Statistics for Social and Behavioral Sciences

For other titles published in this series, go to
http://www.springer.com/series/3463

George T. Duncan · Mark Elliot ·
Juan-José Salazar-González

Statistical Confidentiality

Principles and Practice

Springer

George T. Duncan
Carnegie Mellon University
Santa Fe, NM 87505, USA
gtduncan@gmail.com

Mark Elliot
University of Manchester
Manchester, UK
mark.elliot@manchester.ac.uk

Juan-José Salazar-González
University of La Laguna
La Laguna, 38271 Tenerife, Spain
jjsalaza@ull.es

ISBN 978-1-4419-7801-1 e-ISBN 978-1-4419-7802-8
DOI 10.1007/978-1-4419-7802-8
Springer New York Dordrecht Heidelberg London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Get together with statisticians and you may see a T-shirt emblazoned, "In God we trust, all others bring data." And, beyond doubt, statisticians are bent on getting data. Indeed they are fully employed in the full gamut of sample surveys, government censuses, observational studies, and clinical trials. Observe another T-shirt bannered, "Top 10 Reasons to be a Statistician," listing, "Estimating parameters is easier than dealing with real life." Surely this self-effacing humor makes the contrary point: that, while statisticians do deal with probability models and their parameters, what really fascinates them is how it all relates to life, and especially the real uncertainties of life. They are intrigued with how data can yield information to reduce these uncertainties, and so enable better decisions to be made. Our concern in this book is that data relevant to issues in the public interest, such as a person's medical record or criminal history, are often highly sensitive. Obtaining and using such data forces us to realize that there is tension between, on the one hand, the desire of the individual for a full and free private life and, on the other hand, the needs of the broader community for information that might, say, improve health care or reduce crime. Statistical confidentiality is pivotal in resolving this tension.

This book on statistical confidentiality is written for all involved with personal and proprietary data from empirical studies. Various roles require an understanding of statistical confidentiality. Here are some instances drawn from issues concerning assessment of a drug rehabilitation program:

- You are a researcher. Your undertaking is to seek rich and convincing evidence to assess benefits of the program, both to addicts and to the community.
- You are a statistician. You design a survey of addicts and link the results to administrative data from the program.
- You are a data steward. Your responsibility is to take the statistician's data and build a database useful to researchers *and* acceptable to privacy advocates.
- You are a privacy advocate. You express qualms about the researcher and statistician matching drug-use records to the addict's personal finances and any criminal behavior.
- You are a respondent to the survey. In principle you are happy to provide data about yourself for the good of society, but you are also concerned about what happens to that data once you have handed it over: will your privacy be respected; will your confidentiality be maintained?

So what exactly is statistical confidentiality? Simply put, it is the stewardship of data to be used for statistical purposes. Stewardship, as expressed in statistical confidentiality, is an active embrace of responsibility for both protecting data and ensuring its beneficial use. Explicitly it requires proper practices for both providing and restricting access to data products.

Getting and using data just to support public policy analysis costs a lot of money. Lane (2003) makes this point: "Billions of taxpayer dollars are spent in supporting the collection and dissemination of federal, state and local data, billions of dollars are spent in data analysis, and this, in turn, both informs scientific understanding of core social science issues and guides decision in how to allocate billions of dollars in social programs." Privacy and confidentiality concerns do not come in dollars, pounds or Euros, but quantifying these concerns via Google search on May 19, 2010, yielded more than 1.42 billion hits on "privacy", more than 19.7 million hits on "confidentiality", and more than 55,000 hits on "statistical confidentiality" (including the quotation marks in the query). Just looking at the first few of these hits on "statistical confidentiality" leads us to information on a United Nations work session in Geneva, Switzerland, to a discussion of pertaining laws in France from the National Institute for Statistics and Economic Studies, to the US Federal Register discussion of the Statistical Confidentiality Order, to a research site of the Computational Aspects of Statistical Confidentiality Project based in the Netherlands, to testimony before the US Congress regarding confidentiality and coordination among statistical agencies, and to a discussion of confidentiality in the Japanese 2000 Census of Population.

