# 21

# Sample surveys

"statistical Methods"
Snedecor, George Waddel
Cochran, William Gemmell
ISBN 0813815606
1980 7th Ed. Iowa University Press

**21.1—Introduction.** At the beginning of this book it is pointed out that one of the principal applications of statistical methods is to sample surveys, in which information about a specific population is obtained by selecting and measuring a sample of the members of the population. Three examples were described briefly: the nationwide sample from which the Census Bureau publishes monthly estimates of the number of employed and unemployed members of the labor force; a sample of waybills taken by the Chesapeake & Ohio Railroad in appraising whether sampling could be used to estimate the money due the railroad when the Chesapeake & Ohio is used for only a portion of the freight trip; and a sample of 100 farmers in Boone County, Iowa, taken to estimate the proportion of farmers in the county who had sprayed their cornfields to control the European corn borer.

Statistical bureaus in countries in Western Europe and in this country began to try sampling as a means of saving time and money toward the end of the nineteenth century. Acceptance of sampling took some time, but applications gradually spread as sampling techniques were developed and better understood. Nowadays, most published data except for decennial population counts is collected from samples.

Two simple methods for drawing a sample are introduced in chapter 1. One method is to leave the selection entirely to chance. The members of the population are first listed and numbered. If the members of the sample are to be selected one at a time, use of a table of random digits guarantees that at each draw any member of the population not already drawn has an equal chance of being selected for the sample. The method is called *random sampling without replacement*, or *simple random sampling*.

Simple ran~~dom~~ sampling is intuitively fair and free from distortion—every member of the population is equally likely to appear in the sample. Its weakness is that it does not use any relevant information or judgment that we have about the nature of the population—such as that people in one part of a city are wealthier than those in another or that farmers in the north of the county may be more likely to spray than those in the south. One method of using such knowledge is *stratified random sampling*. From this knowledge we try to divide the population into subpopulations or strata that are internally more homogeneous. Then we draw a sample separately from each stratum. The appeal of early applications of stratification was that it would make the sample more representative of the population. By selecting the same proportion of the members of each stratum, in the method known as *proportional stratification*, we guarantee that the sample has the correct population proportion of rich and poor members instead of leaving this matter to chance, as simple random sampling does. But Neyman (7) showed in 1934 that sometimes the deliberate selection of different proportions in different strata can give more accurate estimates than proportional stratification without introducing bias. As noted in section 1.6, this was the method used in the sample of waybills.

This chapter introduces some of the principal methods used for selecting a sample and estimating population characteristics from the sample data. We begin with examples of simple and stratified random sampling.

**21.2—An example of simple random sampling.** The population consists of $N = 6$ members, denoted by the letters $a$ to $f$. The six values of the quantity being measured are as follows: $a$, 1; $b$, 2; $c$, 4; $d$, 6; $e$, 7; $f$, 16. The total for this population is 36. A sample of three members is to be drawn to estimate this total.

How good an estimate of the population total do we obtain by simple random sampling? We are not quite ready to answer this question. Although we know how the sample is to be drawn, we have not yet discussed how the population total is to be estimated from the results of the sample. Since the sample contains three members and the population contains six members, the simplest procedure is to multiply the sample total by 2, and this procedure will be adopted. Any sampling plan contains two parts—a rule for drawing the sample and a rule for making the estimates from the results of the sample.

We can now write down all possible samples of size 3, make the estimate from each sample, and see how close these estimates lie to the true value of 36. There are 20 different samples. Their results appear in table 21.2.1, where the successive columns show the composition of the sample, the sample total, the

TABLE 21.2.1
Results for All Possible Sample Random Samples of Size Three

| Sample | Sample Total | Estimate of Population Total | Error of Estimate |
|---|---|---|---|
| *abc* | 7 | 14 | −22 |
| *abd* | 9 | 18 | −18 |
| *abe* | 10 | 20 | −16 |
| *abf* | 19 | 38 | + 2 |
| *acd* | 11 | 22 | −14 |
| *ace* | 12 | 24 | −12 |
| *acf* | 21 | 42 | + 6 |
| *ade* | 14 | 28 | − 8 |
| *adf* | 23 | 46 | +10 |
| *aef* | 24 | 48 | +12 |
| *bcd* | 12 | 24 | −12 |
| *bce* | 13 | 26 | −10 |
| *bcf* | 22 | 44 | + 8 |
| *bde* | 15 | 30 | − 6 |
| *bdf* | 24 | 48 | +12 |
| *bef* | 25 | 50 | +14 |
| *cde* | 17 | 34 | − 2 |
| *cdf* | 26 | 52 | +16 |
| *cef* | 27 | 54 | +18 |
| *def* | 29 | 58 | +22 |
| Average | 18 | 36 | 0 |

estimated population total, and the error of estimate (estimate *minus* true value).

Some samples, e.g., *abf* and *cde*, do very well, while others like *abc* give poor estimates. Since we do not know in any individual instance whether we will be lucky or unlucky in the choice of a sample, we appraise any sampling plan by looking at its *average* performance.

The average of the errors of estimate (taking account of their signs) is called the *bias* of the estimate (or, more generally, of the sampling plan). A positive bias implies that the sampling plan gives estimates that are on the whole too high; a negative bias, too low. From table 21.2.1 it is evident that this plan gives unbiased estimates, since the average of the 20 estimates is exactly 36 and consequently the errors of estimate add to 0. With simple random sampling this result holds for any population and any size of sample. Unbiased estimates are a desirable feature of a sampling plan, but a plan that gives a small bias is not ruled out of consideration it has other attractive features.

As a measure of the accuracy of the sampling plan we use the *mean square error* (*MSE*) of the estimates taken about the true population value.

$$MSE = \Sigma(\text{error of estimate})^2/20 = 3504/20 = 175.2$$

The divisor 20 is used instead of 19, because the errors are measured from the true population value. To sum up, this plan gives an estimate of the population total that is unbiased and has a standard error $\sqrt{175.2} = 13.2$. This standard error amounts to 37% of the true population total; evidently the plan is not very accurate for this population.

**21.3—An example of stratified random sampling.** Suppose that before planning the sample we expect that *f* will give a much higher value than any other member in the population. How can we use this information? Clearly the estimate from the sample will depend to a considerable extent on whether *f* falls in the sample. This statement can be verified from table 21.2.1; every sample containing *f* gives an overestimate and every sample without *f* gives an underestimate.

TABLE 21.3.1

RESULTS FOR ALL POSSIBLE STRATIFIED RANDOM SAMPLES WITH THE UNEQUAL SAMPLING FRACTIONS DESCRIBED IN TEXT

| Sample | Sample Total in Stratum II, $T_2$ | Estimate $16 + 2.5T_2$ | Error of Estimate |
|---|---|---|---|
| *abf* | 3 | 23.5 | − 12.5 |
| *acf* | 5 | 28.5 | − 7.5 |
| *adf* | 7 | 33.5 | − 2.5 |
| *aef* | 8 | 36.0 | 0.0 |
| *bcf* | 6 | 31.0 | − 5.0 |
| *bdf* | 8 | 36.0 | 0.0 |
| *bef* | 9 | 38.5 | + 2.5 |
| *cdf* | 10 | 41.0 | + 5.0 |
| *cef* | 11 | 43.5 | + 7.5 |
| *def* | 13 | 48.5 | +12.5 |
| Average | | 36.0 | 0.0 |

The best plan is to be sure that *f* appears in every sample. We can do this by dividing the population into two parts or *strata*. Stratum I, which consists of *f* alone, is completely measured. In stratum II, containing *a*, *b*, *c*, *d*, and *e*, we take a simple random sample of size 2 to keep the total sample size equal to 3.

Some forethought is needed in deciding how to estimate the population total. To use twice the sample total, as was done previously, gives too much weight to *f* and always produces an overestimate of the true total. We can handle this problem by treating the two strata separately. For stratum I we know the correct total, which is 16, since we always measure *f*. For stratum II, where 2 members are measured out of 5, the natural procedure is to multiply the sample total in that stratum by 5/2 or 2.5. Hence the appropriate estimate of the population total is 16 + 2.5 × (sample total in stratum II).

These estimates are shown for the 10 possible samples in table 21.3.1. Again, we note that the estimate is unbiased. Its mean square error is

$$\Sigma(\text{error of estimate})^2/10 = 487.50/10 = 48.75$$

The standard error is 7.0 or 19% of the true total—a marked improvement over the standard error of 13.2 obtained with simple random sampling.

This sampling plan is called *stratified random sampling with unequal sampling fractions*. The last part of the title denotes the fact that stratum I is completely sampled and stratum II is sampled at a rate of 2 units out of 5, or 40%. Stratification allows us to divide the population into subpopulations or strata that are less variable than the original population and to sample different parts of the population at different rates when this seems advisable.

