UNIT-I

BASIC CONCEPT OF SAMPLE SURVEYS

A sample survey is a method of drawing an inference about the characteristics of a population or universe by observing a part of the population. For example, when one has to make an inference about a large lot and is not practicable to examine each individual member of the lot, one always takes help of sample surveys, that is to say one examines only a few member of the lot and, on the basis of this sample information, one makes decisions about the whole lot. Thus, a person wanting to purchase a basket of oranges may examine a few oranges from the basket and on that basis make his decision about the whole basket.

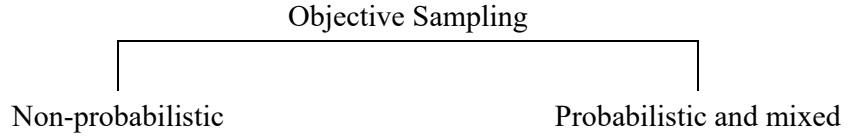
Such methods are extensively used by government bodies throughout the world for assessing, different characteristics of national economy as are required for taking decisions regarding the impositions of taxes, fixation of prices and minimum wages etc. and for planning and projection of future economic structure, for estimation of yield rates and acreages under different crops, number of unemployed persons in the labour forces, construction of cost of living indices for persons in different professions and so on.

Sample survey techniques are extensively used in market research surveys for assessing the preferential pattern of consumers for different types of products, the potential demand for a new product which a company wishes to introduce, scope for any diversification in the production schedule, and so on.

Thus, sampling may become unavoidable because we may have limited resources in terms of money and / or man hours, or it may be preferred because of practical convenience.

Sampling is first broadly classified as **Subjective** and **Objective**.

Any type of sampling which depends upon the personal judgment or discretion of the sampler himself is called **Subjective**. But the sampling method which is fixed by a sampling rule or is independent of the sampler's own judgment is **Objective sampling**.



In non-probabilistic objective sampling, there is a fixed sampling rule but there is no probability attached to the mode of selection, e.g. selecting every $5-$th individual from a list. If, however, the selection of the first individual is made in such a manner that each of the first 10 gets an equal chance of being selected, it becomes a case of mixed sampling, if for each individual there is a definite pre-assigned probability of being selected, the sampling is said to be probabilistic.

Elementary unit or simply unit: It is an element or a group of elements, on which observations can be made or from which the required statistical information can be ascertained according to a well defined procedure, examples of unit are person, family, household, farm, factory, tree, a period of time such as an hour, day etc.

Population: The collection of all units of a specified type in a given region at a particular point or a period of time is termed as a population or inverse. For example, a population of persons, families, farms, cattle, houses or automobiles in a region or a population of trees or a birds in a forest etc.

A population is said to be finite population or an infinite population according to as the number of units in it is finite or infinite.

Sampling units: Elementary units or groups of such units, which, besides being clearly defined, identifiable and observable, are convenient for purposes of sampling, are called sampling units. For example, in a family budget enquiry, usually a family is considered as a sampling unit, since it is formed to be convenient for sampling for ascertaining the required information. In a crop survey, a farm or a group of farms owned or operated by a household may be considered as the sampling units.

Sampling frame: For using sampling methods in the collection of data, it is essential to have a frame of all the sampling units belonging to the population to be studied with their proper identification particulars and such a frame is called the sampling frame. This may be a list of units with their identification particulars.

As the sampling frame forms the basic material from which a sample is drawn, it should be insured that the frame contains all the sampling units of the population under consideration but excludes units of any other population.

Sample: A sample is a subset of a population selected to obtain information concerning the characteristics of the population. In other words, one or more sampling units selected from a population according to some specified procedure are said to constitute a sample.

Random sample: A random or probability sample is a sample drawn in such a manner that each unit in the population has a predetermined probability of selection.

Estimator: An estimator is a statistic obtained by a specified procedure for estimating a population parameter. The estimator is a random variable, as its value differs from sample to sample and the samples are selected with specified probabilities.

The particular value, which the estimator takes for a given sample, is known as an estimate.

The difference between the estimator (t) and the parameter (θ) is called **error**.

An estimator (t) is said to be unbiased estimator for the parameter (θ) if, $E(t) = \theta$, otherwise biased. Thus bias is given by

$$E(t - \theta) = B(t)$$

The mean of squares of error taken from θ is called **mean square error** (MSE). Mathematically it is defined as

$$MSE(t) = E(t - \theta)^2 .$$

The MSE may be considered to be a measure of **accuracy** with which the estimator t estimates the parameter θ.

The expected value of the squared deviation of the estimator from its expected value is termed **sampling variance**. It is a measure of the divergence of the estimator from its expected value and is given by

$$V(t) = E[t - E(t)]^2 .$$

This measure of variability may be termed the **precision of the estimator** t.

The relation between MSE and sampling variance or between accuracy and precision can be obtained as

$$MSE(t) = E(t - \theta)^2 = E[t - E(t) + E(t) - \theta]^2$$

$$= E[t - E(t)]^2 + [E(t) - \theta]^2 = V(t) + [B(t)]^2 \text{, since } E[t - E(t)] = 0 .$$

This shows that MSE of t is the sum of the sampling variance and the square of the bias. However, if t is an unbiased estimator of θ, the MSE and sampling variance are the same.

The square root of the sampling variance is termed as **standard error** of the estimator t.

The ratio of the standard error of the estimator to the expected value of the estimator is known as **relative standard error** or the coefficient of variation of the estimator.

Sample space: The collection of all possible sample, sequence, sets is called the sample space.

Sampling design: The combination of the sample space and the associated probability measure is called a sampling design. For example, let $N = 4$, $n = 2$ and the probability of selection for different samples is

Sample	(1, 2)	(1, 3)	(1, 4)	(2, 3)	(2, 4)	(3, 4)
Probability	1/6	1/6	1/6	1/6	1/6	1/6

The above table gives the sampling design.

Sampling and complete enumeration

The total count of all units of the population for a certain characteristics is known as complete enumeration, also termed census survey. The money, man-power and time required for carrying out complete enumeration will generally be large and there are many situations with limited means where complete enumeration will not be possible, where recourse to selection of a few units will be helpful. When only a part, called sample, is selected from the population and examine, it is called sample enumeration or sample survey.

A sample survey will usually be less expensive then a census survey and the desired information will obtain in less time. This does not imply that economy is the only consideration in conducting a sample survey. It is most important that a degree of accuracy of results is also maintained. Occasionally, the technique of sample survey is applied to verify that the results obtained from the census surveys. The main advantages or merits of sample survey over census survey may be outlined as follows:

i) Reduced cost of survey,

ii) Greater speed of getting results,

iii) Greater accuracy of results,

iv) Greater scope, and

v) Adaptability

Sample survey has its own limitations and the advantages of sampling over complete enumeration can be derived only if

i) the units are drawn in a scientific manner

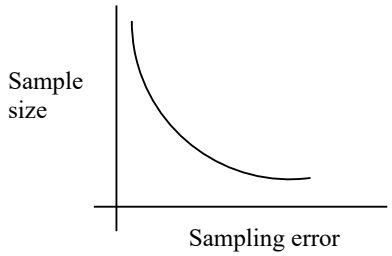
ii) an appropriate sampling technique is used, and

iii) the size of units selected in the sample is adequate.

Sampling and non-sampling errors

The error which arises due to only a sample (a part of population) being used to estimate the population parameters and draw inferences about the population is termed **sampling error** or **sampling fluctuation**. Whatever may be the degree of cautiousness in selecting a sample;

there will always be a difference between the parameter and its corresponding estimate. This error is inherent and unavoidable in any and every sampling scheme. A sample with the smallest sampling error will always be considered a good representative of the population. This error can be reduced by increasing the size of the sample (number of units selected in the sample). In fact, the decrease in sampling error is inversely proportional to the square root of the sample size and the relationship can be examined graphically as below:



When the sample survey becomes a census survey, the sampling error becomes zero.

Non-sampling error

The non-sampling errors primarily arise at the following stages:

i) Failure to measure some of units in the selected sample

i) Observational errors due to defective measurement technique

iii) Errors introduced in editing, coding and tabulating the results.

Non-sampling errors are present in both the complete enumeration survey and the sample survey. In practice, the census survey results may suffer from non-sampling errors although these may be free from sampling error. The non-sampling error is likely to increase with increase in sample size, while sampling error decreases with increase in sample size.

SIMPLE RANDOM SAMPLING

A procedure for selecting a sample of size n out of a finite population of size N in which each of the possible distinct samples has an equal chance of being selected is called **random sampling or simple random sampling.**

We may have two distinct types of simple random sampling as follows:

i) Simple random sampling with replacement $(srswr)$.

ii) Simple random sampling without replacement $(srswor)$.

Simple random sampling with replacement $(srswr)$

In sampling with replacement a unit is selected from the population consisting of N units, its content noted and then returned to the population before the next draw is made, and the process is repeated n times to give a sample of n units. In this method, at each draw, each of the N units of the population gets the same probability $\frac{1}{N}$ of being selected. Here the same unit of the population may occur more than once in the sample (order in which the sample

units are obtained is regarded). There are N^n samples, and each has an equal probability $\frac{1}{N^n}$ of being selected.

Note: If order in which the sample units are obtained is ignored (unordered), then in such case the number of possible samples will be

$${}^{N}C_n + N\,(1+{}^{N-1}C_1+{}^{N-1}C_2+\cdots+{}^{N-1}C_{n-2})\,.$$

Simple random sampling without replacement (*srswor*)

Suppose the population consist of N units, then, in simple random sampling without replacement a unit is selected, its content noted and the unit is not returned to the population before next draw is made. The process is repeated n times to give a sample of n units. In this method at the $r-$th drawing, each of the $N-r+1$ units of the population gets the same probability $\frac{1}{N-r+1}$ of being included in the sample. Here any unit of the population cannot occur more than once in the sample (order is ignored). There are ${}^{N}C_n$ possible samples, and each such sample has an equal probability $\frac{1}{{}^{N}C_n}$ of being selected.

Theory of simple random sampling with replacement

N, population size.

n, sample size.

Y_i, value of the $i-$th unit of the population.

y_i, value of the $i-$th unit of the sample.

$Y=\sum_{i=1}^{N}Y_i$, population total.

$\bar{Y}=\frac{1}{N}\sum_{i=1}^{N}Y_i$, population mean.

$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$, sample mean.

$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(Y_i-\bar{Y})^2=\frac{1}{N}\sum_{i=1}^{N}Y_i^2-\bar{Y}^2$, population variance.

$S^2=\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2=\frac{1}{N-1}\left(\sum_{i=1}^{N}Y_i^2-N\bar{Y}^2\right)$, population mean square.

$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}y_i^2-n\bar{y}^2\right)$, sample mean square.

Theorem: In *srswr*, the sample mean $\bar{y}$ is an unbiased estimate of the population mean $\bar{Y}$ i.e. $E(\bar{y}) = \bar{Y}$ and its variance $V(\bar{y}) = \frac{N-1}{nN} S^2 = \frac{\sigma^2}{n}$.

Corollary: $\hat{Y} = N\bar{y}$ is an unbiased estimate of the population total Y with its variance $V(\hat{Y}) = \frac{N^2\sigma^2}{n} = \frac{N(N-1)}{n} S^2$.

Theorem: In *srswr*, sample mean square s^2 is an unbiased estimate of the population variance σ^2 i.e. $E(s^2) = \sigma^2$.

Theory of simple random sampling without replacement

Theorem: In *srswor*, sample mean $\bar{y}$ is an unbiased estimate of the population mean $\bar{Y}$ i.e. $E(\bar{y}) = \bar{Y}$ and its variance is $V(\bar{y}) = \left(\frac{N-n}{nN}\right) S^2$.

Corollary: $\hat{Y} = N\bar{y}$ is an unbiased estimate of the population total Y with its variance $V(\hat{Y}) = N^2(1-f)S^2/n$.

Theorem: In *srswor*, sample mean square s^2 is an unbiased estimate of the population mean square S^2 i.e. $E(s^2) = S^2$.

Property: $V(\bar{y})$ under *srswor* is less than the $V(\bar{y})$ under *srswr*.

Theorem: Let *srswor* sample of size n is drawn from a population of size N. Let $T = \sum_{i=1}^{n} \alpha_i y_i$ is a class of linear estimator of $\bar{Y}$, where α_i's are coefficient attached to sample values, then,

i) The class T is linear unbiased estimate class if $\sum_{i=1}^{n} \alpha_i = 1$.

ii) The sample mean $\bar{y}$ is the best linear unbiased estimate.

Proof:

i) $E(T) = E\left(\sum_{i=1}^{n} \alpha_i y_i\right) = \sum_{i=1}^{n} \alpha_i E(y_i) = \sum_{i=1}^{n} \alpha_i \bar{Y} = \bar{Y}$, iff $\sum_{i=1}^{n} \alpha_i = 1$.

ii) $V(T) = E\left(\sum_{i=1}^{n} \alpha_i y_i - \bar{Y}\right)^2$, under $\sum_{i=1}^{n} \alpha_i = 1$.

$$= E\left[\left(\sum_{i=1}^{n} \alpha_i y_i\right)^2 - 2\bar{Y}\left(\sum_{i=1}^{n} \alpha_i y_i\right) + \bar{Y}^2\right] = E\left(\sum_{i=1}^{n} \alpha_i y_i\right)^2 - \bar{Y}^2.$$

Consider,

$$E\left(\sum_{i=1}^{n}\alpha_i y_i\right)^2 = \sum_{i=1}^{n}\alpha_i^2\, E(y_i^2) + \sum_{i\neq j}^{n}\alpha_i\alpha_j\, E(y_i y_j) \tag{1.1}$$

Note that

$$V(y_i) = E(y_i^2) - \bar{Y}^2$$

$$\Rightarrow\ E(y_i^2) = \frac{1}{N}(N-1)S^2 + \bar{Y}^2, \text{ since } V(y_i) = \frac{N-1}{N}S^2 \text{ for each } i. \tag{1.2}$$

Now

$$E(y_i y_j) = \sum_{i\neq j}^{N} y_i \Pr(i)\, y_j \Pr(j|i) = \frac{1}{N}\frac{1}{N-1}\sum_{i\neq j}^{N} y_i\, y_j .$$

Note that

$$\left(\sum_{i=1}^{N} y_i\right)^2 = \sum_{i=1}^{N} y_i^2 + \sum_{i\neq j}^{N} y_i y_j = (N-1)S^2 + N\bar{Y}^2 + \sum_{i\neq j}^{N} y_i y_j$$

$$\Rightarrow \sum_{i\neq j}^{N} y_i y_j = N^2\bar{Y}^2 - (N-1)S^2 - N\bar{Y}^2 .$$

Thus

$$E(y_i y_j) = \frac{1}{N}\frac{1}{N-1}[N^2\bar{Y}^2 - (N-1)S^2 - N\bar{Y}^2] = \bar{Y}^2 - S^2/N . \tag{1.3}$$

In view of equations (1.2) and (1.3), equation (1.1) becomes

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

$$= S^2\sum_{i=1}^{n}\alpha_i^2 - \frac{S^2}{N}\sum_{i=1}^{n}\alpha_i^2 + \bar{Y}^2\sum_{i=1}^{n}\alpha_i^2 + \left(1 - \sum_{i=1}^{n}\alpha_i^2\right)\left(\bar{Y}^2 - \frac{S^2}{N}\right)$$

$$= S^2\sum_{i=1}^{n}\alpha_i^2 + \bar{Y}^2 - \frac{S^2}{N}.$$

Therefore,

$$V(T) = S^2\sum_{i=1}^{n}\alpha_i^2 - \frac{S^2}{N}.$$

Since $\sum_{i=1}^{n}\alpha_i^2 = \sum_{i=1}^{n}\left(\alpha_i - \frac{1}{n}\right)^2 + \frac{1}{n}$, under condition $\sum_{i=1}^{n}\alpha_i = 1$, so that

$$V(T) = S^2\left[\sum_{i=1}^{n}\left(\alpha_i - \frac{1}{n}\right)^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\right].$$

We note that $V(T)$ will be minimum, if $\sum_{i=1}^{n}\left(\alpha_i - \frac{1}{n}\right)^2 = 0$, where $\alpha_i = \frac{1}{n}$, for all $i = 1, 2, \cdots, n$, and $T = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$.

OR

To determine α_i such that $V(T)$ is minimum, consider the function

$\phi = V(T) + \lambda\left(\sum_{i=1}^{n}\alpha_i - 1\right)$, where λ is some unknown constant.

Using the calculus method of Lagrange multipliers, we select α_i and the constant λ to minimize ϕ. Differentiating ϕ with respect to α_i and equating to zero, we have

$$\frac{\partial \phi}{\partial \alpha_i} = 0 = 2S^2\alpha_i + \lambda \qquad \text{or} \qquad \alpha_i = -\frac{\lambda}{2S^2} \tag{1.4}$$

Taking summation on both the sides of (1.4), we get

$$\sum_{i=1}^{n}\alpha_i = -\frac{n\lambda}{2S^2} \qquad \Rightarrow \qquad \lambda = -\frac{2S^2}{n} \tag{1.5}$$

Thus, from equations (4) and (5), we have

$\alpha_i = \frac{1}{n}$, for all $i = 1, 2, \cdots, n$, and $T = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$.

Case I) Random sampling with replacement

On replacing $\bar{Y}$ by P, Y by NP, $\bar{y}$ by $p = \frac{a}{n}$, S^2 by $\frac{NPQ}{N-1}$ and σ^2 by PQ in the expressions obtained in expectation and variance of the estimates of population mean and population total, we find

i) $E(p) = E(\bar{y}) = \bar{Y} = P$. This shows that sample proportion p is an unbiased estimate of population proportion P and $V(p) = V(\bar{y}) = \frac{\sigma^2}{n} = \frac{PQ}{n}$.

ii) $E(\hat{A}) = E(Np) = N\,E(p) = NP = A$, means that $Np = \hat{A}$ is an unbiased estimate of $NP = A$ and

$$V(\hat{A}) = V(\hat{Y}) = N^2 V(\bar{y}) = \frac{N^2\sigma^2}{n} = \frac{N^2 PQ}{n}.$$

Theorem: $\hat{V}(p) = v(p) = \frac{pq}{n-1}$ is an unbiased estimate of $V(p) = \frac{PQ}{n}$.

