of Beth Shalom, family members and friends hugged, while hundreds of attendees – who traveled from neighboring communities, New York and overseas – flooded into its second-floor synagogue. Haim Korsia, the chief rabbi of France, attended along with rabbis from Washington, D.C. and other locales. Tree of Life Rabbi Jeffrey Myers was also in attendance.

Fienberg, 75, who was born in Toronto, worked as a researcher at the University of Pittsburgh's Learning Research and Development Center from 1983 until her retirement in 2008. Her husband, Stephen Fienberg, was a statistician at Carnegie Mellon. He died in 2016.

Her family members recalled her as a dedicated attendee at Tree of Life who always put others' needs ahead of her own.

"My heart hurts – my sister is dead, my sister was murdered," said Fienberg's brother, Robert Libman, of Toronto. "She was my role model all my life. Her thoughts and words were of pure intent. She gave life to all she came into contact with. In the end, a life of pure giving is a life well lived. In her death, we owe it to ourselves to try to measure up. Evil tries to shut off the light, but the light refuses to be dimmed.

"The light is still in our hearts – even our broken hearts."

Anthony Fienberg, one of Joyce's two sons, traveled with his family from Paris to attend the funeral. He and his brother, Howard Fienberg, stood together at Beth Shalom's podium while speaking of their mother.

"She will continue to watch over us," Anthony Fienberg said. "Those who crossed her path, it may have been a blessing to them. We miss [her] terribly, but have already started to build on her legacy."

Howard Fienberg, who lives with his family in Virginia, said that both he and his brother had attempted to convince their mother to move closer to one of them after her husband died. But she told them that she planned to remain in Pittsburgh, where she attended Tree of Life and had many long-time friends.

Fienberg's sister-in-law recalled a safari that she took with Joyce and Stephen, while Devorah Kurin – Fienberg's niece – said that her aunt's generosity was unmatched.

She told a story of how her aunt once not only bought season tickets to the Pittsburgh Symphony Orchestra for her and her husband, but she also offered to babysit during every performance, helped her to secure a parking space and sent directions and instructions on what time she should leave to make it to the performance on time.

"She cared about everybody," Kurin said. "She cared and acted on it. She was a doer."

Jonathan Greenblatt, the CEO and national director of the Anti-Defamation League (ADL), said that the ADL has seen a large increase in anti-Semitism over recent years.

"It's an incredibly sad day for the Jewish community of Squirrel Hill," he said outside of Beth Shalom. "Anti-Semitism has been a persistent problem for a long time. It's called the 'oldest hatred.' Our elected officials and political candidates need to stop giving in to their worst impulses."

Rabbi Daniel Yolkut, of the nearby Poale Zedeck synagogue, said that he attended the funeral service to show support for the Tree of Life community and the families of the shooting victims.

"We were all under attack last Saturday," he said. "It could just as easily been one of us. We are no strangers to acts of terror. From across the country, Jews and gentiles have been dropping everything to come to Pittsburgh to help people they never met."

Chapter 35
Reminiscences of Joyce

Margaret Smykla

On Friday, October 26, 2018, Joyce Fienberg attended the inauguration of Carnegie Mellon's tenth president, Farnam Jahanian.

Afterward, she joined the department members at a festive campus-wide picnic, during which she talked about family, the university, and more while enjoying entertainment provided by the Pipe and Drum Band, the Tartan Tuba Band, and CMU Bhangra.

Joyce was especially excited about the "Steve Fienberg Memorial Lecture Series in Advanced Analytics" that Carnegie Mellon's Heinz College of Information Systems and Public Policy was hosting on November 5–8. She told numerous friends and acquaintances she looked forward to attending all of the talks.

Less than 24 hours later, Joyce was dead; 1 of the 11 victims of a gunman in the single worst attack on American Jews in the US history.

Shocked and saddened by Joyce's senseless passing, the Statistics and Data Science faculty, alumni, and friends paid tribute by sharing their memories of a lovely and caring woman whose legendary kindnesses, especially to graduate students from distant lands, made a profound and lasting impact in the lives of everyone with whom she crossed paths.

