Sampling

Design and Analysis

Third Edition

Sharon L. Lohr

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Preface

We rarely have complete information in life. Instead, we make decisions from partial information, often in the form of a sample from the population we are interested in. *Sampling: Design and Analysis* teaches the statistical principles for selecting samples and analyzing data from a sample survey. It shows you how to evaluate the quality of estimates from a survey, and how to design and analyze many different forms of sample surveys.

The third edition has been expanded and updated to incorporate recent research on theoretical and applied aspects of survey sampling, and to reflect developments related to the increasing availability of massive data sets ("big data") and samples selected via the internet. The new chapter on nonprobability sampling tells how to analyze and evaluate information from samples that are not selected randomly (including big data), and contrasts nonprobability samples with low-response-rate probability samples. The chapters on nonsampling errors have been extensively revised to include recent developments on treating nonresponse and measurement errors. Material in other chapters has been revised where there has been new research or I felt I could clarify the presentation of results. Examples retained from the second edition have been updated when needed, and new examples have been added throughout the book to illustrate recent applications of survey sampling.

The third edition has also been revised to be compatible with multiple statistical software packages. Two supplementary books, available for FREE download from the book's companion website (see page xviii for how to obtain the books), provide step-by-step guides of how to use SAS® and R software to analyze the examples in *Sampling: Design and Analysis*. Both books are also available for purchase in paperback form, for readers who prefer a hard copy.

Lohr, S. (2022). *SAS® Software Companion for Sampling: Design and Analysis, Third Edition*. Boca Raton, FL: Chapman & Hall/CRC Press.

Lu, Y. and Lohr, S. (2022). *R Companion for Sampling: Design and Analysis, Third Edition*. Boca Raton, FL: Chapman & Hall/CRC Press.

Instructors can choose which software package to use in the class (SAS software alone, R software alone, or, if desired, both software packages) and have students download the appropriate supplementary book. See the Computing section on page xvi for more information about the supplementary books and about choice of statistical software.

Features of *Sampling: Design and Analysis, Third Edition*

- The book is accessible to students with a wide range of statistical backgrounds, and is flexible for content and level. By appropriate choice of sections, this book can be used for an upper-level undergraduate class in statistics, a first- or second-year graduate class for statistics students, or a class with students from business, sociology, psychology, or biology who want to learn about designing and analyzing data from sample surveys. It is also useful for persons who analyze survey data and want to learn more about the statistical aspects of surveys and recent developments. The book is intended for anyone who is interested in using sampling methods to learn about a population, or who wants to understand how data from surveys are collected, analyzed, and interpreted.

Chapters 1–8 can be read by students who are familiar with basic concepts of probability and statistics from an introductory statistics course, including independence and expectation, confidence intervals, and straight-line regression. Appendix A reviews the probability concepts needed to understand probability sampling. Parts of Chapters 9 to 16 require more advanced knowledge of mathematical and statistical concepts. Section 9.1, on linearization methods for variance estimation, assumes knowledge of calculus. Chapter 10, on categorical data analysis, assumes the reader is familiar with chi-square tests and odds ratios. Chapter 11, on regression analysis for complex survey data, presupposes knowledge of matrices and the theory of multiple regression for independent observations.

Each chapter concludes with a chapter summary, including a glossary of key terms and references for further exploration.

- The examples and exercises feature real data sets from the social sciences, engineering, agriculture, ecology, medicine, business, and a variety of other disciplines. Many of the data sets contain other variables not specifically referenced in the text; an instructor can use these for additional exercises and activities.

  The data sets are available for download from the book's companion website. Full descriptions of the variables in the data sets are given in Appendix A of the supplementary books described above (Lohr, 2022; Lu and Lohr, 2022).

  The exercises also give the instructor much flexibility for course level (see page xv). Some emphasize mastering the mechanics, but many encourage the student to think about the sampling issues involved and to understand the structure of sample designs at a deeper level. Other exercises are open-ended and encourage further exploration of the ideas.

