Estimating Parameters from Simple, Stratified and Cluster Sampling Procedures

Simple Random Sampling

Suppose the observations $y_1, y_2, \ldots y_n$ are to be sampled from a population with mean μ, standard deviation σ, and size *N* in such a way that every possible sample of size *n* has an equal chance of being selected. Then the sample $y_1, y_2, \ldots y_n$ was selected in a simple random sample. If the sample mean is denoted by $\bar{y}$, then we have

$$E(\bar{y}) = \mu$$

and

$$V(\bar{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

The term $\left(\frac{N-n}{N-1}\right)$ in the above expression is known as the finite population correction factor. For the sample variance s^2, it can be shown that

$$E(s^2) = \left(\frac{N}{N-1}\right)\sigma^2.$$

When using s^2 as an estimate of σ^2, we must adjust with $\sigma^2 \approx \left(\frac{N-1}{N}\right)s^2$. Consequently, an unbiased estimator of the variance of the sample mean is given by

$$\hat{V}(\bar{y}) = \frac{\left(\frac{N-1}{N}\right)s^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{s^2}{n}\left(\frac{N-n}{N}\right).$$

As a rule of thumb, the correction factor $\left(\frac{N-n}{N}\right)$ can be ignored if it is greater than 0.9, or if the sample is less than 10% of the population.

As an example, consider the finite population composed of the $N = 4$ elements $\{0, 2, 4, 6\}$. For this population $\mu = 3$ and $\sigma^2 = 5$. Simple random samples, without replacement, of size $n = 2$ are selected from this population. All possible samples along with their summary statistics are listed below.

Sample	Probability	Mean	Variance
$\{0, 2\}$	1/6	1	2
$\{0, 4\}$	1/6	2	8
$\{0, 6\}$	1/6	3	18
$\{2, 4\}$	1/6	3	2
$\{2, 6\}$	1/6	4	8
$\{4, 6\}$	1/6	5	2

(1) The expected value of the sample means is

$$E(\bar{y}) = \sum_{i=1}^{6} \bar{y}_i \cdot p(\bar{y}_i) = \left(\frac{1}{6}\right)(1+2+3+3+4+5) = 3.$$

Notice that $E(\bar{y}) = \mu$.

(2) The variance of the sample means is

$$V(\bar{y}) = E(\bar{y}^2) - (E(\bar{y}))^2 = E(\bar{y}^2) - (3)^2. \quad \text{So}$$

$$E(\bar{y}^2) = \sum_{i=1}^{6} \bar{y}_i^2 \cdot p(\bar{y}_i^2) = \left(\frac{1}{6}\right)(1^2 + 2^2 + 3^2 + 3^2 + 4^2 + 5^2) = \frac{64}{6}$$

and

$$V(\bar{y}) = \frac{64}{6} - 9 = \frac{5}{3}$$

We see in this example that $V(\bar{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \left(\frac{5}{2}\right)\left(\frac{4-2}{4-1}\right) = \left(\frac{5}{2}\right)\left(\frac{2}{3}\right) = \frac{5}{3}$.

(3) The expected value of the sample variances is

$$E(s^2) = \sum_{i=1}^{6} s_i^2 \cdot p(s_i^2) = \left(\frac{1}{6}\right)(2+8+18+2+8+2) = \frac{20}{3}.$$

Again, we see that $E(s^2) = \left(\frac{N}{N-1}\right)\sigma^2 = \left(\frac{4}{3}\right)(5) = \frac{20}{3}$, as the theory states must be true.

Estimation of a Population Mean

If we are interested in estimating a population mean from a simple random sample, we have

$$\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

If we are interested in estimating a population variance from a simple random sample, we have

$$\hat{V}(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$$

where

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

The margin of error is 2 standard errors, so

$$2\sqrt{\hat{V}(\bar{y})} = 2\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)}.$$

Estimation of a Population Proportion

If each observation in the sample is coded 1 for "success" and 0 for "failure", the sample mean becomes the sample proportion. In addition, we have

$$\frac{s^2}{n} = \frac{\hat{p}(1-\hat{p})}{n-1},$$

where $\hat{p}$ denotes the sample proportion. To see this, recall that $s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$, so

$$(n-1)s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\left(y_i^2 - 2y_i\bar{y} + \bar{y}^2\right) = \sum_{i=1}^{n}\left(y_i^2\right) - 2\bar{y}\sum_{i=1}^{n}y_i + \sum_{i=1}^{n}\bar{y}^2.$$

Since $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, we have $n\bar{y} = \sum_{i=1}^{n} y_i$. Also, since each y_i is either 0 or 1, we have $\sum y_i^2 = \sum y_i$ and $\bar{y} = \hat{p}$.

Then

$$\sum_{i=1}^{n}\left(y_i^2\right) - 2\bar{y}\sum_{i=1}^{n}y_i + \sum_{i=1}^{n}\bar{y}^2 = \sum_{i=1}^{n}y_i - 2n\bar{y}^2 + n\bar{y}^2 = n\bar{y} - n\bar{y}^2 = n\hat{p} - n\hat{p}^2 = n\hat{p}(1-\hat{p}).$$

So, we have $(n-1)s^2 = n\hat{p}(1-\hat{p})$ or equivalently, $\frac{s^2}{n} = \frac{\hat{p}(1-\hat{p})}{n-1}$.

Using the formulas for the mean and the equality above, we can determine the estimator of the population proportion, of the variance of $\hat{p}$, and the margin of error for the proportion.