Their reminiscences are listed in alphabetical order by the author's last name.

> *I* grew up thinking that women who were in their husband's shadow were weak. Then I met Joyce and had a complete change of attitude. Joyce was a very competent and impressive woman in her own right who was also absolutely comfortable at times dedicating herself to supporting Steve. It seemed she was able to find her own value in enabling him to achieve his greatest accomplishments. Unbeknownst to her, I've leaned on the lessons she taught me many, many times, and will continue to do so.
>
> *–Kelly Black, President and CEO, Neptune and Co., Inc.*

M. Smykla (✉)
Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: mk74@stat.cmu.edu

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_35

*I*t is heartbreaking to know that Joyce is no longer with us. In my mind, Joyce—the person, the citizen, the mother, and the grandmother—was as far away from violence, terror, and senseless acts as one could be. Completely different words come to my mind when I think about Joyce.

Elegant. The Fienberg family has a photo of her taken by a professional photographer in Toronto when she was a young lady. That photo, which had been published in a local newspaper, and which I have seen on several occasions when I visited Fienberg's home, is engraved in my visual memory. Joyce was amazingly beautiful on that photo, and she remained elegant on every occasion.

Welcoming, warm, and friendly. Joyce was a warm and kind hostess. She had this rare ability to make everyone feel welcome and comfortable in her home, whether it was a new student whom she met for the first time (as was the case for me) or an old family friend. My persona wasn't special. Fienberg's home was open to anyone, no matter the ethnic origin or nationality. And, even though I have never discussed the topic with Joyce, I am sure that she was behind a lion's share of planning, organizing, and executing in order to make their home a welcoming and friendly place, whether the end result was hosting a young faculty member from another country or having students for a sit-down dinner around Thanksgiving. Whenever I casually asked her a question, Joyce was determined to give the best advice possible, even if it meant asking several people to find out—"let's see how the network unfolds."

Interpersonal. Joyce was not only a proud mother of two sons and a grandmother to six grandchildren, but somehow she managed to keep track and remember names of spouses and kids of the many students that her husband had over the years, and send holiday greetings every year. I was looking forward to each year's card, which was usually a reproduction of a tastefully chosen artistic masterpiece and a letter with family news and pictures of the grandkids and others, but almost never pictures of herself. I do not remember exactly when my family started getting these cards, perhaps after graduating, but there was only one pause for a year two years ago when her husband passed away. Joyce shared with me that the letter was written just before Steve's health deteriorated, the cards were not sent out that year, and that she was not sure if she wanted to include it in the next year's card . . .

Perhaps jokingly, but most likely not, Joyce once described herself in an email as a worrier. Indeed, looking back, Joyce seemed to always be worried about others, big things and small. She worried if there were the right kinds of things for breakfast for visitors—"Coffee? Orange juice? Other juice? Toast? Bagel? Cereal?" She was deeply concerned about the impact that losing an advisor prematurely could have on career trajectories of Steve's younger mentees—"What research paths to follow, what conferences to attend, what people to talk with, what other universities have promising programs that one might want to join, what publication outlets are best, where to apply for grant money, what terminology to use in such grant applications, which references to invoke in support of the research idea, etc." And, she was worried about increasing violence around the world . . .

Interpersonal, welcoming, warm, friendly and elegant. I will miss her.

–Elena Erosheva, Professor of Statistics and Social Work, University of Washington

*J*oyce opened her home and her heart to so many of Steve's students. I remember her warmth and kindness both during my time at CMU and after. Just before I defended, after Kevin and I became engaged to marry, she gave me a beautiful tea party bridal shower. It was so sweet, but also had the understated air of sophistication that Joyce seemed to do so well.

–Mary Fowler, Professor of Mathematics, Worcester State University

*E*very year, Joyce and Steve invited students over to their house for Passover. They were so welcoming to me when I first moved to Pittsburgh and later to Jen and I, inviting us into their home to celebrate with their family. It was the first Seder Jen had ever attended. Joyce was incredibly helpful and kind to Jen, making her feel welcome and at ease. I will never

forget those holidays and their kindness. May her memory be for a blessing and may we all strive to show the kindness and love to our neighbors that Joyce and Steve embodied.