  In the exercises, students are asked to design and analyze data from real surveys. Many of the data examples and exercises carry over from chapter to chapter, so students can deepen their knowledge of the statistical concepts and see how different analyses are performed with the sample. Data sets that are featured in multiple chapters are listed in the "Data sets" entry of the Index so you can follow them across chapters.

- *Sampling: Design and Analysis, Third Edition* includes many topics not found in other textbooks at this level. Chapters 7–11 discuss how to analyze complex surveys such as those administered by federal statistical agencies, how to assess the effects of nonresponse and weight the data to adjust for it, how to use computer-intensive methods for estimating variances in complex surveys, and how to perform chi-square tests and regression analyses using data from complex surveys. Chapters 12–14 present methods for two-phase sampling, using a survey to estimate population size, and designing a survey to study a subpopulation that is hard to identify or locate. Chapter 15, new for the third edition, contrasts probability and nonprobability samples, and provides guidance on how to evaluate the quality of nonprobability samples. Chapter 16 discusses a total quality framework for survey design, and presents some thoughts on the future of sampling.

- Design of surveys is emphasized throughout, and is related to methods for analyzing the data from a survey. The book presents the philosophy that the design is by far the most important aspect of any survey: No amount of statistical analysis can compensate for a badly designed survey.

- *Sampling: Design and Analysis, Third Edition* emphasizes the importance of graphing the data. Graphical analysis of survey data is challenging because of the large sizes and complexity of survey data sets but graphs can provide insight into the data structure.

- While most of the book adopts a randomization-based perspective, I have also included sections that approach sampling from a model-based perspective, with the goal of placing sampling methods within the framework used in other areas of statistics. Many important results in survey research have involved models, and an understanding of both approaches is essential for the survey practitioner. All methods for dealing with nonresponse are model-based. The model-based approach is introduced in Section 2.10 and further developed in successive chapters; those sections can be covered while those chapters are taught or discussed at any time later in the course.

Exercises. The book contains more than 550 exercises, organized into four types. More than 150 of the exercises are new to the third edition.

A. Introductory exercises are intended to develop skills on the basic ideas in the book.

B. Working with Survey Data exercises ask students to analyze data from real surveys. Most require the use of statistical software; see section on Computing below.

C. Working with Theory exercises are intended for a more mathematically oriented class, allowing students to work through proofs of results in a step-by-step manner and explore the theory of sampling in more depth. They also include presentations of additional results about survey sampling that may be of interest to more advanced students. Many of these exercises require students to know calculus, probability theory, or mathematical statistics.

D. Projects and Activities exercises contain activities suitable for classroom use or for assignment as a project. Many of these activities ask the student to design, collect, and analyze a sample selected from a population. The activities continue from chapter to chapter, allowing students to build on their knowledge and compare various sampling designs. I always assigned Exercise 35 from Chapter 7 and its continuation in subsequent chapters as a course project, and asked students to write a report with their findings. These exercises ask students to download data from a survey on a topic of their choice and analyze the data. Along the way, the students read and translate the survey design descriptions into the design features studied in class, develop skills in analyzing survey data, and gain experience in dealing with nonresponse or other challenges.

Suggested chapters for sampling classes. Chapters 1–6 treat the building blocks of simple random, stratified, and cluster sampling, as well as ratio and regression estimation. To read them requires familiarity with basic ideas of expectation, sampling distributions, confidence intervals, and linear regression—material covered in most introductory statistics classes. Along with Chapters 7 and 8, these chapters form the foundation of a one-quarter or one-semester course. Sections on the statistical theory in these chapters are marked with asterisks—these require more familiarity with probability theory and mathematical statistics. The material in Chapters 9–16 can be covered in almost any order, with topics chosen to fit the needs of the students.

Sampling: Design and Analysis, Third Edition can be used for many different types of classes, and the choice of chapters to cover can be tailored to meet the needs of the students in that class. Here are suggestions of chapters to cover for four types of sampling classes.

Undergraduate class of statistics students: Chapters 1–8, skipping sections with asterisks; Chapters 15 and 16.

