

WILEY SERIES IN PROBABILITY AND STATISTICS

# Sampling

*Third Edition*



Steven K. Thompson

WILEY



## Sampling

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice,  
Harvey Goldstein, Iain M. Johnstone, Geert Molenberghs, David W. Scott,  
Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Joseph B. Kadane, Jozef L. Teugels*

A complete list of the titles in this series appears at the end of this volume.

# Sampling

Third Edition

STEVEN K. THOMPSON

Simon Fraser University



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2012 by John Wiley & Sons, Inc. All rights reserved.

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

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at [www.copyright.com](http://www.copyright.com). Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

**Limit of Liability/Disclaimer of Warranty:** While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

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](http://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

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

# Contents

<b>Preface</b>	<b>xv</b>
<b>Preface to the Second Edition</b>	<b>xvii</b>
<b>Preface to the First Edition</b>	<b>xix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Basic Ideas of Sampling and Estimation, 2	
1.2 Sampling Units, 4	
1.3 Sampling and Nonsampling Errors, 5	
1.4 Models in Sampling, 5	
1.5 Adaptive and Nonadaptive Designs, 6	
1.6 Some Sampling History, 7	
<b>PART I BASIC SAMPLING</b>	<b>9</b>
<b>2 Simple Random Sampling</b>	<b>11</b>
2.1 Selecting a Simple Random Sample, 11	
2.2 Estimating the Population Mean, 13	
2.3 Estimating the Population Total, 16	
2.4 Some Underlying Ideas, 17	
2.5 Random Sampling with Replacement, 19	
2.6 Derivations for Random Sampling, 20	
2.7 Model-Based Approach to Sampling, 22	
2.8 Computing Notes, 26	
Entering Data in R, 26	
Sample Estimates, 27	

Simulation, 28	
Further Comments on the Use of Simulation, 32	
Exercises, 35	
<b>3 Confidence Intervals</b>	<b>39</b>
3.1 Confidence Interval for the Population Mean or Total, 39	
3.2 Finite-Population Central Limit Theorem, 41	
3.3 Sampling Distributions, 43	
3.4 Computing Notes, 44	
Confidence Interval Computation, 44	
Simulations Illustrating the Approximate Normality of a Sampling Distribution with Small $n$ and $N$ , 45	
Daily Precipitation Data, 46	
Exercises, 50	
<b>4 Sample Size</b>	<b>53</b>
4.1 Sample Size for Estimating a Population Mean, 54	
4.2 Sample Size for Estimating a Population Total, 54	
4.3 Sample Size for Relative Precision, 55	
Exercises, 56	
<b>5 Estimating Proportions, Ratios, and Subpopulation Means</b>	<b>57</b>
5.1 Estimating a Population Proportion, 58	
5.2 Confidence Interval for a Proportion, 58	
5.3 Sample Size for Estimating a Proportion, 59	
5.4 Sample Size for Estimating Several Proportions Simultaneously, 60	
5.5 Estimating a Ratio, 62	
5.6 Estimating a Mean, Total, or Proportion of a Subpopulation, 62	
Estimating a Subpopulation Mean, 63	
Estimating a Proportion for a Subpopulation, 64	
Estimating a Subpopulation Total, 64	
Exercises, 65	
<b>6 Unequal Probability Sampling</b>	<b>67</b>
6.1 Sampling with Replacement: The Hansen–Hurwitz Estimator, 67	
6.2 Any Design: The Horvitz–Thompson Estimator, 69	
6.3 Generalized Unequal-Probability Estimator, 72	



- 6.4 Small Population Example, 73
- 6.5 Derivations and Comments, 75
- 6.6 Computing Notes, 78
  - Writing an R Function to Simulate a Sampling Strategy, 82
  - Comparing Sampling Strategies, 84
- Exercises, 88

## **PART II MAKING THE BEST USE OF SURVEY DATA 91**

### **7 Auxiliary Data and Ratio Estimation 93**

- 7.1 Ratio Estimator, 94
- 7.2 Small Population Illustrating Bias, 97
- 7.3 Derivations and Approximations for the Ratio Estimator, 99
- 7.4 Finite-Population Central Limit Theorem for the Ratio Estimator, 101
- 7.5 Ratio Estimation with Unequal Probability Designs, 102
- 7.6 Models in Ratio Estimation, 105
  - Types of Estimators for a Ratio, 109
- 7.7 Design Implications of Ratio Models, 109
- 7.8 Computing Notes, 110
- Exercises, 112

