

Sampling Techniques 3rd Ed.

William G. Cochran



Sampling Techniques

third edition

WILLIAM G. COCHRAN

*Professor of Statistics, Emeritus
Harvard University*

John Wiley & Sons

New York · Santa Barbara · London · Sydney · Toronto

Copyright © 1977, by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

No part of this book may be reproduced by any means, nor transmitted, nor translated into a machine language without the written permission of the publisher.

Library of Congress Cataloging in Publication Data:

Cochran, William Gemmell, 1909-
Sampling techniques.

(Wiley series in probability and mathematical statistics)
Includes bibliographical references and index.

1. Sampling (Statistics) I. Title.

QA276.6.C6 1977 001.4'222 77-728

ISBN 0-471-16240-X

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Preface

As did the previous editions, this textbook presents a comprehensive account of sampling theory as it has been developed for use in sample surveys. It contains illustrations to show how the theory is applied in practice, and exercises to be worked by the student. The book will be useful both as a text for a course on sample surveys in which the major emphasis is on theory and for individual reading by the student.

The minimum mathematical equipment necessary to follow the great bulk of the material is a familiarity with algebra, especially relatively complicated algebraic expressions, plus a knowledge of probability for finite sample spaces, including combinatorial probabilities. The book presupposes an introductory statistics course that covers means and standard deviations, the normal, binomial, hypergeometric, and multinomial distributions, the central limit theorem, linear regression, and the simpler types of analyses of variance. Since much of classical sample survey theory deals with the distributions of estimators over the set of randomizations provided by the sampling plan, some knowledge of nonparametric methods is helpful.

The topics in this edition are presented in essentially the same order as in earlier editions. New sections have been included, or sections rewritten, primarily for one of three reasons: (1) to present introductions to topics (sampling plans or methods of estimation) relatively new in the field; (2) to cover further work done during the last 15 years on older methods, intended either to improve them or to learn more about the performance of rival methods; and (3) to shorten, clarify, or simplify proofs given in previous editions.

New topics in this edition include the approximate methods developed for the difficult problem of attaching standard errors or confidence limits to nonlinear estimates made from the results of surveys with complex plans. These methods will be more and more needed as statistical analyses (e.g., regressions) are performed on the results. For surveys containing sensitive questions that some respondents are unlikely to be willing to answer truthfully, a new device is to present the respondent with either the sensitive question or an innocuous question; the specific choice, made by randomization, is unknown to the interviewer. In some sampling problems it may seem economically attractive, or essential in countries without full sampling resources, to use two overlapping lists (or frames, as they are called) to cover the complete population. The method of double sampling has been extended to cases where the objective is to compare the means

of a number of subgroups within the population. There has been interesting work on the attractive properties that the ratio and regression estimators have if it can be assumed that the finite population is itself a random sample from an infinite superpopulation in which a mathematical model appropriate to the ratio or regression estimator holds. This kind of assumption is now new—I noticed recently that Laplace used it around 1800 in a sampling problem—but it clarifies the relation between sample survey theory and standard statistical theory.

An example of further work on topics included in previous editions is Chapter 9A, which has been written partly from material previously in Chapter 9; this was done mainly to give a more adequate account of what seem to me the principal methods produced for sampling with unequal probabilities without replacement. These include the similar methods given independently by Brewer, J. N. K. Rao, and Durbin, Murthy's method, the Rao, Hartley, Cochran method, and Madow's method related to systematic sampling, with comparisons of the performances of the methods on natural populations. New studies have been done of the sizes of components of errors of measurement in surveys by repeat measurements by different interviewers, by interpenetrating subsamples, and by a combination of the two approaches. For the ratio estimator, data from natural populations have been used to appraise the small-sample biases in the standard large-sample formulas for the variance and the estimated variance. Attempts have also been made to create less biased variants of the ratio estimator itself and of the formula for estimating its sampling variance. In stratified sampling there has been additional work on allocating sample sizes to strata when more than one item is of importance and on estimating sample errors when only one unit is to be selected per stratum. Some new systematic sampling methods for handling populations having linear trends are also of interest.

