Scheaffer Mendenhall Ott

ELEMENTARY SURVEY SAMPLING

FOURTH EDITION

ELEMENTARY SURVEY SAMPLING

FOURTH EDITION

Richard L. Scheaffer
University of Florida

William Mendenhall
University of Florida, Emeritus

Lyman Ott
Merrell Research Center

Duxbury Press
An Imprint of Wadsworth Publishing Company
Belmont, California

Duxbury Press
An Imprint of Wadsworth Publishing Company
A division of Wadsworth, Inc.

Library of Congress Cataloging-in-Publication Data

Scheaffer, Richard L.
Elementary survey sampling/Richard L. Scheaffer, William Mendenhall, Lyman Ott.—4th ed.
p. cm.
ISBN 0-534-92185-X
1. Sampling (Statistics) I. Mendenhall, William. II. Ott, Lyman. III. Title.
QA276.6.S385 1990
519.5'2—dc20 89–77274
CIP

Sponsoring Editor: Michael Payne
Production Editor: Susan M. C. Caffey
Manufacturing Coordinator: Margaret Sullivan Higgins
Interior and Cover Designer: Catherine L. Johnson
Compositor: The Universities Press Ltd.
Text Printer and Binder: The Maple-Vail Book Manufacturing Group

Printed in the United States of America.

4 5 6 7 8 9 — 94 93

PREFACE

Elementary Survey Sampling is an introductory text on the design and analysis of sample surveys intended for students of business, the social sciences, or natural resource management. The only prerequisite is an elementary course in statistics. The numerous examples, with solutions, also make it suitable for use as a supplemental text for higher-level courses.

Since it is written to appeal to students of limited mathematical background, the text emphasizes the practical aspects of survey problems. Each major chapter introduces a sample survey design or a possible estimation procedure by describing a pertinent practical problem and then explaining the suitability of the methodology proposed. This introduction is followed by the appropriate estimation procedures and a compact presentation of the formulas; then a practical example is worked out. The text is not entirely cookbook in nature. Explanations that appeal to the students' intuition are supplied to justify many of the formulas and to support the choice of particular sample survey designs. Examples and exercises have been selected from many fields of application. Answers, given for selected exercises, may be subject to small rounding errors because of the complexity of some of the formulas.

The "Experiences with Real Data" sections found at the end of most chapters include suggestions on how the student can become involved with real sampling problems. These problems may be large or small projects, with some requiring computations to be handled by a computer, but we have found such projects to be valuable learning experiences for students taking a sampling course. Working on a real project forces students to think about every aspect of the survey and causes them to realize that some ideas that sound simple in the textbook are not so easily carried out in practice.

The text includes a review of elementary concepts (Chapters 1 and 2) and a description of terms pertinent to survey sampling, along with a discussion of

the design of questionnaires and methods of data collection (Chapter 3). Chapters 4, 5, 7, and 8 present the four most common sample survey designs—namely, simple random sampling, stratified random sampling, systematic sampling, and cluster sampling, respectively. Chapter 6 discusses ratio, regression and difference estimation. The remaining chapters deal with two-stage cluster sampling, and other specialized problems that occur in survey sampling.

Practical aspects of conducting sample surveys are emphasized, with sections on sources of errors in surveys, methods of data collection, designing questionnaires, and guidelines for planning surveys. Sampling with probabilities proportional to size is introduced in Chapter 4 and applied to cluster sampling in Chapters 8 and 9. All chapters contain many examples of how the sampling concepts are used in practice.

Topics receiving attention in the fourth edition include poststratification, double sampling, sample size determination in two-stage cluster sampling and estimation of population density. A major emphasis is placed on calculating relative efficiencies for sampling designs so that proper choices for designs can be made in various practical contexts. Since laborious calculations are one of the major deterrents to teaching sample survey design effectively, solutions using standard computer packages (such as Minitab) are introduced throughout the text.

New exercises have been added to most chapters, and Chapter 12 contains a set of exercises that may require some careful thought in the selection of the appropriate analysis. These exercises can serve as a review of the major methods presented in the book. A large, real data set is included in the Appendix, with numerous exercises referring to it. A solutions manual is available from the publisher.

The Appendix also includes the mathematical derivations of many of the main results in the text. The understanding of many of these derivations requires a working knowledge of elementary probability theory.

We wish to express our sincere appreciation to the many people who have helped in the preparation of this text. Particular thanks are due to the reviewers Larry E. Richards, *University of Oregon* and Frederick Williams, *University of Missouri—Columbia* for their helpful comments during the preparation of this manuscript. Survey sampling plays an increasingly important role in the information society of today. The authors hope this book helps students to design better surveys and to understand the subtleties of survey results presented to them. In short, our goal is to improve the flow of knowledge from data.

