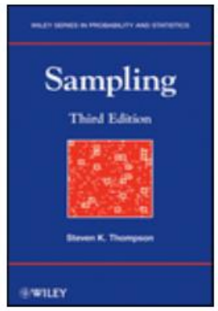
WILEY SERIES IN PROBABILITY AND STATISTICS
Sampling
Third Edition
Steven K. Thompson
WILEY

Table of Contents

Part III: Some Useful Designs

Sampling

Third Edition

STEVEN K. THOMPSON
Simon Fraser University

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Thompson, Steven K., 1945–

Sampling / Steven K. Thompson. – 3rd ed.

p.cm. – (Wiley series in probability and statistics ; 755)
Includes index.
ISBN 978-0-470-40231-3 (hardback)
1 Sampling (Statistics) I. Title.
QA276.6.T58 2012
519.5'2–dc23
2011028944

Preface

One change with this edition of *Sampling* is that I have included sections of computing notes for sample selection, calculation of estimates, and simulations. These computations are illustrated using the statistical programming language R. In doing this I have avoided the use of specialized packages for specific complex designs, choosing instead to show simple calculations and sampling procedures from scratch using a few basic functions. The purpose of these sections is as much for understanding of sampling ideas as for easy ways to select samples and calculate estimates. Other software than R can, of course, be used for the same purpose. The advantages of R include: it is a free and open source, is widely supported by the statistical and other research communities, is available to anyone, and is easily installed on a computer with any of the common operating systems, including Windows, Macintosh OS X, Linux, and other types of Unix. The syntax of R tends to read like generic code and conveys the thinking that goes along with calculations rather than serving as a magic box. R is interactive and has very nice graphics.

Once one learns how to select a sample with a given type of design and to produce various types of estimates using the sample data from the design, it is an easy step to wrap that procedure into a simulation of a sampling strategy. Much of the attention of the computing sections is devoted to the simulation of sampling strategies. The idea is to construct a “population” in the computer as much as possible like the real one which needs to be sampled. With this artificial but more-or-less realistic population, the sampling strategy is then carried out many times. So on each of the runs a sample is selected using the design, and estimates are calculated from the sample data obtained. The distribution of these estimates over the many runs is the sampling distribution. It depends as much on the sampling design and estimation procedure chosen as upon the characteristics of the population. In this way one prospective sampling strategy can be evaluated in comparison to others before committing to one to use in the field. In addition to providing a practical way to evaluate and improve potential sampling strategies, simulations of this kind can give an understanding that is right at the heart of sampling.

Some new examples have been added to this edition. New figures have been added, in particular illustrating the ideas of sampling distributions and the results of various types of simulations. Numerous incremental improvements and the odd new section have been added.

I would like to thank especially the students in my classes and colleagues at other institutions who have helped with corrections of typographical errors and other improvements. I would like to thank Susanne Steitz-Filler and Stephen Quigley at John Wiley & Sons for encouragement in preparation of this edition. Research support for my work in the area of sampling has been provided by the Natural Sciences and Engineering Research Council, the National Center for Health Statistics, Centers for Disease Control and Prevention, the U.S. Census Bureau, the National Institutes of Health, and the National Science Foundation.

Steven K. Thompson

Simon Fraser University

British Columbia

Preface to the Second Edition

The Second Edition retains the general organization of the first, but incorporates new material interspersed throughout the text. For example, model-based ideas and alternatives are included from the earliest chapters, including those on simple random sampling and stratified sampling, rather than suddenly appearing along with ratio and regression estimation methods as has been traditional. Estimation methods deriving from a combination of design and model considerations receive added attention in this edition. Some useful ideas from the ever-developing theory of sampling are briefly described in the chapters on making the most of survey data.

Among the added sections is an expanded description of methods for adjusting for nonsampling errors. A wider discussion of link-tracing designs for sampling hidden human populations—or the Internet—has been added to the chapter on network sampling. New developments in the rapidly expanding field of adaptive sampling are briefly summarized.

Additional numerical examples, as well as exercises, have been added. A number of additional derivations of results have been tucked into the later parts of chapters.

A brief history of sampling has been added to the introduction.

I would like to express my thanks and appreciation to the many people who have so generously shared with me their views on sampling theory and methods in discussions, collaborations, and visits to field sites. They include my colleagues at The Pennsylvania State University and those in the wider research community of sampling and statistics, as well as researchers in other fields such as ecology,

biology, environmental science, computer science, sociology, anthropology, ethnography, and the health sciences. I would like to thank my editor Steve Quigley and editorial program coordinator Heather Haselkorn at John Wiley & Sons for their encouragement and assistance with this project. Research support for my work has been provided by grants from the National Science Foundation (DMS-9626102) and the National Institutes of Health (R01 DA09872).