The estimator of the population proportion is $\hat{p} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

The estimated variance of $\hat{p}$ is $\hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right)$. The margin of error of estimation is $2\sqrt{\hat{V}(\hat{p})} = 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right)}$.

Estimating the Population Total

Finding an estimate of the population total is meaningless for an infinite population. However, for a finite population, the population total is a very important population parameter. For example, we may want to estimate the total yield of corn in Iowa, or the total number of apples in an orchard. If we know the population size *N* and the population mean μ, then the total τ is just $\tau = N\mu$.

So, the estimator of the population total τ is $\hat{\tau} = N\bar{y} = \frac{N\sum_{i=1}^{n} y_i}{n}$. The estimated variance of τ is $\hat{V}(\hat{\tau}) = \hat{V}(N\bar{y}) = N^2 \cdot \hat{V}(\bar{y}) = N^2\left(\frac{s^2}{n}\right)\left(\frac{N-n}{N}\right)$.

Finally, the margin of error of estimation for τ is

$$2\sqrt{\hat{V}(N\bar{y})} = 2\sqrt{N^2\left(\frac{s^2}{n}\right)\left(\frac{N-n}{N}\right)} = 2Ns\sqrt{\frac{1}{n}-\frac{1}{N}}.$$

Sampling with Subsamples

Suppose you require several field workers to perform the sampling or the sampling takes place over several days. There will be variation in the measurements among the field workers or among the days of sampling. The population mean can be estimated using the subsample means of each of the field workers or for each of the days. This is not a stratified sample, but simply breaking up the sample into subsamples. This method of sampling was developed by Edward Deming.

The sample of size *n* is to be divided into *k* subsamples, with each subsample of size *m*. Let $\bar{y}_i$ denote the mean of the i^{th} subsample.

- The estimator of the population mean μ is $\bar{y} = \frac{1}{k}\sum_{i=1}^{k}\bar{y}_i$, the average of the subsample means.
- The estimated variance of $\bar{y}$ is $\hat{V}(\bar{y}) = \left(\frac{N-n}{N}\right)\frac{s_k^2}{k}$ where $s_k^2 = \frac{\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2}{k-1}$ and measures the variation among the subsample means.

Stratified Random Sampling

As described earlier, stratified random sampling produces estimators with smaller variance than those from simple random sampling, for the same sample size, when the measurements under study are homogeneous within strata but the stratum means vary among themselves. The ideal situation for stratified random sampling is to have all measurements within any one stratum equal but have differences occurring as we move from stratum to stratum. To create a stratified random sample, divide the population into subgroups so that every element of the population is in one and only one subgroup (non-overlapping, exhaustive subgroups). Then take a simple random sample within each subgroup.

The reasons one may choose to perform a stratified random sample are

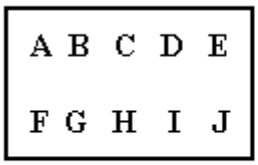
(1) Possible reduction in the variation of the estimators (statistical reason)

(2) Administrative convenience and reduced cost of survey (practical reason)

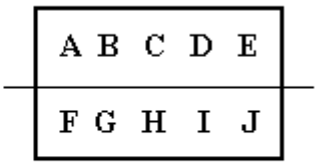
(3) Estimates are often needed for the subgroups of the population

Stratification is a widely used technique as most large surveys have stratification incorporated into the design. Additionally, stratification is one of the basic principles of measuring quality and of quality control. (The noted statistician Edward Deming spent half of his life working in survey sampling and the other half in quality control.) Finally, stratification can substitute for direct control in observational studies.

A stratified sample cannot be a simple random sample. As an example, consider the population of 10 letters given below.



Simple Random Sample



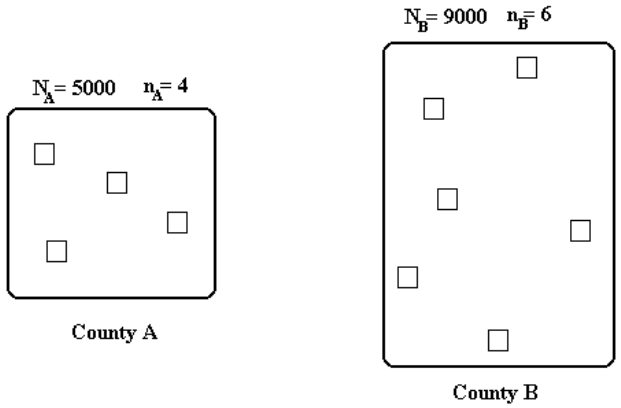
Stratified Random Sample

Take a sample of size 4 from the population on the left. The probability that **A** is in the sample is $P(A)=\frac{4}{10}$. The probability of the sample ABCF (order does not matter) is $P(ABCF)=\frac{1}{\binom{10}{4}}$.

In the stratified population on the right, in which two elements are taken from the first row and two from the second, the probability that **A** is in the sample is still $P(A)=\frac{4}{10}$. However, the probability of achieving the sample ABCF is $P(ABCF)=0$. Even though the probability of any single element being in the sample is the same, all samples of size 4 are not equally likely, and thus, this is not a simple random sample.

Estimating the Population Mean in a Stratified Sample

Suppose we wish to estimate the yield of corn in two counties (A and B) in Iowa. County A has N_A acres of corn and County B has N_B acres of corn. Here, we are assuming that all N_i are sufficiently large so that the finite population correction factor can be ignored. The counties constitute two strata and we will take a simple random sample of size n_A from County A and n_B from County B, as described in the diagram at right.



