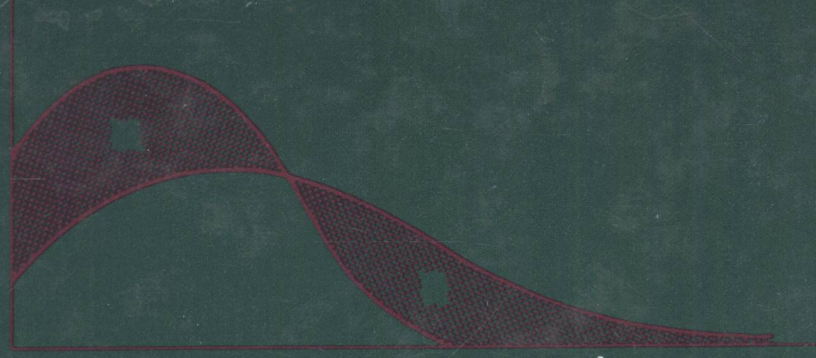
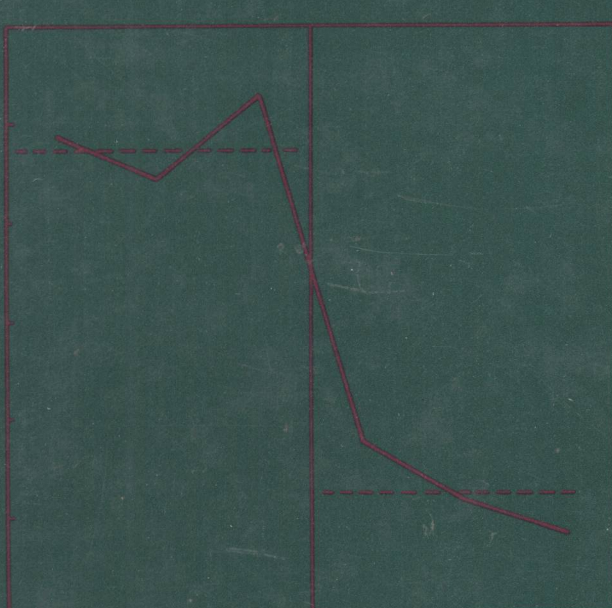
STATISTICS: textbooks and monographs volume 21



CASE STUDIES IN SAMPLE DESIGN



A. C. ROSANDER

CASE STUDIES IN SAMPLE DESIGN

A.C. ROSANDER

MARCEL DEKKER, INC.   New York and Basel

Library of Congress Cataloging in Publication Data

Rosander, Arlyn Custer, 1903-
Case studies in sample design.

(Statistics, textbooks, and monographs ; v. 21)
Includes index.
1. Sampling (Statistics)--Case studies. I. Title.
HA31.2.R67 519.5'2 76-12283
ISBN 0-8247-6467-6

MARCEL DEKKER, INC.

270 Madison Avenue, New York, New York 10016

Current printing (last digit):
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

CASE STUDIES IN SAMPLE DESIGN

STATISTICS

Textbooks and Monographs

A SERIES EDITED BY

OTHER VOLUMES IN PREPARATION

TO MARGARET

PREFACE

This is a book on statistical practice: on the application of the theory of probability sampling to specific management problems. Technical sample design is presented within the context of complete sample studies, which are described as closed-circle processes of six stages. Major difficulties encounted by the practitioner, such as bridging the gap between sampling theory and sampling practice, problem analysis, design of nontechnical aspects, planning and organizing, and managing and controlling, are discussed in some detail.

The case studies illustrate the various difficulties encounted by the practitioner and what was done about them. The cases show to what extent these difficulties were overcome and to what extent they were not. While every attempt was make to find a wide variety of sample cases which would present in some detail all of the major aspects from the initial problem situation to the final results, decisions, and actions, it was found that such presentations are seldom possible. Few companies or agencies want to reveal the detailed operations of a sample study. Some cases came from professional journals, some from legal proceedings, some from published papers in transactions of professional organizations, and some from unpublished papers. Sample studies by the Bureau of the Census have been omitted not only because they are voluminous but because they are described in detail in a large number of books and other publications.