EXAMPLE 21.3.1—In the preceding example, suppose you expect that both *e* and *f* will give high values. You decide that the sample shall consist of *e*, *f*, and one member drawn at random from *a*, *b*, *c*, *d*. Show how to obtain an unbiased estimate of the population total and show that the standard error of this estimate is 7.7. (This sampling plan is not as accurate as the plan in which *f* alone was placed in a separate stratum, because the actual value for *e* is not very high.)

EXAMPLE 21.3.2—If previous information suggests that *f* will be high; *d* and *e* moderate; and *a*, *b*, and *c* small, we might try stratified sampling with three strata. The sample consists of *f*, either *d* or *e*, and one chosen from *a*, *b*, and *c*. Work out the unbiased estimate of the population total for each of the six possible samples and show that its standard error is 3.9, much better than that given by our two strata plan.

**21.4—Probability sampling.** The preceding examples are intended to introduce *probability sampling*. This general name is given to sampling plans in which

(*i*) every member of the population has a known probability of being included in the sample (*ii*) the sample is drawn by some method of random selection consistent with these probabilities

(*iii*) we take account of these probabilities of selection in making the estimates from the sample

Note that the probability of selection need not be equal for all members of the population; it is sufficient that these probabilities be known. In the example in section 21.2, each member of the population had an equal chance of being in the sample and each member of the sample received an equal weight in estimating the population total. But in the example in section 21.3, member *f* was given a probability 1 of appearing in the sample, as against 2/5 for the rest of the population. This inequality in the probabilities of selection was compensated for by assigning a weight 5/2 to the other members when making the estimate. The use of unequal probabilities produces a substantial gain in precision for some types of populations (see section 21.8).

Probability sampling has some important advantages. By probability theory it is possible to study the biases and the standard errors of the estimates from different sampling plans. In this way much has been learned about the scope, advantages, and limitations of each plan. This information helps greatly in selecting a suitable plan for a particular sampling job. As will be seen later, most probability sampling plans also enable the standard error of the estimate and confidence limits for the true population value to be computed from the results of the sample. Thus, when a probability sample has been taken, we have some idea as to how accurate the estimates are.

Probability sampling is by no means the only way of selecting a sample. One alternative method is to ask someone who has studied the population to point out average or typical members and then confine the sample to these members. When the population is highly variable and the sample is small, this method often gives more accurate estimates than probability sampling. Another method is to restrict the sampling to those members that are conveniently accessible. If bales of goods are stacked tightly in a warehouse, it is difficult to get at the inside bales of the pile and one is tempted to confine attention to the outside bales. In many biological problems it is hard to see how a workable probability sample can be devised, for instance, as in estimating the number of houseflies in a town, field mice in a wood, or plankton in the ocean.

One drawback of these alternative methods is that when the sample has been obtained, there is no way to determine how accurate the estimate is. Members of the population picked as typical by an expert may be more or less atypical. Outside bales may or may not be similar to interior bales. Probability sampling formulas for the standard error of the estimate or for confidence limits do not apply to these methods. Consequently, it is wise to use probability sampling unless it is clearly not feasible or prohibitively expensive.

In the following sections we give the formulas for the standard errors of the estimates from simple and stratified random sampling.

**21.5—Standard errors for simple random sampling.** If $Y_i$ ($i = 1, 2, \ldots,$ $N$) denotes the variable being studied, the standard deviation, $S$, of the population is defined as

$$S = \sqrt{\Sigma(Y_i - \bar{Y})^2/(N - 1)} \qquad (21.5.1)$$

where $\bar{Y}$ is the population mean of the $Y_i$ and the sum $\Sigma$ is taken over all sampling units in the population. The symbol $S$ is used instead of $\sigma$ because the population is finite and in (21.5.1) we have divided by $N - 1$ instead of $N$.

Since $\bar{Y}$ denotes the population mean, we shall use $\bar{y}$ to denote the sample mean. In a simple random sample of size $n$, the standard error of $\bar{y}$ is (2):

$$\sigma_{\bar{y}} = (S/\sqrt{n})\sqrt{1 - \phi} \qquad (21.5.2)$$

where $\phi = n/N$ is the *sampling fraction*, i.e., the fraction of the population that is included in the sample.

The term $\sigma/\sqrt{n}$ or in sample survey notation $S/\sqrt{n}$ is already familiar to you; it is the usual formula for the standard error of a sample mean. The factor $\sqrt{1 - \phi}$ is known as the *finite population correction*. It enters because we are sampling from a population of finite size $N$ instead of from an infinite population as assumed in the usual theory. Note that this term makes the standard error zero when $n = N$, as it should do, since we have then measured every unit in the population. In practical applications the finite population correction is close to 1 and can be omitted when $n/N$ is less than 10%, i.e., when the sample includes less than 10% of the population.

This result is very remarkable. In a large population with a fixed amount of variability (a given value of $S$), the standard error of the mean depends mainly on the size of sample and only to a minor extent on the fraction of the population sampled. For a given $S$, the mean of a sample of 100 is almost as precise when the population size is 200,000 as when the population size is 20,000 or 2000. Some people intuitively feel that one cannot possibly get accurate results from a sample of 100 from a population of 200,000, because only a tiny fraction of the population has been measured. Actually, whether the sampling plan is accurate or not depends primarily on the size of $S/\sqrt{n}$. This shows why sampling can bring about a great reduction in the amount of measurement needed.

For the *estimated* standard error of the sample mean we have

$$s_{\bar{y}} = (s/\sqrt{n})\sqrt{1 - \phi} \qquad (21.5.3)$$

where $s$ is the standard deviation of the sample, calculated in the usual way.

If the sample is used to estimate the population *total* of the variable under study, the estimate is $N\bar{y}$ and its estimated standard error is

$$s_{N\bar{y}} = (Ns/\sqrt{n})\sqrt{1 - \phi} \qquad (21.5.4)$$

In simple random sampling for attributes, where every member of the sample is classified into one of two classes, we take

$$s_p = \sqrt{pq/n}\sqrt{1 - \phi} \qquad (21.5.5)$$

vhere $p$ is the proportion of the sample that lies in one of the classes. Suppose hat 50 families are picked at random from a list of 432 families who have elephones and that 10 of the families report they are listening to a certain radio program. Then $p = 0.2$, $q = 0.8$, and

$$s_p = \sqrt{(0.2)(0.8)/50}\sqrt{1 - 50/432} = 0.053$$

f we ignore the finite population correction, we find $s_p = 0.057$.

The formula for $s_p$ holds only *if each sampling unit is classified as a whole ito one of the two classes.* If your sampling unit is a group of elements and you re classifying individual elements within each group, a different formula for $s_p$ iust be used. For instance, in estimating the percentage of diseased plants in a eld from a sample of 360 plants, the formula above holds if the plants were lected independently and at random. To save time in the field, however, we ight have chosen 40 areas, each consisting of 3 plants in each of 3 neighboring ws. With this method the area (a group or cluster of plants) is the sampling iit. If the distribution of disease in the field were extremely patchy, it might ippen that every area had either all plants diseased or no plants diseased. In is event the sample of 40 areas would be no more precise than a sample of 40 lependently chosen plants, and we would be deceiving ourselves if we thought it we had a binomial sample of 360 plants.

The correct procedure for computing $s_p$ in this case is simple. Calculate $p$ parately for each sampling unit and apply formula (21.5.3) for continuous riates to these $p$s. That is, if $p_i$ is the percentage diseased in the $i$th area, the nple standard deviation is

$$s = \sqrt{\Sigma(p_i - p)^2/(n - 1)}$$

ere $n$ is now the number of areas (cluster units). Then, by (21.5.3),

$$s_p = (s/\sqrt{n})\sqrt{1 - \phi}$$

instance, suppose that the numbers of diseased plants in the 40 areas were as en in table 21.5.1. The standard deviation of the numbers of diseased plants in sample is 2.331. Since the *proportions* of diseased plants in the 40 areas are nd by dividing the numbers in table 21.5.1 by 9, the standard deviation of the portions is $s = 2.331/9 = 0.259$. Hence (assuming $N$ large),

$$s_p = s/\sqrt{n} = 0.259/\sqrt{40} = 0.041$$

For comparison, the result given by the binomial formula is worked out.

TABLE 21.5.1
NUMBERS OF DISEASED PLANTS (OUT OF 9) IN EACH OF 40 AREAS

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 5 | 1 | 1 | 1 | 7 | 0 | 0 | 3 | 2 | 3 | 0 | 0 | 0 | 7 | 0 | 4 | 1 | 2 | 6 |
| 0 | 0 | 1 | 4 | 5 | 0 | 1 | 4 | 2 | 6 | 0 | 2 | 4 | 1 | 7 | 3 | 5 | 0 | 3 | 6 |

grand total = 99

From the total in table 21.5.1, $p = 99/360 = 0.275$. The binomial formula is

$$s_p = \sqrt{pq/360} = \sqrt{(0.275)(0.725)/360} = 0.024$$

giving an overoptimistic notion of the precision of $p$.