We want to estimate the total amount of corn for the two counties. If $\bar{y}_A$ is the mean yield of corn per acre for the 4 plots in County A and $\bar{y}_B$ is the mean yield of corn per acre for the 6 plots in County B, then

$$\hat{\tau} = N_A \bar{y}_A + N_B \bar{y}_B$$

is our estimate of the total amount of corn in the two counties.

Our estimate of the mean yield of corn per acre for the two counties is

$$\hat{\mu} = \frac{N_A \bar{y}_A + N_B \bar{y}_B}{N_A + N_B} = \frac{N_A}{N} \bar{y}_A + \frac{N_B}{N} \bar{y}_B,$$

if we let $N = N_A + N_B$ be the total acreage for the two counties. This estimator can be written as a weighted average

$$\hat{\mu} = W_A \bar{y}_A + W_B \bar{y}_B \text{ with } W_A = \frac{N_A}{N} \text{ and } W_B = \frac{N_B}{N}$$

where the weights are the population proportions. The variance of $\hat{\mu}$ is easily computed

$$V(\hat{\mu}) = V(W_A \bar{y}_A + W_B \bar{y}_B) = W_A^2 V(\bar{y}_A) + W_B^2 V(\bar{y}_B) = W_A^2 \frac{\sigma_A^2}{n_A} + W_B^2 \frac{\sigma_B^2}{n_B}.$$

In general, if there are *L* strata of size N_i with $\sum_{i=1}^{L} N_i = N$ with samples of size n_i with $\sum_{i=1}^{L} n_i = n$ taken from each strata, respectively, then:

- the estimator of the total is $\hat{\tau} = \sum_{i=1}^{L} N_i \bar{y}_i$.
- the estimator of the mean is $\hat{\mu} = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i$ or $\hat{\mu} = \sum_{i=1}^{L} W_i \bar{y}_i$ with $W_i = \frac{N_i}{N}$ the population proportion.

We have our estimated mean $\bar{y} = \sum_{i=1}^{L} W_i \bar{y}_i$, so $V(\bar{y}) = \sum_{i=1}^{L} W_i^2 V(\bar{y}_i) = \sum_{i=1}^{L} W_i^2 \frac{\sigma_i^2}{n_i}$.

This last expression can be rewritten using sample proportions as weights $w_i = \frac{n_i}{n}$. So,

$$V(\bar{y}) = \sum_{i=1}^{L} \frac{W_i^2 \sigma_i^2}{n w_i}.$$

The Problems of Sample Size and Allocation

Suppose we want to estimate the mean yield of corn to within 100 bushels/acre. How can we use the equations above to determine the appropriate sample size *n* and the allocations n_i to produce an estimate accurate to a specified tolerance? We will, as usual, use $2\sqrt{V(\bar{y})} = B$ as our margin of error. We require values of *n* and n_i so that $V(\bar{y}) = \frac{B^2}{4} = D$ (called the dispersion). Then $D = \frac{1}{n}\left[\sum_{i=1}^{L} \frac{W_i^2 \sigma_i^2}{w_i}\right]$ and consequently,

$$n = \frac{1}{D}\left[\sum_{i=1}^{L} \frac{W_i^2 \sigma_i^2}{w_i}\right],$$

with $D = \frac{B^2}{4}$ when estimating μ and $D = \frac{B^2}{4N^2}$ when estimating τ.

We know that $W_i = \frac{N_i}{N}$ are population proportions. However, in order to find *n* we must know the weights w_i.

One method for determining the sample proportions w_i is to simply assign them the same values as the population proportions, so $w_i = W_i = \frac{N_i}{N}$. This method is particularly useful when the variances of the strata are similar.

Another standard procedure is to use the weights that minimize the variance. Consider the case when two strata are used. Then

$$V(\bar{y}) = \frac{W_1^2 \sigma_1^2}{n_1} + \frac{W_2^2 \sigma_2^2}{n_2} = \frac{k_1^2}{n_1} + \frac{k_2^2}{n - n_1} \text{ where } k_i^2 = W_i^2 \sigma_i^2 \text{ is a constant.}$$

Now, to find the value of n_1 that minimizes $V(\bar{y})$, we use calculus. So,

$$\frac{d}{dn_1}\left(\frac{k_1^2}{n_1} + \frac{k_2^2}{n - n_1}\right) = \frac{-k_1^2}{n_1^2} + \frac{k_2^2}{(n - n_1)^2} = 0.$$

Solving for n_1, we have

$$\frac{k_2^2}{(n - n_1)^2} = \frac{k_1^2}{n_1^2} \text{ or } \frac{n_1^2}{n_2^2} = \frac{k_1^2}{k_2^2}, \text{ so } \frac{n_1}{n_2} = \frac{k_1}{k_2} = \frac{W_1 \sigma_1}{W_2 \sigma_2}.$$

Then $n = n_1 + n_2 = n_1 + \frac{k_2}{k_1} n_1 = n_1 \left(\frac{k_1 + k_2}{k_1} \right)$. Solving for n_1, we have $n_1 = n \left(\frac{k_1}{k_1 + k_2} \right)$.

In general, we have $n_i = n \left(\frac{k_i}{\sum_{i=1}^{L} k_i} \right) = n \left(\frac{W_i \sigma_i}{\sum_{i=1}^{L} W_i \sigma_i} \right)$.

This last equation indicates that the allocation to region *i* will be large if $W_i = \frac{N_i}{N}$ is large, that is, if it contains a large portion of the population. This should make sense. It also indicates that the allocation to region *i* will be large if there is a lot of variability in the region. If there is little variation in the region, the allocation will be small, since a small sample will give the necessary information. As an extreme example, if there is no variation in a region, a single sample will tell you everything about the region. This optimal allocation was developed by the statistician Jerzy Neyman and is called the Neyman allocation.

