ed Statistics

(10)

#84462

Sampling Techniques

third edition

WILLIAM G. COCHRAN

Professor of Statistics, Emeritus
Harvard University

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

to Betty

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

20 19 18 17 16 15 14 13

ES

ds out the exact volume purchased,
an was delivered. If he has paid for
e fact.
: cost of measuring n logs is cn, find
olume per log may be denoted by S

haracteristics is to be measured on
lation. If P_1, P_2 are the percentages
l and 2, a client wishes to estimate
:ntage points. What sample size do
between 40 and 60% and that the
iits?
acteristics are positively correlated,
al sample of 200, with the following

of units

72
44
14
70
—
00

$P_1 - P_2$) with a standard error $\leq 2\%$?
which is close to equality, and could
two children. Ignoring the small
f factor for a simple random sample

deff factor?

CHAPTER 5

Stratified Random Sampling

5.1 DESCRIPTION

In stratified sampling the population of N units is first divided into subpopulations of $N_1, N_2, \ldots, N_L$ units, respectively. These subpopulations are nonoverlapping, and together they comprise the whole of the population, so that

$$N_1 + N_2 + \cdots + N_L = N$$

The subpopulations are called *strata*. To obtain the full benefit from stratification, the values of the N_h must be known. When the strata have been determined, a sample is drawn from each, the drawings being made independently in different strata. The sample sizes within the strata are denoted by $n_1, n_2, \ldots, n_L$, respectively.

If a simple random sample is taken in each stratum, the whole procedure is described as *stratified random sampling*.

Stratification is a common technique. There are many reasons for this; the principal ones are the following.

1. If data of known precision are wanted for certain subdivisions of the population, it is advisable to treat each subdivision as a "population" in its own right.
2. Administrative convenience may dictate the use of stratification; for example, the agency conducting the survey may have field offices, each of which can supervise the survey for a part of the population.
3. Sampling problems may differ markedly in different parts of the population. With human populations, people living in institutions (e.g., hotels, hospitals, prisons) are often placed in a different stratum from people living in ordinary homes because a different approach to the sampling is appropriate for the two situations. In sampling businesses we may possess a list of the large firms, which are placed in a separate stratum. Some type of area sampling may have to be used for the smaller firms.
4. Stratification may produce a gain in precision in the estimates of characteristics of the whole population. It may possible to divide a heterogeneous population

into subpopulations, each of which is internally homogeneous. This is suggested by the name *strata*, with its implication of a division into layers. If each stratum is homogeneous, in that the measurements vary little from one unit to another, a precise estimate of any stratum mean can be obtained from a small sample in that stratum. These estimates can then be combined into a precise estimate for the whole population.

The theory of stratified sampling deals with the properties of the estimates from a stratified sample and with the best choice of the sample sizes n_h to obtain maximum precision. In this development it is taken for granted that the strata have already been constructed. The problems of how to construct strata and of how many strata there should be are postponed to a later stage (section 5A.7).

5.2 NOTATION

The suffix h denotes the stratum and i the unit within the stratum. The notation is a natural extension of that previously used. The following symbols all refer to stratum h.

N_h	total number of units
n_h	number of units in sample
y_{hi}	value obtained for the ith unit
$W_h = \dfrac{N_h}{N}$	stratum weight
$f_h = \dfrac{n_h}{N_h}$	sampling fraction in the stratum
$\bar{Y}_h = \dfrac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$	true mean
$\bar{y}_h = \dfrac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$	sample mean
$S_h^2 = \dfrac{\sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2}{N_h - 1}$	true variance

Note that the divisor for the variance is $(N_h - 1)$.

5.3 PRO

For the population mean
(*st* for *stratified*), where

where $N = N_1 + N_2 + \cdots + N$
The estimate $\bar{y}_{st}$ is not in
mean, $\bar{y}$, can be written as

The difference is that in $\bar{y}_{st}$ th
correct weights N_h/N. It is evi
stratum

$$\frac{n_h}{n} = \frac{N_h}{N}$$

This means that the sampling
described as stratification w
self-weighting sample. If num
sample is time-saving.

The principal properties o
theorems. The first two theore
not restricted to stratified rand
need not be a simple random

Theorem 5.1. If in every st
an unbiased estimate of the p

Proof.

$$E(\bar{y}_{st})$$

since the estimates are unbiase
$\bar{Y}$ may be written

$$\bar{Y} = \frac{\sum_{h=1}^{L}}{}$$

This completes the proof.