EXAMPLE 21.5.1—If a sample of 4 from the 16 townships of a county has a standard deviation 45, show that the standard error of the mean is 19.5.

EXAMPLE 21.5.2—In the example in section 21.2, $N = 6$, $n = 3$, and the values for the 6 members of the population were 1, 2, 4, 6, 7, and 16. The formula for the true standard error of the estimated population total is

$$\sigma_{N\bar{y}} = (Ns/\sqrt{n})\sqrt{1 - n/N}$$

Verify that this formula agrees with the result, 13.2, which we found by writing down all possible samples.

EXAMPLE 21.5.3—A simple random sample of size 100 is taken to estimate some proportion (e.g., the proportion of males) whose value in the population is close to 1/2. Work out the standard error of the sample proportion $p$ when the size of the population is (*i*) 200, (*ii*) 500, (*iii*) 1000, (*iv*) 10,000, (*v*) 100,000. Note how little the standard error changes for $N$ greater than 1000.

EXAMPLE 21.5.4—Show that the coefficient of variation of the sample mean is the same as that of the estimated population total.

EXAMPLE 21.5.5—In simple random sampling for attributes, show that the standard error of $p$ for given $N$ and $n$ is greatest when $p$ is 50%, but that the coefficient of variation of $p$ is largest when $p$ is very small.

**21.6—Size of sample.** At an early stage in the design of a sample, the question, How large a sample do I need? must be considered. Although a precise answer may not be easy to find (for reasons that will appear), there is a rational method of attack on the problem. At present we assume simple random sampling and ignore $n/N$.

Clearly, we want to avoid making the sample so small that the estimate is too inaccurate to be useful. Equally, we want to avoid taking a sample that is too large so that the estimate is more accurate than we require. Consequently, the first step is to decide how large an error we can tolerate in the estimate. This demands careful thinking about the use to be made of the estimate and the consequences of a sizable error. The figure finally reached may be to some extent arbitrary, yet after some thought samplers often find themselves less hesitant about naming a figure than they expected.

The next step is to express the allowable error in terms of confidence limits. Suppose that $L$ is the allowable error in the sample mean and we are willing to take a 5% chance that the error will exceed $L$. In other words, we want to be reasonably certain that the error will not exceed an amount $\pm L$. Remembering that the 95% confidence limits computed from a sample mean, assumed approximately normally distributed, are

$$\bar{y} \pm 2S/\sqrt{n}$$

we put $L = 2S/\sqrt{n}$. This gives, for the required sample size,

$$n = 4S^2/L^2 \tag{21.6.1}$$

To use this relation, we must have an estimate of the population standard deviation $S$. Often a good guess can be made from the results of previous samplings of this population or of similar populations. For example, an experimental sample was taken in 1938 to estimate the yield per acre of wheat in certain districts of North Dakota (7). For a sample of 222 fields, the variance of the yield per acre from field to field was $s^2 = 90.3$ (in bu$^2$). How many fields are indicated if we wish to estimate the true mean yield within ±1 bu, with a 5% risk that the error will exceed 1 bu? Then

$$n = 4s^2/L^2 = 4(90.3)/1^2 = 361 \text{ fields}$$

If this estimate were being used to plan a sample in a later year, it would be regarded as tentative, since the variance between fields might change from year to year.

In default of previous estimates, Deming (3) has pointed out that $\sigma$ can be estimated from a knowledge of the highest and lowest values in the population and a rough idea of the shape of the distribution. If $h$ = highest − lowest, then $\sigma = 0.29h$ for a uniform (rectangular) distribution, $\sigma = 0.24h$ for a symmetrical distribution shaped like an isosceles triangle, and $\sigma = 0.21h$ for a skew distribution shaped like a right triangle.

If the quantity to be estimated is a binomial proportion, the allowable error $L$ for 95% confidence probability is

$$L = 2\sqrt{pq/n}$$

The sample size required to attain a given limit of error $L$ is therefore

$$n = 4pq/L^2 \tag{21.6.2}$$

In this formula, $p$, $q$, and $L$ may be expressed either as proportions or as percentages, provided they are all expressed in the same units. The result necessitates an advance estimate of $p$. If $p$ is likely to lie between 35% and 65%, the advance estimate can be quite rough, since the product $pq$ varies little for $p$ lying between these limits. If, however, $p$ is near 0% or 100%, accurate determination of $n$ requires a close guess about the value of $p$.

If the computed value of $n$ is found to be more than 10% of the population size $N$, a revised value $n'$ that takes proper account of the finite population fraction is obtained from the relation

$$n' = n/(1 + \phi) \tag{21.6.3}$$

For example, casual inspection of a batch of 480 seedlings indicates that about 15% are diseased. Suppose we wish to know the size of sample needed to determine $p$, the percent diseased, to within ±5%, apart from a 1-in-20 chance.

Formula 21.6.2 gives $n = 4(15)(85)/25 = 204$ seedlings. At this point we might decide that it would be as quick to classify every seedling as to plan a sample that is over 40% of the whole batch. If we decide on sampling, we make a revised estimate $n'$:

$$n' = \frac{n}{1 + \phi} = \frac{204}{1 + 204/480} = 143$$

The preceding formulas assume simple random sampling, which has only limited use in practice. When a more complex plan such as stratified random sampling is employed, a useful quantity known as the design effect of the plan enables simple random sampling formulas to be used more extensively. Kish (5) defines the *design effect* (deff) of a complex plan as the ratio of the variance of the estimate given by the complex plan to the variance of the estimate given by a simple random sample of the same size. The design effects of many plans in common use can be estimated from their sample results. Suppose that a plan has given deff = 2 in recent applications. If we want confidence limits $\pm L$ when using this plan, we first calculate the sample size needed with a simple random sample. Then, if the finite population correction is negligible, we multiply the sample size by 2, or in general by deff, for use with the more complex plan.

EXAMPLE 21.6.1—A simple random sample of houses is to be taken to estimate the percentage of houses that are unoccupied. The estimate is desired to be correct to within ±1%, with 95% confidence. One advance estimate is that the percentage of unoccupied houses will be about 6%; another is that it will be about 4%. What sizes of sample are required on these two forecasts? What size would you recommend?

EXAMPLE 21.6.2—The total number of rats in the residential part of a large city is to be estimated with an error of not more than 20%, apart from a 1-in-20 chance. In a previous survey, the mean number of rats per city block was 9 and the sample standard deviation was 19 (the distribution is extremely skew). Show that a simple random sample of around 450 blocks should suffice.

EXAMPLE 21.6.3—West (9) quotes the following data for 556 full-time farms in Seneca County, New York.

| | Mean | Standard Deviation per Farm |
|---|---|---|
| Acres in corn | 8.8 | 9.0 |
| Acres in small grains | 42.0 | 39.5 |
| Acres in hay | 27.9 | 26.9 |

If a coefficient of variation of up to 5% can be tolerated, show that a random sample of about 240 farms is required to estimate the total acreage of each crop in the 556 farms with this degree of precision. (Note that the finite population correction must be used.) This example illustrates a result that has been reached by several different investigators; with small farm populations such as counties, a substantial part of the whole population must be sampled to obtain accurate estimates.

**21.7—Standard errors for stratified random sampling.** The three steps in stratified random sampling are:

1. The population is divided into a number of parts, called *strata*.
2. A random sample is drawn independently in each part.
3. As an estimate of the population mean, we use

$$\bar{y}_{st} = \Sigma N_h \bar{y}_h / N \tag{21.7.1}$$

where $N_h$ is the total number of sampling units in the $h$th stratum, $\bar{y}_h$ is the sample mean in the $h$th stratum, and $N = \Sigma N_h$ is the size of the population. Note that we must know the values of the $N_h$ (i.e., the sizes of the strata) in order to compute this estimate.

Stratification is commonly employed in sampling plans for several reasons. Differences between the strata means in the population do not contribute to the sampling error of the estimate $\bar{y}_{st}$; it arises solely from variations among sampling units that are in the same stratum. If we can form strata so that a heterogeneous population is divided into parts, each of which is fairly homogeneous, we may expect a substantial gain in precision over simple random sampling. In stratified sampling, we can choose the size of sample to be taken from any stratum. This freedom of choice gives us scope to do an efficient job of allocating resources to the sampling within strata. Furthermore, when different parts of the population present different problems of listing and sampling, stratification enables these problems to be handled separately. For this reason hotels and large apartment houses are frequently placed in a separate stratum in a sample of the inhabitants of a city.