### **8 Regression Estimation 115**

- 8.1 Linear Regression Estimator, 116
- 8.2 Regression Estimation with Unequal Probability Designs, 118
- 8.3 Regression Model, 119
- 8.4 Multiple Regression Models, 120
- 8.5 Design Implications of Regression Models, 123
- Exercises, 124

### **9 The Sufficient Statistic in Sampling 125**

- 9.1 The Set of Distinct, Labeled Observations, 125
- 9.2 Estimation in Random Sampling with Replacement, 126
- 9.3 Estimation in Probability-Proportional-to-Size Sampling, 127
- 9.4 Comments on the Improved Estimates, 128

### **10 Design and Model 131**

- 10.1 Uses of Design and Model in Sampling, 131

10.2	Connections between the Design and Model Approaches,	132
10.3	Some Comments,	134
10.4	Likelihood Function in Sampling,	135
<b>PART III SOME USEFUL DESIGNS</b>		<b>139</b>
<b>11</b>	<b>Stratified Sampling</b>	<b>141</b>
11.1	Estimating the Population Total, 142	
	With Any Stratified Design,	142
	With Stratified Random Sampling,	143
11.2	Estimating the Population Mean, 144	
	With Any Stratified Design,	144
	With Stratified Random Sampling,	144
11.3	Confidence Intervals,	145
11.4	The Stratification Principle,	146
11.5	Allocation in Stratified Random Sampling,	146
11.6	Poststratification,	148
11.7	Population Model for a Stratified Population,	149
11.8	Derivations for Stratified Sampling, 149	
	Optimum Allocation,	149
	Poststratification Variance,	150
11.9	Computing Notes,	151
	Exercises,	155
<b>12</b>	<b>Cluster and Systematic Sampling</b>	<b>157</b>
12.1	Primary Units Selected by Simple Random Sampling, 159	
	Unbiased Estimator,	159
	Ratio Estimator,	160
12.2	Primary Units Selected with Probabilities Proportional to Size, 161	
	Hansen–Hurwitz (PPS) Estimator,	161
	Horvitz–Thompson Estimator,	161
12.3	The Basic Principle,	162
12.4	Single Systematic Sample,	162
12.5	Variance and Cost in Cluster and Systematic Sampling,	163
12.6	Computing Notes,	166
	Exercises,	169

<b>13 Multistage Designs</b>	<b>171</b>
13.1 Simple Random Sampling at Each Stage, 173	
Unbiased Estimator, 173	
Ratio Estimator, 175	
13.2 Primary Units Selected with Probability Proportional to Size, 176	
13.3 Any Multistage Design with Replacement, 177	
13.4 Cost and Sample Sizes, 177	
13.5 Derivations for Multistage Designs, 179	
Unbiased Estimator, 179	
Ratio Estimator, 181	
Probability-Proportional-to-Size Sampling, 181	
More Than Two Stages, 181	
Exercises, 182	
<b>14 Double or Two-Phase Sampling</b>	<b>183</b>
14.1 Ratio Estimation with Double Sampling, 184	
14.2 Allocation in Double Sampling for Ratio Estimation, 186	
14.3 Double Sampling for Stratification, 186	
14.4 Derivations for Double Sampling, 188	
Approximate Mean and Variance: Ratio Estimation, 188	
Optimum Allocation for Ratio Estimation, 189	
Expected Value and Variance: Stratification, 189	
14.5 Nonsampling Errors and Double Sampling, 190	
Nonresponse, Selection Bias, or Volunteer Bias, 191	
Double Sampling to Adjust for Nonresponse:	
Callbacks, 192	
Response Modeling and Nonresponse Adjustments, 193	
14.6 Computing Notes, 195	
Exercises, 197	
<b>PART IV METHODS FOR ELUSIVE AND HARD-TO-DETECT POPULATIONS</b>	<b>199</b>
<b>15 Network Sampling and Link-Tracing Designs</b>	<b>201</b>
15.1 Estimation of the Population Total or Mean, 202	
Multiplicity Estimator, 202	
Horvitz–Thompson Estimator, 204	