5.3 PROPERTIES OF THE ESTIMATES

For the population mean per unit, the estimate used in stratified sampling is $\bar{y}_{st}$ (*st* for *stratified*), where

$$\bar{y}_{st} = \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{N} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{5.1}$$

where $N = N_1 + N_2 + \cdots + N_L$.

The estimate $\bar{y}_{st}$ is not in general the same as the sample mean. The sample mean, $\bar{y}$, can be written as

$$\bar{y} = \frac{\sum_{h=1}^{L} n_h \bar{y}_h}{n} \tag{5.2}$$

The difference is that in $\bar{y}_{st}$ the estimates from the individual strata receive their correct weights N_h/N. It is evident that $\bar{y}$ coincides with $\bar{y}_{st}$ provided that in every stratum

$$\frac{n_h}{n} = \frac{N_h}{N} \quad \text{or} \quad \frac{n_h}{N_h} = \frac{n}{N} \quad \text{or} \quad f_h = f$$

This means that the sampling fraction is the same in all strata. This stratification is described as stratification with *proportional* allocation of the n_h. It gives a *self-weighting* sample. If numerous estimates have to be made, a self-weighting sample is time-saving.

The principal properties of the estimate $\bar{y}_{st}$ are outlined in the following theorems. The first two theorems apply to stratified sampling in general and are not restricted to stratified random sampling; that is, the sample from any stratum need not be a simple random sample.

Theorem 5.1. If in every stratum the sample estimate $\bar{y}_h$ is unbiased, then $\bar{y}_{st}$ is an unbiased estimate of the population mean $\bar{Y}$.

Proof.

$$E(\bar{y}_{st}) = E \sum_{h=1}^{L} W_h \bar{y}_h = \sum_{h=1}^{L} W_h \bar{Y}_h$$

since the estimates are unbiased in the individual strata. But the population mean $\bar{Y}$ may be written

$$\bar{Y} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}}{N} = \frac{\sum_{h=1}^{L} N_h \bar{Y}_h}{N} = \sum_{h=1}^{L} W_h \bar{Y}_h$$

This completes the proof.

Theorem 5.2. If the samples are drawn independently in different strata,

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) \tag{5.3}$$

where $V(\bar{y}_h)$ is the variance of $\bar{y}_h$ over repeated samples from stratum h.

Proof. Since

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{5.4}$$

$\bar{y}_{st}$ is a linear function of the $\bar{y}_h$ with fixed weights W_h. Hence we may quote the result in statistics for the variance of a linear function.

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) + 2 \sum_{h=1}^{L} \sum_{j>h}^{L} W_h W_j \operatorname{Cov}(\bar{y}_h \bar{y}_j) \tag{5.5}$$

But since samples are drawn independently in different strata, all covariance terms vanish. This gives the result (5.3).

To summarize theorems 5.1 and 5.2: if $\bar{y}_h$ is an unbiased estimate of $\bar{Y}_h$ in every stratum, and sample selection is independent in different strata, then $\bar{y}_{st}$ is an unbiased estimate of $\bar{Y}$ with variance $\sum W_h^2 V(\bar{y}_h)$.

The important point about this result is that the variance of $\bar{y}_{st}$ depends only on the variances of the estimates of the individual stratum means $\bar{Y}_h$. If it were possible to divide a highly variable population into strata such that all items had the same value within a stratum, we could estimate $\bar{Y}$ without any error. Equation (5.4) shows that it is the use of the correct stratum weights N_h/N in making the estimate $\bar{y}_{st}$ that leads to this result.

Theorem 5.3. For stratified random sampling, the variance of the estimate $\bar{y}_{st}$ is

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h) \tag{5.6}$$

Proof. Since $\bar{y}_h$ is an unbiased estimate of $\bar{Y}_h$, theorem 5.2 can be applied. Furthermore, by theorem 2.2, applied to an individual stratum,

$$V(\bar{y}_h) = \frac{S_h^2}{n_h}\frac{N_h - n_h}{N_h}$$

By substitution into the result of theorem 5.2, we obtain

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 V(\bar{y}_h) = \frac{1}{N^2} \sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} = \sum W_h^2 \frac{S_h^2}{n_h}(1 - f_h)$$

Some particular cases of this formula are given in the following corollaries.