One-semester graduate class of statistics students: Chapters 1–9, with topics chosen from the remaining chapters according to the desired emphasis of the class.

Two-semester graduate class of statistics students: All chapters, with in-depth coverage of Chapters 1–8 in the first term and Chapters 9–16 in the second term. The exercises contain many additional theoretical results for survey sampling; these can be presented in class or assigned for students to work on.

Students from social sciences, biology, business, or other subjects: Chapters 1–7 should be covered for all classes, skipping sections with asterisks. Choice of other material depends on how the students will be using surveys in the future. Persons teaching classes for social scientists may want to include Chapters 8 (nonresponse), 10 (chi-square tests), and 11 (regression analyses of survey data). Persons teaching classes for biology students may want to cover Chapter 11 and Chapter 13 on using surveys to estimate population sizes. Students who will be analyzing data from large government surveys would want to learn about replication-based variance estimation methods in Chapter 9. Students who may be using nonprobability samples should read Chapter 15.

Any of these can be taught as activity-based classes, and that is how I structured my sampling classes. Students were asked to read the relevant sections of the book at home before class. During class, after I gave a ten-minute review of the concepts, students worked in small groups with their laptops on designing or analyzing survey data from the chapter examples or the "Projects and Activities" section, and I gave help and suggestions as needed. We ended each class with a group discussion of the issues and a preview of the next session's activities.

Computing. You need to use a statistical software package to analyze most of the data sets provided with this book. I wrote *Sampling: Design and Analysis, Third Edition* for use with either SAS or R software. You can choose which software package to use for computations: SAS software alone, R alone, or both, according to your preference. Both software packages are available at no cost for students and independent learners, and the supplementary books tell how to obtain them.

The supplementary books, *SAS® Software Companion for Sampling: Design and Analysis, Third Edition* by Sharon L. Lohr, and *R Companion for Sampling: Design and Analysis, Third Edition* by Yan Lu and Sharon L. Lohr, available for FREE download from the book's companion website, demonstrate how to use SAS and R software, respectively, to analyze the examples in *Sampling: Design and Analysis, Third Edition*. Both books are also available for purchase in paperback form, for readers who prefer hard copies. The two supplementary books are written in parallel format, making it easy to find how a particular example is coded in each software package. They thus would also be useful for a reader who is familiar with one of the software packages but would like to learn how to use the other.

The supplementary books provide the code used to select, or produce estimates or graphs from, the samples used for the examples in Chapters 1–13 of this book. They display and interpret the output produced by the code, and discuss special features of the procedure or function used to produce the output. Each chapter concludes with tips and warnings on how to avoid common errors when designing or analyzing surveys.

Which software package should you use? If you are already familiar with R or SAS software, you may want to consider adopting that package when working through *Sampling: Design and Analysis, Third Edition*. You may also want to consider the following features of the two software packages for survey data.

Features of SAS software for survey data:

- Students and independent learners anywhere in the world can access a FREE, cloud-based version of the software: SAS® OnDemand for Academics (`https://www.sas.com/en_us/software/on-demand-for-academics.html`) contains all of the programs

needed to select samples, compute estimates, and graph data for surveys. Short online videos for instructors show how to create a course site online, upload data that can be accessed by all students, and give students access to the course material. Additional short video tutorials help students become acquainted with the basic features of the system; other videos, and online help materials, introduce students to basic concepts of programming in SAS software.

- Most of the data analyses or sample selections for this book's examples and exercises can be done with short programs (usually containing five or fewer lines) that follow a standard syntax.

  The survey analysis procedures in SAS/STAT® software, which at this writing include the SURVEYMEANS, SURVEYREG, SURVEYFREQ, SURVEYLOGISTIC, and SURVEYPHREG procedures, are specifically designed to produce estimates from complex surveys. The procedures can calculate either linearization-based variance estimates (used in Chapters 1–8) or the replication variance estimates described in Chapter 9, and they will construct replicate weights for a survey design that you specify. They will also produce appropriate survey-weighted plots of the data. The output provides the statistics you request as well as additional information that allows you to verify the design and weighting information used in the analysis. The procedures also print warnings if you have written code that is associated with some common survey data analysis errors.