Example 1. Consider the two counties A and B with $N_A = 5000$ acres and $N_B = 9000$ acres. Suppose we can approximate the variance of the yields for the two counties based on past performance as $\sigma_A \approx 12$ bushels/acre and $\sigma_B \approx 20$ bushels/acre. We want to estimate the mean yield in bushels per acre for the two counties with a margin of error of 5 bushels/acre. What are the values of *n*, n_A, and n_B if

a) we use proportional allocation
b) we allocate samples to minimize the variance (optimal allocation)

a) Here we have $\frac{n_A}{n_B} = \frac{N_A}{N_B} = \frac{5}{9}$. This means that $n_A = \frac{5}{14} n$ and $n_B = \frac{9}{14} n$ and $w_A = \frac{n_A}{n} = \frac{5}{14}$ with $w_B = \frac{9}{14}$. Using the formula derived above,

$$n = \frac{1}{D} \left[\sum_{i=1}^{L} \frac{W_i^2 \sigma_i^2}{w_i} \right] = \frac{1}{D} \left[\frac{W_A^2 \sigma_A^2}{w_A} + \frac{W_B^2 \sigma_B^2}{w_B} \right],$$

we can find the appropriate values of *n*, n_A, and n_B. We know everything except D. To find *D*, we have $B = 5$, so $D = \frac{B^2}{4} = \frac{25}{4}$.

Now,

$$n = \frac{4}{25} \left[\frac{\left(\frac{5}{14} \right)^2 (12)^2}{\left(\frac{5}{14} \right)} + \frac{\left(\frac{9}{14} \right)^2 (20)^2}{\left(\frac{9}{14} \right)} \right] \approx 50$$

So proportional allocation gives $n = 50$, $n_A = \left(\frac{5}{14} \right) 50 \approx 18$ and $n_B = \left(\frac{9}{14} \right) 50 \approx 32$.

b) Optimal allocation requires that

$$n_A = n\left(\frac{W_A\sigma_A}{W_A\sigma_A + W_B\sigma_B}\right) = (n)\frac{\left(\frac{5}{14}\right)(12)}{\left(\frac{5}{14}\right)(12)+\left(\frac{9}{14}\right)(20)} = \frac{1}{4}n$$

and

$$n_B = n\left(\frac{W_B\sigma_B}{W_A\sigma_A + W_B\sigma_B}\right) = (n)\frac{\left(\frac{9}{14}\right)(20)}{\left(\frac{5}{14}\right)(12)+\left(\frac{9}{14}\right)(20)} = \frac{3}{4}n.$$

As before,

$$n = \frac{1}{D}\left[\frac{W_A^2\,\sigma_A^2}{w_A} + \frac{W_B^2\,\sigma_B^2}{w_B}\right],$$

and so,

$$n = \frac{4}{25}\left[\frac{\left(\frac{5}{14}\right)^2(12)^2}{\left(\frac{1}{4}\right)} + \frac{\left(\frac{9}{14}\right)^2(20)^2}{\left(\frac{3}{4}\right)}\right] \approx 47$$

So proportional allocation gives $n = 47$, $n_A = \left(\frac{1}{4}\right)47 \approx 12$ and $n_B = \left(\frac{3}{4}\right)47 \approx 35$.

Notice that, although fewer samples were needed, more samples came from County B, since it had both greater variation and was a larger proportion of the population.

Considering Cost and Finite Population Factor

The equations developed in this section become somewhat more complex if the finite population correction factor must be included in the calculations. In this case, we have

$$n = \frac{\sum_{i=1}^{L} N_i^2 \frac{\sigma_i^2}{w_i}}{N^2 D + \sum_{i=1}^{L} N_i\,\sigma_i^2} \text{ with } D = \frac{B^2}{4} \text{ when estimating } \mu \text{ and } D = \frac{B^2}{4N^2} \text{ when estimating } \tau.$$

The approximate allocation that minimizes total cost for a fixed variance, or minimizes variance for a fixed costs (c_i) is $n_i = n\left(\frac{\frac{N_i\sigma_i}{\sqrt{c_i}}}{\sum_{k=1}^{L}\frac{N_k\sigma_k}{\sqrt{c_k}}}\right)$. Note that n_i is directly proportional to N_i and σ_i and inversely proportional to $\sqrt{c_i}$.

Comparison of Stratified Random Sampling to Simple Random Sampling

Stratification usually produces gains in precision, especially if the stratification is accomplished through a variable correlated with the response. We would like to stratify when the strata are homogeneous and different, that is, we have

1) low variation in the strata
2) differing means among the strata.

The following comparisons apply for situations in which the N_i are all relatively large, so we can replace $\frac{1}{N_i - 1}$ with $\frac{1}{N_i}$. Here we use $f = \frac{n}{N}$ and $W_i = \frac{N_i}{N}$.

The variance of a SRS, denoted V_{SRS}, compared to the variance of a proportional allocation, denoted V_{prop} is described in the equation

$$V_{SRS} - V_{prop} = \frac{1-f}{n} \sum_i W_i \left(\bar{Y}_i - \bar{Y} \right)^2 .$$

From this equation, we see that the proportional allocation will be useful (produce a smaller variance than SRS) when there is a large difference in the means for the different strata.