- 15.2 Derivations and Comments, 207
- 15.3 Stratification in Network Sampling, 208
- 15.4 Other Link-Tracing Designs, 210
- 15.5 Computing Notes, 212
- Exercises, 213

## **16 Detectability and Sampling 215**

- 16.1 Constant Detectability over a Region, 215
- 16.2 Estimating Detectability, 217
- 16.3 Effect of Estimated Detectability, 218
- 16.4 Detectability with Simple Random Sampling, 219
- 16.5 Estimated Detectability and Simple Random Sampling, 220
- 16.6 Sampling with Replacement, 222
- 16.7 Derivations, 222
- 16.8 Unequal Probability Sampling of Groups with Unequal Detection Probabilities, 224
- 16.9 Derivations, 225
- Exercises, 227

## **17 Line and Point Transects 229**

- 17.1 Density Estimation Methods for Line Transects, 230
- 17.2 Narrow-Strip Method, 230
- 17.3 Smooth-by-Eye Method, 233
- 17.4 Parametric Methods, 234
- 17.5 Nonparametric Methods, 237
  - Estimating  $f(0)$  by the Kernel Method, 237
  - Fourier Series Method, 239
- 17.6 Designs for Selecting Transects, 240
- 17.7 Random Sample of Transects, 240
  - Unbiased Estimator, 241
  - Ratio Estimator, 243
- 17.8 Systematic Selection of Transects, 244
- 17.9 Selection with Probability Proportional to Length, 244
- 17.10 Note on Estimation of Variance for the Kernel Method, 246
- 17.11 Some Underlying Ideas about Line Transects, 247
  - Line Transects and Detectability Functions, 247
  - Single Transect, 249
  - Average Detectability, 249

Random Transect, 250	
Average Detectability and Effective Area, 251	
Effect of Estimating Detectability, 252	
Probability Density Function of an Observed Distance, 253	
17.12 Detectability Imperfect on the Line or Dependent on Size, 255	
17.13 Estimation Using Individual Detectabilities, 255	
Estimation of Individual Detectabilities, 256	
17.14 Detectability Functions other than Line Transects, 257	
17.15 Variable Circular Plots or Point Transects, 259	
Exercise, 260	
<b>18 Capture–Recapture Sampling</b>	<b>263</b>
18.1 Single Recapture, 264	
18.2 Models for Simple Capture–Recapture, 266	
18.3 Sampling Design in Capture–Recapture: Ratio Variance Estimator, 267	
Random Sampling with Replacement of Detectability Units, 269	
Random Sampling without Replacement, 270	
18.4 Estimating Detectability with Capture–Recapture Methods, 271	
18.5 Multiple Releases, 272	
18.6 More Elaborate Models, 273	
Exercise, 273	
<b>19 Line-Intercept Sampling</b>	<b>275</b>
19.1 Random Sample of Lines: Fixed Direction, 275	
19.2 Lines of Random Position and Direction, 280	
Exercises, 282	
<b>PART V SPATIAL SAMPLING</b>	<b>283</b>
<b>20 Spatial Prediction or Kriging</b>	<b>285</b>
20.1 Spatial Covariance Function, 286	
20.2 Linear Prediction (Kriging), 286	
20.3 Variogram, 289	
20.4 Predicting the Value over a Region, 291	
20.5 Derivations and Comments, 292	

20.6	Computing Notes, 296	
	Exercise, 299	
<b>21</b>	<b>Spatial Designs</b>	<b>301</b>
21.1	Design for Local Prediction, 302	
21.2	Design for Prediction of Mean of Region, 302	
<b>22</b>	<b>Plot Shapes and Observational Methods</b>	<b>305</b>
22.1	Observations from Plots, 305	
22.2	Observations from Detectability Units, 307	
22.3	Comparisons of Plot Shapes and Detectability Methods, 308	
<b>PART VI</b>	<b>ADAPTIVE SAMPLING</b>	<b>313</b>
<b>23</b>	<b>Adaptive Sampling Designs</b>	<b>315</b>
23.1	Adaptive and Conventional Designs and Estimators, 315	
23.2	Brief Survey of Adaptive Sampling, 316	
<b>24</b>	<b>Adaptive Cluster Sampling</b>	<b>319</b>
24.1	Designs, 321	
	Initial Simple Random Sample without Replacement, 322	
	Initial Random Sample with Replacement, 323	
24.2	Estimators, 323	
	Initial Sample Mean, 323	
	Estimation Using Draw-by-Draw Intersections, 323	
	Estimation Using Initial Intersection Probabilities, 325	
24.3	When Adaptive Cluster Sampling Is Better than Simple Random Sampling, 327	
24.4	Expected Sample Size, Cost, and Yield, 328	
24.5	Comparative Efficiencies of Adaptive and Conventional Sampling, 328	
24.6	Further Improvement of Estimators, 330	
24.7	Derivations, 333	
24.8	Data for Examples and Figures, 336	
	Exercises, 337	
<b>25</b>	<b>Systematic and Strip Adaptive Cluster Sampling</b>	<b>339</b>
25.1	Designs, 341	
25.2	Estimators, 343	