We now consider the estimate from stratified sampling and its standard error. For the population mean, estimate (21.7.1) may be written

$$\bar{y}_{st} = (1/N)\Sigma N_h \bar{y}_h = \Sigma W_h \bar{y}_h$$

where $W_h = N_h/N$ is the relative *weight* attached to the stratum. Note that the sample means $\bar{y}_h$ in the respective strata are weighted by the sizes $N_h$ of the strata. The arithmetic mean of the sample observations is no longer the estimate of the population mean except with proportional stratification. If $n_h/N_h$ = constant = $n/N$, as in proportional stratification, it follows that $W_h = N_h/N = n_h/n$ so that in (21.7.1) the estimate $\bar{y}_{st}$ becomes

$$\bar{y}_{st} = \Sigma W_h \bar{y}_h = \Sigma n_h \bar{y}_h / n = \bar{y} \tag{21.7.2}$$

since $\Sigma n_h \bar{y}_h$ is the total of all observations in the sample.

Since a simple random sample is drawn in each stratum, (21.5.2) gives

$$V(\bar{y}_h) = S_h^2(1 - \phi_h)/n_h \tag{21.7.3}$$

where $\phi_h = n_h/N_h$ is the sampling fraction in the $h$th stratum. Also, since sampling is independent in different strata and the $W_h$ are known numbers,

$$V(\bar{y}_{st}) = \sum_h W_h^2 S_h^2(1 - \phi_h)/n_h \tag{21.7.4}$$

For the estimated standard error of $\bar{y}_{st}$, this gives

$$s(\bar{y}_{st}) = \sqrt{\sum_h W^2 s_h^2(1 - \phi_h)/n_h} \tag{21.7.5}$$

where $s_h^2$ is the sample variance in the $h$th stratum.

In the example to be presented, the $L$ strata were of equal size, so $W_h = 1/L$; and proportional allocation was used, giving $n_h = n/L$, $\phi_h = n/N = \phi$. In this case, (21.7.5) reduces to $\sqrt{\Sigma s_h^2/nL}\,\sqrt{1 - \phi}$, or

$$s(\bar{y}_{st}) = (s_w/\sqrt{n})\sqrt{1 - \phi}$$

where $s_w^2$ is the average within-stratum mean square in the analysis of variance of the sample data as a one-way classification.

The data in table 21.7.1 come from an early investigation by Clapham (1) of the feasibility of sampling for estimating the yields of small cereal plots. A rectangular plot of wheat was divided transversely into three equal strata. Ten samples, each a meter length of a single row, were chosen by simple random sampling from each stratum. The problem is to compute the standard error of the estimated mean yield per meter of row.

In this example, $s_w = \sqrt{240.4} = 15.5$ and $n = 30$. Since the sample is only a negligible part of the whole plot, $n/N$ is negligible and

$$s(\bar{y}_{st}) = s_w/\sqrt{n} = 15.5/\sqrt{30} = 2.83 \text{ g}$$

How effective was the stratification? In the analysis of variance, the mean square between strata is over four times as large as that within strata. This is an indication of real differences in level of yield from stratum to stratum. It is possible to go further and estimate what the standard error of the mean would have been if simple random sampling had been used without any stratification. With simple random sampling, the corresponding formula for the standard error of the mean is

$$s_{\bar{y}} = s/\sqrt{n}$$

where $s$ is the ordinary sample standard deviation. In the sample under discussion, $s$ is $\sqrt{295.3}$ (from the *total* mean square in table 21.7.1). Hence, as an estimate of the standard error of the mean under simple random sampling we might take $s_{\bar{y}} = \sqrt{295.3}/\sqrt{30} = 3.14$ g, as compared with 2.83 g for stratified

TABLE 21.7.1
ANALYSIS OF VARIANCE OF A STRATIFIED RANDOM SAMPLE
(wheat grain yields, g/m)

| Source of Variation | Degrees of Freedom | Sum of Squares | Mean Square |
|---|---|---|---|
| Between strata | 2 | 2073 | 1036.5 |
| Within strata | 27 | 6491 | 240.4 |
| Total | 29 | 8564 | 295.3 |

random sampling. Stratification has reduced the standard error by about 100%. The *design effect* of the stratified plan, as described in section 21.6, is deff = 240.4/295.3 = 0.81.

This comparison is not quite correct for the reason that the value of $s$ was calculated from the results of a stratified sample and not, as it should have been, from the results of a simple random sample. The approximate method that we used is close enough, however, when stratification is proportional and at least ten sampling units are drawn from every stratum.

EXAMPLE 21.7.1—In the example of stratified sampling given in section 21.3, show that the estimate that we used for the population total was $N\bar{y}_{st}$. From (21.7.3), verify that the variance of the estimated population total is 48.75, as found directly in section 21.3. (Note that stratum I makes no contribution to this variance because $n_h = N_h$ in that stratum.)

**21.8—Choice of sample sizes in the strata.** It is sometimes thought that in stratified sampling we should sample the same fraction from every stratum; i.e., we should make $n_h/N_h$ the same in all strata, using proportional allocation. A more thorough analysis of the problem shows, however, that the *optimum* allocation is to take $n_h$ proportional to $N_hS_h/\sqrt{c_h}$, where $S_h$ is the standard deviation of the sampling units in the $h$th stratum and $c_h$ is the cost of sampling per unit in the $h$th stratum. This method of allocation gives the smallest standard error of the estimated mean $\bar{y}_{st}$ for a given total cost of taking the sample. The rule tells us to take a larger sample, as compared with proportional allocation, in a stratum that is unusually variable ($S_h$ large) and a smaller sample in a stratum where sampling is unusually expensive ($c_h$ large). Looked at in this way, the rule is consistent with common sense. The rule reduces to proportional allocation when the standard deviation and the cost per unit are the same in all strata.

To apply the rule, advance estimates of the relative standard deviations and of the relative costs in different strata are needed. These estimates need not be highly accurate; rough estimates often give results satisfactorily near the optimum allocation. When a population is sampled repeatedly, estimates can be obtained from the results of previous samplings. Even when a population is sampled for the first time, it is sometimes obvious that some strata are more accessible to sampling than others. In this event it pays to hazard a guess about the differences in costs. In other situations we are unable to predict with any confidence which strata will be more variable or more costly, or we think that any such differences will be small. Proportional allocation is then used.

Disproportionate sampling pays large dividends when the principal variable being measured has a highly skewed or asymmetrical distribution. Usually such populations contain a few sampling units that have large values for this variable and many units that have small values. Variables that are related to the sizes of economic institutions are often of this type, for instance, the total sales of grocery stores, the number of patients per hospital, the amounts of butter produced by creameries, family incomes, and prices of houses.

With populations of this type, stratification by size of institution is highly effective and the optimum allocation is likely to be much better than proportional allocation. As an illustration, table 21.8.1 shows data for the number of students per institution in a population consisting of the 1019 senior colleges and universities in the United States. The data, which apply mostly to the 1952–1953 academic year, might be used as background information for planning a sample designed to give a quick estimate of total registration in some future year. The institutions are arranged in four strata according to size.

TABLE 21.8.1
DATA FOR TOTAL REGISTRATIONS PER SENIOR COLLEGE OR UNIVERSITY, ARRANGED IN FOUR STRATA

| Stratum: Number of Students per Institution | Number of Institutions $N_h$ | Total Registration for the Stratum | Mean per Institution $\bar{Y}_h$ | Standard Deviation per Institution $S_h$ |
|---|---|---|---|---|
| Less than 1000 | 661 | 292,671 | 443 | 236 |
| 1000–3000 | 205 | 345,302 | 1,684 | 625 |
| 3000–10,000 | 122 | 672,728 | 5,514 | 2,008 |
| Over 10,000 | 31 | 573,693 | 18,506 | 10,023 |
| Total | 1019 | 1,884,394 | | 3,860 |

Note that the 31 largest universities (about 3% in number) have 30% of the students, while the smallest group (which contains 65% of the institutions) contributes only 15% of the students. Note also that the within-stratum standard deviation $S_h$ increases rapidly with increasing size of institution.

Table 21.8.2 shows the calculations needed for choosing the optimum sample sizes within strata. We are assuming equal costs per unit within all strata. The products $N_hS_h$ are formed and added over all strata. Then the relative sample sizes, $N_hS_h/\Sigma N_hS_h$, are computed. These ratios when multiplied by the intended sample size $n$ give the sample sizes in the individual strata.