Case II) Random sampling without replacement

Results are:

i) $E(p) = E(\bar{y}) = \bar{Y} = P$. This shows that sample proportion p is an unbiased estimate of population proportion P and $V(p) = V(\bar{y}) = \frac{N-n}{nN} S^2 = \left(\frac{N-n}{nN}\right)\frac{NPQ}{N-1} = \left(\frac{N-n}{N-1}\right)\frac{PQ}{n}$.

ii) $E(\hat{A}) = E(Np) = N\,E(p) = NP = A$, means that Np is an unbiased estimate of NP and

$$V(\hat{A}) = V(\hat{Y}) = N^2 V(\bar{y}) = N^2\left(\frac{N-n}{nN}\right)S^2 = N^2\left(\frac{N-n}{nN}\right)\frac{NPQ}{N-1} = N^2\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}.$$

Theorem: $\hat{V}(p) = v(p) = \left(\frac{N-n}{n-1}\right)\frac{pq}{N}$ is an unbiased estimate of $V(p) = \left(\frac{N-n}{N-1}\right)\frac{PQ}{n}$.

Corollary: $\hat{V}(\hat{A}) = \hat{V}(Np) = N^2\,\hat{V}(p) = N\left(\frac{N-n}{n-1}\right)pq$ is an unbiased estimate of $V(\hat{A}) = N^2\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}$.

Example: For a population of size $N = 430$ roughly we know that $\bar{Y} = 19$, $S^2 = 85.6$ with srs, what should be the size of sample to estimate $\hat{\bar{Y}}$ with a margin of error 10% of $\bar{Y}$ apart chance is 1 in 20.

Solution: Margin of error in the estimate $\bar{y}$ of $\bar{Y}$ is given, i.e.

$\bar{y} = \bar{Y} \pm 10\%$ of $\bar{Y}$ or $|\bar{y} - \bar{Y}| = 10\%$ of $\bar{Y} = \frac{19}{10} = 1.9$, so that

$$\Pr[|\bar{y} - \bar{Y}| \geq 1.9] = \frac{1}{20} = 0.05, \text{ and } n_0 = \frac{Z_{\alpha/2}^2 S^2}{d^2} = \frac{(1.96)^2 \times 85.6}{(1.9)^2} = 91.091678.$$

Therefore,

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = 75.168 \cong 75.$$

Example: In the population of 676 petition sheets. How large must the sample be if the total number of signatures is to be estimated with a margin of error of 1000, apart from a 1 in 20 chance? Assume that the population mean square to be 229.

Solution: Let Y be the number of signature on all the sheets. Let $\hat{Y}$ is the estimate of Y. Margin of error is specified in the estimate $\hat{Y}$ of Y as

$$|\hat{Y} - Y| = 1000, \text{ so that, } \Pr[|\hat{Y} - Y| \geq 1000] = \frac{1}{20} = 0.05.$$

We know that

$$n = \frac{n_0}{1 + \frac{n_0}{N}}, \text{ here, } n_0 = \left(\frac{N Z_{\alpha/2} S}{d'}\right)^2 = \left(\frac{676 \times 1.96}{1000}\right)^2 229 = 402.01385$$

and hence

$n = 252.09 \cong 252$.

Estimation of sample size for proportion

a) **When precision is specified in terms of margin of error:** Suppose size of the population is N and population proportion is P. Let a *srs* of size n is taken and p be the corresponding sample proportion and d is the margin of error in the estimate p of P. The margin of error can be specified in the form of probability statement as

$$\Pr[|\, p - P \,| \geq d] = \alpha \quad \text{or} \qquad \Pr[|\, p - P \,| \leq d] = 1 - \alpha \tag{1.6}$$

As the population is normally distributed, so $\bar{y} \sim N[P, V(p)]$, then $Z = \dfrac{p - P}{\sqrt{V(p)}} \sim N(0,1)$. For the given value of α we can find a value Z_α of the standard normal variate from the standard normal table by the following relation:

$$\Pr\left[\frac{|\, p - P \,|}{\sqrt{V(p)}} \geq Z_{\alpha/2}\right] = \alpha \qquad \text{or} \qquad \Pr[|\, p - P \,| \geq \sqrt{V(p)}\, Z_{\alpha/2}] = \alpha \tag{1.7}$$

Comparing equation (1.6) and (1.7), the relation which gives the value of n with the required precision of the estimate p of P is given by

$d = Z_{\alpha/2}\sqrt{V(p)}$ or $d^2 = Z^2_{\alpha/2} V(p) = Z^2_{\alpha/2}\left(\dfrac{N-n}{N-1}\right)\dfrac{PQ}{n}$, as sampling is *srswr*.

$$\Rightarrow 1 = \frac{Z^2_{\alpha/2}\, PQ}{d^2}\left(\frac{N-n}{n(N-1)}\right) = n_0 \frac{N-n}{n(N-1)}, \text{ where } n_0 = \frac{Z^2_{\alpha/2}\, PQ}{d^2} = \frac{PQ}{V(p)} \tag{1.8}$$

$$\text{or} \quad \frac{N-1}{n_0} = \frac{N-n}{n} = \frac{N}{n} - 1 \qquad \Rightarrow \qquad \frac{N}{n} = 1 + \frac{N-1}{n_0}$$

$$\text{or} \quad n = \frac{N}{1 + \dfrac{N-1}{n_0}} = \frac{N\, n_0}{n_0 + (N-1)} = \frac{n_0}{\dfrac{n_0}{N} + \dfrac{N-1}{N}} = \frac{n_0}{1 + \dfrac{n_0}{N}} \tag{1.9}$$

If N is sufficiently large, then $n \cong n_0$

b) If precision is specified in terms of $V(p)$ i.e. $V(p) = V$ (given).

Substituting $V(p) = V$ in relation (1.16) we get, $n_0 = \dfrac{PQ}{V}$, and hence n can be obtained by relation (1.17).

c) When precision is given in terms of coefficient of variation of p

Let

$$CV(p) = e = \frac{\sqrt{V(p)}}{P} \quad \Rightarrow \qquad \frac{V(p)}{P^2} = e^2, \quad \text{or} \qquad V(p) = e^2\, P^2 \tag{1.18}$$

Substitute equation (1.18) in relation (1.16), we get,

$n_0 = \frac{PQ}{e^2 P^2} = \frac{Q}{e^2 P} = \frac{1}{e^2}\left(\frac{1}{P} - 1\right)$, and hence n is given by the relation (1.9).

Example: In a population of 4000 people who were called for casting their votes, 50% returned to the poll. Estimate the sample size to estimate this proportion so that the marginal error is 5% with 95% confidence coefficient.

Solution: Margin of error in the estimate p of P is given by

$| p - P | = 0.05$, then $\Pr[| p - P | \geq 0.05] = 0.05$.

We know that

$$n_0 = \frac{Z_{\alpha/2}^2 PQ}{d^2} = \frac{(1.96)^2 \times 0.5 \times 0.5}{0.0025} = 384.16 \cong 384$$

and hence,

$$n = \frac{n_0}{1 + (n_0 / N)} = 350.498 \cong 351.$$

Exercise: In a study of the possible use of sampling to cut down the work in taking inventory in a stock room, a count is made of the value of the articles on each of 36 shelves in the room. The values to the nearest dollar are as follows.

29, 38, 42, 44, 45, 47, 51, 53, 53, 54, 56, 56, 56, 58, 58, 59, 60, 60, 60, 60, 61, 61, 61, 62, 64, 65, 65, 67, 67, 68, 69, 71, 74, 77, 82, 85.

The estimate of total value made from a sample is to be correct within \$200, apart from a 1 in 20 chance. An advisor suggests that a simple random sample of 12 shelves will meet the requirements. Do you agree? $\sum Y_i = 2138$, and $\sum Y_i^2 = 131\,682$.

Solution: It is given that

$\sum_i Y_i = 2138$, $\sum_i Y_i^2 = 131\,682$, and $N = 36$, then

$$S^2 = \frac{1}{N-1}\left[\sum_i Y_i^2 - N\bar{Y}^2\right] = \frac{1}{36-1}\left[131\,682 - 36\left(\frac{2138}{36}\right)^2\right] = 134.5$$

and

$| \hat{Y} - Y | \leq 200$, then, $\Pr[| \hat{Y} - Y | \leq 200] = \frac{1}{20} = 0.05$.

We know that

$$n = \frac{n_0}{1 + \frac{n_0}{N}}, \text{ here } \quad n_0 = \left(\frac{N Z_{\alpha/2}}{d}\right)^2 S = \left(\frac{36 \times 1.96}{200}\right)^2 134.5 = 16.7409$$

and therefore,

$n = 11.42765 \cong 12$.

Exercise: The selling price of a lot of standing timber is UW, where U is the price per unit volume and W is the volume of timber on the lot. The number N of logs on the lot is counted, and the average volume per log is estimated from a simple random sample of n

logs. The estimate is made and paid for by the seller and is provisionally accepted by the buyer. Later, the buyer finds out the exact volume purchased, and the seller reimburses him if he has paid for more than was delivered. If he has paid for less than was delivered, the buyer does not mention the fact.

Construct the seller's loss function. Assuming that the cost of measuring n logs is cn, find the optimum value of n. The standard deviation of the volume per log may be denoted by S and the fpc ignored.

Solution: Let $\hat{W}$ be the estimated total volume of the timber. The error in the estimate is $\hat{W} - W$.

If $\hat{W} - W = z > 0$ sellers loss is zero, i.e. $l(z) = 0$.

If $\hat{W} - W = z < 0$ sellers loss is $-Uz$, i.e. $l(z) = -Uz$.

When fpc is ignored $V(\hat{W}) = N^2 S^2 / n$, then

$$\hat{W} \sim N\left(W, \frac{N^2 S^2}{n}\right), \qquad \text{or} \qquad z = (\hat{W} - W) \sim N\left(0, \frac{N S}{\sqrt{n}}\right), \text{ so that}$$

$$f(z) = \frac{1}{(NS/\sqrt{n})\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z}{NS/\sqrt{n}}\right)^2\right] = \frac{1}{(NS/\sqrt{n})\sqrt{2\pi}} \exp\left(-\frac{n z^2}{2 N^2 S^2}\right)$$

Thus, the expected loss

$$L(n) = \int_{-\infty}^{\infty} l(z)\, f(z)\, dz = \int_{-\infty}^{0} (-Uz) \frac{1}{(NS/\sqrt{n})\sqrt{2\pi}} \exp\left(-\frac{n z^2}{2 N^2 S^2}\right) dz$$

$$= -\int_{-\infty}^{0} Uz \frac{1}{(NS/\sqrt{n})\sqrt{2\pi}} \exp\left(-\frac{n z^2}{2 N^2 S^2}\right) dz$$

$$= \int_{0}^{\infty} Uz \frac{1}{(NS/\sqrt{n})\sqrt{2\pi}} \exp\left(-\frac{n z^2}{2 N^2 S^2}\right) dz$$

Put $\frac{n z^2}{2 N^2 S^2} = t$, then $\frac{2 n z}{2 N^2 S^2} dz = dt$ or $z\, dz = \frac{N^2 S^2}{n} dt$.

Therefore,

$$L(n) = \int_0^{\infty} \frac{U N^2 S^2}{n} \frac{1}{(NS/\sqrt{n})\sqrt{2\pi}} e^{-t}\, dt = \frac{UNS}{\sqrt{2n\pi}} \int_0^{\infty} e^{-t}\, dt = \frac{UNS}{\sqrt{2n\pi}}, \text{ as } \int_0^{\infty} e^{-t}\, dt = 1.$$

To determine the value of n, consider the function

$$\phi(n) = L(n) + C(n) = c\, n + \frac{UNS}{\sqrt{2\pi}} n^{-1/2}.$$

Differentiate this function with respect to n, we get

$$\frac{\partial \phi}{\partial n} = 0 = c - \frac{1}{2}\left(\frac{UNS}{\sqrt{2\pi}}\right) n^{-3/2} \qquad \text{or} \qquad \frac{UNS}{2\sqrt{2\pi}} n^{-3/2} = c$$

or $n^{-3/2} = \dfrac{2c\sqrt{2\pi}}{UNS}$ or $n = \left(\dfrac{UNS}{2c\sqrt{2\pi}}\right)^{2/3}$.

Exercise: With certain populations, it is known that the observations Y_i are all zero on a portion QN of N units $(0 < Q < 1)$. Sometimes with varying expenditure of efforts, these units can be found and listed, so that they need not be sampled. If σ^2 is the variance of Y_i in the original population and σ_0^2 is the variance when all zeros are excluded, then show that $\sigma_0^2 = \dfrac{\sigma^2}{P} - \dfrac{Q}{P^2}\bar{Y}^2$, where $P = 1 - Q$, and $\bar{Y}$ is the mean value of Y_i for the whole population.

Solution: Given $Y_1, Y_2, \cdots, Y_{NP}, Y_{NP+1}, \cdots, Y_N$ (first NP units not zero, and rest NQ units which are all zero). Thus, $\bar{Y} = \dfrac{1}{N}\sum_{i=1}^{N} Y_i$, population mean, and $\bar{Y}_{NP} = \dfrac{1}{NP}\sum_{i=1}^{NP} Y_i$, $\bar{Y}_{NQ} = \dfrac{1}{NQ}\sum_{i=1}^{NQ} Y_i = 0$, also, $\sum_{i=1}^{N} Y_i = \sum_{i=1}^{NP} Y_i$, and $\sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{NP} Y_i^2$, so that $N\bar{Y} = NP\,\bar{Y}_{NP}$,

or $\bar{Y}_{NP} = \dfrac{1}{P}\bar{Y}$. By definition,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \bar{Y}^2, \quad \text{or} \quad N\sigma^2 = \sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2.$$

Similarly, $NP\sigma_0^2 = \sum_{i=1}^{NP} Y_i^2 - NP\bar{Y}_{NP}^2$.

Thus,

$$N(\sigma^2 - P\sigma_0^2) = NP\bar{Y}_{NP}^2 - N\bar{Y}^2 = NP\frac{1}{P^2}\bar{Y}^2 - N\bar{Y}^2 = N\left(\frac{1}{P} - 1\right)\bar{Y}^2 = N\left(\frac{Q}{P}\right)\bar{Y}^2.$$

Therefore,

$$P\sigma_o^2 = \sigma^2 - \left(\frac{Q}{P}\right)\bar{Y}^2 \quad \text{or} \quad \sigma_o^2 = \frac{\sigma^2}{P} - \frac{Q}{P^2}\bar{Y}^2.$$

Exercise: From a random sample of n units, a random sub-sample of n_1 units is drawn without replacement and added to the original sample. Show that the mean based on $(n + n_1)$ units is an unbiased estimator of the population mean, and that ratio of its variance to that of the mean of the original n units is approximately $\dfrac{1 + 3n_1/n}{(1 + n_1/n)^2}$, assuming that the population size is large.

Solution: Let the sample mean based on n, n_1, and $n + n_1$ elements are denoted by $\bar{y}_n$, $\bar{y}_{n_1}$, and $\bar{y}_{n+n_1}$ respectively, and are defined as $\bar{y}_n = \dfrac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{y}_{n_1} = \dfrac{1}{n_1}\sum_{i=1}^{n_1} y_i$, and

$\bar{y}_{n+n_1} = \dfrac{n\,\bar{y}_n + n_1\,\bar{y}_{n_1}}{n+n_1}$. We have to show $E(\bar{y}_{n+n_1}) = \bar{Y}$, in this case the expectation is taken in two stages,

i) when n is fixed

ii) over all expectation

$$E(\bar{y}_{n+n_1}) = \frac{1}{n+n_1} E(n\,\bar{y}_n + n_1\,\bar{y}_{n_1}) = \frac{1}{n+n_1} E[n\,\bar{y}_n + n_1\,E(\bar{y}_{n_1} \mid n)]$$

$$= \frac{1}{n+n_1} E(n\,\bar{y}_n + n_1\,\bar{y}_n)\text{, since } n_1 \text{ is a sub-sample of the sample of size } n.$$

$$= \frac{1}{n+n_1}(n\bar{Y} + n_1\,\bar{Y}) = \bar{Y}.$$

To obtain the variance

$$V(\bar{y}_{n+n_1}) = E(\bar{y}_{n+n_1} - \bar{Y})^2 = E\left(\frac{n\,\bar{y}_n + n_1\,\bar{y}_{n_1}}{n+n_1} - \bar{Y}\right)^2$$

$$= \frac{1}{(n+n_1)^2} E[n\,\bar{y}_n + n_1\,\bar{y}_{n_1} - (n+n_1)\bar{Y}]^2$$

$$= \frac{1}{(n+n_1)^2} E[n\,\bar{y}_n - n\bar{Y} + n_1\,\bar{y}_{n_1} - n_1\,\bar{Y}]^2$$

$$= \frac{1}{(n+n_1)^2} E[n(\bar{y}_n - \bar{Y}) + n_1\bar{y}_{n_1} - n_1\bar{y}_n + n_1\bar{y}_n - n_1\bar{Y}]^2$$

$$= \frac{1}{(n+n_1)^2} E[(n+n_1)(\bar{y}_n - \bar{Y}) + n_1(\bar{y}_{n_1} - \bar{y}_n)]^2$$

$$= \frac{1}{(n+n_1)^2}[(n+n_1)^2\,E(\bar{y}_n - \bar{Y})^2 + n_1^2\,E(\bar{y}_{n_1} - \bar{y}_n)^2],\quad \text{as samples are drawn independently.}$$

$$= \frac{1}{(n+n_1)^2}[(n+n_1)^2\,V(\bar{y}_n) + n_1^2\,E\{E(\bar{y}_{n_1} - \bar{y}_n)^2 \mid n\}]$$

$$= \frac{1}{(n+n_1)^2}\left[(n+n_1)^2\,V(\bar{y}_n) + n_1^2\,E\left\{\left(\frac{1}{n_1} - \frac{1}{n}\right)S_n^2\right\}\right]$$

$$= \frac{1}{(n+n_1)^2}\left[(n+n_1)^2\,V(\bar{y}_n) + n_1^2\left(\frac{n-n_1}{n_1 n}\right)S^2\right]$$

$$= \frac{1}{(n+n_1)^2}\left[(n+n_1)^2\,V(\bar{y}_n) + \frac{n_1(n-n_1)}{n}S^2\right] = V(\bar{y}_n) + \frac{n_1(n-n_1)}{n(n+n_1)^2}S^2.$$

Therefore,

$$\frac{V(\bar{y}_{n+n_1})}{V(\bar{y}_n)} = 1 + \frac{n_1(n-n_1)}{n(n+n_1)^2 V(\bar{y}_n)} S^2 \cong 1 + \frac{n_1(n-n_1)}{n(n+n_1)^2 S^2/n} S^2$$

$$= \frac{(n+n_1)^2 + n_1(n-n_1)}{(n+n_1)^2} = \frac{n^2 + n_1^2 + 2n_1 n + n_1 n - n_1^2}{(n+n_1)^2}$$

$$= \frac{n^2 + 3n_1 n}{(n+n_1)^2} = \frac{1 + (3n_1/n)}{(1+n_1/n)^2}.$$

Exercise: A simple random sample of size $n = n_1 + n_2$ with mean $\bar{y}$ is drawn from a finite population, and a simple random subsample of size n_1 is drawn from it with mean $\bar{y}_1$. Show that

i) $V(\bar{y}_1 - \bar{y}_2) = S^2[(1/n_1) + (1/n_2)]$, where $\bar{y}_2$ is mean of the remaining n_2 units in the sample,

ii) $V(\bar{y}_1 - \bar{y}) = S^2[(1/n_1) - (1/n)]$,

iii) $Cov(\bar{y}, \bar{y}_1 - \bar{y}) = 0$.