Alva L. Finkner and Emil H. Jebe prepared a large part of the lecture notes from which the first edition of this book was written. Some investigations that provided background material were supported by the Office of Naval Research, Navy Department. From discussions of recent developments in sampling or suggestions about this edition, I have been greatly helped by Tore Dalenius, David J. Finney, Daniel G. Horvitz, Leslie Kish, P. S. R. Sambasiva Rao, Martin Sandelius, Joseph Sedransk, Amode R. Sen, and especially Jon N. K. Rao, whose painstaking reading of the new and revised sections of this edition resulted in many constructive suggestions about gaps, weaknesses, obscurities, and selection of topics. For typing and other work involved in production of a typescript I am indebted to Rowena Foss, Holly Grano, and Edith Klotz. My thanks to all.

William G. Cochran

South Orleans, Massachusetts
February, 1977

Contents

CHAPTER	PAGE
1 INTRODUCTION	1
1.1 Advantages of the Sampling Method	1
1.2 Some Uses of Sample Surveys	2
1.3 The Principal Steps in a Sample Survey	4
1.4 The Role of Sampling Theory	8
1.5 Probability Sampling	9
1.6 Alternatives to Probability Sampling	10
1.7 Use of the Normal Distribution	11
1.8 Bias and Its Effects	12
1.9 The Mean Square Error	15
<i>Exercises</i>	16
CHAPTER	
2 SIMPLE RANDOM SAMPLING	18
2.1 Simple Random Sampling	18
2.2 Selection of a Simple Random Sample	19
2.3 Definitions and Notation	20
2.4 Properties of the Estimates	21
2.5 Variances of the Estimates	23
2.6 The Finite Population Correction	24
2.7 Estimation of the Standard Error from a Sample	25
2.8 Confidence Limits	27
2.9 An Alternative Method of Proof	28
2.10 Random Sampling with Replacement	29
2.11 Estimation of a Ratio	30
2.12 Estimates of Means Over Subpopulations	34
2.13 Estimates of Totals Over Subpopulations	35
2.14 Comparisons Between Domain Means	39
2.15 Validity of the Normal Approximation	39
2.16 Linear Estimators of the Population Mean	44
<i>Exercises</i>	45

CHAPTER

3 SAMPLING PROPORTIONS AND PERCENTAGES 50

3.1 Qualitative Characteristics 50

3.2 Variances of the Sample Estimates 50

3.3 The Effect of P on the Standard Errors 53

3.4 The Binomial Distribution 55

3.5 The Hypergeometric Distribution 55

3.6 Confidence Limits 57

3.7 Classification into More than Two Classes 60

3.8 Confidence Limits with More than Two Classes 60

3.9 The Conditional Distribution of p 61

3.10 Proportions and Totals Over Subpopulations 63

3.11 Comparisons Between Different Domains 64

3.12 Estimation of Proportions in Cluster Sampling 64

Exercises 68

CHAPTER

4 THE ESTIMATION OF SAMPLE SIZE 72

4.1 A Hypothetical Example 72

4.2 Analysis of the Problem 73

4.3 The Specification of Precision 74

4.4 The Formula for n in Sampling for Proportions 75

4.5 Rare Items—Inverse Sampling 76

4.6 The Formula for n with Continuous Data 77

4.7 Advance Estimates of Population Variances 78

4.8 Sample Size with More than One Item 81

4.9 Sample Size when Estimates Are Wanted for Subdivisions of the Population 82

4.10 Sample Size in Decision Problems 83

4.11 The Design Effect ($Deff$) 85

Exercises 86

CHAPTER

5 STRATIFIED RANDOM SAMPLING 89

5.1 Description 89

5.2 Notation 90

5.3 Properties of the Estimates 91

5.4 The Estimated Variance and Confidence Limits 95

5.5 Optimum Allocation 96

5.6	Relative Precision of Stratified Random and Simple Random Sampling	99
5.7	When Does Stratification Produce Large Gains in Precision? . . .	101
5.8	Allocation Requiring More than 100 Per Cent Sampling	104
5.9	Estimation of Sample Size with Continuous Data	105
5.10	Stratified Sampling for Proportions	107
5.11	Gains in Precision in Stratified Sampling for Proportions	109
5.12	Estimation of Sample Size with Proportions	110
	<i>Exercises</i>	111