As a consequence of the large standard deviation in the stratum with the largest universities, the rule requires 37% of the sample to be taken from this stratum. Suppose we are aiming at a total sample size of 250. The rule then calls for (0.37)(250) or 92 universities from this stratum, although the stratum contains only 31 universities in all. With highly skewed populations, as here, the optimum allocation may demand 100% sampling, or even more, of the largest institutions. When this situation occurs, a good procedure is to take 100% of the

TABLE 21.8.2
CALCULATIONS FOR OBTAINING THE OPTIMUM SAMPLE SIZES IN INDIVIDUAL STRATA

| Stratum: Number of Students | Number of Institutions $N_h$ | $N_hS_h$ | Relative Sample Sizes $N_hS_h/\Sigma N_hS_h$ | Actual Sample Sizes | Sampling Rate (%) |
|---|---|---|---|---|---|
| Less than 1000 | 661 | 155,996 | 0.1857 | 65 | 10 |
| 1000–3000 | 205 | 128,125 | .1526 | 53 | 26 |
| 3000–10,000 | 122 | 244,976 | .2917 | 101 | 83 |
| Over 10,000 | 31 | 310,713 | 0.3700 | 31 | 100 |
| Total | 1019 | 839,810 | 1.0000 | 250 | |

large stratum and employ the rule $n_h \propto N_hS_h$ to distribute the remainder of the sample over the other strata. Following this procedure, we include in the sample all 31 of the largest institutions, leaving 219 to be distributed among the first three strata. In the first stratum, the size of sample is

$$219[0.1857/(0.1857 + 0.1526 + 0.2917)] = 65$$

The allocations (second column from the right of table 21.8.2) call for over 80% sampling in the second largest group of institutions (101 out of 122) but only a 10% sample of the small colleges. In practice we might decide for administrative convenience to take a 100% sample in the second largest group as well as in the largest.

Is the optimum allocation much superior to proportional allocation? From tables 21.8.1 and 21.8.2 and the sampling error formulas, we can calculate the standard errors of the estimated population totals $N\bar{y}_{st}$ or $N\bar{y}$ by stratification with optimum allocation or with proportional allocation, and by simple random sampling. These standard errors are:

| Sampling Plan | $s(\hat{Y})$ |
|---|---|
| Simple random sampling | 216,000 |
| Stratification, proportional allocation | 107,000 |
| Stratification, optimum allocation | 26,000 |

The reduction in the standard error due to stratification and the further reduction due to optimum allocation are both striking.

If every unit lies in one or the other of two classes, (e.g., sprayed, not sprayed), the estimate of the population proportion $p_{st}$ in one of the classes is

$$p_{st} = \Sigma W_h p_h \qquad (21.8.1)$$

where $p_h$ is the sample proportion in stratum $k$ and $W_h = N_h/N$ as before. To obtain the estimated standard error of $p_{st}$, substitute $p_hq_h$ for $s_h^2$ in (21.7.5).

For the optimum choice of sample sizes within strata, take $n_h$ proportional to $N_h\sqrt{p_hq_h/c_h}$. If $c_h$ is about the same in all strata, this rule implies that the fraction sampled, $n_h/N_h$, should be proportional to $\sqrt{p_hq_h}$. Now the quantity $\sqrt{pq}$ changes little as $p$ ranges from 25% to 75%. Consequently, proportional allocation is nearly optimal if the strata proportions lie in this range.

EXAMPLE 21.8.1—From the data in table 21.8.1, verify the standard error of 107,000 reported for the estimated total registration as given by a stratified random sample with $n = 250$ and proportional allocation.

EXAMPLE 21.8.2—A sample of 692 families in Iowa was taken to determine among other things the percentage of families with vegetable gardens in 1943. The families were classified into three strata—urban, rural nonfarm, and farm—because these groups were expected to show differences in the frequency and size of vegetable gardens. The sampling fraction was roughly the same in all strata, a sample of 1 per 1000 being aimed at. The values of $W_h$, $n_h$, and the numbers and percentages with gardens are as follows:

| Stratum | $W_h$ | Sample Size $n_h$ | Number with Gardens | Percent with Gardens |
|---|---|---|---|---|
| Urban | 0.445 | 300 | 218 | 72.7 |
| Rural nonfarm | 0.230 | 155 | 147 | 94.8 |
| Farm | 0.325 | 237 | 229 | 96.6 |
| Total | 1.000 | 692 | 594 | 85.8 |

The finite population corrections can, of course, be ignored. (*i*) Calculate the estimated population percent $p_{st}$ with gardens and give its standard error. (*ii*) If the costs $c_h$ are constant, find the optimum sample sizes in the strata and the resulting $s(p_{st})$. Assume the sample $p_h$ are the same as those in the population. Note that the optimum $n_h = nW_h\sqrt{p_hq_h}/\Sigma W_h\sqrt{p_hq_h}$, where $q_h = 100 - p_h$ when $p_h$ is expressed in percent. (*iii*) Estimate approximately the value of $s(p)$ given by a simple random sample with $n = 692$. Ans. (*i*) $p_{st} = 85.6\%$, $SE = \pm 1.27\%$. (*ii*) Optimum $n_h = 445, 115, 132$; $SE(p_{st}) = \pm 1.17\%$. (*iii*) $SE \doteq \sqrt{(85.8)(14.2)/692} = \pm 1.33$. The gain in precision due to stratification and the further gain due to optimum allocation are both modest. The deff factors are 0.91 and 0.77 for stratification with proportional and optimal allocation.

**21.9—Systematic sampling.** To draw a 10% sample from a list of 730 cards, we might select a random number between 1 and 10, say 3, and pick every tenth card thereafter, i.e., the cards numbered 3, 13, 23, and so on, ending with card number 723. A sample of this kind is known as a *systematic sample,* since the choice of its first member, 3, determines the whole sample.

Systematic sampling has two advantages over simple random sampling. It is easier to draw, since only one random number is required, and it distributes the sample more evenly over the listed population. It has a built-in stratification. In our example, cards 1–10, 11–20, etc., in effect form strata, one sampling unit being drawn from each stratum. Systematic sampling differs, however, from stratified random sampling in that the unit from the stratum is not drawn at random; in our example, this unit is always in the third position. Systematic sampling often gives substantially more accurate estimates than simple random sampling and has become popular, for example, with samples taken regularly for inspection and control of quality in mass production.

Systematic sampling has one disadvantage and one potential disadvantage. It has no reliable method of estimating the standard error of the sample mean (the formula for stratified sampling cannot be used, since only one unit is drawn per stratum). Some formulas work well for particular types of populations but cannot be trusted for general application. However, systematic sampling is often a part of a more complex sampling plan such as two-stage sampling in which unbiased estimates of the sampling errors can be obtained.

A potential disadvantage is that if the population contains a periodic type of variation and if the interval between successive units in the sample happens to equal the wavelength or a multiple of it, the sample may be badly biased. For instance, a systematic sample of the houses in a city might contain too many (or too few) corner houses; a systematic sample of families listed by name might contain too many male heads of households or too many children. These situations can be guarded against by changing the random start number frequently.

EXAMPLE 21.9.1—In estimating mean per capita income per state, we might list the 48 U.S. states (excluding Hawaii and Alaska) in order from east to west, putting neighboring states near one another in the sequence, and draw a systematic sample of 1 in 4, with $n = 12$ states. For 1976 incomes (in $1000s), the following data give the four systematic samples.

| Sample | | | | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5.4 | 6.5 | 6.5 | 5.4 | 6.4 | 6.1 | 6.4 | 6.3 | 5.4 | 6.2 | 7.3 | 5.6 | 73.5 |
| 2 | 6.0 | 7.4 | 7.3 | 5.4 | 6.3 | 5.1 | 6.2 | 6.2 | 5.4 | 5.2 | 7.2 | 5.7 | 73.4 |
| 3 | 5.5 | 7.1 | 7.0 | 5.1 | 7.4 | 4.6 | 6.5 | 5.4 | 5.1 | 5.8 | 6.8 | 6.7 | 73.0 |
| 4 | 6.6 | 7.3 | 6.3 | 5.6 | 7.0 | 5.4 | 6.0 | 4.8 | 5.7 | 6.5 | 6.3 | 5.5 | 73.0 |

The population total is 292.9, the mean $\overline{Y}$ per state is 6102, and the variance $S^2$ is 0.5632 with 47 df. Compare the standard errors of the estimated mean per capita income per state as given by (*i*) the systematic sample, (*ii*) a simple random sample of 12 states, (*iii*) a stratified random sample with 12 strata and 1 unit per stratum. Note that for the systematic sample, $V(\bar{y}_{sy}) = \Sigma(\bar{y}_{sy} - \overline{Y})^2/4$. Ans. (*i*) $SE(\bar{y}_{sy}) = \pm 0.0190$, (*ii*) $SE(\bar{y}) = \pm 0.1876$, (*iii*) $SE(\bar{y}_{st}) = \pm 0.1212$. Why the systematic sample does so much better than stratified random sampling is puzzling.