Initial Sample Mean, 343	
Estimator Based on Partial Selection Probabilities, 344	
Estimator Based on Partial Inclusion Probabilities, 345	
25.3 Calculations for Adaptive Cluster Sampling Strategies, 347	
25.4 Comparisons with Conventional Systematic and Cluster Sampling, 349	
25.5 Derivations, 350	
25.6 Example Data, 352	
Exercises, 352	
<b>26 Stratified Adaptive Cluster Sampling</b>	<b>353</b>
26.1 Designs, 353	
26.2 Estimators, 356	
Estimators Using Expected Numbers of Initial Intersections, 357	
Estimator Using Initial Intersection Probabilities, 359	
26.3 Comparisons with Conventional Stratified Sampling, 362	
26.4 Further Improvement of Estimators, 364	
26.5 Example Data, 367	
Exercises, 367	
<b>Answers to Selected Exercises</b>	<b>369</b>
<b>References</b>	<b>375</b>
<b>Author Index</b>	<b>395</b>
<b>Subject Index</b>	<b>399</b>





# 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*





# 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.

## 1.1. BASIC IDEAS OF SAMPLING AND ESTIMATION

In the basic sampling setup, the population consists of a known, finite number  $N$  of units—such as people or plots of ground. With each unit is associated a value of a variable of interest, sometimes referred to as the *y-value* of that unit. The *y-value* of each unit in the population is viewed as a fixed, if unknown quantity—not a random variable. The units in the population are identifiable and may be labeled with numbers  $1, 2, \dots, N$ .

Only a sample of the units in the population are selected and observed. The data collected consist of the *y-value* for each unit in the sample, together with the unit's label. Thus, for each hole drilled in the oil reserve, the data not only record how much oil was found but also identify, through the label, the location of the hole. In addition to the variable of interest, any number of auxiliary variables, such as depth and substrate types, may be recorded. In a lichen survey, auxiliary variables recorded could include elevation, presence of other vegetation, or even “eyeball” estimates of the lichen biomass. In an opinion poll, auxiliary variables such as gender, age, or income class may be recorded along with the opinions.

The procedure by which the sample of units is selected from the population is called the *sampling design*. With most well-known sampling designs, the design is determined by assigning to each possible sample  $s$  the probability  $P(s)$  of selecting that sample. For example, in a simple random sampling design with sample size  $n$ , a possible sample  $s$  consists of a set of  $n$  distinct units from the population, and the probability  $P(s)$  is the same for every possible sample  $s$ . In practice, the design may equivalently be described as a step-by-step procedure for selecting units rather than the resulting probabilities for selecting whole samples. In the case of simple random sampling, a step-by-step procedure consists of selecting a unit label at random from  $\{1, 2, \dots, N\}$ , selecting the next unit label at random from the remaining numbers between 1 and  $N$ , and so on until  $n$  distinct sample units are selected.

The entire sequence  $y_1, y_2, \dots, y_N$  of  $y$ -values in the population is considered a fixed characteristic or parameter of the population in the basic sampling view. The usual inference problem in sampling is to estimate some summary characteristic of the population, such as the mean or the total of the  $y$ -values, after observing only the sample. Additionally, in most sampling and estimation situations, one would like to be able to assess the accuracy or confidence associated with estimates; this assessment is most often expressed with a confidence interval.