Steven K. Thompson

University Park, Pennsylvania

Preface to the First Edition

This book covers the basic and standard sampling design and estimation methods and, in addition, gives special attention to methods for populations that are inherently difficult to sample, elusive, rare, clustered, or hard to detect. It is intended as a reference for scientific researchers and others who use sampling and as a textbook for a graduate or upper-level undergraduate course in sampling.

The twenty-six chapters of the book are organized into six parts. Part I covers basic sampling from simple random sampling to unequal probability sampling. Part II treats the use of auxiliary data with ratio and regression estimation and looks at the ideas of sufficient data and of model and design in practical sampling. Part III covers major useful designs including stratified, cluster, systematic, multistage, double, and network sampling. Part IV examines detectability methods for elusive populations: Basic problems in detectability, visibility, and catchability are discussed and specific methods of line transects, variable circular plots, capture–recapture, and line-intercept sampling are covered. Part V concerns spatial sampling, with the prediction or “kriging” methods of geostatistics, considerations of efficient spatial designs, and comparisons of different observational methods including plot shapes and detection aspects. Part VI introduces adaptive sampling designs, in which the sampling procedure depends on what is observed during the survey; for example, sampling effort may be increased in the vicinity of high observed abundance. The adaptive cluster sampling designs described can be remarkably effective for sampling rare, clustered populations, which by conventional methods are notoriously difficult to sample.

Researchers faced with such problems as estimating the abundance of an animal population or an elusive human population, predicting the amount of mineral or fossil-fuel resource at a new site, or estimating the prevalence of a rare disease must be aware that the most effective methods go beyond the material traditionally found in sampling books. At the same time, such researchers may not be aware of the potential usefulness of some of the relatively recent developments in sampling theory and methods—such as network sampling, adaptive sampling designs, and generalized ratio and regression estimation with unequal probability designs. For these reasons, the selection of topics covered in this book is wider than has been traditional for sampling texts.

Some important sampling methodologies have developed largely in particular fields—such as ecology, geology, or health sciences—seemingly in isolation from the mainstream of statistical sampling theory. In the chapters on such methods, I have endeavored to bring out the connections with and the advantages to be gained from basic sampling design, estimation, and prediction results. Thus, for instance, in the chapters on detectability methods associated in particular with ecological sampling, sampling design is emphasized. In the chapter on the prediction or kriging methods associated with geostatistics, the connection to regression estimation results is noted. In the chapter on network sampling, originally associated with epidemiological surveys, the notation has been simplified and connections to basic unequal probability sampling estimators are observed.

Although the range of topics in this book is for the above-noted reasons considerably wider than has been traditional for sampling texts, it has been necessary, in order to keep the book of the desired size, to be selective in what to include. To the reader for whom an additional topic would have been particularly helpful, I can only offer the recompense of the references cited throughout the text to give access to the wider literature in sampling.

My immediate purposes in writing this book were to provide a text for graduate and upper-level undergraduate courses in sampling at the University of Alaska Fairbanks and at the University of Auckland and to provide a manual of useful sampling and estimation methods for researchers with whom I had worked on various projects in a variety of scientific fields. No available manual or text covered the range of topics of interest to these people.

In my experience the backgrounds of the researchers and students interested in sampling topics have been extremely diverse: While some are in statistics or mathematics, many others are in the natural and social sciences and other fields. In writing this book I have assumed the same diversity of backgrounds; the only common factor I feel I can take for granted is some previous course in statistics. The chapters are for the most part organized so that the basic methods and worked examples come first, with generalizations and key derivations following for those interested.

A basic one-semester course in sampling can consist of Chapters 1 through 8 and 11 through 13 or 14, with one or more topics from the remainder of the book added, depending on time and interest. For a graduate class in which many of the students are interested in the special topics of the last three parts of the book, the instructor may wish to cover the basic ideas and methods of the first three parts quite quickly, drawing on them for background later, and spend most of the time on the second half of the book.