CHAPTER

5A	FURTHER ASPECTS OF STRATIFIED SAMPLING	115
5A.1	Effects of Deviations from the Optimum Allocation	115
5A.2	Effects of Errors in the Stratum Sizes	117
5A.3	The Problem of Allocation with More than One Item	119
5A.4	Other Methods of Allocation with More than One Item	121
5A.5	Two-Way Stratification with Small Samples	124
5A.6	Controlled Selection	126
5A.7	The Construction of Strata	127
5A.8	Number of Strata	132
5A.9	Stratification After Selection of the Sample (Poststratification) . . .	134
5A.10	Quota Sampling	135
5A.11	Estimation from a Sample of the Gain Due to Stratification	136
5A.12	Estimation of Variance with One Unit per Stratum	138
5A.13	Strata as Domains of Study	140
5A.14	Estimating Totals and Means Over Subpopulations	142
5A.15	Sampling from Two Frames	144
	<i>Exercises</i>	146

CHAPTER

6	RATIO ESTIMATORS	150
6.1	Methods of Estimation	150
6.2	The Ratio Estimate	150
6.3	Approximate Variance of the Ratio Estimate	153
6.4	Estimation of the Variance from a Sample	155
6.5	Confidence Limits	156
6.6	Comparison of the Ratio Estimate with Mean per Unit	157
6.7	Conditions Under Which the Ratio Estimate Is a Best Linear Unbiased Estimator	158
6.8	Bias of the Ratio Estimate	160

6.9 Accuracy of the Formulas for the Variance and Estimated Variance	162
6.10 Ratio Estimates in Stratified Random Sampling	164
6.11 The Combined Ratio Estimate	165
6.12 Comparison of the Combined and Separate Estimates	167
6.13 Short-Cut Computation of the Estimated Variance	169
6.14 Optimum Allocation with a Ratio Estimate	172
6.15 Unbiased Ratio-type Estimates	174
6.16 Comparison of the Methods	177
6.17 Improved Estimation of Variance	178
6.18 Comparison of Two Ratios	180
6.19 Ratio of Two Ratios	183
6.20 Multivariate Ratio Estimates	184
6.21 Product Estimators	186
<i>Exercises</i>	186

CHAPTER

7 REGRESSION ESTIMATORS 189

7.1 The Linear Regression Estimate	189
7.2 Regression Estimates with Preassigned b	190
7.3 Regression Estimates when b Is Computed from the Sample	193
7.4 Sample Estimate of Variance	195
7.5 Large-Sample Comparison with the Ratio Estimate and the Mean per Unit	195
7.6 Accuracy of the Large-Sample Formulas for $V(\bar{y}_r)$ and $v(\bar{y}_r)$	197
7.7 Bias of the Linear Regression Estimate	198
7.8 The Linear Regression Estimator Under a Linear Regression Model	199
7.9 Regression Estimates in Stratified Sampling	200
7.10 Regression Coefficients Estimated from the Sample	201
7.11 Comparison of the Two Types of Regression Estimate	203
<i>Exercises</i>	203

CHAPTER

8 SYSTEMATIC SAMPLING 205

8.1 Description	205
8.2 Relation to Cluster Sampling	207
8.3 Variance of the Estimated Mean	207
8.4 Comparison of Systematic with Stratified Random Sampling	212
8.5 Populations in "Random" Order	212

8.6	Populations with Linear Trend	214
8.7	Methods for Populations with Linear Trends	216
8.8	Populations with Periodic Variation	217
8.9	Autocorrelated Populations	219
8.10	Natural Populations	221
8.11	Estimation of the Variance from a Single Sample	223
8.12	Stratified Systematic Sampling	226
8.13	Systematic Sampling in Two Dimensions	227
8.14	Summary	229
	<i>Exercises</i>	231

CHAPTER

9 SINGLE-STAGE CLUSTER SAMPLING: CLUSTERS OF EQUAL SIZES 233

9.1	Reasons for Cluster Sampling	233
9.2	A Simple Rule	234
9.3	Comparisons of Precision Made from Survey Data	238
9.4	Variance in Terms of Intracluster Correlation	240
9.5	Variance Functions	243
9.6	A Cost Function	244
9.7	Cluster Sampling for Proportions	246
	<i>Exercises</i>	247