Repeated sampling implies repetition of the drawing of both the sample and subsample.

Solution:

i) In repeated sampling the given procedure is equivalent to draw subsamples of sizes n_1 and n_2 independently, thus

$V(\bar{y}_1 - \bar{y}_2) = V(\bar{y}_1) + V(\bar{y}_2)$, since $Cov(\bar{y}_1, \bar{y}_2) = 0$

$= S^2[(1/n_1) + (1/n_2)]$, ignoring fpc.

ii) $\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} \Rightarrow \quad \bar{y}_1 - \bar{y} = \bar{y}_1 - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$

or $\bar{y}_1 - \bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_1 - n_1\bar{y}_1 - n_2\bar{y}_2}{n_1 + n_2} = \frac{n_2(\bar{y}_1 - \bar{y}_2)}{n}$.

Therefore,

$$V(\bar{y}_1 - \bar{y}) = V\left(\frac{n_2(\bar{y}_1 - \bar{y}_2)}{n}\right) = \frac{n_2^2}{n^2} V(\bar{y}_1 - \bar{y}_2) = \frac{n_2^2}{n^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S^2$$

$$= \frac{n_2^2}{n^2}\left(\frac{n_1 + n_2}{n_1 n_2}\right) S^2 = \frac{n_2}{n_1 n} S^2 = \frac{n - n_1}{n_1 n} S^2 = \left(\frac{1}{n_1} - \frac{1}{n}\right) S^2.$$

iii) $Cov(\bar{y}, \bar{y}_1 - \bar{y}) = E[\bar{y}(\bar{y}_1 - \bar{y})] - E(\bar{y})E(\bar{y}_1 - \bar{y})$

$$= E(\bar{y}\,\bar{y}_1 - \bar{y}^2) - \bar{Y} \times 0 = E(\bar{y}\,\bar{y}_1) - E(\bar{y}^2) \qquad (1)$$

Consider

$$E(\bar{y}\,\bar{y}_1) = E\left(\frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n}\,\bar{y}_1\right) = E\left(\frac{n_1}{n}\,\bar{y}_1^2 + \frac{n_2}{n}\,\bar{y}_1\bar{y}_2\right)$$

$$= \frac{n_1}{n}E(\bar{y}_1^2) + \frac{n_2}{n}E(\bar{y}_1)E(\bar{y}_2)$$

$$= \frac{n_1}{n}[V(\bar{y}_1) + \bar{Y}^2] + \frac{n_2}{n}\bar{Y}^2 = \frac{n_1}{n}\left(\frac{S^2}{n_1} + \bar{Y}^2\right) + \frac{n_2}{n}\bar{Y}^2$$

$$= \frac{S^2}{n} + \frac{n_1}{n}\bar{Y}^2 + \frac{n_2}{n}\bar{Y}^2 = \frac{S^2}{n} + \bar{Y}^2 \qquad (2)$$

Now

$$V(\bar{y}) = E(\bar{y}^2) - \bar{Y}^2 \qquad \text{or} \qquad E(\bar{y}^2) = V(\bar{y}) + \bar{Y}^2 = \frac{S^2}{n} + \bar{Y}^2 \qquad (3)$$

In view of equations (1), (2), and (3), we get

$$Cov(\bar{y},\ \bar{y}_1 - \bar{y}) = \left(\frac{S^2}{n} + \bar{Y}^2\right) - \left(\frac{S^2}{n} + \bar{Y}^2\right) = 0.$$

Exercise: A population has three units U_1, U_2 and U_3 with variates Y_1, Y_2 and Y_3 respectively. It is required to estimate the population total Y by selecting a sample of two units. Let the sampling and estimation procedures be as follows:

Sample (s)	$P(s)$	Estimator t	Estimator t'
(U_1, U_2)	$1/2$	$Y_1 + 2Y_2$	$Y_1 + 2Y_2 + Y_1^2$
(U_1, U_3)	$1/2$	$Y_1 + 2Y_3$	$Y_1 + 2Y_3 - Y_1^2$

Prove that both t and t' are unbiased for Y and find their variances. Comment on the estimators.

Solution: By definition

$$E(t) = \sum_i t_i\, p(t_i) = \frac{1}{2}(Y_1 + 2Y_2 + Y_1 + 2Y_3) = Y.$$

This shows that estimator t is unbiased for Y.

$$E(t^2) = \frac{1}{2}[(Y_1 + 2Y_2)^2 + (Y_1 + 2Y_3)^2] = \frac{1}{2}(Y_1^2 + 4Y_2^2 + 4Y_1Y_2 + Y_1^2 + 4Y_3^2 + 4Y_1Y_3)$$

$$= Y_1^2 + 2Y_2^2 + 2Y_3^2 + 2Y_1Y_2 + 2Y_1Y_3.$$

Therefore,

$$V(t) = E(t^2) - [E(t)]^2 = Y_1^2 + 2Y_2^2 + 2Y_3^2 + 2Y_1Y_2 + 2Y_1Y_3 - (Y_1 + Y_2 + Y_3)^2$$

$$= Y_2^2 + Y_3^2 - 2Y_2Y_3 = (Y_2 - Y_3)^2.$$

Similarly,

$E(t') = \sum_i t_i' \, p(t_i') = \frac{1}{2}(Y_1 + 2Y_2 + Y_1^2 + Y_1 + 2Y_3 - Y_1^2) = Y$, hence, t' is unbiased for Y.

$$E(t'^2) = \frac{1}{2}[(Y_1 + 2Y_2 + Y_1^2)^2 + (Y_1 + 2Y_3 - Y_1^2)^2]$$

$$= \frac{1}{2}(Y_1^4 + 2Y_1^3 + Y_1^2 + 4Y_1^2Y_2 + 4Y_1Y_2 + 4Y_2^2 + Y_1^4 - 2Y_1^3$$

$$+ Y_1^2 - 4Y_1^2Y_3 + 4Y_1Y_3 + 4Y_3^2)$$

$$= Y_1^4 + Y_1^2 + 2Y_1^2Y_2 + 2Y_1Y_2 + 2Y_2^2 - 2Y_1^2Y_3 + 2Y_1Y_3 + 2Y_3^2.$$

Therefore,

$$V(t') = E(t'^2) - [E(t')]^2$$

$$= Y_1^4 + Y_1^2 + 2Y_1^2Y_2 + 2Y_1Y_2 + 2Y_2^2 - 2Y_1^2Y_3 + 2Y_1Y_3 + 2Y_3^2 - (Y_1 + Y_2 + Y_3)^2$$

$$= (Y_2 - Y_3)^2 + Y_1^2 (Y_1^2 + 2Y_2 - 2Y_3)$$

$$= V(t) + Y_1^2 (Y_1^2 + 2Y_2 - 2Y_3).$$

We conclude that both linear estimator t and quadratic estimator t' are unbiased; among which estimator has minimum variance depends on the variate values.

UNIT-II

SRATIFIED RANDOM SAMPLING

The precision of an estimator of the population parameters (mean or total etc.) depends on the size of the sample and the variability or heterogeneity among the units of the population. If the population is very heterogeneous and considerations of cost limit the size of the sample, it may be found impossible to get a sufficiently precise estimate by taking a simple random sample from the entire population. For this, one possible way to estimate the population mean or total with greater precision is to divide the population in several groups (sub-population or classes, these sub-populations are non-overlapping) each of which is more homogenous than the entire population and draw a random sample of predetermined size from each one of the groups. The groups, into which the population is divided, are called **strata** or each group is called **stratum** and the whole procedure of dividing the population into the strata and then drawing a random sample from each one of the strata is called **stratified random sampling**. For example, to estimate the average income per household, it may be appropriate to group the households into two or more groups (strata) according to the rent paid by the households. The households in any stratum so form are likely to be more homogeneous with respect to income as compared to the whole population. Thus, the estimated income per household based on a stratified sample is likely to be more precise than that based on a simple random sample of the same size drawn from the whole population.

Principal reasons for stratification

- To gain in precision, divide a heterogeneous population into strata in such a way that each stratum is internally homogeneous.
- To accommodate administrative convenience (cost consideration), fieldwork is organized by strata, which usually results in saving in cost and effort.
- To obtain separate estimates for strata.
- We can accommodate different sampling plan in different strata.
- We can have data of known precision for certain subdivisions treating each subdivision as a population in its own right.

Notations

Let the population, consisting of N units is first divided into k strata (sub-populations) of size $N_1, N_2, \cdots, N_k$. These sub-populations are non-overlapping such that $N_1 + N_2 + \cdots + N_k = N$. A sample is drawn (by the method of srs) from each stratum (group or sub-population) independently, the sample size within the $i-$th stratum being n_i, $(i = 1, 2, \cdots, k)$ such that $n_1 + n_2 + \cdots + n_k = n$. The following symbols refer to stratum i.

N_i, total number of units.

n_i, number of units in sample.

$f_i = \dfrac{n_i}{N_i}$, sampling fraction in the stratum.

$W_i = \dfrac{N_i}{N}$, stratum weight.

y_{ij}, value of the characteristic under study for the $j-$th unit in the $i-$th stratum, $j = 1,2,\cdots, N_i$.

$\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$, mean based on N_i units (stratum mean).

$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, mean based on n_i units (sample mean).

$\sigma_i^2 = \frac{1}{N_i}\sum_{j=1}^{N_i}(y_{ij} - \bar{Y}_i)^2$, variance based on N_i units (stratum variance).

$S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(y_{ij} - \bar{Y}_i)^2$, mean square based on N_i units (stratum mean square).

$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$, sample mean square based on n_i units.

$Y = \sum_{i=1}^{k}\sum_{j=1}^{N_i} y_{ij} = \sum_{i=1}^{k} N_i \bar{Y}_i$, population total.

$\bar{Y} = \frac{Y}{N} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{Y}_i = \sum_{i=1}^{k} W_i \bar{Y}_i$, over all population mean.

Theorem: For stratified random sampling, *wor*, if in every stratum the sample estimate $\bar{y}_i$ is an unbiased of $\bar{Y}_i$, and samples are drawn independently in different strata, then $\bar{y}_{st} = \sum_{i=1}^{k} W_i \bar{y}_i$ is an unbiased estimate of the over all population mean $\bar{Y}$ and its variance is $V(\bar{y}_{st}) = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 S_i^2$.

Proof: Since sampling within each stratum is simple random sampling, i.e. $E(\bar{y}_i) = \bar{Y}_i$, it follows that

$E(\bar{y}_{st}) = E\left(\sum_{i=1}^{k} W_i \bar{y}_i\right) = \sum_{i=1}^{k} W_i E(\bar{y}_i) = \sum_{i=1}^{k} W_i \bar{Y}_i = \bar{Y}$. To obtain the variance, we have

$$V(\bar{y}_{st}) = E[\bar{y}_{st} - E(\bar{y}_{st})]^2 = E\left[\sum_{i=1}^{k} W_i \bar{y}_i - E\left(\sum_{i=1}^{k} W_i \bar{y}_i\right)\right]^2 = E\left[\sum_{i=1}^{k} W_i \{\bar{y}_i - E(\bar{y}_i)\}\right]^2$$

$$= E\left[\sum_{i=1}^{k} W_i^2 \{\bar{y}_i - E(\bar{y}_i)\}^2\right] + E\left[\sum_{\substack{i,i' \\ i \neq i'}}^{k} W_i W_{i'} \{\bar{y}_i - E(\bar{y}_i)\} \{\bar{y}_{i'} - E(\bar{y}_{i'})\}\right]$$

$$= \sum_{i=1}^{k} W_i^2 V(\bar{y}_i) + \sum_{i=1}^{k} \sum_{i' \neq i=1}^{k} W_i W_{i'} Cov(\bar{y}_i, \bar{y}_{i'}).$$

Since samples are drawn independently in different strata, all covariance terms vanishes, then

$V(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2 V(\bar{y}_i) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 S_i^2$, as *srswor* within each stratum.

Alternative expressions of $V(\bar{y}_{st})$

i) $V(\bar{y}_{st}) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 S_i^2 = \sum_{i=1}^{k} \left(\frac{N_i - n_i}{N_i}\right) \frac{N_i^2}{N^2} S_i^2 / n_i = \frac{1}{N^2} \sum_{i=1}^{k} N_i (N_i - n_i) S_i^2 / n_i.$

ii) $V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} N_i (N_i - n_i) S_i^2 / n_i = \frac{1}{N^2} \sum_{i=1}^{k} N_i^2 \left(1 - \frac{n_i}{N_i}\right) S_i^2 / n_i = \sum_{i=1}^{k} W_i^2 \frac{(1 - f_i) S_i^2}{n_i}.$

Corollary: $\hat{Y}_{st} = N \bar{y}_{st}$ is an unbiased estimate of the population total Y with its variance $V(\hat{Y}_{st}) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) N_i^2 S_i^2$.

Proof: By definition

$E(\hat{Y}_{st}) = N E(\bar{y}_{st}) = N\bar{Y} = Y$, and

$$V(\hat{Y}_{st}) = N^2 V(\bar{y}_{st}) = N^2 \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 S_i^2 = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) N_i^2 S_i^2$$

$$= \sum_{i=1}^{k} N_i (N_i - n_i) S_i^2 / n_i = \sum_{i=1}^{k} N_i^2 (1 - f_i) S_i^2 / n_i.$$

Remarks

a) If N_i are large as compared to n_i (if the sampling fractions $f_i = \frac{n_i}{N_i}$ are negligible in all strata), then,

i) $V(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2 S_i^2 / n_i = \frac{1}{N^2} \sum_{i=1}^{k} N_i^2 S_i^2 / n_i.$

ii) $V(\hat{Y}_{st}) = \sum_{i=1}^{k} N_i^2 S_i^2 / n_i.$

b) If in every stratum $\frac{n_i}{n} = \frac{N_i}{N}$ i.e. $n_i = n\frac{N_i}{N} = nW_i$, the variance of $\bar{y}_{st}$ reduces to

$$V(\bar{y}_{st}) = \sum_{i=1}^{k}\left(\frac{N_i - n_i}{N_i}\right)W_i^2\, S_i^2 / n_i = \sum_{i=1}^{k}\left(\frac{N_i - nW_i}{N_i}\right)W_i S_i^2 / n = \frac{1-f}{n}\sum_{i=1}^{k} W_i\, S_i^2 .$$

c) If in every stratum $\frac{n_i}{n} = \frac{N_i}{N}$, and the variance of $\bar{y}_{st}$ in all strata have the same value S^2 , then the result reduces to $V(\bar{y}_{st}) = \frac{1-f}{n}\sum_{i=1}^{k} W_i S^2 = \frac{1-f}{n}S^2$, since $\sum_{i=1}^{k} W_i = 1.$

Estimation of variance

If a simple random sample is taken within each stratum, then an unbiased estimator of S_i^2, is $s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$, and an unbiased estimator of variance $\bar{y}_{st}$ is

$$\hat{V}(\bar{y}_{st}) = v(\bar{y}_{st}) = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)W_i^2\, s_i^2 = \frac{1}{N^2}\sum_{i=1}^{k} N_i(N_i - n_i)\, s_i^2 / n_i$$

$$= \sum_{i=1}^{k} W_i^2 (1 - f_i)\, s_i^2 / n_i .$$

Alternative form for computing purposes

$$V(\bar{y}_{st}) = \sum_{i=1}^{k}\frac{W_i^2 s_i^2}{n_i} - \sum_{i=1}^{k}\frac{W_i^2 s_i^2}{N_i} = \sum_{i=1}^{k}\frac{W_i^2 s_i^2}{n_i} - \sum_{i=1}^{k}\frac{W_i s_i^2}{N}.$$

Theorem: If stratified random sampling is with replacement, then $\bar{y}_{st} = \sum_{i=1}^{k} W_i\, \bar{y}_i$ is an unbiased estimate of population mean $\bar{Y}$ and its variance is $V(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2\, S_i^2 / n_i$.

Proof: As in stratified random sampling, *wor*, $E(\bar{y}_{st}) = \bar{Y}$, and

$$V(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2\, V(\bar{y}_i) = \sum_{i=1}^{k} W_i^2 \sigma_i^2 / n_i = \sum_{i=1}^{k} W_i^2\left(\frac{N_i - 1}{N_i}\right)S_i^2 / n_i \cong \sum_{i=1}^{k} W_i^2\, S_i^2 / n_i$$

Corollary: $\hat{Y}_{st} = N\,\bar{y}_{st} = N\sum_{i=1}^{k} W_i\,\bar{y}_i = \sum_{i=1}^{k} N_i\,\bar{y}_i$ is an unbiased estimate of the population total Y and its variance is

$$V(\hat{Y}_{st}) = V(N\,\bar{y}_{st}) = N^2\, V(\bar{y}_{st}) = N^2\sum_{i=1}^{k} W_i^2 S_i^2 / n_i = \sum_{i=1}^{k} N_i^2\, S_i^2 / n_i .$$

Choice of sample size in different strata

There are three methods of allocation of sample sizes to different strata in a stratified sampling procedure. These are

i) Equal allocation.

ii) Proportional allocation.

iii) Optimum allocation.