Corollary 1. If the sampl

$V(\bar{y}_{st}$

This is the appropriate formul

Corollary 2. With propor

in (5.6). The variance reduces

$V(\bar{y}_{st}) = \sum$

Corollary 3. If sampling is
same value, S_w^2, we obtain the

Theorem 5.4. If $\hat{Y}_{st} = N\bar{y}_s$

$V($

This follows at once from thec

Example. Table 5.1 shows the
64 large cities in the United Stat
ranked fifth to sixty-eighth in the U
cities are arranged in two strata, th
remaining 48 cities.

The total number of inhabitants
size 24. Find the standard error of
stratified random sample with pro
12 units drawn from each stratum

This population resembles the p
some units—the large cities—con
greater variability than the remai

The stratum totals and sums of s
used in this example: the 1920 da

For the complete population in

$Y =$

The three estimates of Y are de
1. For simple random sampling

$$V(\hat{Y}_{ran}) = \frac{N^2 S^2}{n}$$

ndently in different strata,

) (5.3)

mples from stratum h.

(5.4)

W_h. Hence we may quote the tion.

$_h W_j \operatorname{Cov}(\bar{y}_h \bar{y}_j)$ (5.5)

different strata, all covariance

nbiased estimate of $\bar{Y}_h$ in every different strata, then $\bar{y}_{st}$ is an).

variance of $\bar{y}_{st}$ depends only on stratum means $\bar{Y}_h$. If it were to strata such that all items had $\bar{Y}$ without any error. Equation m weights N_h/N in making the

, the variance of the estimate $\bar{y}_{st}$

$$\sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1-f_h) \qquad (5.6)$$

$_h$, theorem 5.2 can be applied. vidual stratum,

$_h$

e obtain

$$n_h)\frac{S_h^2}{n_h} = \sum W_h^2 \frac{S_h^2}{n_h}(1-f_h)$$

in the following corollaries.

Corollary 1. If the sampling fractions n_h/N_h are negligible in all strata,

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum \frac{N_h^2 S_h^2}{n_h} = \sum \frac{W_h^2 S_h^2}{n_h} \qquad (5.7)$$

This is the appropriate formula when finite population corrections can be ignored.

Corollary 2. With proportional allocation, we substitute

$$n_h = \frac{nN_h}{N}$$

in (5.6). The variance reduces to

$$V(\bar{y}_{st}) = \sum \frac{N_h}{N}\frac{S_h^2}{n}\left(\frac{N-n}{N}\right) = \frac{1-f}{n}\sum W_h S_h^2 \qquad (5.8)$$

Corollary 3. If sampling is proportional and the variances in all strata have the same value, S_w^2, we obtain the simple result

$$V(\bar{y}_{st}) = \frac{S_w^2}{n}\left(\frac{N-n}{N}\right) \qquad (5.9)$$

Theorem 5.4. If $\hat{Y}_{st} = N\bar{y}_{st}$ is the estimate of the population total Y, then

$$V(\hat{Y}_{st}) = \sum N_h(N_h - n_h)\frac{S_h^2}{n_h} \qquad (5.10)$$

This follows at once from theorem 5.3.

Example. Table 5.1 shows the 1920 and 1930 number of inhabitants, in thousands, of 64 large cities in the United States. The data were obtained by taking the cities which ranked fifth to sixty-eighth in the United States in total number of inhabitants in 1920. The cities are arranged in two strata, the first containing the 16 largest cities and the second the remaining 48 cities.

The total number of inhabitants in all 64 cities in 1930 is to be estimated from a sample of size 24. Find the standard error of the estimated total for (1) a simp' andom sample, (2) a stratified random sample with proportional allocation, (3) a stratifi dom sample with 12 units drawn from each stratum.

This population resembles the populations of many types of bus enterprise in that some units—the large cities—contribute very substantially to the total and display much greater variability than the remainder.

The stratum totals and sums of squares are given under Table 5.1. Only the 1930 data are used in this example: the 1920 data appear in a later example.

For the complete population in 1930, we find

$$Y = 19{,}568, \qquad S^2 = 52{,}448$$

The three estimates of Y are denoted by $\hat{Y}_{ran}$, $\hat{Y}_{prop}$, and $\hat{Y}_{equal}$.

