SAMPLING METHODOLOGIES
WITH APPLICATIONS

Poduri S.R.S. Rao

Professor of Statistics
University of Rochester
Rochester, New York

CHAPMAN & HALL/CRC

Boca Raton London New York Washington, D.C.

© 2000 by Chapman & Hall/CRC

Library of Congress Cataloging-in-Publication Data

Rao, Poduri S.R.S.
Sampling methodologies : with applications / Poduri S.R.S. Rao.
p. cm. --(Texts in statistical science)
Includes bibliographical references and index.
ISBN 1-58488-214-X (alk. paper)
1. Social sciences—Statistical methods. 2. Sampling (Statistics) I. Title. II. Series.

HA31.2 .M555 2000
519.5'2—dc21 00-063926

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Neither this book nor any part may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage or retrieval system, without prior permission in writing from the publisher.

The consent of CRC Press LLC does not extend to copying for general distribution, for promotion, for creating new works, or for resale. Specific permission must be obtained in writing from CRC Press LLC for such copying.

Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation, without intent to infringe.

© 2000 by Chapman & Hall/CRC

No claim to original U.S. Government works
International Standard Book Number 1-58488-214-X
Library of Congress Card Number 00-063926
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

Contents

Preface

1 Introduction
1.1 Introduction
1.2 Censuses and surveys
1.3 Types of surveys
1.4 Sampling frame
1.5 Questionnaires, interviews, and sample sizes
Questionnaires
Interviewing methods
Sample sizes
1.6 Probability sampling
1.7 Nonprobability sampling
Purposive sampling
Quota sampling
Other types of surveys
1.8 Sampling in practice
Demographic and economic surveys
Educational surveys
Employment and expenditure surveys
Health and nutrition surveys
Agricultural surveys
Marketing surveys
Election surveys
Public polls and surveys
Campus surveys
Exercises

2 Simple Random Sampling: Estimation of Means and Totals
2.1 Introduction
2.2 Population total, mean, and variance
Total and mean

© 2000 by Chapman & Hall/CRC

Variance and standard deviation (S.D.)
Coefficient of variation (C.V.)
2.3 Sampling without replacement
2.4 Sample mean and variance
2.5 Properties of simple random sampling
2.6 Unbiasedness of the sample mean and variance
Sample mean
Sample variance
2.7 Standard error of the sample mean
2.8 Distribution of the sample mean
2.9 Confidence limits for the mean
2.10 Estimating the total
2.11 Coefficient of variation of the sample mean
2.12 Sample size determination
Precision
Error of estimation
Confidence width
Coefficient of variation
Relative error
2.13 Sample sizes for individual groups
Exercises
Appendix A2
Expected value, variance, and standard deviation
Tschebycheff inequality
Unbiasedness of the sample variance
Variance of the sample mean

3 Simple Random Sampling: Related Topics
3.1 Introduction
3.2 Bias, variance, and mean square error of an estimator
3.3 Precision, accuracy, and consistency
3.4 Covariance and correlation
Ungrouped data
Grouped data
3.5 Linear combination of two characteristics
Linear combination
The sum
The difference
Covariance and correlation of the sample means
3.6 Systematic sampling

© 2000 by Chapman & Hall/CRC

Sample selection
Expectation and variance
Comparison with simple random sampling
Further properties
3.7 Nonresponse
3.8 Inference regarding the means
Population mean
Difference of two population means
Correlated samples
3.9 Sampling with replacement
Exercises
Appendix A3
Joint distribution
Expectation and variance
Covariance and correlation
Conditional and unconditional expectations and variances
Variance of a linear combination
An expression for S^2
Cauchy–Schwartz inequality

4 **Proportions, Percentages, and Counts**
4.1 Introduction
4.2 Two classes
4.3 Total number or count
4.4 Confidence limits
4.5 Sample sizes for specified requirement
4.6 Political and public polls
4.7 Estimation for more than two classes
Probabilities of the outcomes
Multinomial approximation
4.8 Combining estimates from surveys
4.9 Covariances of proportions
4.10 Difference between proportions
4.11 Sample sizes for more than two classes
4.12 Estimating population size
Exercises
Appendix A4
Unbiasedness of the sample proportion and variance
Binomial approximation