–David Friedenberg, Senior Research Scientist, Battelle Memorial Institute

I met Joyce during the sitting shiva period for Steve Fienberg. She was very kind and shared memories of Steve over photo albums full of photos of them traveling together on conferences. Seeing them having had a happy, loving family and academic career together left a big imprint in my life goals. I am grateful to her, and Steve, for that.

–Sangwon (Justin) Hyun, Postdoctoral Research Associate, Dept. of Data Sciences and Operations, University of Southern California

I first met Joyce in August of 1994 as a new master's student in the Statistics Department. Joyce's husband Steve was my adviser and I was also a TA for Steve's intro to stats class. My remembrance of Joyce was always her very warm smile and welcoming nature. Back in those years we had only about 30 graduate students so it was always like one big family. Pretty much every weekend we had a gathering at either a student or faculty home. Though my time at Carnegie Mellon was short as I was only there for my Master's program, the friends that I made have been lifelong. I always took many pictures.

So as I sifted through my old albums I was able to find some lovely pictures of Joyce's warm smile. I am thankful that I did take those pictures. I found pictures of Joyce at various gatherings including when Steve and Joyce had us TA's over for dinner. Joyce was always such a gracious host and genuinely cared about us students. I will always cherish the annual holiday letter I received from Joyce over these 20 plus years telling me all about Howard and Anthony and their families, and all the fun family trips and above all their love for their grandchildren. I would also let Joyce know how things were going on at my end and we continued to stay in touch. As the year draws to a close and the holidays get closer, Joyce will be in my thoughts and I will always cherish those letters I received over these many years.

Joyce, may you rest in peace.

–Cristina Ilangakoon, Independent Consultant

I saw Joyce a few times, including a party in her house in 2016 a couple of months before Steve passed away, at which time I brought my 9-year old daughter with me. Joyce was a very nice and friendly person. My daughter vividly remembers how kind Joyce talked to her and offered her toys during the party. My daughter was very sad when she heard the news about Joyce the other day. Joyce liked travels, and she traveled with Steve to many places. I vividly remember one time Joyce and Steve and I and several others dined at a restaurant. That night, Joyce told a story about a dinner party she and Steve had a few years back with colleagues in Taiwan during a visit there. Obviously, Steve impressed the host by showing that he can drink a great deal of wines and liquors, and won the title of "ocean capacity" (a Chinese saying meaning that a person can drink a lot without getting drunk). The Fienbergs were very happy about the title of "ocean capacity."

–Jiashun Jin, Professor, Dept. of Statistics and Data Science, Carnegie Mellon University

I enjoyed being with Joyce at Bayesian conferences. Although I was a brought up in the Bayesian tradition in the Statistics Department, my career went in the direction of epidemiology and clinical trials. So after one or two Bayesian talks, I became a supportive spouse (to George Duncan) and spent time with Joyce. In Dublin after tea at Bewleys on Grafton Street we went looking for a synagogue where she had a friend of a friend and to a hospital where I had a friend of a friend. It was much better to be a supportive spouse in Peniscola, Spain, since many of those who attended all the sessions developed a strange flu-like malady (Bayesian flu). In Italy near Bologna Joyce and I drank coffee and talked about family, Pittsburgh and religion. Joyce was a gracious host and George and I attended

many parties and receptions on Dunmoyle St. I will remember her spark, her generosity and charm.

– Sheryl Kelsey, Professor Emerita of Epidemiology, University of Pittsburgh

*J*oyce used to send me new year photo cards with letters talking about lives with Steve and her grandchildren. Her photo cards were beautiful, reminding me of the streets in Pittsburgh covered with snow. This made me wait for new year cards from her. She seemed to have got well over the sadness and loss last year since Steve passed away. Her family members were all with her. I was grateful for that. I don't know what to say to her family members, in particular to her children. I wish her loss should be filled with good memories of her in their minds.