Equal allocation: In this method, the total sample size n is divided equally among all the strata, i.e. for $i-$th stratum $n_i = n/k$. In practice, this method is not used except when the strata sizes are almost equal.

Proportional allocation: This procedure of allocation is very common in practice because of its simplicity. When no other information except N_i, the total number of units in the $i-$th stratum, is available, the allocation of a given sample of size n to different strata is done in proportion to their sizes, i.e. in the $i-$th stratum $n_i \propto N_i$ or $n_i = \lambda N_i$, where λ is the constant of proportionality, and

$$\sum_{i=1}^{k} n_i = \lambda \sum_{i=1}^{k} N_i \ , \quad \text{or} \qquad \lambda = \frac{n}{N}, \qquad \Rightarrow \qquad n_i = \frac{n}{N} N_i = nW_i .$$

$V(\bar{y}_{st})$ **Under proportional allocation**

$$V(\bar{y}_{st})_{prop} = \sum_{i=1}^{k} \left(\frac{1}{nW_i} - \frac{1}{N_i} \right) W_i^2 \, S_i^2 = \sum_{i=1}^{k} \left(\frac{W_i}{nW_i} - \frac{W_i}{N_i} \right) W_i S_i^2$$

$$= \left(\frac{1}{n} - \frac{1}{N} \right) \sum_{i=1}^{k} W_i S_i^2 = \frac{1-f}{n} \sum_{i=1}^{k} W_i S_i^2 \ .$$

Note: If the variances in all strata have the same value, S^2 (say), then

$$V(\bar{y}_{st})_{prop} = \frac{1-f}{n} S^2 , \text{ as } \sum_{i=1}^{k} W_i = 1 .$$

Alternative expressions of $V(\bar{y}_{st})_{prop}$

$$V(\bar{y}_{st})_{prop} = \left(\frac{1}{n} - \frac{1}{N} \right) \sum_{i=1}^{k} W_i S_i^2 = \left(\frac{N-n}{nN} \right) \sum_{i=1}^{k} \frac{N_i}{N} \ S_i^2 = \frac{N-n}{nN^2} \sum_{i=1}^{k} N_i S_i^2 \ .$$

Optimum allocation: In this method of allocation the sample sizes n_i in the respective strata are determined with a view to minimize $V(\bar{y}_{st})$ for a specified cost of conducting the sample survey or to minimize the cost for a specified value of $V(\bar{y}_{st})$. The simplest cost function is of the form

Cost $= C = c_0 + \sum_{i=1}^{k} c_i \, n_i$, where the overhead cost c_0 is constant and c_i is the average cost of surveying one unit in the $i-$th stratum

$$C - c_0 = \sum_{i=1}^{k} n_i c_i = C' \text{ (say)} \tag{2.1}$$

and $V(\bar{y}_{st}) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) W_i^2 S_i^2 = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} - \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{N_i}$, so that

$$V(\bar{y}_{st}) + \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{N_i} = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} = V' \text{ (say)} \tag{2.2}$$

where C' and V' are function of n_i. Choosing the n_i to minimize V for fixed C or C for fixed V are both equivalent to minimizing the product

$$V'C' = \left(\sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} \right) \left(\sum_{i=1}^{k} n_i c_i \right)$$

It may be minimized by use of the Cauchy-Schwartz inequality, i.e. if $a_i, b_i,\ i = 1,2,\cdots,k$ are two sets of k positive numbers, then

$\left(\sum_{i=1}^{k} a_i^2 \right) \left(\sum_{i=1}^{k} b_i^2 \right) \geq \left(\sum_{i=1}^{k} a_i b_i \right)^2$, equality holds if and only if $\frac{b_i}{a_i}$ is constant for all i.

Taking $a_i = W_i S_i / \sqrt{n_i} > 0$, and $b_i = \sqrt{n_i c_i} > 0$, then

$$V'C' = \sum_{i=1}^{k} (W_i S_i / \sqrt{n_i})^2 \sum_{i=1}^{k} (\sqrt{n_i c_i})^2 \geq \left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \right)^2 .$$

Thus, no choice of n_i can make $V'C'$ smaller than $\left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \right)^2$. This minimum value occurs when $\frac{b_i}{a_i} = \text{constant, say } \lambda$.

$$\Rightarrow \qquad \frac{b_i}{a_i} = \sqrt{n_i c_i} \left(\frac{\sqrt{n_i}}{W_i S_i} \right) = \frac{n_i \sqrt{c_i}}{W_i S_i} = \lambda \qquad \text{or} \qquad n_i = \lambda \frac{W_i S_i}{\sqrt{c_i}} \tag{2.3}$$

$\Rightarrow \quad n_i \propto W_i S_i / \sqrt{c_i}$, this allocation is known as **optimum allocation**.

Taking summation on both the sides of equation (2.3), we get

$$\sum_{i=1}^{k} n_i = \lambda \sum_{i=1}^{k} \frac{W_i S_i}{\sqrt{c_i}} \qquad \text{or} \qquad \lambda = \frac{n}{\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}}, \text{ and hence,}$$

$$n_i = n \frac{W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}} = n \frac{N_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} N_i S_i / \sqrt{c_i}} . \tag{2.4}$$

Alternative method

To determine n_i such that $V(\bar{y}_{st})$ is minimum and cost C is fixed, consider the function

$$\phi = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} - \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{N_i} + \lambda \left(c_0 + \sum_{i=1}^{k} c_i n_i - C \right), \quad \text{where} \quad \lambda \quad \text{is some unknown constant.}$$

Using the calculus method of Lagrange multipliers, we select n_i, and the constant λ to minimize ϕ. Differentiating ϕ with respect to n_i, and equating to zero, we have

$$\frac{\partial \phi}{\partial n_i} = 0 = -\frac{W_i^2 S_i^2}{n_i^2} + \lambda c_i \qquad \text{or} \qquad n_i = \frac{1}{\sqrt{\lambda}} \frac{W_i S_i}{\sqrt{c_i}} \tag{23a}$$

$$\Rightarrow \qquad n_i \propto W_i S_i / \sqrt{c_i} \qquad \text{or} \qquad n_i \propto N_i S_i / \sqrt{c_i} \ .$$

Taking summation on both the sides of equation (2.3a), we get

$$\sum_{i=1}^{k} n_i = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{k} W_i S_i / \sqrt{c_i} \qquad \text{or} \qquad \frac{1}{\sqrt{\lambda}} = \frac{n}{\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}}$$

$$\Rightarrow \qquad n_i = n \frac{W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}} = n \frac{N_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} N_i S_i / \sqrt{c_i}} \tag{2.4a}$$

The total sample size n required for the optimum sample sizes within strata. The solution for the value of n depends on whether the sample is chosen to meet a specified total cost C or to give a specified variance V for $\bar{y}_{st}$.

i) **If cost is fixed**, substitute the optimum values of n_i in (cost function) equation (2.1) and solve for n as

$$C - c_0 = \sum_{i=1}^{k} c_i n_i = \sum_{i=1}^{k} n \frac{W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}} c_i = n \sum_{i=1}^{k} \frac{W_i S_i \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}}$$

$$\Rightarrow \quad n = \frac{C - c_0}{\sum_{i=1}^{k} W_i S_i \sqrt{c_i}} \sum_{i=1}^{k} W_i S_i / \sqrt{c_i} \ .$$

Hence,

$$n_i = \frac{C - c_0}{\sum_{i=1}^{k} W_i S_i \sqrt{c_i}} \sum_{i=1}^{k} W_i S_i / \sqrt{c_i} \times \frac{W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}} = \frac{(C - c_0) W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i \sqrt{c_i}} .$$

$V(\bar{y}_{st})$ **Under optimum allocation for fixed Cost**

$$V(\bar{y}_{st})_{opt} = \sum_{i=1}^{k}\left[\frac{1}{(C-c_0)\,W_i\,S_i}\left(\sqrt{c_i}\sum_{i=1}^{k}W_i\,S_i\sqrt{c_i}\right)-\frac{1}{N_i}\right]W_i^2\,S_i^2$$

$$=\sum_{i=1}^{k}\left[\frac{1}{C-c_0}\left(\sum_{i=1}^{k}W_i\,S_i\sqrt{c_i}\right)W_iS_i\sqrt{c_i}-\frac{W_i^2\,S_i^2}{N_i}\right]$$

$$=\frac{1}{C-c_0}\left(\sum_{i=1}^{k}W_i\,S_i\sqrt{c_i}\right)^2-\sum_{i=1}^{k}\frac{W_i\,S_i^2}{N}.$$

ii) **If *V* is fixed**, substitute the optimum n_i in equation (2.2), we get

$$V(\bar{y}_{st})+\frac{1}{N}\sum_{i=1}^{k}W_iS_i^2=\frac{1}{n}\sum_{i=1}^{k}\frac{W_i^2S_i^2\sum_{i=1}^{k}W_iS_i/\sqrt{c_i}}{W_iS_i/\sqrt{c_i}}=\frac{1}{n}\sum_{i=1}^{k}W_iS_i\sqrt{c_i}\left(\sum_{i=1}^{k}W_iS_i/\sqrt{c_i}\right).$$

Thus,

$$n=\frac{1}{V+\frac{1}{N}\sum_{i=1}^{k}W_iS_i^2}\sum_{i=1}^{k}W_iS_i\sqrt{c_i}\left(\sum_{i=1}^{k}W_iS_i/\sqrt{c_i}\right), \text{ and hence,}$$

$$n_i=\frac{1}{V+\frac{1}{N}\sum_{i=1}^{k}W_iS_i^2}\left[(W_i\,S_i/\sqrt{c_i})\sum_{i=1}^{k}W_iS_i\sqrt{c_i}\right].$$

Optimum cost for fixed variance

$$C-c_0=\sum_{i=1}^{k}c_i\left[\frac{(W_i\,S_i/\sqrt{c_i})\sum_{i=1}^{k}W_iS_i\sqrt{c_i}}{V+\frac{1}{N}\sum_{i=1}^{k}W_iS_i^2}\right]=\frac{\left(\sum_{i=1}^{k}W_iS_i\sqrt{c_i}\right)\sum_{i=1}^{k}W_iS_i\sqrt{c_i}}{V+\frac{1}{N}\sum_{i=1}^{k}W_iS_i^2}$$

$$=\frac{1}{V+\frac{1}{N}\sum_{i=1}^{k}W_iS_i^2}\left(\sum_{i=1}^{k}W_iS_i\sqrt{c_i}\right)^2.$$

Remark

An important special case arises if $c_i = c$, that is, if the cost per unit is the same in all strata. The cost becomes $C = c_0 + \sum_{i=1}^{k} c\,n_i = c_0 + cn$, and optimum allocation for fixed cost reduces to optimum allocation for fixed sample size. The result in this case is as follows:

In stratified random sampling $V(\bar{y}_{st})$ is minimized for a fixed total size of sample n if

$$n_i = n\frac{W_iS_i}{\sum_{i=1}^{k} W_iS_i} = n\frac{N_iS_i}{\sum_{i=1}^{k} N_iS_i} \quad \Rightarrow \quad n_i \propto W_iS_i \quad \text{or} \quad n_i \propto N_iS_i,$$ and is called **Neyman allocation** and $V(\bar{y}_{st})$ under optimum allocation for fixed n or Neyman allocation.

$$V(\bar{y}_{st})_{opt} = \sum_{i=1}^{k}\left(\frac{1}{nW_iS_i}\sum_{i=1}^{k} W_iS_i - \frac{1}{N_i}\right) W_i^2S_i^2 = \sum_{i=1}^{k}\left[\frac{1}{n}\left(\sum_{i=1}^{k} W_iS_i\right)W_iS_i - \frac{1}{N_i}\left(\frac{N_i}{N}\right) W_iS_i^2\right]$$

$$= \frac{1}{n}\left(\sum_{i=1}^{k} W_iS_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} W_iS_i^2 .$$

Note: If N is large, $V(\bar{y}_{st})_{opt}$ reduces to $V(\bar{y}_{st})_{opt} = \frac{1}{n}\left(\sum_{i=1}^{k} W_iS_i\right)^2$.

Relative precision of stratified with simple random sampling

Here, we shall make a comparative study of the usual estimators under simple random sampling, without stratification and stratified random sampling employing various schemes of allocation i.e. proportional and optimum allocations. This comparison shows how the gain due to stratification is achieved.

Consider the variances of these estimators of population mean, which are as follows.

$$V_{ran} = \frac{1-f}{n}S^2$$

$$V_{prop} = \frac{1-f}{n}\sum_{i=1}^{k} W_iS_i^2 = \frac{1}{n}\sum_{i=1}^{k} W_iS_i^2 - \frac{1}{N}\sum_{i=1}^{k} W_iS_i^2 .$$

$$V_{opt} = \frac{1}{n}\left(\sum_{i=1}^{k} W_iS_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} W_iS_i^2 .$$

Now

$$(N-1)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(y_{ij}-\bar{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(y_{ij}-\bar{Y}_i+\bar{Y}_i-\bar{Y})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{N_i}(\bar{Y}_i-\bar{Y})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{N_i}(y_{ij}-\bar{Y}_i)(\bar{Y}_i-\bar{Y})$$

$$= \sum_{i=1}^{k}(N_i-1)S_i^2 + \sum_{i=1}^{k}N_i(\bar{Y}_i-\bar{Y})^2 + 2\sum_{i=1}^{k}(\bar{Y}_i-\bar{Y})\left\{\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y}_i)\right\}$$

$$= \sum_{i=1}^{k}(N_i - 1)S_i^2 + \sum_{i=1}^{k} N_i(\bar{Y}_i - \bar{\bar{Y}})^2 \text{ , as sum of the deviations from their}$$

mean is zero.

or $S^2 = \sum_{i=1}^{k}\left(\frac{N_i - 1}{N - 1}\right)S_i^2 + \sum_{i=1}^{k}\frac{N_i}{N-1}(\bar{Y}_i - \bar{\bar{Y}})^2$

For large N ,

$\frac{1}{N} \to 0$, so that, $\frac{N_i - 1}{N - 1} = \frac{(N_i / N) - (1/N)}{1 - (1/N)} \cong W_i$

and

$\frac{N_i}{N-1} = \frac{(N_i / N)}{1 - (1/N)} \cong W_i$,

so that

$S^2 = \sum_{i=1}^{k} W_i S_i^2 + \sum_{i=1}^{k} W_i(\bar{Y}_i - \bar{\bar{Y}})^2$.

Hence,

$$V_{ran} = \frac{1-f}{n}S^2 = \frac{1-f}{n}\sum_{i=1}^{k} W_i S_i^2 + \frac{1-f}{n}\sum_{i=1}^{k} W_i(\bar{Y}_i - \bar{\bar{Y}})^2$$

$$= V_{prop} + \frac{1-f}{n}\sum_{i=1}^{k} W_i(\bar{Y}_i - \bar{\bar{Y}})^2 = V_{prop} + \text{positive quantity.}$$

Thus, $V_{ran} \geq V_{prop}$. (2.5)

Further, consider

$$V_{prop} - V_{opt} = \frac{1}{n}\sum_{i=1}^{k} W_i S_i^2 - \frac{1}{N}\sum_{i=1}^{k} W_i S_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} W_i S_i\right)^2 + \frac{1}{N}\sum_{i=1}^{k} W_i S_i^2$$

$$= \frac{1}{n}\left[\sum_{i=1}^{k} W_i S_i^2 - \left(\sum_{i=1}^{k} W_i S_i\right)^2\right] = \frac{1}{n}\left[\sum_{i=1}^{k} W_i S_i^2 + \left(\sum_{i=1}^{k} W_i S_i\right)^2 - 2\left(\sum_{i=1}^{k} W_i S_i\right)^2\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{k} W_i S_i^2 + \left(\sum_{i=1}^{k} W_i S_i\right)^2 \sum_{i=1}^{k} W_i - 2\left(\sum_{i=1}^{k} W_i S_i\right)\left(\sum_{i=1}^{k} W_i S_i\right)\right],$$

as $\sum_{i=1}^{k} W_i = 1$

$$= \frac{1}{n}\left[\sum_{i=1}^{k} W_i \left\{ S_i^2 + \left(\sum_{i=1}^{k} W_i S_i\right)^2 - 2 S_i \left(\sum_{i=1}^{k} W_i S_i\right)\right\}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{k} W_i \left(S_i - \sum_{i=1}^{k} W_i S_i \right)^2 = +\text{ve quantity.}$$

$$\Rightarrow V_{prop} = V_{opt} + \frac{1}{n}\sum_{i=1}^{k} W_i \left(S_i - \sum_{i=1}^{k} W_i S_i \right)^2 .$$

Thus,

$$V_{prop} \geq V_{opt} . \tag{2.6}$$

From equation (2.5) and (2.6), we get

$$V_{ran} \geq V_{prop} \geq V_{opt} .$$

Also,

$$V_{ran} = V_{opt} + \frac{1}{n}\sum_{i=1}^{k} W_i \left(S_i - \sum_{i=1}^{k} W_i S_i \right)^2 + \frac{1-f}{n}\sum_{i=1}^{k} W_i (\bar{Y}_i - \bar{Y})^2 .$$

Remark

In comparing the precision of stratified with un-stratified random sampling, it was assumed that the population values of stratum means and variances were known.