CHAPTER

9A SINGLE-STAGE CLUSTER SAMPLING: CLUSTERS OF UNEQUAL SIZES 249

9A.1	Cluster Units of Unequal Sizes	249
9A.2	Sampling with Probability Proportional to Size	250
9A.3	Selection with Unequal Probabilities with Replacement	252
9A.4	The Optimum Measure of Size	255
9A.5	Relative Accuracies of Three Techniques	255
9A.6	Sampling with Unequal Probabilities Without Replacement	258
9A.7	The Horvitz-Thompson Estimator	259
9A.8	Brewer's Method	261
9A.9	Murthy's Method	263
9A.10	Methods Related to Systematic Sampling	265
9A.11	The Rao, Hartley, Cochran Method	266
9A.12	Numerical Comparisons	267
9A.13	Stratified and Ratio Estimates	270
	<i>Exercises</i>	272

CHAPTER

10 SUBSAMPLING WITH UNITS OF EQUAL SIZE 274

10.1	Two-Stage Sampling	274
10.2	Finding Means and Variances in Two-Stage Sampling	275
10.3	Variance of the Estimated Mean in Two-Stage Sampling	276
10.4	Sample Estimation of the Variance	278
10.5	The Estimation of Proportions	279
10.6	Optimum Sampling and Subsampling Fractions	280
10.7	Estimation of m_{opt} from a Pilot Survey	283
10.8	Three-Stage Sampling	285
10.9	Stratified Sampling of the Units	288
10.10	Optimum Allocation with Stratified Sampling	289
	<i>Exercises</i>	290

CHAPTER

11 SUBSAMPLING WITH UNITS OF UNEQUAL SIZES 292

11.1	Introduction	292
11.2	Sampling Methods when $n = 1$	293
11.3	Sampling with Probability Proportional to Estimated Size	297
11.4	Summary of Methods for $n = 1$	299
11.5	Sampling Methods When $n > 1$	300
11.6	Two Useful Results	300
11.7	Units Selected with Equal Probabilities: Unbiased Estimator	303
11.8	Units Selected with Equal Probabilities: Ratio to Size Estimate	303
11.9	Units Selected with Unequal Probabilities with Replacement: Unbiased Estimator	306
11.10	Units Selected Without Replacement	308
11.11	Comparison of the Methods	310
11.12	Ratios to Another Variable	311
11.13	Choice of Sampling and Subsampling Fractions. Equal Probabilities	313
11.14	Optimum Selection Probabilities and Sampling and Subsampling Rates	314
11.15	Stratified Sampling. Unbiased Estimators	316
11.16	Stratified Sampling. Ratio Estimates	317
11.17	Nonlinear Estimators in Complex Surveys	318
11.18	Taylor Series Expansion	319
11.19	Balanced Repeated Replications	320
11.20	The Jackknife Method	321
11.21	Comparison of the Three Approaches	322
	<i>Exercises</i>	324

CHAPTER

12	DOUBLE SAMPLING	327
12.1	Description of the Technique	327
12.2	Double Sampling for Stratification	327
12.3	Optimum Allocation	331
12.4	Estimated Variance in Double Sampling for Stratification	333
12.5	Double Sampling for Analytical Comparisons	335
12.6	Regression Estimators	338
12.7	Optimum Allocation and Comparison with Single Sampling	341
12.8	Estimated Variance in Double Sampling for Regression	343
12.9	Ratio Estimators	343
12.10	Repeated Sampling of the Same Population	344
12.11	Sampling on Two Occasions	346
12.12	Sampling on More than Two Occasions	348
12.13	Simplifications and Further Developments	351
	<i>Exercises</i>	355

CHAPTER

13	SOURCES OF ERROR IN SURVEYS	359
13.1	Introduction	359
13.2	Effects of Nonresponse	359
13.3	Types of Nonresponse	361
13.4	Call-backs	365
13.5	A Mathematical Model of the Effects of Call-backs	367
13.6	Optimum Sampling Fraction Among the Nonrespondents	370
13.7	Adjustments for Bias Without Call-backs	374
13.8	A Mathematical Model for Errors of Measurement	377
13.9	Effects of Constant Bias	379
13.10	Effects of Errors that Are Uncorrelated Within the Sample	380
13.11	Effects of Intrasample Correlation Between Errors of Measurement	383
13.12	Summary of the Effects of Errors of Measurement	384
13.13	The Study of Errors of Measurement	384
13.14	Repeated Measurement of Subsamples	386
13.15	Interpenetrating Subsamples	388
13.16	Combination of Interpenetration and Repeated Measurement	391
13.17	Sensitive Questions: Randomized Responses	392
13.18	The Unrelated Second Question	393
13.19	Summary	395
	<i>Exercises</i>	396