**21.10—Two-stage sampling.** Consider the following miscellaneous group of sampling problems: (1) a study of the vitamin A content of butter produced by creameries, (2) a study of the protein content of wheat in the wheat fields in an area, (3) a study of red blood cell counts in a population of men aged 20–30, (4) a study of insect infestation of the leaves of the trees in an orchard, and (5) a study of the number of defective teeth in third-grade children in the schools of a large city. What do these investigations have in common? First, in each study an appropriate sampling unit suggests itself naturally—the creamery, the field of wheat, the individual man, the tree, and the school. Second, and this is the important point, in each study the chosen sampling units can be *subsampled* instead of being measured completely. Indeed, subsampling is essential in the first three studies. Obviously, we cannot examine *all* the blood in a man in order to make a complete count of his red cells. In the insect infestation study, it might be feasible, although tedious, to examine *all* leaves on any selected tree. If the insect distribution is spotty, however, we would probably decide to take only a small sample of leaves from any selected tree so as to include more trees.

This type of sampling is called *two-stage sampling.* The first stage is the selection of a sample of *primary sampling units*—the creameries, wheat fields, and so on. The second stage is the taking of a subsample of *second-stage units,* or *subunits,* from each selected primary unit.

As illustrated by these examples, the two-stage method is sometimes the only practicable way the sampling can be done. Even with a choice between subsampling the units and measuring them completely, two-stage sampling gives the sampler greater scope, since both the size of the sample of primary units and the size of the sample that is taken from a primary unit can be chosen by the sampler. In some applications an important advantage of two-stage sampling is that it facilitates the problem of listing the population. Often it is relatively easy to obtain a list of the primary units but difficult or expensive to list all the subunits.

Listing is an important problem that we have not discussed. To use probability sampling, we must have in effect a complete list of the sampling units in the population in order to select a sample according to our randomized plan. In the national sample mentioned in section 1.3 for estimating unemployment, the primary unit in urban areas is what is called a standard metropolitan area; in rural areas it is a county or a group of small contiguous counties. These units have all been defined and listed. No list of all the households or families in the country exists. To list the trees in an orchard and draw a sample of them is usually simple, but the problem of making a random selection of the leaves on a tree may be very troublesome. With two-stage sampling this problem is faced only for those trees that are in the sample. No complete listing of all leaves in the orchard is required.

Assume for simplicity that the primary units are of equal size. The population contains $N_1$ primary units, each containing $N_2$ second-stage units or subunits. A random sample of $n_1$ primary units is drawn. From each selected primary unit, $n_2$ subunits are drawn at random. If the sampling fractions $n_1/N_1$ and $n_2/N_2$ are small, we can apply to our results the random effects model (section 13.3) for a one-way classification, the primary units being the classes. Considered as an estimate of the population mean, the observation on any subunit is the sum of two independent terms. One term, associated with the primary unit, has the same value for all subunits in the primary unit and varies from one primary unit to another with variance $s_1^2$. The second term measures differences between subunits in the same primary unit and has variance $s_2^2$.

The sample as a whole contains $n_1$ independent values of the first term and $n_1 n_2$ independent values of the second term. Hence the variance of the sample mean per subunit is

$$V(\bar{y}) = s_1^2/n_1 + s_2^2/(n_1 n_2) \tag{21.10.1}$$

Furthermore, as shown in section 13.3, the two components of variance $s_1^2$ and $s_2^2$ can be estimated from an analysis of variance of the sample results, as given in table 21.10.1. It follows from table 21.10.1 that an unbiased sample estimate of $V(\bar{y})$ in (21.10.1) is

$$\hat{V}(\bar{y}) = s_1^2/(n_1 n_2) = \Sigma(\bar{y}_i - \bar{y})^2/[n_1(n_1 - 1)] \tag{21.10.2}$$

When $n_1/N_1$ is negligible, it can be shown that this very simple result holds also (*i*) if the second-stage sampling fraction $n_2/N_2$ is not negligible; (*ii*) if the second-stage variance differs from one primary unit to another; and (*iii*) if the second-stage samples are drawn systematically, provided that the random start is chosen independently in each sample primary unit.

As pointed out in section 13.3, the analysis of variance (table 21.10.1) is

TABLE 21.10.1
ANALYSIS OF VARIANCE FOR A TWO-STAGE SAMPLE (SUBUNIT BASIS)

| Source of Variation | df | Mean Square | Expected Value |
|---|---|---|---|
| Between primary units (p.u.) | $n_1 - 1$ | $s_1^2$ | $S_2^2 + n_2 S_1^2$ |
| Between subunits within p.u. | $n_1(n_2 - 1)$ | $s_2^2$ | $S_2^2$ |
| | $\hat{S}_1^2 = (s_1^2 - s_2^2)n_1$ | $\hat{S}_2^2 = s_2^2$ | |

TABLE 21.10.2

ANALYSIS OF VARIANCE OF SUGAR PERCENT OF BEETS (ON A SINGLE BEET BASIS)

| Source of Variation | df | Mean Square | Expected Value |
|---|---|---|---|
| Between plots (primary units) | 80 | 2.9254 | $S_2^2 + 10S_1^2$ |
| Between beets (subunits) within plots | 900 | 2.1374 | $S_2^2$ |
| $\hat{S}_1^2 = (2.9254 - 2.1374)/10 = 0.0788$ | | $\hat{S}_2^2 = 2.1374$ | |

useful as a guide in choosing values of $n_1$ and $n_2$ for future samples. Table 21.10.2 gives the analysis of variance in a study by Immer (6), whose object was to develop a sampling technique for the determination of the sugar percentage in sugar beets in field experiments. Ten beets were chosen from each of 100 plots in a uniformity trial; the plots were the primary units. The sugar percentage was obtained separately for each beet. To simulate conditions in field experiments, the between-plots mean square was computed as the mean square between plots within blocks of 5 plots. This mean square gives the experimental error variance that would apply in a randomized blocks experiment with five treatments.

Hence, if a new experiment is to consist of $n_1$ replications with $n_2$ beets sampled from each plot, the predicted variance of a treatment mean is, from the variance estimates in table 21.10.2

$$s_{\bar{y}}^2 = 0.0788/n_1 + 2.1374/(n_1 n_2)$$

We shall illustrate three of the questions that can be tackled from these data. How accurate are the treatment means in an experiment with 6 replications and 5 beets per plot? For this experiment we would expect

$$s_{\bar{y}} = \sqrt{0.0788/6 + 2.1374/30} = 0.29\%$$

The sugar percentage figure for a treatment mean would be correct to within $\pm(2)(0.29)$ or 0.58%, with 95% confidence, assuming $\bar{y}$ approximately normally distributed.

If the standard error of a treatment mean is not to exceed 0.2%, what combinations of $n_1$ and $n_2$ are allowable? We must have

$$0.0788/n_1 + 2.1374/(n_1 n_2) \leq 0.2^2 = 0.04$$

You can verify that with 4 replications ($n_1 = 4$), there must be 27 beets per plot; with 8 replications, 9 beets per plot are sufficient; and with 10 replications, 7 beets per plot. As one would expect, the intensity of subsampling decreases as the intensity of sampling is increased. The total size of sample also decreases from 108 beets when $n_1 = 4$ to 70 beets when $n_1 = 10$.

Some surveys entail a cost $c_1$ of selecting and getting access to a primary unit to sample it and a cost $c_2$ of selecting and measuring each sample subunit. Thus, apart from overhead cost, the cost of taking the sample is

$$C = c_1 n_1 + c_2 n_1 n_2 \qquad (21.10.3)$$

In section 13.3 it is noted that for a given total cost, the value of $n_2$ that minimizes $V(\bar{y})$ is

$$n_2 = \sqrt{c_1 S_2^2/(c_2 S_1^2)} \qquad (21.10.4)$$

With the sugar beets, $\sqrt{S_2^2/S_1^2} = \sqrt{2.1374/0.0788} = 5.2$, giving $n_2 = 5.2\sqrt{c_1/c_2}$.

In this study, cost data were not reported. If $c_1$ were to include the cost of the land and the field operations required to produce one plot, it would be much greater than $c_2$. Evidently a fairly large number of beets per plot would be advisable. In practice, factors other than the sugar percentage determinations must also be considered in deciding on costs and number of replications in sugar beet experiments.

**21.11—Selection with probability proportional to size.** In many important sampling problems the natural primary sampling units are of unequal sizes. Schools, hospitals, and factories all contain different numbers of individuals. A sample of the houses in a town may use blocks as first-stage units, the number of houses per block ranging from 0 to around 40. In national surveys the primary unit is often an administrative area—a county or a metropolitan district. A relatively large unit of this type cuts down travel costs and makes supervision and control of the field work more manageable.

When primary units vary in size, Hansen and Hurwitz (8) pointed out the advantages of selecting primary units with probabilities proportional to their sizes. To illustrate, consider a population with $N = 3$ schools having 600, 300, and 100 children. A 5% sample of 50 children is to be taken to estimate the population mean per child for some characteristic. The means per child in the three schools are $\bar{Y}_1 = 2$, $\bar{Y}_2 = 4$, $\bar{Y}_3 = 1$. Hence, the population mean per child is

$$\bar{\bar{Y}} = [(600)(2) + (300)(4) + (100)(1)]/1000 = 2.5 \qquad (21.11.1)$$

For simplicity, suppose that only one school is chosen, the 50 children are drawn at random from the selected school, and the variation in $Y$ between children in the same school is negligible. Thus the mean $\bar{y}$ of any sample is equal to the mean of the school from which it is drawn.