© 2000 by Chapman & Hall/CRC

5 Stratification
5.1 Introduction
5.2 Notation
5.3 Estimation for a single stratum
5.4 Estimation of the population mean and total
5.5 Confidence limits
5.6 Proportions and totals
5.7 Population total and proportion
5.8 Proportional and equal division of the sample
Proportional allocation
Equal distribution
5.9 Neyman allocation
5.10 Gains from Neyman allocation
5.11 Summary on the precision of the allocations
5.12 Sample size allocation to estimate proportions
Proportional allocation
Neyman allocation
5.13 Sample sizes to estimate means and totals
5.14 Sample sizes to estimate proportions
5.15 Sample sizes for minimizing variance or cost
5.16 Further topics
Poststratification
Estimating differences of the stratum means
Two-phase sampling for stratification
Sample size determination for two or more characteristics
Strata formation
Exercises
Appendix A5
An expression for the variance
Variance of the estimator for the mean
Neyman allocation
Optimum allocation

6 Subpopulations
6.1 Introduction
6.2 Totals, means, and variances
6.3 Estimation of the means and their differences
Means
Difference between two groups
6.4 Totals of subpopulations
Known sizes
Estimated sizes

© 2000 by Chapman & Hall/CRC

6.5 Sample sizes for estimating the means and totals
6.6 Proportions and counts
6.7 Subpopulations of a stratum
6.8 Totals and means of subpopulations in strata
6.9 Stratification: Proportions and counts of subpopulations
Exercises

7 Cluster Sampling
7.1 Introduction
7.2 Clusters of equal sizes
7.3 Estimation of the means
7.4 Comparison with simple random sampling
7.5 Estimation of the standard error
7.6 Optimum cluster and sample sizes
7.7 Clusters of unequal size
7.8 Alternative estimation with unequal sizes
7.9 Proportions and percentages
Average of the proportions
Proportion of the totals or means
7.10 Stratification
7.11 Unequal probability selection
A method of selecting the cluster units
Probabilities of selecting the units
7.12 Horvitz–Thompson estimator
Variance of the estimator
Variance estimators
Estimators for the mean per unit and the population mean
7.13 Alternative approaches
Probabilities proportional to the sum of sizes
Randomization approach
Selection with probabilities proportional to size and replacement
Exercises
Appendix A7
Alternative expression for S^2
Estimator for S^2
Bias of the ratio estimator for the mean
Variance of the ratio estimato

© 2000 by Chapman & Hall/CRC

Expectation and variance of the Horvitz–Thompson estimator
Expectation and variance of the PPSS estimator
Variance of pps estimator and the variance estimator

8 **Sampling in Two Stages**
8.1 Introduction
8.2 Equal size first-stage units
8.3 Estimating the mean
Estimator for the mean and its variance
Estimation of the variance
8.4 Sample size determination
8.5 Unequal size primary units
Estimators for the population total and mean
Estimation of the variance
8.6 Ratio adjustment for the sizes
8.7 Proportions and totals
Equal size PSUs
Unequal size PSUs
An alternative estimator
8.8 Unequal probability selectio
Horvitz-Thompson estimator
Selection with unequal probabilities and replacement
Exercises
Appendix A8
Expected value and variance of the estimator for the mean
Expectation and variance of the estimator for the total
Estimation of the variance—unequal cluster sizes
Variance of the ratio estimator for the total
Expectation and variance of $\hat{Y}_{HT}$
Estimator for $V(\hat{Y}_{HT})$
Variance of $\hat{Y}_{pps}$ and its estimator

9 **Ratios and Ratio Estimators**
9.1 Introduction

© 2000 by Chapman & Hall/CRC

9.2 Bias and variance of the sample ratio
9.3 Confidence limits for the ratio of means
9.4 Ratio estimators for the mean and total
9.5 Confidence limits for the mean and total
9.6 Differences between ratios, means, or totals
A common supplementary variable
Different supplementary variables
9.7 Regression through the origin and the BLUEs
Least squares
Superpopulation model for evaluations and estimation
9.8 Ratio estimation vs. stratification
9.9 Ratio estimation with stratification
Single stratum
The separate ratio estimator
The combined ratio estimator
Statistical test for the equality of the slopes
9.10 Bias reduction
9.11 Two-phase or double sampling ratio estimators
9.12 Ratio estimator with unequal probability selection
9.13 Multivariate ratio estimator
Exercises
Appendix A9
Bias of the ratio of two sample means
MSE of the ratio of two sample means
Variance of the combined ratio estimator