References	400
Answers to Exercises	412
Author Index	419
Subject Index	422

CHAPTER 1

Introduction

1.1 ADVANTAGES OF THE SAMPLING METHOD

Our knowledge, our attitudes, and our actions are based to a very large extent on samples. This is equally true in everyday life and in scientific research. A person's opinion of an institution that conducts thousands of transactions every day is often determined by the one or two encounters he has had with the institution in the course of several years. Travelers who spend 10 days in a foreign country and then proceed to write a book telling the inhabitants how to revive their industries, reform their political system, balance their budget, and improve the food in their hotels are a familiar figure of fun. But in a real sense they differ from the political scientist who devotes 20 years to living and studying in the country only in that they base their conclusions on a much smaller sample of experience and are less likely to be aware of the extent of their ignorance. In science and human affairs alike we lack the resources to study more than a fragment of the phenomena that might advance our knowledge.

This book contains an account of the body of theory that has been built up to provide a background for good sampling methods. In most of the applications for which this theory was constructed, the aggregate about which information is desired is finite and delimited—the inhabitants of a town, the machines in a factory, the fish in a lake. In some cases it may seem feasible to obtain the information by taking a complete enumeration or census of the aggregate. Administrators accustomed to dealing with censuses were at first inclined to be suspicious of samples and reluctant to use them in place of censuses. Although this attitude no longer persists, it may be well to list the principal advantages of sampling as compared with complete enumeration.

Reduced Cost

If data are secured from only a small fraction of the aggregate, expenditures are smaller than if a complete census is attempted. With large populations, results accurate enough to be useful can be obtained from samples that represent only a small fraction of the population. In the United States the most important recurrent surveys taken by the government use samples of around 105,000

persons, or about one person in 1240. Surveys used to provide facts bearing on sales and advertising policy in market research may employ samples of only a few thousand.

Greater Speed

For the same reason, the data can be collected and summarized more quickly with a sample than with a complete count. This is a vital consideration when the information is urgently needed.

Greater Scope

In certain types of inquiry highly trained personnel or specialized equipment, limited in availability, must be used to obtain the data. A complete census is impracticable: the choice lies between obtaining the information by sampling or not at all. Thus surveys that rely on sampling have more scope and flexibility regarding the types of information that can be obtained. On the other hand, if accurate information is wanted for many subdivisions of the population, the size of sample needed to do the job is sometimes so large that a complete enumeration offers the best solution.

Greater Accuracy

Because personnel of higher quality can be employed and given intensive training and because more careful supervision of the field work and processing of results becomes feasible when the volume of work is reduced, a sample may produce more accurate results than the kind of complete enumeration that can be taken.

1.2 SOME USES OF SAMPLE SURVEYS

To an observer of developments in sampling over the last 25 years the most striking feature is the rapid increase in the number and types of surveys taken by sampling. The Statistical Office of the United Nations publishes reports from time to time on "Sample Surveys of Current Interest" conducted by member countries. The 1968 report lists surveys from 46 countries. Many of these surveys seek information of obvious importance to national planning on topics such as agricultural production and land use, unemployment and the size of the labor force, industrial production, wholesale and retail prices, health status of the people, and family incomes and expenditures. But more specialized inquiries can also be found: for example, annual leave arrangements (Australia), causes of divorce (Hungary), rural debt and investment (India), household water consumption (Israel), radio listening (Malaysia), holiday spending (Netherlands), age structure of cows (Czechoslovakia), and job vacancies (United States).

Sampling has come to play a prominent part in national decennial censuses. In the United States a 5% sample was introduced into the 1940-Census by asking

extra questions about occupation, parentage, fertility, and the like, of those persons whose names fell on two of the 40 lines on each page of the schedule. The use of sampling was greatly extended in 1950. From a 20% sample (every fifth line) information was obtained on items such as income, years in school, migration, and service in armed forces. By taking every sixth person in the 20% sample, a further sample of $3\frac{1}{3}\%$ was created to give information on marriage and fertility. A series of questions dealing with the condition and age of housing was split into five sets, each set being filled in at every fifth house. Sampling was also employed to speed up publication of the results. Preliminary tabulations for many important items, made on a sample basis, appeared more than a year and half before the final reports.