Richard L. Scheaffer
William Mendenhall
Lyman Ott

CONTENTS

CHAPTER 6

RATIO, REGRESSION, AND DIFFERENCE ESTIMATION 151

CHAPTER 7

SYSTEMATIC SAMPLING 205

CHAPTER 8

CLUSTER SAMPLING 243

CHAPTER 1

INTRODUCTION

Introductory courses stress that modern statistics is a theory of information with inference as its objective. The target of our curiosity is a set of measurements, a *population,* that exists in fact or may be generated by repeated experimentation. The medium of inference is the *sample,* which is a subset of measurements selected from the population. We wish to make an inference about the population on the basis of characteristics of the sample—or, equivalently, the information contained in the sample.

For example, suppose that a chain of department stores maintains customer charge accounts. The amount of money owed the company will vary from day to day as new charges are made and some accounts are paid. Indeed, the set of amounts due the company on a given day represents a population of measurements of considerable interest to the management. The population characteristic of interest is the total of all measurements in the population or, equivalently, the daily total credit load.

Keeping track of the daily total credit associated with change accounts may seem to be a simple task for an electronic computer. However, the data must be updated daily, and updating takes time. A simpler method for determining

the total credit load associated with the charge accounts is to randomly sample the population of accounts on a given day, estimate the average amount owed per account, and multiply by the number of accounts. In other words, we employ a statistical estimator to make an inference about the population total. Elementary statistics tells us that this estimate can be made as accurate as we wish simply by increasing the sample size. The resulting estimate either is accompanied by a bound on the error of estimation (Mendenhall, 1987, Chapter 8) or is expressed as a confidence interval. Thus information in the sample is used to make an inference about the population.

Information from sample surveys affects almost every facet of our daily lives. Such information determines government policies on, for example, the control of the economy and the promotion of social programs. Opinion polls are the basis of much of the news reported by the various news media. Ratings of television shows determine which shows are to be available for viewing in the future.

One usually thinks of the U.S. Census as contacting every household in the country. Actually, in the 1980 census only 14 questions were asked of all households. Information on an additional 42 questions was obtained from only a sample of households. The resulting information is used by the federal government in determining allocations of funds to states and cities. It is used by business to forecast sales, to manage personnel, and to establish future site locations. It is used by urban and regional planners to plan land use, transportation networks, and energy consumption. It is used by social scientists to study economic conditions, racial balance, and other aspects of the quality of life.

The U.S. Bureau of Labor Statistics routinely conducts over twenty surveys. Some of the best known and most widely used are the surveys that establish the consumer price index (CPI). The CPI is a measure of price change for a fixed market basket of goods and services over time. It is used as a measure of inflation and serves as an economic indicator for government policies. Businesses have wage rates and pension plans tied to the CPI. Federal health and welfare programs, as well as many state and local programs, tie their bases of eligibility to the CPI. Escalator clauses in rents and mortgages are based on the CPI. So we can see that this one index, determined on the basis of sample surveys, plays a fundamental role in our society.

Many other surveys at the Bureau of Labor Statistics (BLS) are crucial to society. The monthly Current Population Survey establishes basic information on the labor force, employment, and unemployment. The consumer expenditure surveys collect data on family expenditures for goods and services used in day-to-day living. The Establishment Survey collects information on employment hours and earnings for nonagricultural business establishments. The survey on occupational outlook provides information on future employment opportunities for a variety of occupations, projecting to approximately ten

years ahead. Other activities of the BLS are addressed in the *BLS Handbook of Methods* (1982).

Opinion polls are constantly in the news, and the names of Gallup and Harris have become well known to everyone. These polls, or sample surveys, reflect the attitudes and opinions of citizens on everything from politics and religion to sports and entertainment. The Nielsen ratings determine the success or failure of TV shows.

Businesses conduct sample surveys for their internal operations, in addition to using government surveys for crucial management decisions. Auditors estimate account balances and check on compliance with operating rules by sampling accounts. Quality control of manufacturing processes relies heavily on sampling techniques.

One particular area of business activity that depends on detailed sampling activities is marketing. Decisions on which products to market, where to market them, and how to advertise them are often made on the basis of sample survey data. The data may come from surveys conducted by the firm that manufactures the product or may be purchased from survey firms that specialize in marketing data. The activities of three such firms are outlined next.

The Nielsen retail index is less famous than the Nielsen television ratings, but it is very important to firms marketing products for retail sale. This index furnishes continuous sales data on foods, cosmetics, pharmaceuticals, beverages, and many other classes of products. It can provide estimates of total sales for a product class, sales for a client's particular brand, sales for a competing brand, retail and wholesale price data, and the percentage of stores stocking a particular item. The data comes from auditing inventories and sales in 1600 stores across the United States every 60 days.