The variance of proportional allocation compared to the variance of an optimal Neyman allocation, denoted V_{opt} is described in the equation

$$V_{prop} - V_{opt} = \frac{1}{n} \sum_i W_i \left(S_i - \bar{S} \right)^2 ,$$

where S_i is a measure of the random variation of the population strata and $\bar{S} = \sum_i W_i S_i$. From this equation, we see that the optimal allocation is an improvement over proportional allocation when there is a large difference in the variation among the strata.

In summary, one should attempt to construct strata so that the strata means differ. If strata variances do not differ much, use proportional allocation. If strata variances differ greatly, use optimum Neyman allocation.

A Word on Post Stratification

At times, we wish to stratify a sample after a simple random sample has been taken. For example, suppose you wish to stratify on gender based on a telephone poll, where you can't know the gender of the respondent until after the SRS is taken. What penalty do we pay if we decide to stratify after selecting a simple random sample? It is possible to show that the estimated variance, $\hat{V}_p \left(\bar{y} \right)$, is given by

$$\hat{V}_p \left(\bar{y} \right) = \left(\frac{N-n}{Nn} \right) \sum_{i=1}^{L} W_i \, s_i^2 + \frac{1}{n^2} \sum_{i=1}^{L} \left(1 - W_i \right) s_i^2 .$$

The first term is what you would expect from a stratified sample mean using proportional allocation, so the second term is the price paid for stratifying after the fact. Notice that the term $\frac{1}{n^2}$ reduces the penalty as n increases. Post-stratification produces good results when *n* is large and all n_i are large as well.

Ratio Estimation

Ratio estimation is an important issue in cluster sampling. We will develop the principles of ratio estimation and then proceed to cluster sampling.

How do you determine the mpg for your car? One way would be to note the miles driven and the number of gallons of gas used each time you fill up the gas tank. This will produce a set of ordered pairs, each of which can be used to estimate your mpg. What is the best estimate you can make from this information?

miles	y_1	y_2	y_3	$\cdots$	y_n
gallons	x_1	x_2	x_3	$\cdots$	x_n

We can compute all *n* ratios $\frac{y_i}{x_i}$ and find the average value $\frac{1}{n}\sum\left(\frac{y_i}{x_i}\right)$. Unfortunately, $E\left(\frac{y_i}{x_i}\right) \neq \frac{\mu_y}{\mu_x}$. Each division of $\frac{y_i}{x_i}$ produces some bias, so we want to perform as few divisions as possible.

The best estimator of the population ratio $R = \frac{\mu_y}{\mu_x}$ is $r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$.

The estimated variance of *r* can be approximated by

$$\hat{V}(r) = \hat{V}\left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\right) = \left(\frac{N-n}{N}\right)\left(\frac{1}{\mu_x^2}\right)\left(\frac{s_r^2}{n}\right),$$

where $s_r^2 = \frac{\sum_{i=1}^{n}\left(y_i - r x_i\right)^2}{n-1}$. The estimated variance of *r* is similar to the formula for the variance of a sample mean, but has the additional $\left(\frac{1}{\mu_x^2}\right)$ term. The value of s_r^2 is similar to the variance of residuals.

If we plot the ordered pairs (x_i, y_i), we are comparing these points to the line $y = r\,x$.

Our estimate of the ratio r allows us to make estimates of the population mean, $\hat{\mu}_y$, and the population total, $\hat{\tau}_y$. If $\frac{\mu_y}{\mu_x}$ is estimated by $\frac{\overline{y}}{\overline{x}}$, then we should be able to estimate μ_y with $\hat{\mu}_y = \frac{\overline{y}}{\overline{x}}\mu_x = r\,\mu_x$. The estimated variance of μ_y is $\hat{V}(\hat{\mu}_y) = \mu_x^2 \hat{V}(r) = \left(\frac{N-n}{N}\right)\frac{s_r^2}{n}$.

Similarly, the ratio estimator of the population total, τ_y, is $\hat{\tau}_y = \frac{\overline{y}}{\overline{x}}\tau_x = r\,\tau_x$.

The estimated variance of τ_y is

$$\hat{V}(\hat{\tau}_y) = \tau_x^2 \hat{V}(r) = \tau_x^2 \left(\frac{N-n}{N}\right)\left(\frac{1}{\mu_x^2}\right)\frac{s_r^2}{n}.$$

Note that we do not need to know τ_x or N to estimate μ_y when using the ratio procedure. However, we must know μ_x.

Example (Adapted from Scheaffer, et al, *Elementary Survey Sampling, 5th Edition*, page 205-206):

In Florida, orange farmers are paid according to the sugar content in their oranges. How much should a farmer be paid for a truckload of oranges? A sample is taken, and the total amount of sugar in the truckload can estimated using the ratio method.

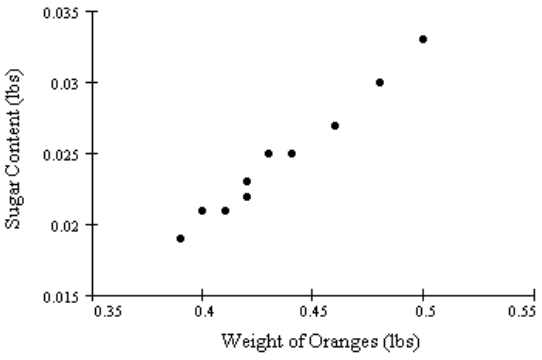
Suppose 10 oranges were selected at random from the truckload to be tested for sugar content. The truck was weighed loaded and unloaded to determine the weight of the oranges. In this case, there were 1800 pounds of oranges. Larger oranges have more sugar, so we want to know the sugar content per pound for the truckload and use this to estimate the total sugar content of the load.