Estimation of the gain in precision due to stratification

It is sometimes of interest to examine, from a survey, whether the mode of stratification has been effective in estimating the population mean with increased gain in precision relative to simple random sampling without replacement. The data available from the sample are the value $N_i, n_i, \bar{y}_i$, and s_i^2. An unbiased estimator of the variance of $\bar{y}_{st}$ is given by

$$\hat{V}(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2 s_i^2 / n_i - \sum_{i=1}^{k} W_i \, s_i^2 / N .$$

The problem is to compare this variance with an unbiased estimate of $V(\bar{y}_{sr})$ based on the given stratified sample. For estimation of $V(\bar{y}_{sr})$, note that

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

We shall first estimate S^2, when $\bar{y}_i$ and s_i^2 are available for all the strata. Consider, the relation

$$(N-1) S^2 = \sum_{i=1}^{k} (N_i - 1) S_i^2 + \sum_{i=1}^{k} N_i (\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{k} (N_i - 1) S_i^2 + N \sum_{i=1}^{k} W_i (\bar{Y}_i - \bar{Y})^2 .$$

$$= \sum_{i=1}^{k} (N_i - 1) S_i^2 + N \left(\sum_{i=1}^{k} W_i \bar{Y}_i^2 - \bar{Y}^2 \right).$$

To get the estimate of S^2, we need the estimates of S_i^2, $\bar{Y}_i^2$ and $\bar{Y}^2$. As sampling is simple random *wor* within each stratum, so s_i^2 is unbiased estimate of S_i^2. Note that

$$V(\bar{y}_i) = E(\bar{y}_i - \bar{Y}_i)^2, \qquad \Rightarrow \qquad \bar{Y}_i^2 = E(\bar{y}_i^2) - V(\bar{y}_i), \text{ and } \quad \hat{\bar{Y}}_i^2 = \bar{y}_i^2 - \left(\frac{1}{n_i} - \frac{1}{N_i} \right) s_i^2$$

Similarly, after noting that $V(\bar{y}_{st}) = E(\bar{y}_{st} - \bar{Y})^2 \quad \Rightarrow$

$\hat{\bar{Y}}^2 = \bar{y}_{st}^2 - \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) W_i^2 s_i^2$.

Thus,

$$(N-1)\hat{S}^2 = \sum_{i=1}^{k} (N_i - 1) s_i^2 + N \left[\sum_{i=1}^{k} W_i \left\{ \bar{y}_i^2 - \left(\frac{1}{n_i} - \frac{1}{N_i} \right) s_i^2 \right\} - \left\{ \bar{y}_{st}^2 - \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) W_i^2 s_i^2 \right\} \right]$$

$$= \sum_{i=1}^{k} (N_i - 1) s_i^2 + N \left[\sum_{i=1}^{k} W_i \bar{y}_i^2 - \bar{y}_{st}^2 - \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) W_i s_i^2 + \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) W_i^2 s_i^2 \right]$$

$$= \sum_{i=1}^{k} (N_i - 1) s_i^2 + N \left[\sum_{i=1}^{k} W_i (\bar{y}_i - \bar{y}_{st})^2 - \sum_{i=1}^{k} W_i (1 - W_i) \left(\frac{1}{n_i} - \frac{1}{N_i} \right) s_i^2 \right]$$

$$= N \left[\frac{1}{N} \sum_{i=1}^{k} (N_i - 1) s_i^2 + \sum_{i=1}^{k} W_i (\bar{y}_i - \bar{y}_{st})^2 - \sum_{i=1}^{k} W_i (1 - W_i) \left(\frac{1}{n_i} - \frac{1}{N_i} \right) s_i^2 \right].$$

Therefore,

$$\hat{V}(\bar{y}_{sr}) = \frac{N-n}{nN} \left[\frac{N}{N-1} \left(\frac{1}{N} \sum_{i=1}^{k} N_i s_i^2 - \frac{1}{N} \sum_{i=1}^{k} s_i^2 + \sum_{i=1}^{k} W_i (\bar{y}_i - \bar{y}_{st})^2 \right.\right.$$

$$\left.\left. - \sum_{i=1}^{k} W_i (1 - W_i) s_i^2 / n_i + \sum_{i=1}^{k} W_i s_i^2 / N_i - \sum_{i=1}^{k} W_i^2 s_i^2 / N_i \right) \right].$$

Put $N_i = N W_i$

$$\hat{V}(\bar{y}_{sr}) = \frac{N-n}{n(N-1)} \left(\frac{1}{N} \sum_{i=1}^{k} N W_i s_i^2 - \frac{1}{N} \sum_{i=1}^{k} s_i^2 + \sum_{i=1}^{k} W_i (\bar{y}_i - \bar{y}_{st})^2 - \sum_{i=1}^{k} W_i (1 - W_i) s_i^2 / n_i \right.$$

$$\left. + \sum_{i=1}^{k} W_i s_i^2 / N W_i - \sum_{i=1}^{k} W_i^2 s_i^2 / N W_i \right)$$

$$= \frac{N-n}{n(N-1)}\left(\sum_{i=1}^{k} W_i\, s_i^2 + \sum_{i=1}^{k} W_i (\bar{y}_i - \bar{y}_{st})^2 - \sum_{i=1}^{k} W_i (1-W_i)\, s_i^2 / n_i - \frac{1}{N}\sum_{i=1}^{k} W_i\, s_i^2 \right)$$

$$= \frac{N-n}{n(N-1)}\left(1-\frac{1}{N}\right)\sum_{i=1}^{k} W_i s_i^2 + \frac{N-n}{n(N-1)}\left(\sum_{i=1}^{k} W_i (\bar{y}_i - \bar{y}_{st})^2 - \sum_{i=1}^{k} W_i (1-W_i)\, s_i^2 / n_i \right)$$

$$= \frac{N-n}{nN}\sum_{i=1}^{k} W_i s_i^2 + \frac{N-n}{n(N-1)}\left(\sum_{i=1}^{k} W_i (\bar{y}_i - \bar{y}_{st})^2 - \sum_{i=1}^{k} W_i (1-W_i)\, s_i^2 / n_i \right).$$

The estimate of the relative gain in precision due to stratification is thus obtained by

$$\frac{\hat{V}(\bar{y}_{sr}) - \hat{V}(\bar{y}_{st})}{\hat{V}(\bar{y}_{st})}.$$

Alternative result

$$\hat{V}(\bar{y}_{sr}) = \frac{N-n}{n(N-1)}\left(\frac{1}{N}\sum_{i=1}^{k} N W_i\, s_i^2 - \frac{1}{N}\sum_{i=1}^{k} s_i^2 + \sum_{i=1}^{k} W_i (\bar{y}_i - \bar{y}_{st})^2 - \sum_{i=1}^{k} W_i (1-W_i)\, s_i^2 / n_i \right.$$

$$\left. + \sum_{i=1}^{k} W_i\, s_i^2 / N W_i - \sum_{i=1}^{k} W_i^2\, s_i^2 / N W_i \right)$$

$$= \frac{N-n}{n(N-1)}\left(\sum_{i=1}^{k} W_i\, s_i^2 + \sum_{i=1}^{k} W_i (\bar{y}_i - \bar{y}_{st})^2 - \sum_{i=1}^{k} W_i\, s_i^2 / n_i + \sum_{i=1}^{k} W_i^2 s_i^2 / n_i - \frac{1}{N}\sum_{i=1}^{k} W_i\, s_i^2 \right)$$

$$= \frac{N-n}{n(N-1)}\left[\sum_{i=1}^{k} W_i (\bar{y}_i - \bar{y}_{st})^2 + \sum_{i=1}^{k} W_i\, s_i^2 \left(1 - \frac{1}{n_i} + \frac{W_i}{n_i} - \frac{1}{N} \right) \right].$$

Exercise: In a population with $N = 6$ and $k = 2$ the values of y_{ij} are 0, 1, 2 in stratum 1 and 4, 6, 11 in stratum 2. A sample with $n = 4$ is to be taken.

i) Show that the optimum n_i under Neyman allocation are $n_1 = 1$ and $n_2 = 3$.

ii) Compute the estimate $\bar{y}_{st}$ for every possible sample under optimum allocation and proportion allocation. Show that the estimates are unbiased. Hence find $V(\bar{y}_{st})$ directly under optimum and proportion allocation and verify that $V(\bar{y}_{st})$ under optimum agrees with the formula $V(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k} W_i S_i \right)^2 - \frac{1}{N}\sum_{i=1}^{k} W_i S_i^2 = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i} \right) W_i^2\, S_i^2$ and $V(\bar{y}_{st})$ under proportion agrees with the formula $V(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N} \right)\sum_{i=1}^{k} W_i\, S_i^2$.

Solution: Given $N = 6$, $n = 4$, $k = 2$, and $N_1 = N_2 = 3$, also $\bar{Y}_1 = 1$, and $\bar{Y}_2 = 7$.

i) Under Neyman allocation,

$n_i = n\dfrac{N_i S_i}{\sum_{i=1}^{k} N_i S_i}$, where $S_i^2 = \dfrac{1}{N_i - 1}\sum_{j=1}^{N_i}(y_{ij} - \bar{Y}_i)^2$, so that,

$S_1^2 = \dfrac{1}{N_1 - 1}\sum_{j=1}^{3}(y_{1j} - \bar{Y}_1)^2 = 1$, and $S_2^2 = \dfrac{1}{N_2 - 1}\sum_{j=1}^{3}(y_{2j} - \bar{Y}_2)^2 = 13$.

Therefore,

$n_1 = n\dfrac{N_1 S_1}{\sum_i N_i S_i} \cong 1$, and $n_2 = n\dfrac{N_{21} S_2}{\sum_i N_i S_i} \cong 3$.

ii) Possible samples under optimum allocation will be ${}^3C_1 \times {}^3C_3 = 3$, since $n_1 = 1$, $n_2 = 3$ and $N_1 = 3$, $N_2 = 3$

Samples		Means		
I	II	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_{st}$
0	(4, 6, 11)	0	7	3.5
1	(4, 6, 11)	1	7	4.0
2	(4, 6, 11)	2	7	4.5

$E(\bar{y}_{st}) = (3.5 + 4.0 + 4.5)/3 = 4 = \bar{Y}$, thus $\bar{y}_{st}$ is unbiased estimate of $\bar{Y}$ under optimum allocation.

$V(\bar{y}_{st}) = [(3.5 - 4)^2 + (4.0 - 4)^2 + (4.5 - 4)^2]/3 = 0.1667$.

$V(\bar{y}_{st}) = \sum_{i=1}^{k}\left(\dfrac{1}{n_i} - \dfrac{1}{N_i}\right)W_i^2 S_i^2 = \left(\dfrac{1}{n_1} - \dfrac{1}{N_1}\right)W_1^2 S_1^2 + \left(\dfrac{1}{n_2} - \dfrac{1}{N_2}\right)W_2^2 S_2^2 = 0.1667$.

$V(\bar{y}_{st}) = \dfrac{1}{n}\left(\sum_{i=1}^{k} W_i S_i\right)^2 - \dfrac{1}{N}\sum_{i=1}^{k} W_i S_i^2 = 0.1667$.

Possible samples under proportional allocation will be ${}^3C_2 \times {}^3C_2 = 9$, since $n_i = nW_i$, so that $n_1 = 2$, $n_2 = 2$.

Samples		Means		
I	II	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_{st}$
(0, 1)	(4, 6)	0.5	5.0	2.75
(0, 1)	(4, 11)	0.5	7.5	4.00
(0, 1)	(6, 11)	0.5	8.5	4.50
(0, 2)	(4, 6)	1.0	5.0	3.00
(0, 2)	(4, 11)	1.0	7.5	4.25
(0, 2)	(6, 11)	1.0	8.5	4.75

(1, 2)	(4, 6)	1.5	5.0	3.25
(1, 2)	(4, 11)	1.5	7.5	4.50
(1, 2)	(6, 11)	1.5	8.5	5.00

$$E(\bar{y}_{st}) = \frac{1}{9}(2.75 + 4.00 + 4.50 + 3.00 + 4.25 + 4.75 + 3.25 + 4.50 + 5.00) = 4 = \bar{Y}$$

Therefore, $\bar{y}_{st}$ is unbiased estimate of $\bar{Y}$ under proportion allocation.

$$V(\bar{y}_{st}) = \frac{1}{9}[(2.75 - 4)^2 + (4.00 - 4)^2 + \cdots + (5.00 - 4)^2] = 0.583.$$

By formula

$$V(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i S_i^2 = 0.583.$$

Exercise: The households in a town are to be sampled in order to estimate the average amount of assets per household. The households are stratified into a high-rent and low-rent stratum. A house in the high-rent stratum is thought to have about nine times as much assets as one in the low-rent stratum, and S_i is expected to be proportional to the square root of the stratum mean. There are 4000 households in the high-rent stratum and 20, 000 in the low-rent stratum.

i) Distribute a sample of 1000 households between the two strata.

ii) If the object is to estimate the difference between assets per household in the two strata, obtain the optimum sample sizes to be distributed in two strata such that $n_1 + n_2 = 1000$.

Solution:

Given $N_1 = 4000$, $N_2 = 20,000$, $W_1 = \frac{1}{6}$, and $W_2 = \frac{5}{6}$.

Also,

$$\bar{Y}_1 = 9\bar{Y}_2, \qquad S_1 \propto \sqrt{\bar{Y}_1}, \quad \Rightarrow \quad S_1 = A\sqrt{\bar{Y}_1}$$

and $S_2 \propto \sqrt{\bar{Y}_2}$, $\Rightarrow$ $S_2 = A\sqrt{\bar{Y}_2}$.

i) Since total sample size is fixed i.e. $n = 1000$, then the optimum value (under Neyman allocation) $n_i = n\frac{W_i S_i}{\sum_{i=1}^{k} W_i S_i}$, so that

$$n_1 = n\frac{W_1 S_1}{W_1 S_1 + W_2 S_2} = 1000 \times \frac{1/6(3A\sqrt{\bar{Y}_2})}{1/6(3A\sqrt{\bar{Y}_2}) + 5/6(A\sqrt{\bar{Y}_2})} = 375, \text{ and } \quad n_2 = 625.$$

ii) Unbiased estimate of $(\bar{Y}_1 - \bar{Y}_2)$ is $(\bar{y}_1 - \bar{y}_2)$, therefore,

$V(\bar{y}_1 - \bar{y}_2) = V(\bar{y}_1) + V(\bar{y}_2) - 0$, as sampling from strata are independent.

$$= \left(\frac{1}{n_1} - \frac{1}{N_1}\right) S_1^2 + \left(\frac{1}{n_2} - \frac{1}{N_2}\right) S_2^2 = \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right) + \text{ terms independent of } n_1 \text{ and } n_2 .$$

Now our problem is to find n_1 and n_2 such that variance of the estimate is minimum subject to condition $n_1 + n_2 = 1000$.

To determine the optimum value of n_i, consider the function

$$\phi = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + \lambda (n_1 + n_2 - 1000) . \tag{1}$$

where λ is some unknown constant. Using the calculus method of Lagrange multipliers, we select n_i and the constant λ to minimize ϕ .

Differentiating equation (1) with respect to n_i, we have

$$\frac{\partial \phi}{\partial n_1} = 0 = -\frac{S_1^2}{n_1^2} + \lambda \quad \Rightarrow \quad \lambda = \frac{S_1^2}{n_1^2} \tag{2}$$

$$\frac{\partial \phi}{\partial n_2} = 0 = -\frac{S_2^2}{n_2^2} + \lambda \quad \Rightarrow \quad \lambda = \frac{S_2^2}{n_2^2} \tag{3}$$

In view of equations (2) and (3), we get

$$\frac{S_1^2}{n_1^2} = \frac{S_2^2}{n_2^2} \qquad \Rightarrow \qquad \frac{S_1^2}{S_2^2} = \frac{n_1^2}{n_2^2} \qquad \text{and} \qquad \frac{S_1}{S_2} = \frac{n_1}{n_2} .$$

But from given values, we have

$$\frac{S_1}{S_2} = \frac{3A\sqrt{\bar{Y}_2}}{A\sqrt{\bar{Y}_2}} = 3 \qquad \Rightarrow \quad S_1 = 3S_2 \text{, and hence,}$$

$$\frac{n_1}{n_2} = \frac{3S_2}{S_2} = 3 \quad \Rightarrow \quad n_1 = 3n_2 .$$

Therefore,

$$3n_2 + n_2 = 1000 \qquad \Rightarrow \quad n_2 = 250 \quad \text{and} \quad n_1 = 750 .$$

Exercise: A sampler has two strata with relative sizes W_1, W_2. He believes that S_1, S_2 can be taken as equal but thinks that c_2 may be between $2c_1$ and $4c_1$. He would prefer to use proportional allocation but does not wish to incur a substantial increase in variance compared with optimum allocation. For a given cost $C = c_1 n_1 + c_2 n_2$, ignoring the *fpc*, show that

$$\frac{V(\bar{y}_{st})_{prop}}{V(\bar{y}_{st})_{opt}} = \frac{W_1 c_1 + W_2 c_2}{(W_1\sqrt{c_1} + W_2\sqrt{c_2})^2} .$$

If $W_1 = W_2$, compute the relative increases in variance from using proportional allocation when $c_2 / c_1 = 2, 4$.

Solution: We know that $V(\bar{y}_{st})$ under proportional allocation, ignoring *fpc* is

$$V(\bar{y}_{st})_{prop} = \frac{1}{n}\sum_{i=1}^{k} W_i S_i^2 = \frac{1}{n}(W_1 S_1^2 + W_2 S_2^2) = \frac{1}{n}S^2, \text{ as } S_1 = S_2 = S \text{ (say), and}$$

$$W_1 + W_2 = 1.$$

Under proportional allocation

$n_1 = nW_1$, and $n_2 = nW_2$, then $C = nW_1c_1 + nW_2c_2 = n(W_1c_1 + W_2c_2)$. So that

$n = \dfrac{C}{W_1c_1 + W_2c_2}$, and $V(\bar{y}_{st})_{prop} = \dfrac{1}{C}(W_1c_1 + W_2c_2)S^2$.