  The SURVEYSELECT procedure will draw every type of probability sample discussed in this book, again with output that confirms the procedure used to draw the sample.

- SAS software is designed to allow you to manipulate and manage large data sets (some survey data sets contain tens of thousands or even millions of records), and compute estimates for those data sets using numerically stable and efficient algorithms. Many large survey data sets (such as the National Health and Nutrition Examination Survey data discussed in Chapter 7) are distributed as SAS data sets; you can also import files from spreadsheet programs, comma- or tab-delimited files, and other formats.

- The software is backward compatible—that is, code written for previous versions of the software will continue to work with newer versions. All programs are thoroughly tested before release, and the customer support team resolves any problems with the software that users might discover after release (they do not answer questions about how to do homework problems, though!). Appendix 5 of SAS Institute Inc. (2020) describes the methods used to quality-check and validate statistical procedures in SAS software.

- You do not need to learn computer programming to perform standard survey data analyses with SAS software. But for advanced users, the software offers the capability to write programs in SAS/IML® software or use macros. In addition, many user-contributed macros that perform specialized analyses of survey data have been published.

Features of the R statistical software environment for survey data:

- The software is available FREE from `https://www.r-project.org/`. It is open-source software, which means anyone can use it without a license or fee. Many tutorials on how to use R are available online; these tell you how to use the software to compute statistics and to create customized graphics.

- Base R contains functions that will select and analyze data from simple random samples. To select and analyze data from other types of samples, however—those discussed

after Chapter 2 of this book—R users must either (1) write their own R functions or (2) use functions that have been developed by other R users and made available through a contributed package. As of September 2020, the Comprehensive R Archive Network (CRAN) contained more than 16,000 contributed packages. If a statistical method has been published, there is a good chance that someone has developed a contributed package for R that performs the computations.

Contributed packages for R are not peer-reviewed or quality-checked unless the package authors arrange for such review. Functions in base R and contributed packages can change at any time, and are not always backward compatible.

But the open-source nature of R means that other users can view and test the functions in the packages. The book by Lu and Lohr (2022) makes use of functions in two popular contributed packages that have been developed for survey data by Lumley (2020) and by Tillé and Matei (2021). These functions will compute estimates and select samples for every type of probability sampling design discussed in *Sampling: Design and Analysis, Third Edition*.

- You need to learn how to work with functions in R in order to use it to analyze or select surveys. After you have gained experience with R, however, you can write functions to produce estimates for new statistical methods or to conduct simulation studies such as that requested in Exercise 21 of Chapter 4.

Software packages other than SAS and R can also be used with the book, as long as they have programs that correctly calculate estimates from complex survey data. Brogan (2015) illustrated the errors that result when non-survey software is used to analyze data from a complex survey. Software packages with survey data capabilities include SUDAAN® (RTI International, 2012), Stata® (Kolenikov, 2010), SPSS® (Zou et al., 2020), Mplus® (Muthén and Muthén, 2017), WesVar® (Westat, 2015), and IVEware (Raghunathan et al., 2016). See West et al. (2018) for reviews of these and other packages. New computer programs for analyzing survey data are developed all the time; the newsletter of the International Association of Survey Statisticians (`http://isi-iass.org`) is a good resource for updated information.

Website for the book. The book's companion website can be reached from either of the following addresses:

`https://www.sharonlohr.com`
`https://www.routledge.com/`.

It contains links to:

- Downloadable pdf files for the supplementary books *SAS® Software Companion for Sampling: Design and Analysis, Third Edition* and *R Companion for Sampling: Design and Analysis, Third Edition*. The pdf files are identical to the published paperback versions of the books.
- All data sets referenced in the book. These are available in comma-delimited (.csv), SAS, or R format. The data sets in R format are also available in the R contributed package *SDAResources* (Lu and Lohr, 2021).
- Other resources related to the book.

A solutions manual for the book is available (for instructors only) from the publisher at `https://www.routledge.com/`.