10 Regression Estimation
10.1 Introduction
10.2 The regression estimator
10.3 Estimation from the sample
10.4 Classical linear regression
10.5 Difference between regression estimators
Common supplementary variable
Different supplementary variables
10.6 Regression estimation vs. stratification
10.7 Stratification and regression estimator
Separate estimator
Combined regression estimator
Preliminary analysis and tests of hypotheses

© 2000 by Chapman & Hall/CRC

10.8 Multiple regression estimator
10.9 Double sampling regression estimator
10.10 Generalized regression and calibration estimators
Exercises

11 Nonresponse and Remedies
11.1 Introduction
11.2 Effects of survey topics and interviewing methods
Subject matter
Types of interviewing
Types of respondents
11.3 Response rates
11.4 Bias and MSE
11.5 Estimating proportions
11.6 Subsampling the nonrespondents
The procedure of Hansen and Hurwitz
Estimator and its variance
Optimum sample sizes
Variance estimation
Ratio and regression estimators
11.7 Estimating the missing observations
Least squares estimation
Alternative predictions
11.8 Ratio and regression estimation
Ratio estimation
Regression estimation
11.9 Poststratification and weighting
Estimators for the mean and total
Subpopulations
11.10 Response probabilities and weighting
11.11 Imputation
Mean imputation
Random duplication
Variance estimation
Confidence limits
Multiple imputation
Hot- and cold-deck imputation
11.12 Related topics
Efforts to increase response rates
Imputation, estimation, and analysis
Models for prediction, poststratification, and weighting
Qualitative and categorical observations

© 2000 by Chapman & Hall/CRC

11.13 Bayesian procedures
11.14 Census undercount
Exercises
Appendix A11
Variance of the Hansen–Hurwitz estimator
Variance estimator for the Hansen–Hurwitz estimator

12 Further Topics
12.1 Introduction
12.2 Linearization
12.3 The jackknife
Bias reduction
Variance estimation
12.4 The bootstrap
Infinite population
Finite population
Bootstrap for estimating the variance of the sample ratio
12.5 Balanced repeated replication (BRR)
12.6 Small-area estimation
Sample estimates
Synthetic and composite estimators
Best linear unbiased predictor (BLUP)
Applications, extensions, and evaluations
12.7 Complex surveys
Exercises
Appendix A12
The Bayesian approach

Appendix Tables

Solutions to Exercises

References

© 2000 by Chapman & Hall/CRC

List of tables

2.1 SAT verbal and math scores

2.2 All possible samples of size two; means and variances of the math scores

2.3 Distribution of the sample mean

2.4 Frequency distribution

3.1 Joint distribution of age and education in 1995 in the 50 states and DC

3.2 Arrangement of 40 population units into four systematic samples

3.3 Systematic samples of the savings ($1000) of 40 employees

4.1 Presidential election—Percentages of voters for the three candidates

4.2 Public surveys—Percentages of responses for different items

4.3 Classification of voters' participation in the primaries on Super Tuesday, March, 2000 (percentages)

5.1 Rice production (in million metric tons)

5.2 Strata and samples

5.3 Variances for simple random sampling and proportional and Neyman allocations; relative precision of Neyman allocation

6.1 Means (top) and standard deviations (bottom) for a sample of 40 hospitals

© 2000 by Chapman & Hall/CRC

6.2 Household composition

6.3 Usage of athletic facilities

7.1 Employment in 54 counties of New York State, divided into nine clusters of size six each; number of establishments (1000s) and number of persons employed (1000s)

7.2 Means and variances of the nine clusters of Table 7.1

7.3 Unequal size clusters: totals and variances

7.4 Composition of the 54 counties

7.5 Probabilities of selection

8.1 A sample of three clusters of equal sizes ($M = 6$) and samples of three counties from the selected clusters

8.2 Sample of clusters of unequal size and sample of counties

9.1 Heights, weights, and blood pressures

9.2 Wheat production

11.1 Survey of families

11.2 Two types of stratification

11.3 Part-time employment of students

12.1 Jackknife calculations for college enrollment

T1 1920 random numbers

T2 SAT scores of 30 candidates

T3 Persons over age 25 in the U.S. (1000s), 1995

T4 Ten states with the largest college enrollments (1000s)