This process continued in the 1960 and 1970 Censuses. Except for certain basic information required from every person for constitutional or legal reasons, the whole census was shifted to a sample basis. This change, accompanied by greatly increased mechanization, resulted in much earlier publication and substantial savings.

In addition to their use in censuses, continuing samples are employed by government bureaus to obtain current information. In the United States, examples are the Current Population Survey, which provides monthly data on the size and composition of the labor force and on the number of unemployed, the National Health Survey, and the series of samples needed for the calculation of the monthly Consumer Price Index.

On a smaller scale, local governments—city, state, and county—are making increased use of sample surveys to obtain information needed for future planning and for meeting pressing problems. In the United States most large cities have commercial agencies that make a business of planning and conducting sample surveys for clients.

Market research is heavily dependent on the sampling approach. Estimates of the sizes of television and radio audiences for different programs and of newspaper and magazine readership (including the advertisements) are kept continually under scrutiny. Manufacturers and retailers want to know the reactions of people to new products or new methods of packaging, their complaints about old products, and their reasons for preferring one product to another.

Business and industry have many uses for sampling in attempting to increase the efficiency of their internal operations. The important areas of quality control and acceptance sampling are outside the scope of this book. But, obviously, decisions taken with respect to level or change of quality or to acceptance or rejection of batches are well grounded only if results obtained from the sample data are valid (within a reasonable tolerance) for the whole batch. The sampling of records of business transactions (accounts, payrolls, stock, personnel)—usually much easier than the sampling of people—can provide serviceable information quickly and economically. Savings can also be made through sampling in the estimation of inventories, in studies of the condition and length of the life of equipment, in the

inspection of the accuracy and rate of output of clerical work, in investigating how key personnel distribute their working time among different tasks, and, more generally, in the field known as operations research. The books by Deming (1960) and Slonim (1960) contain many interesting examples showing the range of applications of the sampling method in business.

Opinion, attitude, and election polls, which did much to bring the technique of sampling before the public eye, continue to be a popular feature of newspapers. In the field of accounting and auditing, which has employed sampling for many years, a new interest has arisen in adapting modern developments to the particular problems of this field. Thus, Neter (1972) describes how airlines and railways save money by using samples of records to apportion income from freight and passenger service. The status of sample surveys as evidence in lawsuits has also been subject to lively discussion. Gallup (1972) has noted the major contribution that sample surveys can make to the process of informed government by determining quickly people's opinions on proposed or new government programs and has stressed their role as sources of information in social science.

Sample surveys can be classified broadly into two types—*descriptive* and *analytical*. In a descriptive survey the objective is simply to obtain certain information about large groups: for example, the numbers of men, women, and children who view a television program. In an analytical survey, comparisons are made between different subgroups of the population, in order to discover whether differences exist among them and to form or to verify hypotheses about the reasons for these differences. The Indianapolis fertility survey, for instance, was an attempt to determine the extent to which married couples plan the number and spacing of children, the husband's and wife's attitudes toward this planning, the reasons for these attitudes, and the degree of success attained (Kiser and Whelpton, 1953).

The distinction between descriptive and analytical surveys is not, of course, clear-cut. Many surveys provide data that serve both purposes. Along with the rise in the number of descriptive surveys, there has, however, been a noticeable increase in surveys taken primarily for analytical purposes, particularly in the study of human behavior and health. Surveys of the teeth of school children before and after fluoridation of water, of the death rates and causes of death of people who smoke different amounts, and the huge study of the effectiveness of the Salk polio vaccine may be cited. The study by Coleman (1966) on equality of educational opportunity, conducted on a national sample of schools, contained many regression analyses that estimated the relative contributions of school characteristics, home background, and the child's outlook to variations in exam results.

1.3 THE PRINCIPAL STEPS IN A SAMPLE SURVEY

As a preliminary to a discussion of the role that theory plays in a sample survey, it is useful to describe briefly the steps involved in the planning and execution of a