Sampling procedure.

Two stage sampling or sub-sampling which consists in first selecting the clusters and then selecting a specified number of elements from each selected cluster. In such sampling designs, clusters which form the units of sampling at the first stage are called first stage units (fsu) or primary sampling units (psu) and the elements within clusters are called second-stage units (ssu). This procedure can be generalized to three or more stages and is termed multi-stage sampling.

For example, In crop surveys for estimating yield of a crop in a district, a block may be considered a primary sampling unit, the villages the second stage units, the crop fields the third stage units, and a plot of fixed size the ultimate unit of sampling.

Two stage sampling with equal first-stage units

Estimation of Mean and its Variance

Since in two-stage sampling, the units are selected in stages by considering a probability structure at each stage, the selecting procedures at both stages are to be considered in deriving the expected value and the variance of an estimator based on the number of observations taken on a sample of ssu's. For getting the expected value and sampling variance of estimators based on units selected through randomization at two stages, we may follow results.

$$E(t) = E_1 E_2(t) \quad \rightarrow (1)$$

$$V(t) = V_1 E_2(t) + E_1 V_2(t) \quad \rightarrow (2)$$

Where $E_1 \Rightarrow$ Expectation over the first stage

$V_1 \Rightarrow$ Variance over the first stage

$E_2 \Rightarrow$ conditional expectation over the second stage for a given sample of fsu's

$V_2 \Rightarrow$ variance over the second stage for a given sample of fsu's

Let us assume that the population consists of NM elements grouped into N fsu's of M ssu's each. Let n be the number of fsu's in the sample and m the number of ssu's to be selected from each sampled first stage unit.

Also we assume that the units at each stage are selected with equal probability. The following notations are used.

y_{ij} = The value obtained for the j^{th} ssu in the i^{th} fsu

$\bar{y}_i = \frac{\sum_i^M y_{ij}}{M}$ = Mean per element in the i^{th} fsu

$\bar{\bar{y}} = \sum_i^N \bar{y}_i$ = Mean per element in the population

$S_b^2 = \frac{\sum_i^N (\bar{y}_{i\cdot} - \bar{\bar{y}})^2}{(N-1)}$ = True variance between first stage unit means

$S_w^2 = \frac{\sum_i^N \sum_i^M (y_{ij} - \bar{y}_{i\cdot})^2}{N(M-1)}$ = True variance within first stage units

$\bar{y}_{i\cdot} = \frac{\sum_j^m y_{ij}}{m}$ = sample mean per ssu in the i^{th} fsu

$\bar{\bar{y}} = \frac{\sum_i^n \bar{y}_i}{n}$ = overall sample mean per element

Theorem

If the n fsu's and the m ssu's from each chosen fsu are selected by simple random sampling, wor, $\bar{y}$ is an unbiased estimator of $\bar{Y}$ with sampling variance

$$V(\bar{y}) = \frac{(N-n)}{N}\frac{S_b^2}{n} + \frac{(M-m)}{M}\frac{S_w^2}{mn} \rightarrow (3)$$

Proof

Applying relation (1) for getting expectation, we have

$$E(\bar{y}) = E_1E_2(\bar{y}_i/i) = E_1(\bar{Y}_i)$$

$$= \bar{Y}$$

It shows that the sample mean of all elements in the sample gives an unbiased estimator of the population mean

To obtain the variance of the estimator by relation (2), we have

$$V(\bar{y}) = V_1[E_2(\bar{y}/i)] + E_1[V_2(\bar{y}/i)]$$

$$= V_1(\bar{Y}_i) + E\left[\frac{1}{n^2}\sum_i\left(\frac{1}{m} - \frac{1}{M}\right)S_i^2\right]$$

$$= \frac{(N-n)}{nN}S_b^2 + \frac{(M-m)}{mM}\frac{S_w^2}{n}$$

$$S_w^2 = \frac{1}{N}\sum_i^N S_i^2$$

If $f_1 = n/N$ and $f_2 = m/M$ are the sampling fractions in the first and second stages, the result can be written as

$$V(\bar{y}) = \frac{(1-f_1)}{n} S_b^2 + \left(\frac{1-f_2}{nm}\right) S_w^2 \quad \rightarrow (4)$$

The variance given in (3) in two-stage sampling is made up of two components. One component comes from the variability of ssu's within fsu's and the second one arises from the variance of fsu's. If the selected fsu's are completely enumerated or, in other words $n = M$, the variance of the sample mean will be given by the first component only. If $n = N$ or, in other words, every fsu in the population is included in the sample, then this case corresponds to stratified sampling with fsu's as strata and a simple random sampling of m ssu's is drawn from each of the strata.