–Sung-Ho Kim, Professor of Statistics, Korea Advanced Institute of Science and Technology

*T*his has been a difficult time for Pittsburgh. Something like this should never have happened. I never met Joyce Fienberg, but knew Steve. When Steve died in 2016, Joyce donated hundreds of books to the Library of Alexandria in Egypt to help build research equity in Arab and African countries. It seems very sad that Mrs. Fienberg has helped so many young people in Arab countries, and something like this happens.

–Ronald LaPorte, Professor Emeritus, Dept. of Epidemiology, University of Pittsburgh

I didn't know Joyce as well as did Steve's advisees, who were invited often to events at his home. But Steve often spoke of her, as people do of their spouses. The way he talked about her made it obvious how much he cared for her. My best memory of Joyce comes from our daughter Allison's birth during my last year as a PhD student at CMU. Joyce offered to lend us their crib, which, she said, she would not be needing until their grandchildren arrived. As I hear stories from other people who knew Joyce better than I did, I realize how characteristic this generosity is of Joyce. I am shocked and heartbroken that this horrific event has befallen my hometown and my old neighborhood, a mere half mile from where we lived when Allison was born. It gives me a tiny bit of comfort to know that my daughter slept in the same crib as Joyce's sons and grandchildren.

–Kathryn Laskey, Professor of Systems Engineering and Operations Research, George Mason University

I had the pleasure of spending some time with Joyce at occasional social events while I was Steve's student in the 1990s. Despite only a handful of interactions 20 years ago, I still vividly remember Joyce as vibrant, with a great sense of humor, who was excellent at eliciting interesting personal anecdotes from others. I can still visualize her miming smoking in a very exaggerated sophisticated manner as she explained to graduate students that she was never cool enough to pull off being a smoker.

I felt I knew Joyce even better than suggested by these brief experiences because of stories Steve would tell about her. I remember him reminiscing about bringing home sandwiches when she was pregnant with their first child, how he had been careful to bring her something bland to avoid upsetting her stomach, but had not thought about his own pungent sandwich, which sent her careening for the bathroom as soon as he unwrapped it. For the life of me I cannot imagine how this came up in conversations between a statistics professor and a dissertation student, but the fact it did is just evidence of his devotion to her and that little things would bring to mind his experiences with her.

Since hearing the tragic news this weekend, I find myself looking at a portrait of my daughter (now 17) taken when she was about a year old, which sits on my dresser. The outfit she is wearing was a baby gift from Joyce and Steve. Even though she only knew me a bit as one of Steve's students, it was clear that she had taken care to select an outfit that was unique and special, yet practical and comfortable. So when my in-laws scheduled my daughter for portraits, and went looking through her wardrobe, that is the outfit they selected.

Joyce was the type of woman who made a memorable impression even on someone like me who did not have the benefit of a longer, closer relationship with her. I can only imagine what her loss means to those who knew her well, and send my deepest, most sincere condolences.

–Terra McKinnish, Professor of Economics, University of Colorado

*M*any years ago Joyce hosted a baby shower for me, even though we hardly knew each other. She knew that I was new in town, and a woman in a man's world without a support group. Because she was generous of heart, she stepped up to the task and extended her hand in friendship.

–Kathryn Roeder, University Professor and the UPMC Professor of Statistics and Life Sciences, Dept. of Statistics and Data Science, Carnegie Mellon University

*J*oyce was a gem! I remember her and Steve having my family over for dinners when Richard was a baby. They would gush over him as if he were their own, and they spoke so fondly and lovingly of their own grandchildren. Fast forward all of these years later . . . I got to see, hug, and converse with Joyce at the JSM 2018 memorial session in Steve's honor. She asked me about Richard and couldn't believe that he was now 16 years old! I shared pictures of him, and she marveled over how much he'd grown and matured. It was so wonderful to see her! I will forever cherish that time together.