1. For simple random sampling

$$V(\hat{Y}_{ran}) = \frac{N^2 S^2}{n}\frac{N-n}{N} = \frac{(64)^2(52{,}448)}{24}\left(\frac{40}{64}\right) = 5{,}594{,}453$$

TABLE 5.1

SIZES OF 64 CITIES (IN 1000's) IN 1920 AND 1930

1920 Size (x_{hi})				1930 Size (y_{hi})			
Stratum $h = 1$	2			Stratum 1	2		
797	314	172	121	900	364	209	113
773	298	172	120	822	317	183	115
748	296	163	119	781	328	163	123
734	258	162	118	805	302	253	154
588	256	161	118	670	288	232	140
577	243	159	116	1238	291	260	119
507	238	153	116	573	253	201	130
507	237	144	113	634	291	147	127
457	235	138	113	578	308	292	100
438	235	138	110	487	272	164	107
415	216	138	110	442	284	143	114
401	208	138	108	451	255	169	111
387	201	136	106	459	270	139	163
381	192	132	104	464	214	170	116
324	180	130	101	400	195	150	122
315	179	126	100	366	260	143	134

Note. Cities are arranged in the same order in both years.

Totals and sums of squares

	1920		1930	
Stratum	$\sum(x_{hi})$	$\sum(x_{hi}^2)$	$\sum(y_{hi})$	$\sum(y_{hi}^2)$
1	8,349	4,756,619	10,070	7,145,450
2	7,941	1,474,871	9,498	2,141,720

from theorem 2.2, corollary 2. The standard error is

$$\sigma(\hat{Y}_{ran}) = 2365$$

2. For the individual strata the variances are

$$S_1^2 = 53{,}843, \qquad S_2^2 = 5581$$

Note that the stratum with the largest cities has a variance nearly 10 times that of the other stratum.

In proportional allocation, v
have

$$V(\hat{Y}_{prop}) = \frac{N}{}$$

$$= \frac{40}{24}$$

$$\sigma(\hat{Y}_{prop}) = 1$$

3. For $n_1 = n_2 = 12$ we use th

$$V(\hat{Y}_{equal}) = \sum N$$

$$= \frac{(16}{}$$

$$\sigma(\hat{Y}_{equal}) = 104$$

In this example equal sample
allocation. Both are greatly sup

5.4 THE I C

If a simple random sample i
S_h^2 (from theorem 2.4) is

$$s_h$$

Hence we obtain the followi

Theorem 5.5. With strati
variance of $\bar{y}_{st}$ is

$$v(\bar{y}_{st}) =$$

An alternative form for co

$$s^2(\bar{y}$$

The second term on the right
In order to compute this es
every stratum. Estimation of t
at which only one unit is cho
The formulas for confiden

Population mean:

Population total:

20 AND 1930

Size (y_{hi})

tratum

2

364	209	113
317	183	115
328	163	123
302	253	154
288	232	140
291	260	119
253	201	130
291	147	127
308	292	100
272	164	107
284	143	114
255	169	111
270	139	163
214	170	116
195	150	122
260	143	134

oth years.

ares

1930

(y_{hi})	$\sum (y_{hi}^2)$
070	7,145,450
498	2,141,720

581

:e nearly 10 times that of the other

In proportional allocation, we have $n_1 = 6$, $n_2 = 18$. From (5.7), multiplying by N^2, we have

$$V(\hat{Y}_{prop}) = \frac{N-n}{n}\sum N_h S_h^2$$

$$= \tfrac{40}{24}[(16)(53{,}843) + (48)(5581)] = 1{,}882{,}293$$

$$\sigma(\hat{Y}_{prop}) = 1372$$

3. For $n_1 = n_2 = 12$ we use the general formula (5.9):

$$V(\hat{Y}_{equal}) = \sum N_h(N_h - n_h)\frac{S_h^2}{n_h}$$

$$= \frac{(16)(4)(53{,}843)}{12} + \frac{(48)(36)(5581)}{12} = 1{,}090{,}827$$

$$\sigma(\hat{Y}_{equal}) = 1044$$

In this example equal sample sizes in the two strata are more precise than proportional allocation. Both are greatly superior to simple random sampling.

5.4 THE ESTIMATED VARIANCE AND CONFIDENCE LIMITS

If a simple random sample is taken within each stratum, an unbiased estimate of S_h^2 (from theorem 2.4) is

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2 \tag{5.11}$$

Hence we obtain the following.