Orange	1	2	3	4	5	6	7	8	9	10
Sugar Content (lbs)	0.021	0.030	0.025	0.022	0.033	0.027	0.019	0.021	0.023	0.025
Wt of Orange (lbs)	0.40	0.48	0.43	0.42	0.50	0.46	0.39	0.41	0.42	0.44

The scatterplot above shows a strong linear relationship between the two variables, so a ratio estimate is appropriate. Using the formula $\hat{\tau}_y = \frac{\overline{y}}{\overline{x}}\tau_x = r\,\tau_x$ we estimate

$$\hat{\tau}_y = \frac{0.0246}{0.4350}(1800) = (0.05655)(1800) = 101.8$$

Pounds of sugar in the truckload. A bound on the error of estimation can be found as well.



We have $\hat{V}\left(\hat{\tau}_y\right) = \tau_x^2 \hat{V}\left(r\right) = \tau_x^2 \left(\frac{N-n}{N}\right)\left(\frac{1}{\mu_x^2}\right)\frac{s_r^2}{n}$, but in this case, we know neither N nor μ_x. Since N is large (a truckload of oranges will be at least 4,000 oranges), so the finite population correction $\left(\frac{N-n}{N}\right)$ is essentially 1. We will use $\bar{x}$ as an estimate of μ_x. With these modifications, we can compute

$$2\sqrt{\hat{V}\left(\hat{\tau}_y\right)} = 2\sqrt{\tau_x^2\left(\frac{1}{\bar{x}^2}\right)\frac{s_r^2}{n}} = 2\sqrt{(1800)^2\left(\frac{1}{0.435^2}\right)\frac{0.0024^2}{10}} = 6.3$$

Our estimate of the total sugar content of the truckload of oranges is 101.8 ± 6.3 pounds.

If the population size *N* is know, we could also use the estimator $N\bar{y}$ instead of $r\tau_x$ to estimate the total. Generally, the estimator $r\tau_x$ has a smaller variance than $N\bar{y}$ when there is a strong positive correlation between *x* and *y*. As a rule of thumb, if $\rho > \frac{1}{2}$, the ratio estimate should be used. This decrease in variance results from taking advantage of the additional information provided by the subsidiary variable *x* in our calculations with the ratio estimation.

Relative Efficiency of Estimators

Suppose there are two unbiased (or nearly unbiased) estimators, E_1 and E_2, for the same parameter. The relative efficiency of the two estimators is measured by the ratio of the reciprocals of their variances. That is,

$$RE\left(\frac{E_1}{E_2}\right) = \frac{V\left(E_2\right)}{V\left(E_1\right)}.$$

If $RE\left(\frac{E_1}{E_2}\right) > 1$, estimator E_1 will be more efficient. If the sample sizes are the same, the variance of E_1 will be smaller. Another way to view this is that estimator E_1 will produce the same variance as E_2 with a smaller sample size.

We can compute the relative efficiency of μ_y and $\bar{y}$. Here, we have

$$RE\left(\frac{\hat{\mu}_y}{\bar{y}}\right) = \frac{V\left(\bar{y}\right)}{V\left(\hat{\mu}_y\right)} = \frac{s_y^2}{s_r^2}.$$

Both variances have the same values of *N* and *n*, so the finite population correction factor divides out. The variance of $\hat{\mu}_y$ can be re-written in terms of the predicted correlation $\hat{\rho}$ so that

$$RE\left(\frac{\hat{\mu}_y}{\bar{y}}\right) = \frac{s_y^2}{s_y^2 + r^2 s_x^2 - 2r\hat{\rho}\, s_x s_y}.$$

If $RE\left(\frac{\hat{\mu}_y}{\bar{y}}\right) > 1$ then $\hat{\mu}_y$ is a more efficient estimator. To determine when $RE\left(\frac{\hat{\mu}_y}{\bar{y}}\right) > 1$, we consider $\frac{s_y^2}{s_y^2 + r^2 s_x^2 - 2r\hat{\rho}\, s_x s_y} > 1$. Then $s_y^2 > s_y^2 + r^2 s_x^2 - 2r\hat{\rho}\, s_x s_y$, or $2\hat{\rho}\, s_x s_y > r\, s_x^2$. If $\rho > 0$, then $\hat{\rho} > \frac{r\, s_x^2}{2 s_x s_y} = \frac{1}{2}\left(\frac{s_x/\bar{x}}{s_y/\bar{y}}\right)$. As is often the case in ratio estimation, $\frac{s_x}{\bar{x}} \approx \frac{s_y}{\bar{y}}$, we see that $\hat{\mu}_y$ is a more efficient estimator than $\bar{y}$ when $\hat{\rho} > \frac{1}{2}$.

Cluster Sampling

Sometimes it is impossible to develop a frame for the elements that we would like to sample. We might be able to develop a frame for clusters of elements, though, such as city blocks rather than households or clinics rather than patients. If each element within a sampled cluster is measured, the result is a **single-stage cluster sample**. A cluster sample is a probability sample in which each sampling unit is a collection, or cluster, of elements. Cluster sampling is less costly than simple or stratified random sampling if the cost of obtaining a frame that lists all population elements is very high or if the cost of obtaining observations increases as the distance separating the elements increases.

To illustrate, suppose we wish to estimate the average income per household in a large city. If we use simple random sampling, we will need a frame listing all households (elements) in the city, which would be difficult and costly to obtain. We cannot avoid this problem by using stratified random sampling because a frame is still required for each stratum in the population. Rather than draw a simple random sample of *elements,* we could divide the city into regions such as blocks (or clusters of elements) and select a simple random sample of blocks from the population. This task is easily accomplished by using a frame that lists all city blocks. Then the income of every household within each sampled block could be measured.