–Kimberly F. Sellers, Associate Professor of Statistics, Department of Mathematics and Statistics, Georgetown University

*W*hen my husband Boaz and I moved to Pittsburgh from Israel, Joyce and Steve graciously hosted us at their home on holidays, introducing us to their family and friends. Joyce was such a warm and loving person, always with a sincere broad smile even to newcomers she hardly knew! She took interest in our lives and shared hers. She had an incredible ability to start and maintain genuine and close relationships over long periods of time. Since we left Pittsburgh in 2002, every Hanukkah we'd have a delightful Happy Hanukkah card from Joyce and Steve, surprising us in the various remote locations we lived in. Joyce wrote long letters in these cards, sharing the Fienberg family's travels, events, joys and sorrows in the recent year. Joyce would make us feel part of her family through these wonderful, thoughtful, and lively cards. She even sent a card last Hanukkah, with the sad news of Steve's passing away, and we were glad to hear that she'd continued her visits with family and friends, her loving spirit still kindled.

We are horrified and deeply sad about the violence and hatred that caused her death: the exact opposite of Joyce's extraordinary gentle, loving, and caring spirit.

–Galit Shmueli, Tsing Hua Distinguished Professor, Director, Center for Service Innovation and Analytics, College of Technology Management, National Tsing Hua University, Taiwan

*W*e were on our way to Hong Kong when at the airport in Washington we heard about the senseless event at the Synagogue. We were concerned because I knew that many colleagues at CMU were members.

Joyce simply radiated all the humanity that a human being can possibly possess. Both Joyce and Steve visited us in Hong Kong several years ago and we experienced their delightful company. That was the last time we saw them together.

Joyce's disposition and kindness are legendary. Indeed, what I remember often is that when I was confined in bed due to an automobile accident, Joyce visited us to cheer me up with a ton load of goodies. It made for a lasting impression on us, and Norah always made an effort to get together with Joyce whenever an opportunity arose; they were close friends. Violence is something we feel happens to others. But this time it is humanity itself, that has been thoughtlessly struck by it!

–Nozer (and Norah) Singpurwalla, Emeritus Professor, The George Washington University

*M*y last email exchange with Joyce was exactly three months prior to her tragic passing. I smile and cry every time I read that now, as it reminds me of her kindness, thoughtfulness, and grace, of her attentiveness and openness, of how special she and Steve were to me, but also of how precious our time together really was. After missing out on meeting Joyce in Vancouver at the end of July, I was hoping to see her in early November to hear more about her volunteering, to see her smile when she speaks proudly of her grandchildren, to reminisce about my sojourn in their home on Dunmoyle St., or a beautiful flamenco dancer we saw in Minneapolis, or hers and Steve's visit to Penn State . . . I miss her, and Steve, dearly.

–Aleksandra (Sesa) Slavkovic, Professor, Departments of Statistics and Public Health Sciences, Associate Dean for Graduate Education, Eberly College of Science, Pennsylvania State University

*J*oyce Fienberg radiated warmth, grace, and kindness. She had an instinctive way of putting people at ease, even on first acquaintance. When I reflect on the times I spent in her company over the past 35 years, two distinct memories span that period.

During my graduate studies at CMU, I received an invitation from Joyce and Steve to attend their youngest son Howard's bar mitzvah. I felt honored they had reached out across lines of faith and asked me to be present for this significant event in their family life. Joyce organized the day beautifully, ensuring that those of us unfamiliar with Jewish religious traditions would be comfortable during the ceremony and subsequent celebratory meal. With this thoughtful gesture of inclusion, Joyce and Steve broadened the life experience of a young man from a small-town Christian upbringing and served by their example as role models on how to engage the world with constructive, positive energy.

Three months ago, at the Joint Statistical Meetings in Vancouver, I attended a special memorial session organized in Steve's honor. During the scheduled talks I was delighted to notice Joyce in the audience, and even more pleased when she ascended the podium to offer some concluding thoughts. She spoke with composure, clarity, and gentle humor. Her closing words were that Steve had had a great life, and she had been privileged to share so much in that life. When I greeted her after the session and asked about her days now, Joyce paused long enough to acknowledge the weight of loss while she continued to look forward. Now those of us who knew Joyce must confront another, less comprehensible loss. In these sad days, I find some comfort in the belief that, if asked directly, Joyce would have expressed that she, too, had had a great life. We will honor her memory best in embracing and advancing the values by which Joyce lived so well.