T5 Ten states with the smallest college enrollments (1000s)

T6 Blood pressures of 20 persons (mm Hg)

T7 Physicians in the U.S. in 1995 and hospital expenditure for a sample of ten states

T8 Wheat production in 25 countries (million metric tons)

T9 Energy production and consumption in 1996 in 25 countries

© 2000 by Chapman & Hall/CRC

List of figures

2.1 Approximate distribution of the sample mean for the math scores

3.1 Joint distribution of the number of persons over 25 years of age and the percentage of college graduates in the 50 states and DC

5.1 Precision of Neyman allocation relative to simple random sampling (top) and proportional allocation (bottom)

6.1 Usage of the athletic facility by the male and female groups

9.1 Regression through the origin of weight on height

10.1 Regression of 1997 wheat production on 1995 production

10.2 Regression of the wheat productions for the two strata

11.1 Regressions of income on husband's age and wife's age

© 2000 by Chapman & Hall/CRC

To Durgi, Ann, and Pal

© 2000 by Chapman & Hall/CRC

Preface

This book contains the methodologies and applications of commonly used procedures of sampling from finite populations. Easy access of the different topics to students and their quick availability to practitioners have been the objectives in preparing this book. With these objectives, the main results along with their logical explanations are presented in the text, but the related derivations are placed in the appendixes to the chapters.

The introductory chapter describes the differences between probability and nonprobability sampling, and presents illustrations of several types of national and international surveys. Properties of simple random sampling along with the estimation of the population mean, total, and variance are considered in the second chapter. Unbiasedness of the sample mean, its variance, and standard error are defined and illustrated. Sample sizes needed to estimate the population mean or total with specified criteria are also examined in this chapter.

The third chapter contains a variety of topics of importance in sample surveys and for statistical analysis in general. Definitions of the bias, variance, and mean square error of an estimator and their relevance to its precision and accuracy are presented and illustrated first. Further topics include the covariance and correlation between two random variables and their sample means, tests of hypotheses related to the means, and the comparison of simple random and systematic random sampling procedures. The effect of the bias arising from the nonresponse in a survey is briefly examined in this chapter, and this topic is continued in the exercises of some of the following chapters and finally in detail in Chapter 11. The appendix to this chapter contains the basic results on conditional and unconditional expectation and variance, and other topics of importance.

Chapters 4 through 11 contain detailed presentations of the estimation of percentages and counts, stratification, subpopulations, single and two-stage cluster sampling with equal and unequal probabilities,

© 2000 by Chapman & Hall/CRC

ratio and regression methods of estimation, and the problem of nonresponse and its remedies. The final chapter presents in detail linearization, jackknife, bootstrap, and balanced repeated replication procedures. Major results on small-area estimation and complex surveys appear in this chapter.

Each topic is presented with illustrations, followed by examples and exercises. They are constructed from data on everyday practical situations covering a wide variety of subjects ranging from scholastic aptitude tests to health care expenditures and presidential elections. The examples and exercises are interwoven throughout different chapters.

For the sake of good comprehension of the results on unbiasedness, standard error, and mean square error of the estimators, some of the exercises are constructed as projects that require the selection of all the possible samples of a specified size from a finite population. For some of the exercises and projects, Minitab, Excel, and similar computer programs are quite adequate. The methodological type of questions at the end of the exercises along with the derivations in the appendixes should be of particular interest to advanced students. Solutions to the odd-numbered exercises are presented at the end of the book. Detailed solutions to all the exercises appear in the *Teacher's Manual.*

This book can be recommended as a text for a one-semester or two-quarter course for students in statistics and also in business, political and social sciences, and other areas. One or two courses in basic theoretical and applied statistical concepts would provide the required preparation. This book can also serve as a reference guide for survey practitioners, political and public pollsters, and researchers in some industries.

I am grateful to Dr. Ramana L. Rao for patiently assembling the entire manuscript on the word processor. I would like to thank Bob Stern for his interest in this project and his editorial suggestions. Thanks to Megan Brooks, Sharon Gresham, and Chris Petrosino for checking the solutions to some of the exercises and their suggestions for presenting the solutions to the examples. I would also like to thank my wife Durgi and children Ann and Pal for their encouragement and patience throughout the years spent preparing this manuscript.

Poduri S.R.S. Rao
University of Rochester, New York

© 2000 by Chapman & Hall/CRC