Each case study was designed for a specific purpose--to solve a specific problem of management. None is a general-purpose sample of the census type. Meaning is built into the data right from the start in terms of problem analysis; the data will make sense only if the technical and nontechnical designs are sound and carried out according to specifications. At least two cases show where questions, classifications, and definitions on the data sheet or questionnaire could be greatly improved. This, however, is one object of this book--to show the wide variety of pitfalls in a sample study, and how a sound technical sample design is a necessary but not a sufficient condition for an acceptable sample study.

Acknowledgments are made to the following persons for assistance and materials, or for their role in a case study: the late Herman Guterman and A. J. McKeon of the Internal Revenue Service; Matthew Paolo, Robert Rhodes, Samuel Towne, H. J. Day, Sidney Fine, Seymour Etkin, and the late James M. George, Jr., of the Interstate Commerce Commission; Milton Schwartz of the Post Office Department; Z. L. Pearson and Henry Meyer of the Rocky Mountain Motor Tariff Bureau; Homer Carpenter of Rice, Carpenter, and Carraway; Herman Matthei of the New England Motor Rate Bureau; James Henry of the Eastern Central Motor Carrier Association; F. M. Connell of Blue Cross; Charles McMurdo of the Chesapeake and Potomac Telephone Company of Virginia; Nathan Keiler of the Department of the Navy; Gordon Richardson of the Great Northern Railroad. Special acknowledgment goes to Dr. W. Edwards Deming, who read the entire manuscript, and whose work on the nationwide continuous motor carrier freight bill sample study and on sampling as applied to rail traffic diversion in connection with proposed mergers has been used extensively.

Numerous others contributed to the success of these sample studies, and indirectly to other chapters in this book, and I am indebted to them. I refer to those who established the foundations of modern statistical and sampling theory; to those officials who gave us the green light so that statistical practice could get off the ground and prove that statistics was a versatile and powerful applied science not only in research but in operations and in management; and to those unsung clerks, typists, stenographers, supervisors, and professional persons whose supporting work helped the statistician to conduct successful sample studies.

CONTENTS

APPENDIX

PART I

FOUNDATIONS

Chapter 1

BASIC TERMS AND IDEAS

Sample study. A sample study means an entire study in which probability sampling is used to collect data, and probability and mathematical statistics are used to analyze and interpret the data. It includes all six phases to be described in Chapter 3, because the use of probability sampling affects planning, technical design, specifications, collection methods, implementation of the plans, tabulation of data, analysis, interpretation of results, and the scope of application. It affects also how the study is administered and managed, because the success or failure of the sample study depends on human factors. A technical design is no better than the persons who plan it, design it, put it into effect, analyze the data, and explain what they mean. The sample study may go for nought if it is not properly staffed, if the sample plan is not technically sound, if the operations are not adequately designed, if the study is not carefully managed. The technical design of the probability sample, while of crucial importance, is still only one part of the sample study. Other factors, if not handled properly, may invalidate even the best technically designed sample.

Purpose of the sample. The nature and direction of the entire study, test, audit, or experiment is determined by what is expected from the sample. The purpose, or goal should be explored and analyzed until it is very clear what is wanted. A sample study may be used to estimate a level, a difference, a relation; to discover the existence or absence of a condition; to control an operation or human performance; to make a comparison; to test some hypothesis; to appraise a program; to make a decision. These different purposes usually require different sampling plans and specifications. The problem situation needs to be analyzed until it is clear what kinds and amounts of information the sample must deliver.

When we do this we build quality into the sample data and need not be concerned with "how to make sense out of the data" because we build sense into the data right from the start. Neither need we be concerned with large masses of data because we collect only the data needed for the problem situation. The waste inherent in collecting large masses of data which are not problem-oriented is eliminated.

Probability sampling. The term "sample" means probability sampling in which every sampling unit in the frame is given a known nonzero probability of being selected into the sample. One form of probability sampling is to give every sampling element an equal chance of being selected. In another form of probability sampling, the probability of selection is proportional to the size of each sampling unit. In still another form of sampling, where stratification of the frame is used, the probability of selection is usually different in the different strata.

If a probability sample is properly designed and managed, it has at least ten merits, which account for the tremendous power and versatility of probability sampling:

1. It provides a quantitative measure of the extent of variation due to random effects.
2. It provides data of known quality.
3. It provides data in a timely fashion.
4. It provides acceptable quality data at minimum cost.
5. It has a built-in method of estimation.