–Duane Steffey, Principal Scientist and Director, Statistical and Data Sciences, Exponent

I am profoundly saddened by the news of Joyce's passing away. In my memory, Joyce was a vivacious, kind and caring person who, together with Steve, loved to share the joy in their life with the others. I remembered vividly once Joyce and Steve threw a wonderful party, showcasing the newly modeled bathroom to everyone. It was such a fun and joyful time! I will miss Joyce. May she rest in peace in heaven!

–Feng Tang, Statistician, Medtronic

I remember Joyce as very kind and welcoming, and am devastated that she died this way. My one specific memory about her is the same as what I said about Steve for his emeritus celebration—that she and Steve invited my husband and me to Tree of Life for high holiday services and hosted us at their home to break the Yom Kippur fast—a really lovely thing to do for people who are new to a place.

–Norma Terrin, Professor, Tufts University School of Medicine

*J*oyce sent me a beautiful Christmas card the year after Steve passed away and thanked me for seeing Yu-Xiang Wang (Steve's student, who I advised after Steve passed away) through to graduation. We ended up emailing a bit back and forth and talked about family—

coincidentally, her sister-in-law knows my mom quite well (they went to school together). I was just struck by what a sweet and kind woman she was.

–Ryan Tibshirani, Associate Professor, Dept. of Statistics and Data Science, Carnegie Mellon University

Chapter 36
Photos

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_36

Fig. 36.1 Steve and Joyce's wedding; Toronto, late 1960s

Fig. 36.2 Steve at the Tanur's house in Montauk; Summer, 1989

Fig. 36.3 Left to right: Joyce, Bill Eddy, Connie Eddy, and Steve Fienberg; Spring, 1996

Fig. 36.4 Gathering at the Fienbergs' home with statistics graduate students; Spring, 1996

Fig. 36.5 Left to right: Joyce, Steve, Cristina Ilangakoon, and son Howard Fienberg; Cristina was graduating with her master's degree; Spring, 1996

Fig. 36.6 Left to right: Steve, faculty John Lehoczky, Cristina Ilangakoon, and faculty Jay Kadane; Spring, 1996

Fig. 36.7 Left to right: Alicia Carriquiry, Steve, Julia Mortera, and Dale Poirier in Cape Town, South Africa; Fall, 1996

Fig. 36.8 Stephen Stigler and Steve, 2003

Fig. 36.9 Steve, son Anthony Fienberg, and Joyce; Paris, November 2003

Fig. 36.10 Joyce and Steve at the wedding of the daughter of Gaea and Sam Leinhardt; 2003

Fig. 36.11 Joyce and Steve in Nice, France; February 2004

Fig. 36.12 Joyce in Italy; July 2005

Fig. 36.13 Joyce and Steve on the Mt. Washington overlook in Pittsburgh, 2005

Fig. 36.14 Minneapolis, February 2007

Fig. 36.15 Joyce's 65th birthday party, Pittsburgh; February 9, 2008

Fig. 36.16 Alum Sung-Ho Kim, Joyce, and Steve in a Pittsburgh restaurant when Sung-Ho visited in April 2008

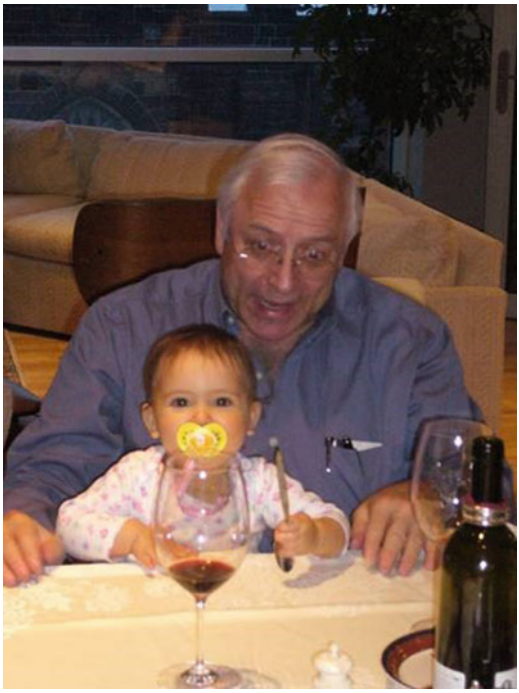


Fig. 36.17 Steve and granddaughter, Sophia, in Steve/Joyce's apartment; April 2009

Fig. 36.18 National Academies' annual gala, date unknown.