In the basic sampling view, if the sample size were expanded until all  $N$  units of the population were included in the sample, the population characteristic of interest would be known exactly. The uncertainty in estimates obtained by sampling thus stems from the fact that only part of the population is observed. While the population characteristic remains fixed, the estimate of it depends on which sample is selected. If for every possible sample the estimate is quite close to the true value of the population characteristic, there is little uncertainty associated with the sampling strategy; such a strategy is considered desirable. If, on the other hand, the value of the estimate varies greatly from one possible sample to another, uncertainty is associated with the method. A trick performed with many of the most useful sampling designs—cleverer than it may appear at first glance—is that this variability from sample to sample is estimated using only the single sample selected.

With careful attention to the sampling design and using a suitable estimation method, one can obtain estimates that are unbiased for population quantities, such as the population mean or total, without relying on any assumptions about the population itself. The estimate is unbiased in that its expected value over all possible samples that might be selected with the design equals the actual population value. Thus, through the design and estimation procedure, an unbiased estimate of lichen biomass is obtained whether lichens are evenly distributed throughout the study area or are clumped into a few patches. Additionally, the random or probability selection of samples removes recognized and unrecognized human sources of bias, such as conscious or unconscious tendencies to select units with larger (or smaller) than average values of the variable of interest. Such a procedure is especially desirable when survey results are relied on by persons with conflicting sets of interests—a fish population survey that will be used by fishery managers, commercial fishermen, and environmentalists, for instance. In such cases, it is unlikely that all parties concerned could agree on the purposive selection of a “representative” sample.

A probability design such as simple random sampling thus can provide unbiased estimates of the population mean or total and also an unbiased estimate of variability, which is used to assess the reliability of the survey result. Unbiased estimates and estimates of variance can also be obtained from unequal probability designs, provided that the probability of inclusion in the sample is known for each unit and for pairs of units.

Along with the goal of unbiased or nearly unbiased estimates from the survey come goals of precise or low-variance estimates and procedures that are convenient or cost-effective to carry out. The desire to satisfy as many of these goals as possible under a variety of circumstances has led to the development of widely used

sampling designs and estimation methods, including simple random and unequal probability sampling; the use of auxiliary information; stratified, systematic, cluster, multistage, and double sampling; and other techniques.

## 1.2. SAMPLING UNITS

With many populations of people and institutions, it is straightforward to identify the type of units to be sampled and to conceive of a list or frame of the units in the population, whatever the practical problems of obtaining the frame or observing the selected sample. The units may be people, households, hospitals, or businesses. A complete list of the people, households, medical institutions, or firms in the target population would provide an ideal frame from which the sample units could be selected. In practice, it is often difficult to obtain a list that corresponds exactly to the population of interest. A telephone directory does not list people without telephones or with unlisted numbers. The set of all possible telephone numbers, which may be sampled by random dialing, still does not include households without telephones. A list of public or private institutions may not be up-to-date.

With many other populations, it is not so clear what the units should be. In a survey of a natural resource or agricultural crop in a region, the region may be divided into a set of geographic units (*plots* or *segments*) and a sample of units may be selected using a map. However, one is free to choose alternative sizes and shapes of units, and such choices may affect the cost of the survey and the precision of estimators. Further, with a sampling procedure in which a point location is chosen at random in a study region and sample units are then centered around the selected points, the sample units can potentially overlap, and hence the number of units in the population from which the sample is selected is not finite.

For an elusive population with detectability problems, the role of units or plots may be superseded by that of detectability functions, which are associated with the methods by which the population is observed and the locations are selected for making the observations. For example, in selecting the locations of line transects in a bird survey and choosing the speed at which they are traversed, one determines the “effective areas” observed within the study area in place of traditional sampling units or plots.

In some sampling situations the variable of interest may vary continuously over a region. For example, in a survey to assess the oil reserves in a region, the variable measured may be the depth or core volume of oil at a location. The value of such a variable is not necessarily associated with any of a finite set of units in the region, but rather, may be measured or estimated either at a point or as a total over a subregion of any size or shape.

Although the foregoing sampling situations go beyond the framework of a population divided uniquely into a finite collection of units from which the sample is selected, basic sampling design considerations regarding random sampling, stratified sampling, and other designs, and estimation results on design-unbiased estimation, ratio estimation, and other methods still apply.