Note that, $V(\bar{y}_{st})$ under optimum allocation, ignoring *fpc* is

$$V(\bar{y}_{st})_{opt} = \frac{1}{C}\left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i}\right)^2 = \frac{1}{C}(W_1 S_1\sqrt{c_1} + W_2 S_2\sqrt{c_2})^2 = \frac{1}{C}(W_1\sqrt{c_1} + W_2\sqrt{c_2})^2 S^2.$$

Therefore,

$$\frac{V(\bar{y}_{st})_{prop}}{V(\bar{y}_{st})_{opt}} = \frac{\frac{1}{C}(W_1c_1 + W_2c_2)S^2}{\frac{1}{C}(W_1\sqrt{c_1} + W_2\sqrt{c_2})^2 S^2} = \frac{W_1c_1 + W_2c_2}{(W_1\sqrt{c_1} + W_2\sqrt{c_2})^2}.$$

The relative increase in variance from using proportional allocation is given by

$$RI = \frac{V(\bar{y}_{st})_{prop} - V(\bar{y}_{st})_{opt}}{V(\bar{y}_{st})_{opt}} = \frac{V(\bar{y}_{st})_{prop}}{V(\bar{y}_{st})_{opt}} - 1 = \frac{W_1c_1 + W_2c_2}{(W_1\sqrt{c_1} + W_2\sqrt{c_2})^2} - 1.$$

If $W_1 = W_2$, we have $W_1 = W_2 = 0.5$, since $W_1 + W_2 = 1$. Thus

$$RI = \frac{0.5c_1 + 0.5c_2}{(0.5\sqrt{c_1} + 0.5\sqrt{c_2})^2} - 1 = \frac{c_1 + c_2}{0.5(\sqrt{c_1} + \sqrt{c_2})^2} - 1.$$

i) When $\dfrac{c_2}{c_1} = 2$ or $c_2 = 2c_1$, then

$$RI = \frac{c_1 + 2c_1}{0.5(\sqrt{c_1} + \sqrt{2c_1})^2} - 1 = \frac{3c_1}{0.5c_1(1+\sqrt{2})^2} - 1 = 0.029437.$$

ii) When $\dfrac{c_2}{c_1} = 4$ or $c_2 = 4c_1$, then

$$RI = \frac{c_1 + 4c_1}{0.5(\sqrt{c_1} + 2\sqrt{c_1})^2} - 1 = \frac{5c_1}{0.5c_1(1+2)^2} - 1 = 0.11111.$$

Exercise: A sampler proposes to take a stratified random sample. He expects that his field costs will be of the form $\sum c_i n_i$. His advance estimates of relevant quantities for two strata are as follows:

Stratum	W_i	S_i	c_i
1	0.4	10	4
2	0.6	20	9

i) Find the values of $\frac{n_1}{n}$ and $\frac{n_2}{n}$ that minimize the total cost for a given value of $V(\bar{y}_{st})$.

ii) Find the sample size required, under this optimum allocation, to make $V(\bar{y}_{st}) = 1$, if *fpc* is ignored.

iii) Obtain the total fixed cost.

Solution:

i) The optimum value of n_i for given variance when cost is minimum are given by

$$n_i = n\frac{W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}} \Rightarrow \frac{n_i}{n} = \frac{W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}}, \text{ then}$$

$$\frac{n_1}{n} = \frac{W_1 S_1 / \sqrt{c_1}}{W_1 S_1 / \sqrt{c_1} + W_2 S_2 / \sqrt{c_2}} = \frac{1}{3}$$

and

$$\frac{n_2}{n} = \frac{W_2 S_2 / \sqrt{c_2}}{W_1 S_1 / \sqrt{c_1} + W_2 S_2 / \sqrt{c_2}} = \frac{2}{3}.$$

ii) We know that

$$V(\bar{y}_{st}) = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 S_i^2, \text{ if } fpc \text{ is ignored, then } V(\bar{y}_{st}) \text{ reduces to}$$

$$V(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} = \frac{W_1^2 S_1^2}{n_1} + \frac{W_2^2 S_2^2}{n_2} = \frac{0.16 \times 100 \times 3}{n} + \frac{0.36 \times 400 \times 3}{2n} = \frac{264}{n}.$$

It is given that $V(\bar{y}_{st}) = 1$, so that, $n = 264$.

Therefore,

$$n_1 = \frac{n}{3} = 88, \text{ and } n_2 = 176.$$

Or

We know that the optimum value of n_i for given variance are

$$n_i = \frac{(W_i S_i / \sqrt{c_i}) \sum_i W_i S_i \sqrt{c_i}}{V + \frac{1}{N}\sum_i W_i S_i^2}.$$

For large N, and $V(\bar{y}_{st}) = 1$, it reduces to

$$n_i = (W_i S_i / \sqrt{c_i}) \sum_i W_i S_i \sqrt{c_i} \ .$$

Therefore, after simplification, $n_1 = 88$, and $n_2 = 176$.

iii) Cost function is given as $\sum c_i n_i$, then, $\sum c_i n_i = c_1 n_1 + c_2 n_2 = 1936$.

Exercise: After the sample in previous exercise is taken, the sampler finds that his field costs were actually \$2 per unit in stratum 1 and \$12 in stratum 2.

i) How much greater is the field cost than anticipated?

ii) If he had known the correct field costs in advance, could he have attained $V(\bar{y}_{st}) = 1$ for the original estimated field cost in previous exercise?

Solution:

i) The correct field cost $= c_1 n_1 + c_2 n_2 = 2 \times 88 + 12 \times 176 = 2288$.

ii) By Cauchy-Schwartz inequality

$$V'C' \geq \left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \right)^2 \text{, where} \qquad V' = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} \text{, and } C' = \sum_{i=1}^{k} n_i c_i$$

Thus, to get $V' = 1$, the minimum cost will be

$$C' = (W_1 S_1 \sqrt{c_1} + W_2 S_2 \sqrt{c_2})^2 = (0.4 \times 10\sqrt{2} + 0.6 \times 20\sqrt{12})^2 \cong 2230 \ .$$

Or

Note that, optimum cost for fix variance, ignoring *fpc*, is

$$C = \frac{1}{V} \left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \right)^2 = \left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \right)^2 \cong 2230 \text{, as } V(\bar{y}_{st}) = 1 \ .$$

Exercise: With two strata, a sampler would like to have $n_1 = n_2$ (equal allocation) for administrative convenience, instead of using the values given by Neyman allocation. If $V(\bar{y}_{st})$, and $V(\bar{y}_{st})_{opt}$ denotes the variances of equal allocation and the Neyman allocation, respectively, show that the fractional increase in variance $\dfrac{V(\bar{y}_{st}) - V(\bar{y}_{st})_{opt}}{V(\bar{y}_{st})_{opt}} = \left(\dfrac{r-1}{r+1} \right)^2$, where $r = \dfrac{n_1}{n_2}$ as given by Neyman allocation i.e. $r = \left(\dfrac{n_1}{n_2} \right)_{opt}$.

Solution: We know that $V(\bar{y}_{st}) = \sum_{i=1}^{k} \left(\dfrac{1}{n_i} - \dfrac{1}{N_i} \right) W_i^2 \, S_i^2$, then variance of equal allocation $(n_1 = n_2 = n')$, if *fpc* is ignored, for two strata reduces as

$$V(\bar{y}_{st}) = \frac{W_1^2 \, S_1^2}{n'} + \frac{W_2^2 \, S_2^2}{n'} = \frac{1}{n'} (W_1^2 \, S_1^2 + W_2^2 \, S_2^2)$$

and variance under Neyman allocation (for fixed n), is

$$V(\bar{y}_{st})_{opt} = \frac{1}{n}\left(\sum_{i=1}^{k} W_i S_i\right)^2 = \frac{1}{2n'}(W_1 S_1 + W_2 S_2)^2 .$$

For fixed n optimum allocation reduces Neyman allocation, so that

$$n_i = 2n' \frac{W_i S_i}{\sum_{i=1}^{k} W_i S_i}, \text{ so } n_1 = 2n' \frac{W_1 S_1}{\sum_{i=1}^{k} W_i S_i}, \text{ and } n_2 = 2n' \frac{W_2 S_2}{\sum_{i=1}^{k} W_i S_i}, \text{ then,}$$

$$\left(\frac{n_1}{n_2}\right)_{opt} = \frac{W_1 S_1}{W_2 S_2} = r \text{ (given).}$$

Therefore,

$$\frac{V(\bar{y}_{st}) - V(\bar{y}_{st})_{opt}}{V(\bar{y}_{st})_{opt}} = \frac{\frac{1}{n'}(W_1^2 S_1^2 + W_2^2 S_2^2) - \frac{1}{2n'}(W_1 S_1 + W_2 S_2)^2}{\frac{1}{2n'}(W_1 S_1 + W_2 S_2)^2}$$

$$= \frac{2(W_1^2 S_1^2 + W_2^2 S_2^2) - (W_1 S_1 + W_2 S_2)^2}{(W_1 S_1 + W_2 S_2)^2}$$

$$= \frac{2W_1^2 S_1^2 + 2W_2^2 S_2^2 - W_1^2 S_1^2 - W_2^2 S_2^2 - 2W_1 W_2 S_1 S_2}{(W_1 S_1 + W_2 S_2)^2}$$

$$= \frac{(W_1 S_1 - W_2 S_2)^2}{(W_1 S_1 + W_2 S_2)^2} = \frac{\left(\frac{W_1 S_1}{W_2 S_2} - 1\right)^2}{\left(\frac{W_1 S_1}{W_2 S_2} + 1\right)^2} = \frac{(r-1)^2}{(r+1)^2} = \left(\frac{r-1}{r+1}\right)^2 .$$

Exercise: If the cost function is the form $C = c_0 + \sum_{i=1}^{k} c_i \sqrt{n_i}$, where c_0 and c_i are known numbers, then

i) Show that in order to minimize $V(\bar{y}_{st})$ for fixed total cost, n_i must be proportional to $\left(\frac{W_i^2 S_i^2}{c_i}\right)^{2/3}$.

ii) Find the n_i for a sample of size 1000 under the following conditions:

Stratum	W_i	S_i	c_i
1	0.4	4	1
2	0.3	5	2
3	0.2	6	4

Solution:

i) We have $V(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} - \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{N_i}$

To determine n_i such that $V(\bar{y}_{st})$ is minimum, and cost $C = c_0 + \sum_{i=1}^{k} c_i \sqrt{n_i}$ is fixed (given), we consider the function

$$\phi = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} - \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{N_i} + \lambda \left(c_0 + \sum_{i=1}^{k} c_i \sqrt{n_i} - C \right). \quad (1)$$

where λ is some unknown constant. Using the calculus method of Lagrange multipliers, we select n_i and the constant λ to minimize ϕ.

Differentiating equation (1) with respect to n_i, we have

$$\frac{\partial \phi}{\partial n_i} = 0 = -\frac{W_i^2 S_i^2}{n_i^2} + \frac{1}{2} \lambda c_i (n_i)^{-1/2} \Rightarrow \frac{W_i^2 S_i^2}{n_i^2} = \frac{1}{2} \lambda c_i (n_i)^{-1/2}$$

or $(n_i)^{3/2} = \frac{2 W_i^2 S_i^2}{\lambda c_i}$ or $n_i = \left(\frac{2}{\lambda} \right)^{2/3} \left(\frac{W_i^2 S_i^2}{c_i} \right)^{2/3}$

and hence,

$$n_i \propto \left(\frac{W_i^2 S_i^2}{c_i} \right)^{2/3}, \text{ since } \left(\frac{2}{\lambda} \right)^{2/3} \text{ is constant.}$$

ii) We have $n_i = \left(\frac{2}{\lambda} \right)^{2/3} \left(\frac{W_i^2 S_i^2}{c_i} \right)^{2/3}$ (2)

Taking summation over all strata, we get

$$\sum_{i=1}^{k} n_i = \left(\frac{2}{\lambda} \right)^{2/3} \sum_{i=1}^{k} \left(\frac{W_i^2 S_i^2}{c_i} \right)^{2/3} \Rightarrow \left(\frac{2}{\lambda} \right)^{2/3} = \frac{n}{\sum_{i=1}^{k} \left(\frac{W_i^2 S_i^2}{c_i} \right)^{2/3}} \quad (3)$$

Substitute equation (3) in equation (2), we get

$$n_i = \frac{n}{\sum_{i=1}^{k} \left(\frac{W_i^2 S_i^2}{c_i} \right)^{2/3}} \left(\frac{W_i^2 S_i^2}{c_i} \right)^{2/3}.$$

Therefore,

$$n_1 = \frac{1000}{(2.56)^{2/3} + (1.125)^{2/3} + (0.36)^{2/3}} \times (2.56)^{2/3} = 541, \; n_2 = 313, \text{ and } n_3 = 146.$$

Stratified random sampling for proportion

Theory for estimating population mean $\bar{Y}$ or population total Y on the basis of stratified sampling with *srswor* and *srswr* in the strata can easily be applied to the estimation of a population proportion say P by taking the population values of y_{ij} as 1 or 0 according as the unit belong to that class or possesses a particular character C, then

$\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij} = P_i$, proportion based on N_i units (stratum proportion)

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} y_{ij} = \frac{1}{N}\sum_{i=1}^{k} N_i P_i = \sum_{i=1}^{k} W_i P_i = P$, over all population proportion

$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = p_i$, sample proportion based on n_i units

$\sigma_i^2 = \frac{1}{N_i}\sum_{j=1}^{N_i}(y_{ij} - P_i)^2 = P_i - P_i^2 = P_i Q_i$, stratum variance of proportion based on N_i units

$S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(y_{ij} - P_i)^2 = \frac{N_i}{N_i - 1} P_i Q_i$, stratum mean square of proportion based on N_i units

$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - p_i)^2 = \frac{n_i}{n_i - 1} p_i q_i$, sample mean square of proportion based on n_i units

Theorem: In stratified random sampling, *wor*, an unbiased estimate of the over all population proportion is given by $p_{st} = \sum_{i=1}^{k} W_i p_i$ with its variance $V(p_{st}) = \sum_{i=1}^{k} W_i^2 \left(\frac{N_i - n_i}{N_i - 1}\right)\frac{P_i Q_i}{n_i}$, where p_i is the sample estimate of proportion P_i in the $i-$th stratum.

Proof: Since sampling within each stratum is simple random sampling, so that $E(p_i) = P_i$, it follows that

$E(p_{st}) = \sum_{i=1}^{k} W_i E(p_i) = \sum_{i=1}^{k} W_i P_i = P$. To obtain the variance, we have

$V(p_{st}) = E[p_{st} - E(p_{st})]^2 = \sum_{i=1}^{k} W_i^2 V(p_i) = \sum_{i=1}^{k} W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right)\frac{N_i}{N_i - 1} P_i Q_i$, as sampling is *srwor* within each stratum.

$= \sum_{i=1}^{k} W_i^2 \left(\frac{N_i - n_i}{n_i N_i}\right)\frac{N_i}{N_i - 1} P_i Q_i = \sum_{i=1}^{k} W_i^2 \left(\frac{N_i - n_i}{N_i - 1}\right) P_i Q_i / n_i$.

Corollary: If stratified random sampling is with replacement, then the variance is $V(p_{st}) = \sum_{i=1}^{k} W_i^2 P_i Q_i / n_i$.

Theorem: In stratified random sampling, *wor*, an unbiased estimate of $V(p_{st}) = \sum_{i=1}^{k} W_i^2 \left(\frac{N_i - n_i}{N_i - 1} \right) \frac{P_i Q_i}{n_i}$ is $\hat{V}(p_{st}) = v(p_{st}) = \frac{1}{N} \sum_{i=1}^{k} (N_i - n_i) W_i \frac{p_i q_i}{n_i - 1}$.

Proof: $E[\hat{V}(p_{st})] = E\left[\frac{1}{N} \sum_{i=1}^{k} (N_i - n_i) W_i \frac{p_i q_i}{n_i - 1} \right] = \frac{1}{N} E\left[\sum_{i=1}^{k} (N_i - n_i) \frac{W_i}{n_i} \frac{n_i p_i q_i}{n_i - 1} \right]$

$$= \frac{1}{N} \sum_{i=1}^{k} (N_i - n_i) \frac{W_i}{n_i} E\left(\frac{n_i p_i q_i}{n_i - 1} \right)$$

$$= \frac{1}{N} \sum_{i=1}^{k} (N_i - n_i) \frac{W_i}{n_i} \frac{N_i P_i Q_i}{N_i - 1}, \text{ since } E(s_i^2) = S_i^2 \text{ with } srswor .$$

$$= \sum_{i=1}^{k} \left(\frac{N_i - n_i}{N_i - 1} \right) W_i^2 \frac{P_i Q_i}{n_i} .$$

Corollary: With stratified random sampling, *wr*, an unbiased estimate of $V(p_{st}) = \sum_{i=1}^{k} W_i^2 P_i Q_i / n_i$ is $\hat{V}(p_{st}) = \sum_{i=1}^{k} W_i^2 \; p_i q_i / n_i - 1$.

\

UNIT-III

CLUSTER SAMPLING

In random sampling, it is presumed (to suppose) that the population has been divided into a finite number of distinct and identifiable units called the **sampling units**. The smallest units into which the population can be divided are called the **elements** of the population, and a group of such elements is known as a **cluster**. After dividing the population into specified cluster (as a simple rule, the number of elements in a cluster should be small and the number of cluster should be large), the required number of clusters are then obtained either by the method of equal or unequal probabilities of selection, such procedure, when the sampling units is a cluster, is called **cluster sampling**. If the entire area containing the population under study is subdivided into smaller area segments, and each element in the population is associated with one and only one such area segment, the procedure is alternatively called **area sampling**. There are two main reasons for using cluster as a sampling unit.

i) Usually a complete list of the population units is not available and therefore the use of individual unit as sampling unit is not feasible.

ii) Even when a complete list of the population units is available, by using cluster as sampling unit the cost of sampling can be reduced considerably.

For instance, in a population survey it may be cheaper to collect data from all persons in a sample of households than from a sample of the same number of persons selected directly from all the persons. Similarly, it would be operationally more convenient to survey all households situated in a sample of areas such as villages than to survey a sample of the same number of households selected at random from a list of all households. Another example of the utility of cluster sampling is provided by crop survey, where locating a randomly selected farms or plot requires a considerable part the total time taken for the survey, but once the plot is located, the time taken for identifying and surveying a few neighbouring plots will generally be only marginal.

Theory of equal clusters

Suppose the population consists of N clusters, each of M elements. A sample of n clusters is drawn by the method of simple random sampling and every unit in the selected clusters is enumerated. Let us denote by

y_{ij}, value of the $j-$th element in the $i-$th cluster, $j=1,\cdots,M$; $i=1,\cdots,N$.