I would like to give my thanks to the many people who have influenced and enriched the contents of this book through conversations, joint work, and other interactions on sampling and statistics. In particular, I would like to express appreciation to Fred Ramsey, P. X. Quang, Dana Thomas, and Lyle Calvin. Also, I am grateful to Lyman McDonald, David Siegmund, Richard Cormack, Stephen Buckland, Bryan Manly, Scott Overton, and Tore Schweder for enlightening conversations on statistical sampling methods. I would like to thank my colleagues at Auckland—George Seber, Alastair Scott, Chris Wild, Chris Triggs, Alan Lee, Peter Danaher, and Ross Ihaka—for the benefits of our collaborations, discussions, and daily interactions through which my awareness of relevant and interesting issues in sampling has been increased. I thank my sabbatical hosts at the Institute of Mathematical Statistics at the University of Copenhagen, where some of the sampling designs of this book were first seen as sketches on napkins in the lunch room: Søren Johansen, Tue Tjur, Hans Brøns, Martin Jacobsen, Inge Henningsen, Søren Tolver Jensen, and Steen Andersson. Among the many friends and associates around Alaska who have shared their experiences and ideas on sampling to the benefit of this book are Pat Holmes, Peter Jackson, Jerry McCrary, Jack Hodges, Hal Geiger, Dan Reed, Earl Becker, Dave Bernard, Sam Harbo, Linda Brannian, Allen Bingham, Alan Johnson, Terry Quinn, Bob Fagen, Don Marx, and Daniel Hawkins. Questions and comments leading to rethinking and rewriting of sampling topics have been contributed by many students, to each of whom I offer my thanks and among whom I would particularly like to mention Cheang Wai Kwong, Steve Fleischman, Ed Berg, and Heather McIntyre.

I would like to give a special thanks to my editor, Kate Roach, at John Wiley & Sons for her encouragement and enthusiasm. Research support provided by two grants from the National Science Foundation (DMS-8705812, supported by the Probability and Statistics Program and DMS-9016708, jointly supported by the Probability and Statistics Program and the Environmental Biology Division) resulted in a better book than would have otherwise been possible. I wish to thank Mary for, among many other things, her supportive sense of humor; when on a trip through Norway I could not find a certain guide book after ransacking the luggage jumble from one end of our vehicle to the other, she reminded me to “use adaptive sampling” and, starting with the location of another book randomly discovered amidst the chaos, soon produced the wanted volume. Finally, I thank Jonathan, Lynn, Daniel, and Christopher for an environment of enthusiasm and innovativeness providing inspiration all along the way.

Steven K. Thompson

Auckland, New Zealand

Chapter 1

Introduction

Sampling consists of selecting some part of a population to observe so that one may estimate something about the whole population. Thus, to estimate the amount of lichen available as food for caribou in Alaska, a biologist collects lichen from selected small plots within the study area. Based on the dry weight of these specimens, the available biomass for the whole region is estimated. Similarly, to estimate the amount of recoverable oil in a region, a few (highly expensive) sample holes are drilled. The situation is similar in a national opinion survey, in which only a sample of the people in the population is contacted, and the opinions in the sample are used to estimate the proportions with the various opinions in the whole population. To estimate the prevalence of a rare disease, the sample might consist of a number of medical institutions, each of which has records of patients treated. To estimate the abundance of a rare and endangered bird species, the abundance of birds in the population is estimated based on the pattern of detections from a sample of sites in the study region. In a study of risk behaviors associated with the transmission of the human immunodeficiency virus (HIV), a sample of injecting drug users is obtained by following social links from one member of the population to another.

Some obvious questions for such studies are how best to obtain the sample and make the observations and, once the sample data are in hand, how best to use them to estimate the characteristic of the whole population. Obtaining the observations involves questions of sample size, how to select the sample, what observational methods to use, and what measurements to record. Getting good estimates with observations means picking out the relevant aspects of the data, deciding whether to use auxiliary information in estimation, and choosing the form of the estimator.

Sampling is usually distinguished from the closely related field of *experimental design*, in that in experiments one deliberately perturbs some part of a population in order to see what the effect of that action is. In sampling, more often one likes to find out what the population is like without perturbing or disturbing it. Thus, one hopes that the wording of a questionnaire will not influence the respondents' opinions or that observing animals in a population will not significantly affect the distribution or behavior of the population.

Sampling is also usually distinguished from *observational studies*, in which one has little or no control over how the observations on the population were obtained. In sampling one has the opportunity to deliberately select the sample, thus avoiding many of the factors that make data observed by happenstance, convenience, or other uncontrolled means "unrepresentative."

More broadly, the field of sampling concerns every aspect of how data are selected, out of all the possibilities that might have been observed, whether the selection process has been under the control of investigators or has been determined by nature or happenstance, and how to use such data to make inferences about the larger population of interest. Surveys in which there is some control over the procedure by which the sample is selected turn out to have considerable advantages for purposes of inference about the population from which the sample comes.