In selecting the school with probability proportional to size (pps), the three schools receive probabilities 0.6, 0.3, and 0.1, respectively, of being drawn. We shall compare the mean square error of the estimate given by this method with that given by selecting the schools with equal probabilities. Table 21.11.1 contains the calculations.

If the first school is selected, the sample estimate is in error by $2.0 - 2.5 = -0.5$, and so on. These errors and their squares appear in the two right-hand columns of table 21.11.1. In repeated sampling with probability proportional to size, the first school is drawn 60% of the time, the second school 30%, and the third school 10%. The mean square error of the estimate is therefore

$$MSE_{pps} = (0.6)(0.25) + (0.3)(2.25) + (0.1)(2.25) = 1.05$$

TABLE 21.11.1
SELECTION OF A SCHOOL WITH PROBABILITY PROPORTIONAL TO SIZE

| School | Children | Probability of Selection $\pi_i$ | Mean per Child $\overline{Y}_i$ | Error of Estimate $\overline{Y}_i - \overline{\overline{Y}}$ | $(\overline{Y}_i - \overline{\overline{Y}})^2$ |
|---|---|---|---|---|---|
| 1 | 600 | 0.6 | 2 | −0.5 | 0.25 |
| 2 | 300 | 0.3 | 4 | +1.5 | 2.25 |
| 3 | 100 | 0.1 | 1 | −1.5 | 2.25 |
| Population | 1000 | 1.0 | 2.5 | | |

If, alternatively, the schools are drawn with equal probability, the mean square error is

$$MSE_{eq} = (1/3)(0.25 + 2.25 + 2.25) = 1.58$$

which is about 50% higher than that given by pps selection.

The reason it usually pays to select large units with higher probabilities is that the population mean depends more on the means of the large units than on those of the small units, as (21.11.1) shows. The large units are therefore likely to give better estimates when most of the variation is between primary units.

You may ask, Does the result in our example depend on the choice or the order of the means (2, 4, 1) assigned to schools 1, 2, and 3? The answer is yes. With means 4, 2, 1, you will find $MSE_{pps}$ = 1.29 and $MSE_{eq}$ = 2.14, the latter being 66% higher. Over the six possible orders of the numbers 1, 2, 4, the ratio $MSE_{eq}/MSE_{pps}$ varies from 0.93 to 2.52. However, the ratio of the averages $\overline{MSE}_{eq}/\overline{MSE}_{pps}$ taken over all six possible orders does not depend on the numbers 1, 2, 4. With $N_1$ primary units in the population, the ratio is

$$\frac{\overline{MSE}_{eq}}{\overline{MSE}_{pps}} = \frac{(N_1 - 1) + N_1 \sum^{N} (\pi_i - \overline{\pi})^2}{(N_1 - 1) - N_1 \sum^{N} (\pi_i - \overline{\pi})^2} \tag{21.11.2}$$

where $\pi_i$ is the probability of selection (relative size) of the $i$th school. Clearly, the ratio exceeds one unless all $\pi_i$ are equal, that is, all schools are the same size. In our example, this ratio is found to equal 1.47.

In two-stage sampling with primary units of unequal sizes, a simple method is to select $n_1$ primary units with probability proportional to size and take *an equal number of subunits* (e.g., children) in every selected primary unit, as in our illustration. This method gives every subunit in the population an equal chance of being in the sample. The method used in the sample (section 1.3) from which unemployment figures are estimated is an extension of this method to more than two stages of sampling. The sample mean $\overline{y}$ per subunit is an unbiased estimate of the population mean. If the $n_1$ primary units are drawn with

replacement, an unbiased estimate of the variance of $\overline{y}$ is

$$s_{\overline{y}}^2 = \sum^{n_1} (\overline{y}_i - \overline{y})^2/[n_1(n_1 - 1)] \tag{21.11.3}$$

where $\overline{y}_i$ is the mean of the subsample from the $i$th primary unit.

We have illustrated only the simplest case. Some complications arise when we select units without replacement. Often the sizes of the units are not known exactly and have to be estimated in advance. Considerations of cost or of the structure of variability in the population may lead to the selection of units with unequal probabilities that are proportional to some quantity other than sizes. For details, see the references. In extensive surveys, multistage sampling with unequal probabilities of selection of primary units is the commonest method in current practice.

**21.12—Ratio estimates.** The *ratio estimate* is a different way of estimating population totals (or means) that is useful in many sampling problems. Suppose that you have taken a sample to estimate the population total $Y$ of a variable $y_i$ and that a complete count of the population was made on some previous occasion. Let $x_i$ denote the value of the variable on the previous occasion. You might then compute the ratio

$$\hat{R} = \Sigma y_i/\Sigma x_i = \overline{y}/\overline{x}$$

where the sums are taken over the sample. This ratio is an estimate of the present level of the variate relative to that on the previous occasion. On multiplying the ratio by the known population total $X$ on the previous occasion (8), you obtain the ratio estimate $\hat{Y}_R = RX = (\overline{y}/\overline{x})X$ of the population total of $Y$. Clearly, if the relative change is about the same on all sampling units, the estimate of the population total will be a good one.

The ratio estimate can also be used when $x_i$ is some other kind of supplementary variable. The conditions for a successful application of this estimate are that the ratio $y_i/x_i$ should be relatively constant over the population and the population total $X$ should be known. Consider an estimate of the total amount of a crop, just after harvest, made from a sample of farms in a region. For each farm in the sample we record the total yield $y_i$ and the total acreage $x_i$ of that crop. In this case the ratio $\hat{R} = \Sigma y_i/\Sigma x_i$ is the sample estimate of the mean yield per acre. This is multiplied by the total acreage of the crop in the region, which would have to be known accurately from some other source. This estimate will be precise if the mean yield per acre varies little from farm to farm.

In large samples the estimated standard error of the ratio estimate $\hat{Y}_R$ of the population total from a simple random sample of size $n$ is approximately

$$s(\hat{Y}_R) = N \sqrt{\frac{\Sigma(y_i - \hat{R}x_i)^2}{n(n - 1)}} \sqrt{1 - \phi} \tag{21.12.1}$$

TABLE 21.12.1
1970 AND 1960 POPULATIONS (MILLIONS) OF SIX LARGE CITIES

| | 1 | 2 | 3 | 4 | 5 | 6 | Total |
|---|---|---|---|---|---|---|---|
| 1970 population, $y_i$ | 3.36 | 2.82 | 7.90 | 1.95 | 1.51 | 1.23 | $18.77 = Y$ |
| 1960 population, $x_i$ | 3.55 | 2.48 | 7.78 | 2.00 | 1.67 | 0.94 | $18.42 = X$ |

The ratio estimate is not always more precise than the simpler estimate $N\bar{y}$ (number of units in population × sample mean). It has been shown that in large samples the ratio estimate is more precise only if $\rho$, the correlation coefficient between $Y$ and $X$, exceeds $cv(x)/[2cv(y)]$. Consequently, ratio estimates must not be used indiscriminately, although in appropriate circumstances they produce large gains in precision.

Sometimes the purpose of the sampling is to estimate a ratio, e.g., ratio of dry weight to total weight or ratio of clean wool to total wool. The estimated standard error of the estimate in large samples is then

$$s(R) = \frac{1}{\bar{x}}\sqrt{\frac{\Sigma(y_i - \hat{R}x_i)^2}{n(n-1)}}\sqrt{1-\phi}$$

This formula has already been given (in a different notation) at the end of section 19.8 in fitting an asymptotic regression.

As an illustration in which the ratio estimate works well, table 21.12.1 shows the 1970 and 1960 populations of the six U.S. cities with 1970 populations over 1 million.

Suppose that we have to estimate the 1970 total population of $N = 6$ cities from a simple random sample of $n = 2$ cities. The 1970 populations range from 1.23 million to 7.90 million; but while some cities have declined and some increased since 1960, the 1970/1960 ratios are relatively stable.

The estimate based on the sample mean is $\hat{Y} = N\bar{y} = 6\bar{y}$. From (21.5.2) for the standard error of $\bar{y}$, the variance of $\hat{Y}$ is

$$V(\hat{Y}) = N^2s^2(1-\phi)/n = (36)(6.1057)/3 = 73.268$$

The standard error of this estimate is ±8.56, giving a coefficient of variation of ±46%—very inaccurate.