6. It forces a sharper analysis of the problem and a clearer definition of terms.
7. It forces better control over nonsampling sources of error, that is, over bias.
8. Mathematical statistics and probability can be applied to analyze and interpret the data.
9. It can yield better quality data than a census or 100% tabulation.
10. It provides within itself a basis for progressive improvement.

Why does probability sampling work? One way to explain sampling is in terms of the redundancy of numbers, in a fashion analogous to the role of redundancy in the theory of communication. For example, if all sampling units in a frame are identical in characteristics, then a sample of one will give all the information possible about the characteristics of the group. Furthermore, it makes no difference how this sample of one is selected, since we will always obtain the correct answers. In this extreme case the variability of the characteristics in the frame is zero, and the redundancy is maximum (unity or 100%).

Sampling works because characteristics of all of the various types of sampling units in the real world have a variability much closer to zero than to plus infinity, and hence have high degrees of redundancy. If redundancy was zero, that is, variability was infinite, sampling would be of no value.

This is illustrated by the different random sample sizes required to estimate with the same precision the average height of adult males in the United States and the average annual income. Since more than 99% of the heights of 50 million adult males in the United States fall between 60 and 80 inches, a range of only 20 inches, very large numbers of heights are for all practical purposes identical. Therefore a random sample of only about 150 out of 50 million will give an estimate of average height with a margin of error due to random sampling of about one-half inch 95% of the time in a large number of such random samples. (Of course the real problem is that of drawing 150 at random from 50 million.)

By contrast, obtaining the same precision in estimating average annual income using 80 million federal individual income tax returns as the frame would require a random sample of about 100,000. The difference in sample sizes is due to differences in variability in the differences in the redundancy of the two characteristics. Whereas height ranges over only 20 inches, individual income ranges over more than $5 million.

It is easy to see why probability sampling is used instead of judgment sampling, in which sampling units are selected on the basis of personal judgment, often by experts or specialists in the field in question. In judgment sampling an effort is made to select "typical" or "representative" units, or to select units spread over the entire frame. Since it has none of the merits of probability sampling listed above and is not recognized in the theory of sampling, its use is not recommended. Only in the field of exploratory work and pretesting have judgment samples of a wide variety of conditions, persons, and procedures found a place in probability sample studies.

Population or universe. The population or universe establishes the limits of the study whether sampling is used or not. The two terms are used synonymously in both sampling theory and practice. We use "population." Usually the definition and analysis of the major problem to be solved, or the basic question to be answered, leads directly to the identification of the population. In practice the population is usually an ideal rather than a reality because it is not completely available, it is hard to circumscribe, it is mobile or constantly changing, or it is otherwise imperfect or inaccessible. Hence we use a "frame" to stand for the population.

An exception is time, where a population of days in a year, hours in a day, and minutes in a day are known exactly. The population includes a wide variety of things: time such as minutes in a working year, areas such as farms or city blocks, objects such as transistors and freight cars, space such as offices in an office building, operations such as truck trips to and from a terminal, conditions such as families below the poverty level or renting families, places such as households, groups such as families, air at a given place at a given time, water in a given stream at a specified place at a given time. Obviously air and water raise special problems when defining a population, not only because they are not discrete objects or entities, but because they cannot be readily bounded. Only finite populations and frames are described in the case studies.

Frame. The frame is a set of symbols which is available, or which can be constructed, and which is used to stand for objects, operations, time, space, or other elements in the population. The frame consists of the total number of sampling units from which the sample is selected. A frame may consist of all of the truck registrations in a state as of July 1 of a specified year, all of the working days in a calendar year, all of the offices in an office building, all of the households listed in a telephone directory as of a certain date, all of the employees in a factory, a list of railroad stations and offices where orders are taken for freight cars, all of the freight bills for a truck company for one calendar year, all of the waybills prepared by Class I railroads for one calendar year, all of the pickup and delivery trips made at a truck terminal during 12 months.

A time frame has to be specified. When the frame consists of a continuous set of records such as invoices, registrations, licenses, tax returns, revenue bills, household moving bills, purchase orders, and the like, time is automatically taken care of by sampling the files or record books for the year desired, or for some shorter period of time if justified. Where no such records exist, as in many operations, time itself must be sampled, such as minutes or days. Examples are sampling parcel post, truck trips, loading and unloading at rail or truck platforms, employees at work, or machinery in operation. Time and place sampling of air and water are even more difficult because no finite population or frame exists. These areas are outside the scope of this book, although one such study is described.