EXHIBIT J

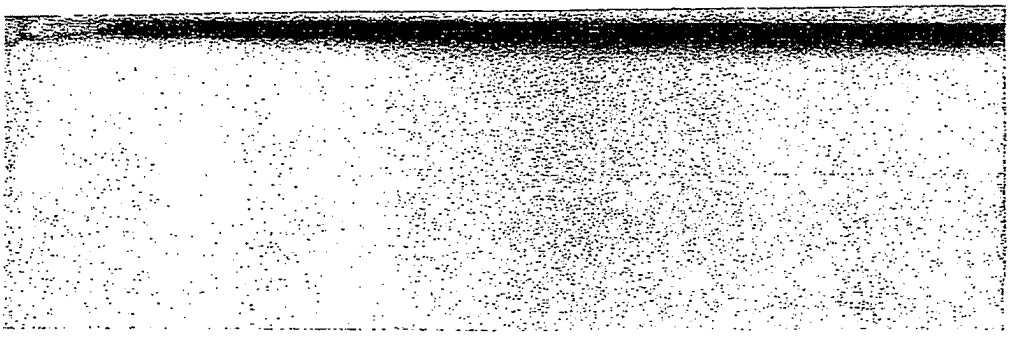
Sampling Techniques

third edition

WILLIAM G. COCHRAN

Professor of Statistics, Emeritus
Harvard University

JOHN WILEY & SONS
New York • Chichester • Brisbane • Toronto • Singapore



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

40 39 38 37 36

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

sampler learns to recognize mistakes in execution and to see that they do not occur in future surveys.

1.4 THE ROLE OF SAMPLING THEORY

This list of the steps in a sample survey has been given in order to emphasize that sampling is a practical business, which calls for several different types of skill. In some of the steps—the definition of the population, the determination of the data to be collected and of the methods of measurement, and the organization of the field work—sampling theory plays at most a minor role. Although these topics are not discussed further in this book, their importance should be realized. Sampling demands attention to all phases of the activity: poor work in one phase may ruin a survey in which everything else is done well.

The purpose of sampling theory is to make sampling more efficient. It attempts to develop methods of sample selection and of estimation that provide, at the lowest possible cost, estimates that are precise enough for our purpose. This principle of specified precision at minimum cost recurs repeatedly in the presentation of theory.

In order to apply this principle, we must be able to predict, for any sampling procedure that is under consideration, the precision and the cost to be expected. So far as precision is concerned, we cannot foretell exactly how large an error will be present in an estimate in any specific situation, for this would require a knowledge of the true value for the population. Instead, the precision of a sampling procedure is judged by examining the frequency distribution generated for the estimate if the procedure is applied again and again to the same population. This is, of course, the standard technique by which precision is judged in statistical theory.

A further simplification is introduced. With samples of the sizes that are common in practice, there is often good reason to suppose that the sample estimates are approximately normally distributed. With a normally distributed estimate, the whole shape of the frequency distribution is known if we know the mean and the standard deviation (or the variance). A considerable part of sample survey theory is concerned with finding formulas for these means and variances.

There are two differences between standard sample survey theory and the classical sampling theory as taught in books on mathematical statistics. In classical theory the measurements that are made on the sampling units in the population are usually assumed to follow a frequency distribution, for example, the normal distribution, of known mathematical form apart from certain population parameters such as the mean and variance whose values have to be estimated from the sample data. In sample survey theory, on the other hand, the attitude has been to assume only very limited information about this frequency distribution. In particular, its mathematical form is not assumed known, so that the approach might be described as model-free or distribution-free. This attitude is natural for

CHAPTER 2

Simple Random Sampling

2.1 SIMPLE RANDOM SAMPLING

Simple random sampling is a method of selecting n units out of the N such that every one of the ${}_NC_n$ distinct samples has an equal chance of being drawn. In practice a simple random sample is drawn unit by unit. The units in the population are numbered from 1 to N. A series of random numbers between 1 and N is then drawn, either by means of a table of random numbers or by means of a computer program that produces such a table. At any draw the process used must give an equal chance of selection to any number in the population *not already drawn*. The units that bear these n numbers constitute the sample.

It is easily verified that all ${}_NC_n$ distinct samples have an equal chance of being selected by this method. Consider one distinct sample, that is, one set of n specified units. At the first draw the probability that some one of the n specified units is selected is n/N. At the second draw the probability that some one of the remaining $(n-1)$ specified units is drawn is $(n-1)/(N-1)$, and so on. Hence the the probability that all n specified units are selected in n draws is

$$\frac{n}{N}\cdot\frac{(n-1)}{(N-1)}\cdot\frac{(n-2)}{(N-2)}\cdots\frac{1}{(N-n+1)}=\frac{n!(N-n)!}{(N)!}=\frac{1}{{}_NC_n} \qquad 2.1$$

Since a number that has been drawn is removed from the population for all subsequent draws, this method is also called random sampling *without replacement*. Random sampling *with replacement* is entirely feasible: at any draw, all N members of the population are given an equal chance of being drawn, no matter how often they have already been drawn. The formulas for the variances and estimated variances of estimates made from the sample are often simpler when sampling is with replacement than when it is without replacement. For this reason sampling with replacement is sometimes used in the more complex sampling plans, although at first sight there seems little point in having the same unit two or more times in the sample.