Selling Areas—Marketing, Inc. (SAMI), collects information on the movement of products from warehouses and wholesalers. Data is obtained in 36 major television market areas, containing 74% of national food sales, and covers 425 product categories.

The Market Research Corporation of America provides many types of marketing data through the use of surveys, but some of the more interesting results come from its National Menu Census. This survey samples families and observes their eating patterns for two weeks. As many as four thousand families may participate during a year. Data are obtained on the number of times a particular food item is served, how it is served, how many persons eat the item, and many other details, including what happens to the leftovers. Such details are important for product development and advertising.

Many interesting examples of the practical uses of statistics in general and sampling in particular can be found in *Statistics: A Guide to the Unknown* (see the references in the Appendix). In this book you might want to look at some of the methods and uses of opinion polling discussed in the articles "Opinion

Polling in a Democracy" by George Gallup and "Election Night on Television" by R. F. Link. Those interested in wildlife ecology should read "The Plight of the Whales" by D. G. Chapman. Find out how interrailroad and interairline billing is handled economically through sampling by reading "How Accountants Save Money by Sampling" by John Neter.

Since the objective of modern statistics is inference, you may question what particular aspect of statistics will be covered in a course on sample survey design. The answer to this question is twofold. First, we will focus on the economics of purchasing a specific quantity of information. More specifically, how can we design sampling procedures that reduce the cost of a fixed quantity of information? Although introductory courses in statistics acknowledge the importance of this subject, they place major emphasis on basic concepts and on how to make inferences in specific situations *after* the data have been collected. The second distinguishing feature of our topic is that it is aimed at the particular types of sampling situations and inferential problems most frequently encountered in business, the social sciences, and natural resource management (timber, wildlife, and recreation) rather than in the physical sciences.

Even the terminology of the social scientist differs from that of the physical scientist. Social scientists conduct *surveys* to collect a sample, while physical scientists perform *experiments*. Thus we acknowledge that differences exist from one field of science to another in the nature of the populations and the manner in which a sample can be drawn. For example, populations of voters, financial accounts, or animals of a particular species may contain only a small number of elements. In contrast, the conceptual population of responses generated by measuring the yield of a chemical process is very large indeed. (You may recall that the properties of estimators and test statistics covered in most introductory courses assume that the population of interest is large relative to the sample.) Limitations placed on the sampling procedure also vary from one area of science to another. Sampling in the biological and physical sciences can frequently be performed under controlled experimental conditions. Such control is frequently impossible in the social sciences, business, and natural resource management. For example, a medical researcher might compare the growth of rats subjected to two different drugs. For this experiment the initial weights of the rats and the daily intake of food could be controlled to reduce unwanted variation in the experiment. In contrast, very few variables can be controlled in comparing the effect of two different television advertisements on sales for a given product; no control is possible when studying the effect of environmental conditions on the number of seals in the North Pacific Ocean.

In summary, this text is concerned with the peculiarities of sampling and inference commonly encountered in business, the social sciences, and natural resource management. Specifically, we will consider methods for actually selecting the sample from an existing population and ways of circumventing

various difficulties that arise. Methods for designing surveys that capitalize on characteristics of the population will be presented along with associated estimators to reduce the cost for acquiring an estimate of specified accuracy.

Chapter 2 reviews some of the basic concepts encountered in introductory statistics, including the fundamental role that probability plays in making inferences. Chapter 3 presents some of the basic terminology of sampling, as well as a discussion of problems arising in sample survey design. Simple random sampling, familiar to the beginning student, is carefully presented in Chapter 4; it includes physical procedures for actually selecting the sample. Following chapters cover economical methods for selecting a sample and associated methods for estimating population parameters.

In reading this text, keep in mind that the ultimate objective of each chapter is *inference*. Identify the sampling procedure associated with each chapter, the population parameters of interest, their estimators, and the associated bounds on the errors of estimation. Develop an intuitive understanding of and appreciation for the benefits to be derived from specialized sampling procedures. Focus on the broad concepts, and do not become hypnotized by the formulas for estimators and variances that sometimes are unavoidably complicated. In short, focus on the forest rather than the trees. Work some exercises, and the details will fall into place.

CHAPTER 2

A REVIEW OF SOME BASIC CONCEPTS

2.1

INTRODUCTION

Knowledge of the basic concepts of statistics is a prerequisite for a study of sample survey design. Thus in this chapter we will review some of these basic concepts.

The ultimate objective of statistics is to make inferences about a population from information contained in a sample. The target of our inference, the population, is a set of measurements, finite or infinite, existing or conceptual. Hence the first step in statistics is to find a way to phrase an inference about a population or, equivalently, to describe a set of measurements. Thus frequency distributions and numerical descriptive measures are the first topics of our review.

The second step in statistics is to consider how inferences can be made about the population from information contained in a sample. For this step we must consider probability distributions of sample quantities, or sampling distribu-