Cluster sampling is an effective design for obtaining a specified amount of information at minimum cost under the following conditions:

1. A good frame listing population elements either is not available or is very costly to obtain, while a frame listing clusters is easily obtained.

2. The cost of obtaining observations increases as the distance separating the elements increases.

Elements other than people are often sampled in clusters. An automobile forms a nice cluster of four tires for studies of tire wear and safety. A circuit board manufactured for a computer forms a cluster of semiconductors for testing. An orange tree forms a cluster of oranges for investigating an insect infestation. A plot in a forest contains a cluster of trees for estimating timber volume or proportions of diseased trees.

Notice the main difference between the optimal construction of strata and the construction of clusters. Strata are to be as homogeneous (alike) as possible within, but one stratum should differ as much as possible from another with respect to the characteristic being measured. Clusters, on the other hand, should be as heterogeneous (different) as possible within, and one cluster should look very much like another in order for the economic advantages of cluster sampling to pay off.

Estimation of a Population Mean and Total

Cluster sampling is simple random sampling with each sampling unit containing a collection or cluster of elements. Hence, the estimators of the population mean μ and total τ are similar to those for simple random sampling. In particular, the sample mean $\bar{y}$ is a good estimator of the population mean μ. The following notation is used in this section:

N = the number of clusters in the population

n = the number of clusters selected in a simple random sample

m_i = the number of elements in cluster i, $i = 1, \ldots, N$

$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i =$ the average cluster size for the sample

$M = \sum_{i=1}^{n} m_i$ = the number of elements in the population

$\bar{M} = \frac{M}{N}$ = the average cluster size for the population

y_i = the total of all observations in the ith cluster

y_{ij} = the measure for the jth element in the ith cluster

The estimator of the population mean μ is the sample mean $\bar{y}$, which is given by

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}.$$

Since both y_i and m_i are random variables, $\bar{y}$ is a ratio estimator, so the formulas developed earlier will apply. We simply replace x_i with m_i.

The estimated variance of $\bar{y}$ is

$$\hat{V}(\bar{y}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{\bar{M}^2}\right)\left(\frac{s_r^2}{n}\right)$$

where

$$s_r^2 = \frac{\sum_{i=1}^{n} \left(y_i - \bar{y} m_i \right)^2}{n-1}.$$

If $\bar{M}$ is unknown, it can be estimated by $\bar{m}$. This estimated variance is biased and will be a good estimate of $V(\bar{y})$ only if *n* is large. A rule of thumb is to require $n \geq 20$. The bias disappears if all m_i are equal.

Example 8.2 (Scheaffer, et al, page 294)

A city is to be divided into 415 clusters. Twenty-five of the clusters will be sampled, and interviews are conducted at every household in each of the 25 blocks sampled. The data on incomes are presented in the table below. Use the data to estimate the per-capita income in the city and place a bound on the error of estimation.

Cluster i	Number of Residents, m_i	Total income per cluster, y_i	Cluster i	Number of Residents, m_i	Total income per cluster, y_i
1	8	$96,000	14	10	$49,000
2	12	121,000	15	9	53,000
3	4	42,000	16	3	50,000
4	5	65,000	17	6	32,000
5	6	52,000	18	5	22,000
6	6	40,000	19	5	45,000
7	7	75,000	20	4	37,000
8	5	65,000	21	6	51,000
9	8	45,000	22	8	30,000
10	3	50000	23	7	39,000
11	2	85,000	24	3	47,000
12	6	43.000	25	8	41,000
13	5	54,000			

Here we have $\sum_{i=1}^{n} m_i = 151$, $\sum_{i=1}^{n} y_i = 1{,}329{,}000$, and $s_r = 25{,}189$.

Solution: The best estimate of the population mean μ is $\bar{y} = \frac{\$1{,}329{,}000}{151} = \$8801$. The estimate of per capita income is $8801. Since *M* is not known, $\bar{M}$ must be estimated by $\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n} = \frac{151}{25} = 6.04$. Since there were at total of 415 clusters, $N = 415$. So,

$$\hat{V}(\bar{y}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{\bar{M}^2}\right)\left(\frac{s_r^2}{n}\right) = \left(\frac{415-25}{415}\right)\left(\frac{1}{6.04^2}\right)\left(\frac{25189^2}{25}\right) = 653,785$$

Thus, the estimate of μ with a bound on the error of estimation is given by

$$\bar{y} \pm 2\sqrt{\hat{V}(\bar{y})} = 8801 \pm 2\sqrt{653,785} = 8801 \pm 1617$$

The best estimate of the average per-capita income is \$8801, and the error of estimation should be less than \$1617 with probability close to 0.95. This bound on the error of estimation is rather large; it could be reduced by sampling more clusters and, consequently, increasing the sample size.

Comparing Cluster Sampling and Stratified Sampling

It is advantageous to use a cluster sample when the individual clusters contain as much within cluster variability as possible, but the clusters themselves are as similar as possible. This can be seen in the computation of the variation,

$$s_r^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}m_i\right)^2}{n-1} = \frac{\sum_{i=1}^{n} m_i^2\left(\bar{y}_i - \bar{y}\right)^2}{n-1},$$

which will be small when the $\bar{y}_i$'s are similar in value. For cluster sampling, the differences are found within the clusters and the similarity between the clusters.