2.5 VARIANCES OF THE ESTIMATES

The variance of the y_i in a finite population is usually defined as

$$\sigma^2=\frac{\sum_1^N (y_i-\bar{Y})^2}{N} \tag{2.6}$$

As a matter of notation, results are presented in terms of a slightly different expression, in which the divisor $(N-1)$ is used instead of N. We take

$$S^2=\frac{\sum_1^N (y_i-\bar{Y})^2}{N-1} \tag{2.7}$$

This convention has been used by those who approach sampling theory by means of the analysis of variance. Its advantage is that most results take a slightly simpler form. Provided that the same notation is maintained consistently, all results are equivalent in either notation.

We now consider the variance of $\bar{y}$. By this we mean $E(\bar{y}-\bar{Y})^2$ taken over all ${}_NC_n$ samples.

Theorem 2.2. The variance of the mean $\bar{y}$ from a simple random sample is

$$V(\bar{y})=E(\bar{y}-\bar{Y})^2=\frac{S^2}{n}\frac{(N-n)}{N}=\frac{S^2}{n}(1-f) \tag{2.8}$$

where $f=n/N$ is the sampling fraction.

Proof.

$$n(\bar{y}-\bar{Y})=(y_1-\bar{Y})+(y_2-\bar{Y})+\cdots+(y_n-\bar{Y}) \tag{2.9}$$

By the argument of symmetry used in relation (2.5), it follows that

$$E[(y_1-\bar{Y})^2+\cdots+(y_n-\bar{Y})^2]=\frac{n}{N}[(y_1-\bar{Y})^2+\cdots+(y_N-\bar{Y})^2] \tag{2.10}$$

and also that

$$\begin{aligned}E[(y_1-\bar{Y})(y_2-\bar{Y})+(y_1-\bar{Y})(y_3-\bar{Y})+\cdots+(y_{n-1}-\bar{Y})(y_n-\bar{Y})]\\ =\frac{n(n-1)}{N(N-1)}[(y_1-\bar{Y})(y_2-\bar{Y})+(y_1-\bar{Y})(y_3-\bar{Y})\\ +\cdots+(y_{N-1}-\bar{Y})(y_N-\bar{Y})]\end{aligned} \tag{2.11}$$

In (2.11) the sums of products extend over all pairs of units in the sample and population, respectively. The sum on the left contains $n(n-1)/2$ terms and that on the right contains $N(N-1)/2$ terms.

Now square (2.9) and average over all simple random samples. Using (2.10) and (2.11), we obtain

$$n^2E(\bar{y}-\bar{Y})^2=\frac{n}{N}\Big\{(y_1-\bar{Y})^2+\cdots+(y_N-\bar{Y})^2 + \frac{2(n-1)}{N-1}[(y_1-\bar{Y})(y_2-\bar{Y})+\cdots+(y_{N-1}-\bar{Y})(y_N-\bar{Y})]\Big\}$$

Completing the square on the cross-product term, we have

$$n^2E(\bar{y}-\bar{Y})^2=\frac{n}{N}\Big\{\Big(1-\frac{n-1}{N-1}\Big)[(y_1-\bar{Y})^2+\cdots+(y_N-\bar{Y})^2] + \frac{n-1}{N-1}[(y_1-\bar{Y})+\cdots+(y_N-\bar{Y})]^2\Big\}$$

The second term inside the curly bracket vanishes, since the sum of the y_i equals $N\bar{Y}$. Division by n^2 gives

$$V(\bar{y})=E(\bar{y}-\bar{Y})^2=\frac{N-n}{nN(N-1)}\sum_{i=1}^{N}(y_i-\bar{Y})^2=\frac{S^2}{n}\frac{(N-n)}{N}$$

This completes the proof.

Corollary 1. The standard error of $\bar{y}$ is

$$\sigma_{\bar{y}}=\frac{S}{\sqrt{n}}\sqrt{(N-n)/N}=\frac{S}{\sqrt{n}}\sqrt{1-f} \tag{2.12}$$

Corollary 2. The variance of $\hat{Y}=N\bar{y}$, as an estimate of the population total Y, is

$$V(\hat{Y})=E(\hat{Y}-Y)^2=\frac{N^2S^2}{n}\frac{(N-n)}{N}=\frac{N^2S^2}{n}(1-f) \tag{2.13}$$

Corollary 3. The standard error of $\hat{Y}$ is

$$\sigma_{\hat{Y}}=\frac{NS}{\sqrt{n}}\sqrt{(N-n)/N}=\frac{NS}{\sqrt{n}}\sqrt{1-f} \tag{2.14}$$