The ratio estimate $\hat{Y}_R = (\bar{y}/\bar{x})X = 18.42\bar{y}/\bar{x}$ is slightly biased. Since no exact formula for its mean square error is known, table 21.12.2 presents the estimates from all 15 simple random samples of size 2. From these we calculate the mean square error as $\Sigma(\hat{Y}_R - Y)^2/15$. The 15 estimates $N\bar{y}$ are also shown for comparison.

Note that the bias in $\hat{Y}_R$, +0.05, is trivial and that the 15 ratio estimates $\hat{Y}_R$ range from 17.18 to 21.81, as against a range from 8.22 to 33.78 for the 15 unbiased estimates $N\bar{y}$. You may verify that $MSE(\hat{Y}_R) = 1.201$, giving a coefficient of variation of 5.8% and a deff value of only 0.016 relative to $\hat{Y}$.

TABLE 21.12.2
RATIO ESTIMATES $\hat{Y}_R = (\bar{y}/\bar{x})X$ AND ESTIMATES $\hat{Y} = N\bar{y}$

| Sample Units | Sample Totals | | | | Sample Units | Sample Totals | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1960 | 1970 | $\hat{Y}_R$ | $\hat{Y}$ | | 1960 | 1970 | $\hat{Y}_R$ | $\hat{Y}$ |
| 1, 2 | 6.03 | 6.18 | 18.88 | 18.54 | 2, 6 | 3.42 | 4.05 | 21.81 | 12.15 |
| 1, 3 | 11.33 | 11.26 | 18.31 | 33.78 | 3, 4 | 9.78 | 9.85 | 18.55 | 29.55 |
| 1, 4 | 5.55 | 5.31 | 17.62 | 15.93 | 3, 5 | 9.45 | 9.41 | 18.34 | 28.23 |
| 1, 5 | 5.22 | 4.87 | 17.18 | 14.61 | 3, 6 | 8.72 | 9.13 | 19.29 | 27.39 |
| 1, 6 | 4.49 | 4.59 | 18.83 | 13.77 | 4, 5 | 3.67 | 3.40 | 17.37 | 10.38 |
| 2, 3 | 10.26 | 10.72 | 19.25 | 32.16 | 4, 6 | 2.94 | 3.18 | 19.92 | 9.54 |
| 2, 4 | 4.48 | 4.77 | 18.38 | 14.31 | 5, 6 | 2.61 | 2.74 | 19.34 | 8.22 |
| 2, 5 | 4.15 | 4.33 | 19.22 | 12.99 | Mean | | | 18.82 | 18.77 |

**21.13—Nonsampling errors.** In many surveys, especially surveys dealing with human subjects and institutions, sources of error other than those due to sampling affect the estimates. The most common are probably missing data. In a survey taken by mail, only 30% of those to whom questionnaires are sent may reply. In an interview survey made by visiting a sample of names and addresses, perhaps 10% of the people may not be home and a further 4% refuse or are unable to answer the questions.

With missing data our sample is smaller than planned, but a bigger problem is that we often have reason to believe that the misses (the nonrespondents) differ systematically from the respondents. Consequently, our sample of respondents is biased, though evidence about the size of this bias may naturally be hard to obtain. To illustrate from an oversimplified model, suppose that our field method (mail, telephone, household interview) can reach a proportion $w_1$ of the population but fails to get replies from a proportion $w_0$; and that the two subpopulations have means $\bar{Y}_1$, $\bar{Y}_0$ for the variable being measured. Our sample of size $n_1$ is a random sample of the respondents. The mean square error of the sample mean $\bar{y}_1$ is then

$$MSE(\bar{y}_1) = E(\bar{y}_1 - \bar{Y})^2 = E(\bar{y}_1 - w_1\bar{Y}_1 - w_0\bar{Y}_0)^2$$
$$= E[(\bar{y}_1 - \bar{Y}_1) + w_0(\bar{Y}_1 - \bar{Y}_0)]^2 = s_1^2/n_1 + w_0^2(\bar{Y}_1 - \bar{Y}_0)^2$$

ignoring the finite population correction. With large samples, the bias term may dominate this mean square error and our sampling error formulas may seriously underestimate it and our real errors.

There are two strategies for attacking this problem. One is to use field methods that reduce $w_0$, for instance by insisting that at least three or four attempts be made to reach and obtain answers from any sample member. Alternatively, if supplementary information can be obtained about nonrespondents that indicates to some extent how they differ from respondents, another strategy is to use a different estimate that takes this knowledge into account.

For example, suppose that males differ markedly from females in their replies to one question. A planned sample of 1000 has 487 males and 513 females, this being the proportion in the population. Responses are obtained from 410 males (84%) and 492 females (96%). Instead of the sample mean, we use the estimate

$$\hat{\bar{Y}} = 0.487\bar{y}_m + 0.513\bar{y}_f$$

If the males who did not respond have relatively little bias as compared with those who did respond, this estimate should be almost free from bias. Another approach uses available knowledge about nonrespondents to substitute or *impute* estimates of the responses that they would have given. The assumptions in this approach are similar to those made in substituting for a missing value in randomized blocks from knowledge of the treatment and block corresponding to the missing observation.

Errors of measurement, including those introduced in classifying and coding the responses for analysis, are another source of inaccuracy. Sometimes the question is poorly worded and has different meanings for different subjects. Pretests and revisions of the questionnaire help here. In summary, the objective in planning and conducting a survey should be to minimize the total error, not just the sampling error. This involves the difficult job of allocating resources among reduction of sampling errors, missing data, and errors of measurement, and of deciding from what we know how best to use these resources for each purpose.

**21.14—Further reading.** The general books on sample surveys that have become standard (2, 3, 4, 5, 10) involve roughly the same level of mathematical difficulty and knowledge of statistics. Reference (3) is oriented toward applications in business and (10) toward those in agriculture. Another good book for agricultural applications, at a lower mathematical level, is (11).

Useful short books are (12), an informal, popular account of some of the interesting applications of survey methods; (13), which conducts the reader painlessly through the principal results in probability sampling at about the mathematical level of this chapter; and (14), which discusses the technique of constructing interview questions.

Books and papers have also begun to appear on some of the common specific types of application. For sampling a town under U.S. conditions, with the block as primary sampling unit, references (15) and (16) are recommended. Reference (17), intended primarily for surveys by health agencies to check on the immunization status of children, gives instructions for the sampling of attributes in local areas, while (18) deals with the sampling of hospitals and patients. Much helpful advice on the use of sampling in agricultural censuses is found in (19), and (20) presents methods for reducing errors of measurement.

## TECHNICAL TERMS

design effect (deff)
finite population correlation
imputing
listing
nonsampling errors
probability proportional to size (pps)

probability sampling
proportional stratification
ratio estimates
stratified random sampling
systematic sampling
two-stage sampling

## REFERENCES

1. Clapham, A. R. *J. Agric. Sci.* 19 (1929):214.
2. Cochran, W. G. 1967. *Sampling Techniques*, 3rd ed. Wiley, New York.
3. Deming, W. Edwards. 1960. *Sample Design in Business Research.* Wiley, New York.
4. Hansen, M. H.; Hurwitz, W. N.; and Madow, W. G. 1953. *Sample Survey Methods and Theory.* Wiley, New York.
5. Kish, L. 1965. *Survey Sampling.* Wiley, New York.
6. Immer, F. R. *J. Agric. Res.* 44 (1932):633.
7. Neyman, J. *J. R. Stat. Soc.* 97 (1934):558.
8. Hansen, M. H., and Hurwitz, W. N. *Ann. Math. Stat.* 14 (1943):333.
9. West, Q. M. 1951. Mimeographed Report. Cornell Univ. Agric. Exp. Stn.
10. Yates, F. 1960. *Sampling Methods for Censuses and Surveys*, 3rd ed. Griffin, London.
11. Sampford, M. R. 1962. *An Introduction to Sampling Theory.* Oliver & Boyd, Edinburgh.
12. Slonim, M. J. 1960. *Sampling in a Nutshell.* Simon & Schuster, New York.
13. Stuart, A. 1962. *Basic Ideas of Scientific Sampling.* Griffin, London.
14. Payne, S. L. 1951. *The Art of Asking Questions.* Princeton Univ. Press.
15. Woolsey, T. D. 1956. *Sampling Methods for a Small Household Survey.* Public Health Monographs 40.
16. Kish, L. *Am. Soc. Rev.* 17 (1952):761.
17. Serfling, R. E., and Sherman, I. L. 1965. *Attribute Sampling Methods.* U.S. Government Printing Office, Washington, D.C.
18. Hess, I.; Riedel, D. C.; and Fitzpatrick, T. B. 1975. *Probability Sampling of Hospitals and Patients*, 2nd ed. Univ. Michigan, Ann Arbor.
19. Zarcovich, S. S. 1965. *Sampling Methods and Censuses.* FAO, Rome.
20. Zarcovich, S. S. 1966. *Quality of Statistical Data.* FAO, Rome.