$\bar{y}_{i.}=\frac{1}{M}\sum_{j=1}^{M}y_{ij}$, mean per element of the $i-$th cluster.

$\bar{Y}_N=\frac{1}{N}\sum_{i=1}^{N}\bar{y}_{i.}$, mean of cluster means in the population of N clusters.

$\bar{Y}=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}$, mean per element in the population.

$\bar{y}_n=\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.}$, mean of cluster means in a sample of n clusters.

$\bar{y} = \frac{1}{nM}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij}$, mean per element in the sample.

Note: $\bar{Y}_N = \bar{Y}$, and $\bar{y}_n = \bar{y}$, if size of clusters are same.

$S_i^2 = \frac{1}{M-1}\sum_{j=1}^{M}(y_{ij} - \bar{y}_{i.})^2$, mean square between elements within the $i-$th cluster.

$S_w^2 = \frac{1}{N}\sum_{i=1}^{N} S_i^2$, mean square within clusters.

$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_{i.} - \bar{Y}_N)^2$, mean square between cluster means in the population.

$S^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{Y})^2$, mean square between elements in the population.

$$\rho = \frac{E(y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{E(y_{ij} - \bar{Y})^2} = \frac{\frac{1}{NM(M-1)}\sum_{i=1}^{N}\sum_{j\neq k}^{M}(y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{Y})^2}$$

$= \frac{\sum_{i=1}^{N}\sum_{j\neq k}^{M}(y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{(M-1)(NM-1)S^2}$, intracluster correlation coefficient between elements with in clusters.

Theorem: A simple random sample, *wor*, of n clusters each having M elements is drawn from a population of N clusters, the sample mean $\bar{y}_n$ is an unbiased estimator of population mean $\bar{Y}$ and its variance is $V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right)S_b^2 = \frac{1-f}{n}S_b^2$.

Proof: We have,

$$E(\bar{y}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.}\right) = \frac{1}{n}\sum_{i=1}^{n}E(\bar{y}_{i.}) = \frac{1}{N}\sum_{i=1}^{N}\bar{y}_{i.} = \bar{Y}_N = \bar{Y}\,.$$

To obtain the variance, we have, by definition

$$V(\bar{y}_n) = E(\bar{y}_n - \bar{Y}_N)^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.} - \frac{n}{n}\bar{Y}_N\right)^2 = \frac{1}{n^2}E\left(\sum_{i=1}^{n}(\bar{y}_{i.} - \bar{Y}_N)\right)^2$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n}E(\bar{y}_{i.} - \bar{Y}_N)^2 + \sum_{i\neq i'}^{n}E(\bar{y}_{i.} - \bar{Y}_N)(\bar{y}_{i'.} - \bar{Y}_N)\right)$$

Consider

$$E(\bar{y}_{i.}-\bar{Y}_N)^2=\frac{1}{N}\sum_{i=1}^{N}(\bar{y}_{i.}-\bar{Y}_N)^2=\frac{N-1}{N}S_b^2\ . \tag{3.1}$$

and

$$E(\bar{y}_{i.}-\bar{Y}_N)(\bar{y}_{i'.}-\bar{Y}_N)=\frac{1}{N(N-1)}\sum_{i\neq i'}^{N}(\bar{y}_{i.}-\bar{Y}_N)(\bar{y}_{i'.}-\bar{Y}_N)$$

$$=\frac{1}{N(N-1)}\left[\sum_{i=1}^{N}(\bar{y}_{i.}-\bar{Y}_N)\left\{\sum_{i'=1}^{N}(\bar{y}_{i'.}-\bar{Y}_N)-(\bar{y}_{i.}-\bar{Y}_N)\right\}\right]$$

$$=\frac{1}{N(N-1)}\left[\sum_{i=1}^{N}(\bar{y}_{i.}-\bar{Y}_N)\sum_{i'=1}^{N}(\bar{y}_{i'.}-\bar{Y}_N)-\sum_{i=1}^{N}(\bar{y}_{i.}-\bar{Y}_N)^2\right]$$

$$=-\frac{1}{N(N-1)}\sum_{i=1}^{N}(\bar{y}_{i.}-\bar{Y}_N)^2=-\frac{1}{N}S_b^2 \tag{3.2}$$

In view of equations (3.1) and (3.2), $V(\bar{y}_n)$ reduces to

$$V(\bar{y}_n)=\frac{1}{n^2}\left[\sum_{i=1}^{n}\frac{N-1}{N}S_b^2+\sum_{i\neq i'}^{n}\left(-\frac{1}{N}S_b^2\right)\right]=\frac{1}{n^2}\left[\frac{n(N-1)}{N}S_b^2-\frac{n(n-1)}{N}S_b^2\right]$$

$$=\frac{N-n}{nN}S_b^2=\frac{1-f}{n}S_b^2\ .$$

Note: For large N, $V(\bar{y}_n)=\frac{1}{n}S_b^2$.

Alternative expression of $V(\bar{y}_n)$ interms of correlation coefficient

Consider the intracluster correlation coefficient between elements within clusters and is defined as

$$\rho=\frac{E(y_{ij}-\bar{Y})(y_{ik}-\bar{Y})}{E(y_{ij}-\bar{Y})^2}=\frac{\sum_{i=1}^{N}\sum_{j\neq k}^{M}(y_{ij}-\bar{Y})(y_{ik}-\bar{Y})}{(M-1)(NM-1)S^2}$$

$$\Rightarrow \quad \sum_{i=1}^{N}\sum_{j\neq k}^{M}(y_{ij}-\bar{Y})(y_{ik}-\bar{Y})=(M-1)(NM-1)\rho S^2\ .$$

By definition,

$$V(\bar{y}_n)=\frac{1-f}{n}S_b^2=\frac{1-f}{n(N-1)}\sum_{i=1}^{N}(\bar{y}_{i.}-\bar{Y}_N)^2 \tag{3.3}$$

Consider

$$\sum_{i=1}^{N}(\bar{y}_{i.}-\bar{Y}_N)^2=\sum_{i=1}^{N}\left(\frac{1}{M}\sum_{j=1}^{M}y_{ij}-\frac{M}{M}\bar{Y}_N\right)^2=\frac{1}{M^2}\sum_{i=1}^{N}\left(\sum_{j=1}^{M}(y_{ij}-\bar{Y}_N)\right)^2$$

$$=\frac{1}{M^2}\left(\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{Y})^2+\sum_{i=1}^{N}\sum_{j\neq k}^{M}(y_{ij}-\bar{Y})(y_{ik}-\bar{Y})\right), \text{ as } \bar{Y}_N=\bar{Y} \qquad (3.4)$$

$$=\frac{1}{M^2}[(NM-1)S^2+(M-1)(NM-1)\rho S^2]$$

$$=\frac{(NM-1)S^2}{M^2}[1+(M-1)\rho] \qquad (3.5)$$

Substitute the values of equation (3.5) in equation (3.3), we get

$$V(\bar{y}_n)=\frac{1-f}{n}\left(\frac{(NM-1)S^2}{M^2(N-1)}\right)[1+(M-1)\rho].$$

Note: For large N, $\frac{1}{N}\to 0$, so that $(1-f)\to 1$, and $\frac{NM-1}{M^2(N-1)}=\frac{N(M-1/N)}{NM^2(1-1/N)}=\frac{1}{M}$.

Hence,

$$V(\bar{y}_n)=\frac{S^2}{nM}[1+(M-1)\rho].$$

Corollary: $\hat{Y}=NM\,\bar{y}_n$ is an unbiased estimate of the population total Y, and its variance

$$V(\hat{Y})=N^2M^2\left(\frac{1-f}{n}\right)S_b^2=N^2\left(\frac{1-f}{n}\right)\frac{(NM-1)S^2}{N-1}[1+(M-1)\rho]$$

$$\cong N^2M\left(\frac{1-f}{n}\right)S^2[1+(M-1)\rho], \text{ for large } N.$$

Estimation of variance $V(\bar{y}_n)$

Define,

$$s_b^2=\frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_{i.}-\bar{y}_n)^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}\bar{y}_{i.}^{\,2}-n\bar{y}_n^{\,2}\right), \text{ then}$$

$$E(s_b^2)=\frac{1}{n-1}\left(\sum_{i=1}^{n}E(\bar{y}_{i.}^{\,2})-nE(\bar{y}_n^{\,2})\right)$$

Note that,

$$V(\bar{y}_{i.})=E(\bar{y}_{i.}^{\,2})-\bar{Y}_N^{\,2}, \text{ so that}$$

$$E(\bar{y}_{i.}^{\,2})=\left(\frac{N-1}{N}\right)S_b^2+\bar{Y}_N^{\,2}. \qquad (3.6)$$

and

$V(\bar{y}_n) = E(\bar{y}_n{}^2) - \bar{Y}_N{}^2$, so that

$$E(\bar{y}_n{}^2) = \left(\frac{N-n}{nN}\right) S_b^2 + \bar{Y}_N{}^2. \tag{3.7}$$

In view of equations (3.7), and (3.6), $E(s_b^2)$ reduces as

$$E(s_b^2) = \frac{1}{n-1}\left[n\left(\frac{N-1}{N}\right) S_b^2 - n\left(\frac{N-n}{nN}\right) S_b^2 \right] = \frac{1}{n-1}\left(\frac{nN-n-N+n}{N}\right) S_b^2 = S_b^2.$$

This shows that s_b^2 is an unbiased estimate of S_b^2. Hence $v(\bar{y}_n) = \frac{1-f}{n} s_b^2$ is an unbiased estimator of $V(\bar{y}_n) = \frac{1-f}{n} S_b^2$.

Relative efficiency (RE) of cluster sampling

In sampling of nM elements from the population by simple random sampling, wor, the variance of the sample mean $\bar{y}$ is given by

$$V(\bar{y}_{sr}) = \left(\frac{NM-nM}{NM}\right)\frac{S^2}{nM} = \frac{1-f}{nM} S^2, \text{ and } \quad V(\bar{y}_n) = \frac{1-f}{n} S_b^2.$$

Thus, the relative efficiency of cluster sampling compared with simple random sampling is given by

$RE = \frac{V(\bar{y}_{sr})}{V(\bar{y}_n)} = \frac{S^2}{M\, S_b^2}$. This shows that the efficiency of cluster sampling increases as the mean square between clusters means S_b^2 decreases.

Note: For large N, the relative efficiency of cluster sampling in terms of intracluster correlation coefficient ρ is given by

$$RE = \frac{V(\bar{y}_{sr})}{V(\bar{y}_n)} = \frac{1}{1+(M-1)\rho}.$$

It can be seen that the relative efficiency depends on the value of ρ, if

i) $\rho = 0$, then $V(\bar{y}_{sr}) = V(\bar{y}_n)$, i.e. both methods are equally precise.

ii) $\rho > 0$, then $V(\bar{y}_{sr}) < V(\bar{y}_n)$, i.e. simple random sampling is more precise.

iii) $\rho < 0$, then $V(\bar{y}_{sr}) > V(\bar{y}_n)$, i.e. cluster sampling is more precise.

Estimation of relative efficiency of cluster sampling

We have,

$Est.(RE) = \dfrac{Est.S^2}{M\ Est.S_b^2}$, here s^2 will not be a unbiased estimate of S^2 i.e. $E(s^2) \neq S^2$, because a sample of nM elements is not taken randomly from the population of NM elements. To find unbiased estimate of S^2, consider

$$(NM-1)S^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{Y})^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.}+\bar{y}_{i.}-\bar{Y})^2$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M}[(y_{ij}-\bar{y}_{i.})^2 + (\bar{y}_{i.}-\bar{Y})^2 + 2(y_{ij}-\bar{y}_{i.})(\bar{y}_{i.}-\bar{Y})]$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})^2 + M\sum_{i=1}^{N}(\bar{y}_{i.}-\bar{Y})^2 + 0 = (M-1)\sum_{i=1}^{N}S_i^2 + M(N-1)S_b^2$$

$$= N(M-1)S_w^2 + M(N-1)S_b^2. \qquad (3.8)$$

It can be seen that in a random sample of n clusters, s_b^2 and s_w^2 will provide unbiased estimates of S_b^2 and S_w^2, respectively.

Define,

$$s_w^2 = \frac{1}{n(M-1)}\sum_{i=1}^{n}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})^2 \text{, and} \qquad s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_{i.}-\bar{y}_n)^2.$$

Consider

$$s_w^2 = \frac{1}{n(M-1)}\sum_{i=1}^{n}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})^2 = \frac{1}{n(M-1)}\left(\sum_{i=1}^{n}\sum_{j=1}^{M}y_{ij}^2 - M\sum_{i=1}^{n}\bar{y}_{i.}^2\right), \text{ so that}$$

$$E(s_w^2) = \frac{1}{n(M-1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{M}E(y_{ij}^2) - M\sum_{i=1}^{n}E(\bar{y}_{i.}^2)\right]$$

Note that

$V(y_{ij}) = E(y_{ij}^2) - \bar{Y}_N^2$, then

$E(y_{ij}^2) = \dfrac{(NM-1)}{NM}S^2 + \bar{Y}_N^2$. Similarly, we can see, $E(\bar{y}_{i.}^2) = \dfrac{(N-1)}{N}S_b^2 + \bar{Y}_N^2$.

Therefore,

$$E(s_w^2) = \frac{1}{n(M-1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{M}\left\{\frac{(NM-1)}{NM}S^2 + \bar{Y}_N^2\right\} - M\sum_{i=1}^{n}\left\{\frac{(N-1)}{N}S_b^2 + \bar{Y}_N^2\right\}\right]$$

$$= \frac{1}{n(M-1)}\left[nM\,\frac{(NM-1)}{NM} S^2 + nM\,\bar{Y}_N^2 - nM\,\frac{(N-1)}{N} S_b^2 - nM\,\bar{Y}_N^2 \right]$$

$$= \frac{1}{N(M-1)}[(NM-1)S^2 - M(N-1)S_b^2]$$

$$= \frac{1}{N(M-1)}[N(M-1)S_w^2] = S_w^2, \text{ by using relation, which is given in equation (3.8).}$$

and

$$E(s_b^2) = S_b^2, \text{ as } n \text{ clusters are drawn under } srswor.$$

Thus, an unbiased estimate of S^2 will be

$$\hat{S}^2 = \frac{1}{NM-1}[N(M-1)s_w^2 + M(N-1)s_b^2].$$

Therefore,

$$Est(RE) = \frac{\frac{1}{NM-1}[N(M-1)s_w^2 + M(N-1)s_b^2]}{M s_b^2}.$$

Note: For large N,

$$Est.(RE) = \frac{\frac{1}{N(M-1/N)}[N(M-1)s_w^2 + M(N-1)s_b^2]}{M s_b^2}$$

$$= \frac{\frac{1}{NM}[N(M-1)s_w^2 + NM(1-1/N)s_b^2]}{M s_b^2} = \frac{(M-1)s_w^2 + M s_b^2}{M^2 s_b^2}.$$

Estimation of ρ

For large N, $RE = \frac{1}{1+(M-1)\rho} = E$ (say), so that

$$\hat{E} + (M-1)\hat{E}\hat{\rho} = 1, \text{ where } \hat{E} = \frac{(M-1)s_w^2 + M s_b^2}{M^2 s_b^2}$$

or
$$\hat{\rho} = \frac{1-\hat{E}}{(M-1)\hat{E}} = \frac{1 - \frac{1}{M^2 s_b^2}[(M-1)s_w^2 + M s_b^2]}{(M-1)\left(\frac{1}{M^2 s_b^2}[(M-1)s_w^2 + M s_b^2]\right)} = \frac{M^2 s_b^2 - (M-1)s_w^2 - M s_b^2}{(M-1)[(M-1)s_w^2 + M s_b^2]}$$

$$= \frac{M(M-1)s_b^2 - (M-1)s_w^2}{(M-1)[(M-1)s_w^2 + M s_b^2]} = \frac{M s_b^2 - s_w^2}{(M-1)s_w^2 + M s_b^2}.$$

Alternative method

We have,

$$\rho = \frac{\frac{1}{M-1}\sum_{i=1}^{N}\sum_{j\neq k}^{M}(y_{ij}-\bar{Y})(y_{ik}-\bar{Y})}{(NM-1)S^2}, \text{and } (NM-1)S^2 = N(M-1)S_w^2 + M(N-1)S_b^2$$

.

Note that, from equation (3.4)

$$M^2\sum_{i=1}^{N}(\bar{y}_{i.}-\bar{Y}_N)^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{Y})^2 + \sum_{i=1}^{N}\sum_{j\neq k}^{M}(y_{ij}-\bar{Y})(y_{ik}-\bar{Y})$$

or $$\sum_{i=1}^{N}\sum_{j\neq k}^{M}(y_{ij}-\bar{Y})(y_{ik}-\bar{Y}) = M^2\sum_{i=1}^{N}(\bar{y}_{i.}-\bar{Y})^2 - \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{Y})^2$$

$$= M^2(N-1)S_b^2 - (NM-1)S^2 = M^2(N-1)S_b^2 - N(M-1)S_w^2 - M(N-1)S_b^2$$

$$= M(N-1)S_b^2(M-1) - N(M-1)S_w^2.$$

Hence,

$$\rho = \frac{M(N-1)S_b^2 - NS_w^2}{M(N-1)S_b^2 + N(M-1)S_w^2}.$$

It can be seen that in a random sample of n clusters, s_b^2, and s_w^2 will provide unbiased estimate of S_b^2, and S_w^2 respectively. Therefore, an estimator of ρ will be

$$\hat{\rho} = \frac{M(N-1)s_b^2 - Ns_w^2}{M(N-1)s_b^2 + N(M-1)s_w^2}, \text{ and for large } N, \hat{\rho} = \frac{Ms_b^2 - s_w^2}{Ms_b^2 + (M-1)s_w^2}.$$

Determination of optimum cluster size

The best size of the cluster to use depends on the cost of collecting information from clusters and the resulting variance. Regarding the variance function, it is found that variability between elements within clusters increases as the size of cluster increases (this means that large clusters are found to be more heterogeneous than small clusters) and decreases with increasing number of clusters. On the other hand, the cost decreases as the size of cluster increases and increases with the number of clusters increases. Hence, it is necessary to determine a balancing point by finding out the optimum cluster size and the number of clusters in the samples, which can minimize the sampling variance for a given cost or, alternatively, minimize the cost for a fixed variance.

i) The cost of a survey, apart from overhead cost, will be made up of two components.

ii) Cost due to expenses in enumerating the elements in the sample and in travelling within the cluster, which is proportional to the number of elements in the sample.

iii) Cost due to expenses on travelling between clusters, which is proportional to the distance to be travelled between clusters. It has been shown empirically that the

expected value of minimum distance between n points located at random is proportional to $\sqrt{n}$.