The issues cited above illustrate the scope of statistical confidentiality. By absorbing the ideas in this book, you will gain understanding in both the principles and practice of this important field that can benefit your work:

- As a researcher, you will understand why an agency holding statistical data does not respond well to approaching them saying, "Just give me the data; I'm only going to do good things with it," and appreciate why you need to learn about what motivates statistical confidentiality and how it works in practice.
- As a statistician, you will incorporate the requirements of statistical confidentiality into your methodologies for data collection and analysis.
- As a data steward, you are caught between those eager for data and those who worry about confidentiality in its dissemination. Fortunately, using the tools of statistical confidentiality you will progress toward satisfying both groups.
- As a privacy advocate, you will comprehend how confidentiality can be protected even though statistical data are, and should be, made available.
- As a respondent, you will have a better understanding of why your data are needed, how they will be used, and how they will be protected.

We have organized this book into eight chapters. In Chapter 1 we motivate and define the study of statistical confidentiality, laying out the dilemma of data stewardship organizations (we will call them DSOs; important examples are statistical agencies) in resolving the tension between protecting data from snoopers and providing data to legitimate users. We identify the stakeholders in the statistical process, show why statistical data are so useful and in such demand, and explain why DSOs are so concerned about confidentiality. We explore the concept of disclosure risk in terms of an attack by a data snooper and show the basic ways statistical confidentiality can be protected. In Chapter 2 we lay out the fundamental concepts of statistical confidentiality, develop conceptual models of disclosure risk and data utility, identify ways risk can be assessed and controlled, and explore the types of attack that a data snooper might mount. From a rational decision-making perspective, Chapter 3 presents the methodologies of disclosure risk assessment, including a variety of useful metrics for risk assessment. Chapter 4 gives techniques for statistical disclosure limitation of aggregate data, specifically data in tabular form. We present an appropriate definition of disclosure-limited tabular output. We develop deterministic methods, especially through mathematical programming, and stochastic methods, such as cyclic perturbation, for statistical disclosure limitation of tabular data. Chapter 5 gives techniques for disclosure limitation of microdata, that is, data in original record form. We affirm the value of microdata and clarify what users need from such data. We identify the concerns that a DSO has in satisfying the ethical, pragmatic, and legal considerations that motivate their confidentiality promises to data providers. We point out the characteristics of microdata that make it vulnerable to confidentiality attack, and explore various masking methods. We introduce the idea of synthetic data, that is, data that are stochastically generated from a model inferred from the source data. In Chapter 6 we give measures of the impact of disclosure limitation on data utility, and develop the methodology of R-U confidentiality maps and their empirical analog. Chapter 7 provides restricted access methods as a body of administrative procedures for disclosure control. We explore the issues a DSO must face in deciding who can have access, where can access be obtained, what analysis is permitted, and what modes of access should be allowed. Finally, Chapter 8 explores the future of statistical confidentiality. We address a number of important questions: Will privacy and statistical confidentiality have new meanings? Who will care about statistical data? What new forms of DSO will develop? Will statistical data remain valuable? Will there be new issues for statistical confidentiality? Will there be new forms of data snooping? What new strategies of disclosure limitation should be developed?

Plainly, statistical confidentiality is a large and important issue of concern. Research and work in statistical confidentiality is growing rapidly, as is the number of people working on this problem. Many people are screaming that their privacy and confidentiality are vanishing. It is the job of DSOs to provide as much quality data as possible without violating confidentiality laws and promises. This book provides a reader with a comprehensive understanding of the principles and practice of statistical confidentiality. It balances methods and ideas with specific examples. We have written it to be accessible to those just entering the field. Much of the material requires no specific background in mathematics or statistics. Those sections which are more technical are prefaced with explanations of the general ideas without complicated equations.

Our thanks go to our many colleagues who helped us comprehend the need for new ways of dealing with statistical confidentiality, collaborated in our research on this topic over many years, and inspired this book.

Santa Fe, New Mexico, USA	George T. Duncan
Manchester, UK	Mark Elliot
La Laguna, Spain	Juan-José Salazar-González

Contents