Fig. 36.19 Steve and granddaughter, Sophia, in Manchester, NH, at Steve's sister-in-law's adult bat mitzvah; June 1, 2013

Fig. 36.20 Steve, granddaughter Sophie, and daughter-in-law Marnie skating at PPG Place in Pittsburgh; November 2013

Fig. 36.21 Steve with sons Howard and Anthony in Lake Forest, IL, at a family wedding on August 15, 2015

Fig. 36.22 Steve with his five French grandchildren. From the left: Selena, Tiffany, Adam, Victoria, Steve, and Juliana; Pittsburgh, Summer 2015

Chapter 37
Stephen Fienberg's Doctoral Advisees and Co-Advisees by Institution and Graduation Year

Shelby J. Haberman	University of Chicago	1970
F. Kinley Larntz	University of Chicago	1971
(Timothy) Tar Chen	University of Chicago	1971
Howard Hochman	University of Chicago	1971
Ying Shang Lin	University of Minnesota	1975
S. Keith Lee	University of Minnesota	1975
Stephen S. Brier	University of Minnesota	1979
Michael M. Meyer	University of Minnesota	1981
Diane G. Saphire	Carnegie Mellon University	1983
Elizabeth A. Stasny	Carnegie Mellon University	1983
Sherryl May	Carnegie Mellon University	1984
Syni-an Huang	Carnegie Mellon University	1984
Mary Santi	Carnegie Mellon University	1988
Arieh Epstein	Carnegie Mellon University	1989
Sung-Ho Kim	Carnegie Mellon University	1989
Yi Ding	Carnegie Mellon University	1991
Terra G. McKinnish	Carnegie Mellon University	1999
Daniel Cork	Carnegie Mellon University	2000
Jan A. van den Brakel	Erasmus Universiteit Rotterdam	2001
Adrian Dobra	Carnegie Mellon University	2002
Stella Maris Salvatierra	Carnegie Mellon University	2002

(continued)

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0_37

Elena Erosheva	Carnegie Mellon University	2002
Amelia Haviland	Carnegie Mellon University	2003
Mario Trottini	Carnegie Mellon University	2003
Ana Maria Araneda	Carnegie Mellon University	2004
Aleksandra Slavkovic	Carnegie Mellon University	2004
Sinjini Mitra	Carnegie Mellon University	2005
Alessandro Rinaldo	Carnegie Mellon University	2005
Ana Maria Sfer	Universidad Nacional de Tucuman	2005
Edoardo Maria Airoldi	Carnegie Mellon University	2006
Jason Connor	Carnegie Mellon University	2006
Justin Gross	Carnegie Mellon University	2010
Daniel Heinz	Carnegie Mellon University	2010
Daniel Manrique	Carnegie Mellon University	2010
Anne-Sophie Charest	Carnegie Mellon University	2012
April Galyardt	Carnegie Mellon University	2012
Di Liu	Carnegie Mellon University	2012
Robert Hall	Carnegie Mellon University	2012
Seungil Huh	Carnegie Mellon University	2012
Fei Yu	Carnegie Mellon University	2015
Mauricio Sadinle	Carnegie Mellon University	2015
Xiaolin Yang	Carnegie Mellon University	2015
Jana Asher	Carnegie Mellon University	2016
Kirstin Early	Carnegie Mellon University	2017
Yu-Xiang Wang	Carnegie Mellon University	2017
Maria Cuellar	Carnegie Mellon University	2017

Index

A. L. Carriquiry et al. (eds.), *Statistics in the Public Interest*, Springer Series in the Data Sciences, https://doi.org/10.1007/978-3-030-75460-0

C