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

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 12 DOUBLE SAMPLING 327

CHAPTER 13 SOURCES OF ERROR IN SURVEYS 359

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 survey. Surveys vary greatly in their complexity. To take a sample from 5000 cards, neatly arranged and numbered in a file, is an easy task. It is another matter to sample the inhabitants of a region where transport is by water through the forests, where there are no maps, where 15 different dialects are spoken, and where the inhabitants are very suspicious of an inquisitve stranger. Problems that are baffling in one survey may be trivial or nonexistent in another.

The principal steps in a survey are grouped somewhat arbitrarily under 11 headings.

Objectives of the Survey

A lucid statement of the objectives is most helpful. Without this, it is easy in a complex survey to forget the objectives when engrossed in the details of planning, and to make decisions that are at variance with the objectives.

Population to be Sampled

The word *population* is used to denote the aggregate from which the sample is chosen. The definition of the population may present no problem, as when sampling a batch of electric light bulbs in order to estimate the average length of life of a bulb. In sampling a population of farms, on the other hand, rules must be set up to define a farm, and borderline cases arise. These rules must be usable in practice: the enumerator must be able to decide in the field, without much hesitation, whether or not a doubtful case belongs to the population.

The population to be sampled (the *sampled* population) should coincide with the population about which information is wanted (the *target* population). Sometimes, for reasons of practicability or convenience, the sampled population is more restricted than the target population. If so, it should be remembered that conclusions drawn from the sample apply to the sampled population. Judgment about the extent to which these conclusions will also apply to the target population must depend on other sources of information. Any supplementary information that can be gathered about the nature of the differences between sampled and target population may be helpful.

Data to be Collected

It is well to verify that all the data are relevant to the purposes of the survey and that no essential data are omitted. There is frequently a tendency, particularly with human populations, to ask too many questions, some of which are never subsequently analyzed. An overlong questionnaire lowers the quality of the answers to important as well as unimportant questions.

Degree of Precision Desired

The results of sample surveys are always subject to some uncertainty because only part of the population has been measured and because of errors of measurement. This uncertainty can be reduced by taking larger samples and by using