It is advantageous to use stratified sampling when elements within each strata are as similar as possible, but the strata themselves are as different as possible. Here, the differences are found between the strata and the similarity within the strata. Two examples will help illustrate this distinction.

Example 1 Suppose you want to take a sample of a large high school and you must use classes to accomplish your sampling. In this school, students are randomly assigned to homerooms, so each homeroom has a mixture of students from all grade-levels (Freshman-Senior). Also, in this school, the study halls are grade-level specific, so all of the students in a large study hall are from the same grade. If you believe that students in the different grade-levels will have different responses, you want to be assured that each grade-level is represented in the sample.

You could perform a cluster sample by selecting *n* homerooms at random and surveying everyone in those homerooms. You would not use the homerooms as strata, since there would be no advantage over a simple random sample.

You could perform a stratified sample using study halls as your strata. Randomly select *k* students from study halls for each grade-level. Study halls would make a poor cluster, since the responses from all of the students are expected to be similar.

Example 2 We would like to estimate the number of diseased trees in the forest represented below. The diseased trees are indicated with a D, while the trees free of disease are represented by F. Consider the rows and columns of the grid.

(a) If a cluster sample is used, should the rows or columns be used as a cluster?

(b) If a stratified sample is used, should the rows or columns be used as strata?

Row	C1	C2	C3	C4	C5
1	F	F	F	D	D
2	F	F	D	D	D
3	F	F	F	F	F
4	F	F	D	F	D
5	F	F	F	F	D
6	D	F	D	F	F
7	F	F	D	F	D
8	F	D	D	F	D
9	F	F	F	D	D
10	F	F	F	D	D
11	F	F	F	D	F
12	F	D	D	D	D
13	F	D	F	D	D
14	F	F	F	D	D
15	F	D	F	D	D
16	F	F	D	D	D
17	F	F	D	D	D
18	F	F	F	D	D
19	F	F	D	D	D
20	F	F	F	F	F
21	D	F	F	D	F
22	F	D	F	F	D
23	F	F	D	D	F
24	F	F	F	D	D
25	F	F	F	D	D
26	F	D	F	F	D
27	F	F	D	F	D
28	D	F	F	F	D
29	F	F	F	F	D
30	F	F	D	D	D

It appears that there are more diseased trees in the right-most columns, however, there does not appear to be a difference among the rows. If we wanted a sample of size 25, we could obviously select a simple random sample, but we might miss the concentration of diseased trees in C4 and C5 just by chance. We want to insure that C4 and C5 show up in the sample. We have two choices:

- For a cluster sample, we should use the rows as clusters. We could select 5 rows at random, and consider every tree in each of those clusters (rows).
- For a stratified sample, we could use the columns as strata. We would select 5 elements from each of the 5 strata (columns) to consider.

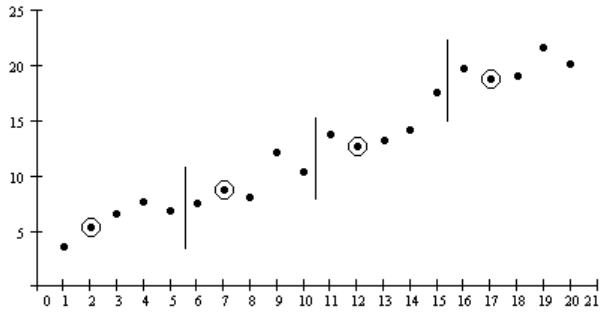
Systematic Sampling

Suppose the population elements are on a list or come to the investigator sequentially. It is convenient to find a starting point near the beginning of the list and then sample every k^{th} element thereafter. If the starting point is random, this is called a 1-in-*k* systematic sample.

If the population elements are in random order, systematic sampling is equivalent to simple random sampling. If the population elements have trends or periodicities, systematic sampling may be better or worse than simple random sampling depending on how information on population structure is used. Many estimators of variance have been proposed to handle various population structures.

Repeated Systematic Sampling

In the 1-in-*k* systematic sample, there is only one randomization, which limits the analysis. The randomness in the systematic sample can be improved by choosing more than one random start. For example, instead of selecting a random number between 1 and 4 to start and then picking every 4^{th} element, you could select



2 numbers at random between 1 and 8, and then selecting those elements in each group of 8.

Relationship to Stratified and Cluster Sampling

Recall that if the elements are in random order, we have no problem with systematic sampling. If there is some structure to the data, as shown below, we can compare systematic sampling to stratified and cluster samples.

Systematic sampling is closely related to

- stratified sampling with one sample element per stratum
- cluster sampling with the sample consisting of a single cluster

As a stratified sample, we think of having 4 different strata, each with 5 elements. The elements of the strata are similar and the means of the strata are different, so this fits the requirements for a stratified sample. We take one element from each stratum (in this illustration, the second in each stratum). We have lost some randomness, since the second item is taken from all strata rather than a random element from each stratum.

As a cluster sample, we think of the 5 possible clusters. Cluster 1 contains all of the first elements, cluster 2 (the one selected) contains all the second elements, etc. Here we have surveyed all elements in one cluster (cluster 2). In this case, the clusters contain as much variation as possible with similar means, so the cluster process is appropriate. Since we have only one cluster, we have no estimate of the variance. A repeated systematic sample (taking clusters 2 and 5, for example) would eliminate this difficulty.

If the structure of the data is periodic, it is important that the systematic sample not mimic the periodic behavior. In the diagram below, the circles begin at the 3^{rd} element and select every 8^{th} element. Since this matches closely the period of data, we select only values in the upper range. If we begin at the 3^{rd} element and select every 5^{th} element, we are able to capture data across the full range.