2.6 THE FINITE POPULATION CORRECTION

For a random sample of size n from an infinite population, it is well known that the variance of the mean is σ^2/n. The only change in this result when the population is finite is the introduction of the factor $(N-n)/N$. The factors $(N-n)/N$ for the variance and $\sqrt{(N-n)/N}$ for the standard error are called the *finite population corrections* (fpc). They are given with a divisor $(N-1)$ in place of N by writers who present results in terms of σ. Provided that the sampling fraction n/N remains low, these factors are close to unity, and the size of the population as

such has no direct effect on the standard error of the sample mean. For instance, if S is the same in the two populations, a sample of 500 from a population of 200,000 gives almost as precise an estimate of the population mean as a sample of 500 from a population of 10,000. Persons unfamiliar with sampling often find this result difficult to believe and, indeed, it is remarkable. To them it seems intuitively obvious that if information has been obtained about only a very small fraction of the population, the sample mean cannot be accurate. It is instructive for the reader to consider why this point of view is erroneous.

In practice the fpc can be ignored whenever the sampling fraction does not exceed 5% and for many purposes even if it is as high as 10%. The effect of ignoring the correction is to overestimate the standard error of the estimate $\bar{y}$.

The following theorem, which is an extension of theorem 2.2, is not required for the discussion in this chapter, but it is proved here for later reference.

Theorem 2.3. If y_i, x_i are a pair of variates defined on every unit in the population and $\bar{y}$, $\bar{x}$ are the corresponding means from a simple random sample of size n, then their *covariance*

$$E(\bar{y}-\bar{Y})(\bar{x}-\bar{X})=\frac{N-n}{nN}\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X}) \tag{2.15}$$

This theorem reduces to theorem 2.2 if the variates y_i, x_i are equal on every unit.

Proof. Apply theorem 2.2 to the variate $u_i = y_i + x_i$. The population mean of u_i is $\bar{U} = \bar{Y} + \bar{X}$, and theorem 2.2 gives

$$E(\bar{u}-\bar{U})^2=\frac{N-n}{nN}\frac{1}{N-1}\sum_{i=1}^{N}(u_i-\bar{U})^2$$

that is

$$E[(\bar{y}-\bar{Y})+(\bar{x}-\bar{X})]^2=\frac{N-1}{nN}\frac{1}{N-1}\sum_{i=1}^{N}[(y_i-\bar{Y})+(x_i-\bar{X})]^2 \tag{2.16}$$

Expand the quadratic terms on both sides. By theorem 2.2,

$$E(\bar{y}-\bar{Y})^2=\frac{N-n}{nN}\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})^2$$

with a similar relation for $E(\bar{x}-\bar{X})^2$. Hence these two terms cancel on the left and right sides of (2.16). The result of the theorem (equation 2.15) follows from the cross-product terms.

2.7 ESTIMATION OF THE STANDARD ERROR FROM A SAMPLE

The formulas for the standard errors of the estimated population mean and total are used primarily for three purposes: (1) to compare the precision obtained by simple random sampling with that given by other methods of sampling, (2) to

estimate the size of the sample needed in a survey that is being planned, and (3) to estimate the precision actually attained in a survey that has been completed. The formulas involve S^2, the population variance. In practice this will not be known, but it can be estimated from the sample data. The relevant result is stated in theorem 2.4.

Theorem 2.4. For a simple random sample

$$s^2 = \frac{\sum_1^n (y_i - \bar{y})^2}{n-1}$$

is an unbiased estimate of

$$S^2 = \frac{\sum_1^N (y_i - \bar{Y})^2}{N-1}$$

Proof. We may write

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} [(y_i - \bar{Y}) - (\bar{y} - \bar{Y})]^2 \tag{2.17}$$

$$= \frac{1}{n-1} \left[\sum_{i=1}^{n} (y_i - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2 \right] \tag{2.18}$$

Now average over all simple random samples of size n. By the argument of symmetry used in theorem 2.2,

$$E\left[\sum_{i=1}^{n} (y_i - \bar{Y})^2 \right] = \frac{n}{N} \sum_{i=1}^{N} (y_i - \bar{Y})^2 = \frac{n(N-1)}{N} S^2$$

by the definition of S^2. Furthermore, by theorem 2.2,

$$E[n(\bar{y} - \bar{Y})^2] = \frac{N-n}{N} S^2$$

Hence

$$E(s^2) = \frac{S^2}{(n-1)N} [n(N-1) - (N-n)] = S^2 \tag{2.19}$$

Corollary. Unbiased estimates of the variances of $\bar{y}$ and $\hat{Y} = N\bar{y}$ are

$$v(\bar{y}) = s_{\bar{y}}^2 = \frac{s^2}{n} \left(\frac{N-n}{N} \right) = \frac{s^2}{n} (1-f) \tag{2.20}$$

$$v(\hat{Y}) = s_{\hat{Y}}^2 = \frac{N^2 s^2}{n} \left(\frac{N-n}{N} \right) = \frac{N^2 s^2}{n} (1-f) \tag{2.21}$$