### 1.3. SAMPLING AND NONSAMPLING ERRORS

The basic sampling view assumes that the variable of interest is measured on every unit in the sample without error, so that errors in the estimates occur only because just part of the population is included in the sample. Such errors are referred to as *sampling errors*. But in real survey situations, nonsampling errors may arise also. Some people in a sample may be away from home when phoned or may refuse to answer a question on a questionnaire, and such nonrespondents may not be typical of the population as a whole, so that the sample tends to be unrepresentative of the population and the estimates are biased. In a fish survey, some selected sites may not be observed due to rough weather conditions; sites farthest from shore, which may not be typical of the study region as a whole, are the most likely to have such weather problems.

The problem of nonresponse is particularly pronounced in a survey with a very low response rate, in which the probability of responding is related to the characteristic to be measured—magazine readership surveys of sexual practices exemplify the problem. The effect of the nonresponse problem may be reduced through additional sampling effort to estimate the characteristics of the nonresponse stratum of the population, by judicious use of auxiliary information available on both responding and nonresponding units, or by modeling of the nonresponse situation. But perhaps the best advice is to strive to keep nonresponse rates as low as possible.

Errors in measuring or recording the variable of interest may also occur. Quality-control effort throughout every stage of a survey is needed to keep errors to a minimum. In some situations, it may be possible to model measurement errors separately from sampling issues in order to relate the observations to population characteristics.

Detectability problems are a type of nonsampling error that occurs with a wide range of elusive populations. On a bird survey, the observer is typically unable to detect every individual of the species in the vicinity of a sampling site. In a trawl survey of fish, not every fish in the path of the net is caught. Nor is every homeless person in a society counted in a census. A number of special techniques, including line transect, capture–recapture, and related methods, have been developed for estimating population quantities when detectability problems are a central issue.

### 1.4. MODELS IN SAMPLING

In the basic sampling view the population is a finite set of units, each with a fixed value of the variable of interest, and probability enters only through the design, that is, the procedure by which the sample of units is selected. But for some populations it may be realistic and of practical advantage to consider a probability model for the population itself. The model might be based on knowledge of the natural phenomena influencing the distribution of the type of population or on a pragmatic statistical model summarizing some basic characteristics of such populations.

For example, a regression model may empirically describe a relationship between a variable of interest, the yield of a horticultural crop, say, with an

auxiliary variable, such as the median level of an air pollutant. The model relating the variable of interest with the auxiliary variable has implications both for how to design the survey and how to make estimates.

In spatial sampling situations, the existence of correlations between values of the variable of interest at different sites, depending on the distance between the sites, has implications for choices regarding sampling design, estimation or prediction, and observational method. A model-based approach utilizing such correlation patterns has been particularly influential in geological surveys of mineral and fossil-fuel resources. In ecological surveys, such correlation patterns have implications not only for the spatial selection of observational sites, but for the observational methods (including plot shapes) used.

Ideally, one would like to be able to use a model of the population without having all conclusions of the survey depend on the model's being exactly true. A "robust" approach to sampling uses models to suggest efficient procedures while using the design to protect against departures from the model.

## 1.5. ADAPTIVE AND NONADAPTIVE DESIGNS

Surveys of rare, clustered populations motivate a further advance beyond the basic view of a sampling design. In adaptive sampling designs, the procedure for selecting sites or units on which to make observations may depend on observed values of the variable of interest. For example, in a survey for estimating the abundance of a natural resource, additional sites may be added to the sample during the survey in the vicinity of high observed abundance. Such designs have important applications to surveys of animal, plant, mineral, and fossil-fuel resources and may also have applications to other fields such as epidemiology and quality control.

The main purpose of adaptive procedures is to achieve gains in precision or efficiency, compared to conventional designs of equivalent sample size, by taking advantage of observed characteristics of the population. Adaptive procedures include such procedures as sequential stopping rules and sequential allocation among strata—procedures that have been rather heavily studied outside the finite-population context in the field of sequential analysis. With the population units identifiable as in the sampling situation, the possibilities for adaptive procedures are even greater, since it is possible to decide during a survey not just how many units to sample next but exactly which units or group of units to sample next.

In adaptive cluster sampling, whenever an observed value of the variable of interest satisfies a given criterion—for example, high abundance of animals observed at a site—units in the neighborhood of that unit (site) are added to the sample. A number of variations on this type of design are described in the final chapters of this book. For some populations, the designs produce remarkable increases in efficiency and appear to be particularly effective for sampling rare, clustered populations.

The sampling design is given for a conventional or nonadaptive design by a probability  $P(s)$  of selecting any particular sample  $s$ . For an adaptive design, the