The cost of a survey can be, therefore expressed as

$$C = c_1 nM + c_2\sqrt{n},$$

where c_1 is the cost of collecting information from an element within the cluster and c_2 is the cost per unit distance travelled between clusters. In various agricultural surveys it has been observed that S_w^2 is related to M by the relation $S_w^2 = a M^g$, $g > 0$, where a and g are positive constant, then

$$S_b^2 = \frac{(NM-1)S^2 - N(M-1)aM^g}{M(N-1)} = S^2 - (M-1)aM^{g-1}, \text{ for large } N.$$

Thus, the variance $V(\bar{y}_n)$ for large N, reduces as

$$V(\bar{y}_n) = \frac{1}{n}[S^2 - (M-1)aM^{g-1}].$$

The problem is to determine n and M such that for specified cost, the variance of $\bar{y}_n$ is a minimum. Using calculus methods we form

$$\phi = V(\bar{y}_n) + \lambda(c_1 nM + c_2\sqrt{n} - C),$$

where λ is an unknown constant. Differentiating with respect to n and M respectively, and equating the results to zero, we obtain

$$\frac{\partial\phi}{\partial n} = 0 = -\frac{1}{n^2}[S^2 - (M-1)aM^{g-1}] + \lambda\left(c_1 M + \frac{c_2}{2\sqrt{n}}\right), \text{ so that}$$

$$\frac{1}{n}V(\bar{y}_n) = \lambda\left(c_1 M + \frac{c_2}{2\sqrt{n}}\right) \tag{3.9}$$

and

$$\frac{\partial\phi}{\partial M} = 0 = \frac{\partial}{\partial M}V(\bar{y}_n) + \lambda c_1 n, \text{ so that}$$

$$\frac{\partial}{\partial M}V(\bar{y}_n) = -\lambda c_1 n. \tag{3.10}$$

On eliminating λ from equation (3.9) and (3.10), we have

$$\frac{1}{\frac{1}{n}V(\bar{y}_n)}\frac{\partial}{\partial M}V(\bar{y}_n) = -\frac{c_1 n}{\left(c_1 M + \frac{c_2}{2\sqrt{n}}\right)} \quad \text{or}$$

$$\frac{1}{V(\bar{y}_n)}\frac{\partial}{\partial M}V(\bar{y}_n) = -\frac{c_1}{c_1 M\left(1 + \frac{c_2}{2c_1 M\sqrt{n}}\right)}$$

or $$\frac{M}{V(\bar{y}_n)}\frac{\partial}{\partial M}V(\bar{y}_n) = -\frac{1}{1+\frac{c_2}{2c_1M\sqrt{n}}}$$

Now solving, $c_1 n M + c_2\sqrt{n} - C = 0$ as a quadratic in $\sqrt{n}$, we have

$$\sqrt{n} = \frac{-c_2 \pm \sqrt{c_2^2 + 4c_1MC}}{2c_1M} \text{ or } 2c_1M\sqrt{n} = -c_2 \pm c_2\sqrt{1+\frac{4c_1MC}{c_2^2}} = c_2\left(\sqrt{1+\frac{4c_1MC}{c_2^2}} - 1\right)$$

Hence,

$$\frac{M}{V(\bar{y}_n)}\frac{\partial}{\partial M}V(\bar{y}_n) = -\frac{1}{1+\frac{c_2}{c_2\left(\sqrt{1+\frac{4c_1MC}{c_2^2}} - 1\right)}} = \left(1+\frac{4c_1MC}{c_2^2}\right)^{-1/2} - 1. \qquad (3.11)$$

Now, solve LHS of equation (6.11), we have

$$\frac{M}{V(\bar{y}_n)}\frac{\partial}{\partial M}V(\bar{y}_n) = \frac{M}{nV(\bar{y}_n)}\frac{\partial}{\partial M}[S^2 - (M-1)aM^{g-1}]$$

$$= \frac{1}{nV(\bar{y}_n)}[-agM^g + a(g-1)M^{g-1}].$$

Therefore,

$$\frac{aM^{g-1}[gM-(g-1)]}{nV(\bar{y}_n)} = 1 - \left(1+\frac{4c_1MC}{c_2^2}\right)^{-1/2} \qquad (3.12)$$

It is difficult to get an explicit expression for M. However, M can be obtained by the iterative method (trial and error method). On substituting the value of M thus obtained in equation (3.12), we can obtain the optimum value of n.

It is evident from equation (3.12) that the optimum size of the unit becomes smaller when

i) c_1 increases i.e. time of measurement increases.

ii) c_2 decreases i.e. travel become cheaper.

iii) total cost of survey C increases.

Cluster sampling for proportion

If it is desired to estimate the proportion P of elements belonging to a specified category A when the population consists of N clusters, each of size M and a random sample, wor, of n clusters is selected. Defining y_{ij} as 1 if the $j-$th element of the $i-$th cluster belongs to the class A and 0 otherwise, it is easy to note that $a_i = \sum_{j=1}^{M} y_{ij}$ gives the total number of

elements in the $i-$th cluster that belong to class A, and $p_i = \frac{a_i}{M}$ is the proportion in the $i-$th cluster. Hence the proportion P is

$$P = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} = \frac{1}{NM}\sum_{i=1}^{N} a_i = \frac{1}{N}\sum_{i=1}^{N} p_i \ .$$

An unbiased estimate of P is $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} p_i = p$

and

$$V(p) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}(p_i - P)^2 \cong \frac{N-n}{N^2 n}\sum_{i=1}^{N}(p_i - P)^2 \text{ , for large } N \ .$$

As an estimate of $V(p)$ we may use $\hat{V}(p) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i=1}^{n}(p_i - p)^2$.

Alternatively, if we take a simple random sample, *wor* of nM elements from the population of size, NM, the variance of p is $V(p) = \left(\frac{NM - nM}{NM - 1}\right)\frac{PQ}{nM} = \left(1 - \frac{n}{N}\right)\frac{PQ}{nM}$, for large N .

Theory of unequal clusters

There are a number of situations where the cluster size vary from cluster to cluster, for example, villages or urban blocks which are groups of households, and households, which are groups of persons are usually considered as clusters for purposes of sampling, because of operational convenience.

Suppose the population, consisting N clusters of size $M_1, M_2, \cdots, M_N$ such that $\sum_{i=1}^{N} M_i = M_0$. A sample of n clusters is drawn by the method of simple random sampling, *wor*, and all elements of the clusters surveyed. Let us denote by

y_{ij} , value of the $j-$th element in the $i-$th cluster, $j = 1, 2, \cdots, M_i$; $\quad i = 1, 2, \cdots, N$.

$\bar{y}_{i.} = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij}$, mean per element of the $i-$th cluster.

$\bar{Y}_N = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_{i.}$, mean of the cluster means in the population of N clusters.

$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{i.}$, mean of the cluster means in the sample of n clusters.

$\bar{Y} = \frac{1}{\sum_{i=1}^{N} M_i}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = \frac{1}{M_0}\sum_{i=1}^{N} M_i \bar{y}_{i.}$, mean per element in the population.

$\overline{M} = \frac{1}{N}\sum_{i=1}^{N} M_i = \frac{M_0}{N}$, mean of cluster size.

Three estimators of population mean $\overline{Y}$, that are in common use may be considered.

1st estimate: It is defined by the sample mean of clusters means as $\bar{y}_I = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{i.} = \bar{y}_n$.

By definition,

$$E(\bar{y}_I) = E\left(\frac{1}{n}\sum_{i=1}^{n} \bar{y}_{i.}\right) = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_{i.} = \overline{Y}_N \neq \overline{Y}$$, as the sampling is sr.

Thus, $\bar{y}_I$ is biased estimator of the population mean $\overline{Y}$.

The bias of the estimator is given as

$$B = E(\bar{y}_I) - \overline{Y} = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_{i.} - \frac{1}{M_0}\sum_{i=1}^{N} M_i\, \bar{y}_{i.} = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_{i.} - \frac{1}{N\overline{M}}\sum_{i=1}^{N} M_i\, \bar{y}_{i.}$$

$$= \frac{1}{N\overline{M}}\left(\overline{M}\sum_{i=1}^{N} \bar{y}_{i.} - \sum_{i=1}^{N} M_i\, \bar{y}_{i.}\right) = -\frac{1}{N\overline{M}}\sum_{i=1}^{N} (M_i - \overline{M})\, \bar{y}_{i.}$$

$$= -\frac{1}{N\overline{M}}\left[\sum_{i=1}^{N} (M_i - \overline{M})(\bar{y}_{i.} - \overline{Y}_N + \overline{Y}_N)\right]$$

$$= -\frac{1}{N\overline{M}}\left[\sum_{i=1}^{N} (M_i - \overline{M})(\bar{y}_{i.} - \overline{Y}_N)\right] - \frac{1}{N\overline{M}}\sum_{i=1}^{N} (M_i - \overline{M})\overline{Y}_N$$

$$= -\frac{1}{\overline{M}} Cov(\bar{y}_{i.}, M_i).$$

This shows that bias is expected to be small when M_i and $\bar{y}_{i.}$ are not highly correlated. In such a case, it is advisable to use this estimator.

Its variance is given by

$$V(\bar{y}_I) = E(\bar{y}_I - \overline{Y}_N)^2 = \frac{1-f}{n} S_b^2$$, where $S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N} (\bar{y}_{i.} - \overline{Y}_N)^2$

and an unbiased estimator of $V(\bar{y}_I)$ is

$$v(\bar{y}_I) = \frac{1-f}{n} s_b^2$$, where $s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\bar{y}_{i.} - \bar{y}_I)^2$.

2nd estimate: It is defined as $\bar{y}_{II} = \frac{1}{n\overline{M}}\sum_{i=1}^{n} M_i\, \bar{y}_{i.}$.

By definition,

$$E(\bar{y}_{II}) = \frac{1}{n\overline{M}} \sum_{i=1}^{n} E(M_i\, \bar{y}_{i.}) = \frac{1}{\overline{M}} \left(\frac{1}{N} \sum_{i=1}^{N} M_i\, \bar{y}_{i.} \right) = \frac{1}{N\overline{M}} \sum_{i=1}^{N} M_i\, \bar{y}_{i.} = \bar{Y}, \qquad \text{as } \textit{srwor}.$$

This shows that $\bar{y}_{II}$ is unbiased estimate of $\bar{Y}$. Its variance is given by

$$V(\bar{y}_{II}) = V\left(\frac{1}{n\overline{M}} \sum_{i=1}^{n} M_i\, \bar{y}_{i.} \right) = V\left(\frac{1}{n} \sum_{i=1}^{n} \frac{M_i\, \bar{y}_{i.}}{\overline{M}} \right).$$

Define, a variate

$$u_i = \frac{M_i \bar{y}_{i.}}{\overline{M}},\ i = 1, 2, \cdots, N.$$

Let $\bar{u}$ and $\overline{U}$ be the sample and population means of variable u, respectively, where,

$$\bar{u} = \frac{1}{n} \sum_{i=1}^{n} \frac{M_i\, \bar{y}_{i.}}{\overline{M}} = \bar{y}_{II} \quad \text{, and} \qquad \overline{U} = \frac{1}{N} \sum_{i=1}^{N} \frac{M_i\, \bar{y}_{i.}}{\overline{M}} = \frac{1}{M_0} \sum_{i=1}^{N} M_i\, \bar{y}_{i.} = \bar{Y}.$$

Therefore,

$$V(\bar{y}_{II}) = V(\bar{u}) = \frac{1-f}{n} S_b'^2, \text{ as clusters are randomly drawn } \textit{wor}.$$

where, $S_b'^2 = \frac{1}{N-1} \sum_{i=1}^{N} (u_i - \overline{U})^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{M_i\, \bar{y}_{i.}}{\overline{M}} - \bar{Y} \right)^2$

and an unbiased estimator of $V(\bar{y}_{II})$ is

$$v(\bar{y}_{II}) = \frac{1-f}{n}\, s_u^2, \text{ where} \qquad s_u^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{M_i\, \bar{y}_{i.}}{\overline{M}} - \bar{y}_{II} \right)^2.$$

3rd estimator: It is defined as $\bar{y}_{III} = \frac{1}{\sum_i M_i} \sum_{i=1}^{n} M_i\, \bar{y}_{i.}$. This estimate is a ratio estimate of the form $\hat{R} = \frac{1}{\sum_i x_i} \sum_i y_i$, and its variance is given by replacing x_i by M_i and y_i by $M_i \bar{y}_{i.}$ in the variance of ratio estimator, where, $V(\hat{R}) = \frac{1-f}{n(N-1)\,\overline{X}^2} \sum_{i=1}^{N} (y_i - R\, x_i)^2$, and $\overline{X}^2 = \left(\frac{1}{N} \sum_{i=1}^{N} M_i \right)^2 = \overline{M}^2$. Hence,

$$V(\bar{y}_{III}) = \frac{1-f}{n(N-1)\,\overline{M}^2} \sum_{i=1}^{N} \left[M_i\, \bar{y}_{i.} - \left(\frac{1}{\sum_{i=1}^{N} M_i} \sum_{i=1}^{N} M_i\, \bar{y}_{i.} \right) M_i \right]^2$$

$$= \frac{1-f}{n(N-1)\bar{M}^2}\sum_{i=1}^{N}(M_i\bar{y}_{i.} - \bar{Y}M_i)^2$$

$$= \frac{1-f}{n(N-1)}\sum_{i=1}^{N}\left[\frac{M_i}{\bar{M}}(\bar{y}_{i.} - \bar{Y})\right]^2 = \frac{1-f}{n}S_b''^2,$$

where $S_b''^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left[\frac{M_i}{\bar{M}}(\bar{y}_{i.} - \bar{Y})\right]^2$.

An unbiased estimate of $V(\bar{y}_{III})$ is given by

$$v(\bar{y}_{III}) = \frac{1-f}{n}s_b''^2, \text{ where } \quad s_b''^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left[\frac{M_i}{\bar{M}}(\bar{y}_{i.} - \bar{y}_{III})\right]^2.$$

Cluster sampling with varying probabilities and with replacement

Theorem: If a sample of n clusters is drawn with probabilities proportional to size, i.e. $p_i \propto M_i$ or $p_i = \frac{M_i}{M_0}$ and with replacement, then an unbiased estimate of $\bar{Y}$ is given by $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.}$ with variance $V(\bar{y}_n) = \frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{M_0}(\bar{y}_{i.} - \bar{Y})^2$.

Proof: By definition,

$$E(\bar{y}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.}\right) = \frac{1}{n}\sum_{i=1}^{n}E(\bar{y}_{i.}) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{i=1}^{N}p_i\,\bar{y}_{i.}\right) = \frac{1}{M_0}\sum_{i=1}^{N}M_i\bar{y}_{i.} = \bar{Y}.$$

This shows that $\bar{y}_n$ is an unbiased estimator of $\bar{Y}$.

To obtain the variance of $\bar{y}_n$, we have

$$V(\bar{y}_n) = E[\bar{y}_n - E(\bar{y}_n)]^2 = E(\bar{y}_n^2) - \bar{Y}^2. \tag{3.13}$$

Consider

$$E(\bar{y}_n^2) = E\left(\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.}\right)^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n}E(\bar{y}_{i.}^2) + \sum_{i=1}^{n}\sum_{i'\neq i=1}^{n}E(\bar{y}_{i.})E(\bar{y}_{i'.})\right)$$

$$= \frac{1}{n^2}\left(n\sum_{i=1}^{N}\frac{M_i}{M_0}\bar{y}_{i.}^2 + n(n-1)\bar{Y}^2\right), \quad \text{since } i-\text{th cluster is drawn with}$$

probability $\frac{M_i}{M_0}$, and sampling of clusters are wr, i.e. $E(\bar{y}_{i.}) = \bar{Y} = E(\bar{y}_{i'.})$.

$$= \frac{1}{n}\left(\sum_{i=1}^{N}\frac{M_i}{M_0}\bar{y}_{i.}^2 + (n-1)\bar{Y}^2\right). \tag{3.14}$$

In view of equations (3.14) and (3.13), we get

$$V(\bar{y}_n) = \frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{M_0}\bar{y}_{i.}^2 + (n-1)\bar{Y}^2 - \bar{Y}^2 = \frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{M_0}(\bar{y}_{i.} - \bar{Y})^2 = \frac{1}{n}\sigma_b^2 \text{, (say).}$$

Estimation of $V(\bar{y}_n)$

Define,

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_{i.} - \bar{y}_n)^2 \text{, then}$$

$$E\,s_b^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}E(\bar{y}_{i.}^2) - nE(\bar{y}_n^2)\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\sum_{i=1}^{N}\frac{M_i}{M_0}\bar{y}_{i.}^2 - nV(\bar{y}_n) - n\bar{Y}^2\right)\right]$$

$$= \frac{1}{n-1}\left[n\left(\sum_{i=1}^{N}\frac{M_i}{M_0}\bar{y}_{i.}^2 - n\bar{Y}^2\right) - nV(\bar{y}_n)\right]$$

$$= \frac{1}{n-1}\left(n\sum_{i=1}^{N}\frac{M_i}{M_0}(\bar{y}_{i.} - \bar{Y})^2 - n\frac{\sigma_b^2}{n}\right) = \frac{1}{n-1}(n\sigma_b^2 - \sigma_b^2) = \sigma_b^2.$$

This shows that s_b^2 is an unbiased estimate of σ_b^2. Therefore, $\hat{V}(\bar{y}_n) = \frac{1}{n}s_b^2$ is an unbiased estimate of $V(\bar{y}_n) = \sigma_b^2/n$.