Theorem 5.5. With stratified random sampling, an unbiased estimate of the variance of $\bar{y}_{st}$ is

$$v(\bar{y}_{st}) = s^2(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h} \tag{5.12}$$

An alternative form for computing purposes is

$$s^2(\bar{y}_{st}) = \sum_{h=1}^{L}\frac{W_h^2 s_h^2}{n_h} - \sum_{h=1}^{L}\frac{W_h s_h^2}{N} \tag{5.13}$$

The second term on the right represents the reduction due to the fpc.

In order to compute this estimate, there must be at least two units drawn from every stratum. Estimation of the variance when stratification is carried to the point at which only one unit is chosen per stratum is discussed in section 5A.12.

The formulas for confidence limits are as follows.

Population mean: $$\bar{y}_{st} \pm ts(\bar{y}_{st}) \tag{5.14}$$

Population total: $$N\bar{y}_{st} \pm tNs(\bar{y}_{st}) \tag{5.15}$$

These formulas assume that $\bar{y}_{st}$ is normally distributed and that $s(\bar{y}_{st})$ is well determined, so that the multiplier t can be read from tables of the normal distribution.

If only a few degrees of freedom are provided by each stratum, the usual procedure for taking account of the sampling error attached to a quantity like $s(\bar{y}_{st})$ is to read the t-value from the tables of Student's t instead of from the normal table. The distribution of $s(\bar{y}_{st})$ is in general too complex to allow a strict application of this method. An approximate method of assigning an effective number of degrees of freedom to $s(\bar{y}_{st})$ is as follows (Satterthwaite, 1946).

We may write

$$s^2(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} g_h s_h^2, \quad \text{where } g_h = \frac{N_h(N_h - n_h)}{n_h}$$

The effective number of degrees of freedom n_e is

$$n_e = \frac{\left(\sum g_h s_h^2\right)^2}{\sum \frac{g_h^2 s_h^4}{n_h - 1}} \tag{5.16}$$

The value of n_e always lies between the smallest of the values $(n_h - 1)$ and their sum. The approximation takes account of the fact that S_h^2 may vary from stratum to stratum. It requires the assumption that the y_{hi} are normal, since it depends on the result that the variance of s_h^2 is $2\sigma_h^4/(n_h - 1)$. If the distribution of y_{hi} has positive kurtosis, the variance of s_h^2 will be larger than this and formula 5.16 overestimates the effective degrees of freedom.

5.5 OPTIMUM ALLOCATION

In stratified sampling the values of the sample sizes n_h in the respective strata are chosen by the sampler. They may be selected to minimize $V(\bar{y}_{st})$ for a specified cost of taking the sample or to minimize the cost for a specified value of $V(\bar{y}_{st})$.

The simplest cost function is of the form

$$\text{cost} = C = c_0 + \sum c_h n_h \tag{5.17}$$

Within any stratum the cost is proportional to the size of sample, but the cost per unit c_h may vary from stratum to stratum. The term c_0 represents an overhead cost. This cost function is appropriate when the major item of cost is that of taking the measurements on each unit. If travel costs between units are substantial, empirical and mathematical studies suggest that travel costs are better represented by the expression $\Sigma t_h \sqrt{n_h}$ where t_h is the travel cost per unit (Beardwood et al., 1959). Only the linear cost function (5.17) is considered here.

Theorem 5.6. In stratifie
form (5.17), the variance of
cost C, and the cost is a mi
proportional to $W_h S_h/\sqrt{c_h}$.

Proof. We have

$$V = V(\bar{y}_{st}) = \sum_{h=1}^{L}$$

Our problems are either (1) to
(2) to choose the n_h so as to mi
their final steps, the proble
minimize V for fixed C or C
product

$$V'C'$$

Stuart (1954) has noted th
Cauchy–Schwarz inequality.
inequality comes from the id

$$\left(\sum a_h^2\right)\left(\sum b_h\right.$$

It follows from (5.20) that

$$\left(\sum\right.$$

equality occurring if and only

$$a_h = \frac{W_h S_h}{\sqrt{n_h}},$$

The inequality (5.21) gives

$$V'C' = \left(\sum \frac{W_h^2 S_h^2}{n_h}\right)\left($$

Thus, no choice of the n_h