Corollary

1. Under the conditions of Theorem 4.2, an unbiased estimator of $v(\bar{y})$ is given by

$$v(\bar{y}) = \frac{(1-f_1)}{n} s_b^2 + \frac{(1-f_2)}{nm} s_w^2 \quad \rightarrow (5)$$

where $$s_b^2 = \frac{\sum_i^n (\bar{y}_{i\cdot} - \bar{y})^2}{(n-1)}$$

$$s_w^2 = \frac{\sum_i^n \sum_j^m (y_{ij} - \bar{y}_{i\cdot})^2}{n(m-1)}$$

2. Show that an unbiased estimator of S_b^2 is given by

$$\hat{S}_b^2 = s_b^2 - \frac{(1-f_2)}{m} s_w^2$$

3. If the n fsu's are selected randomly with replacement and the m ssu's from each chosen unit are selected by simple random sampling wor, $\bar{y}$ is an unbiased estimator of $\bar{Y}$ with sampling variance

$$V(\bar{y}) = \frac{S_b^2}{n} + (1-f_2)\frac{S_w^2}{mn}$$

4. If the n fsu's are selected randomly, wor and the m ssu's from each chosen unit are selected randomly, wor, $\bar{y}$, is an unbiased estimator of $\bar{Y}$ with its variance

$$V(\bar{y}) = (1-f_1)\frac{S_b^2}{n} + \frac{S_w^2}{mn}$$

5. If the n fsu's and m ssu's from each chosen unit are selected by simple random sampling wr, $\bar{y}$ is an unbiased estimator of $\bar{Y}$ with its variance,

$$V(\bar{y}) = \frac{S_b^2}{n} + \frac{S_w^2}{mn}$$

6. If the n fsu's and m ssu's from each chosen unit are selected by simple random sampling, wor, the estimator

$$\hat{Y} = NM \sum_i \frac{\bar{y_i}}{n}$$

is an unbiased estimator of the population total Y and its sampling variance is given by,

$$V(\hat{Y}) = N^2M^2(1-f_1)\frac{S_b^2}{n} + N^2N^2(1-f_2)\frac{S_w^2}{mn}$$

Two stage sampling with unequal First stage units

Let the population under consideration consists of N first-stage units. The i^{th} fsu consists M_i second stage units. Further, units are selected without replacement with equal or unequal probabilities. A sample of n fsu's is selected and from the i^{th} selected fsu, a sample of m_i ssu's is selected,

Let us denote

M_i - the number of ssu's in the i^{th} fsu,

$M_0 = \sum_i^N M_i$ - The total number of ssu's in the population

m_i = The number of ssu's to be selected from the ith fsu included in the sample.

$m_0 = \sum_i^n m_i$ = The total number of ssu's in the sample.

$\bar{Y}_i = \sum_j^{M_i} y_{ij}/M_i$ = The i^{th} fsu mean.

$\bar{Y}_N = \sum_i^N \bar{Y}_i/N$ = The overall mean of fsu means

$\bar{Y} = \dfrac{\sum_i^N M_i \bar{Y}_i}{\sum_i^N M_i} = \sum_i^N w_i \bar{Y}_i$ = The mean per ssu or the population mean per element.

There are several estimators of the population mean $\bar{y}$ but we propose only to study some of the practical methods which are.

$$\bar{y} = \frac{1}{n} \sum_{i}^{n} v_i \bar{y}_i = \frac{\sum_{i}^{n} M_i \bar{y}_i}{n\bar{M}} \quad \rightarrow (1)$$

$$\bar{y}_1 = \frac{\sum_{i}^{n} \bar{y}_i}{n} \quad \rightarrow (2)$$

$$\bar{y}_2 = \frac{\sum_{i}^{n} M_i \bar{y}_i}{\sum_{i}^{n} M_i} \quad \rightarrow (3)$$

$$\bar{y}_i = \sum_{j}^{m_i} \frac{y_{ij}}{m_i}, \quad \bar{M} = \frac{M_0}{N}, \text{ and } v_i = \frac{M_i}{M}$$

Theorem 1

Show that the estimator is given by relation (1) is unbiased and its sampling variance is given by

$$V(\bar{y}) = (1-f_1)\frac{S_b^2}{n} + \sum_{i}^{N} \frac{M_i^2}{nNM^2}(1-f_2)\frac{S_{iw}^2}{m_i} \quad \rightarrow (4)$$

where $S_b^2 = \frac{\sum_{i}^{N} (v_i \bar{y}_i - \bar{Y})^2}{(N-1)}$

$S_{iw}^2 = \sum_{j}^{M_i} \frac{(y_{ij} - \bar{Y}_i)^2}{(M_i - 1)}$

Proof

To prove that $\bar{y} = \frac{1}{n} \sum_i^n v_i \bar{y}_i$ is an unbiased estimator, we can write

$$E(\bar{y}) = E_1 \left[\frac{1}{n} \sum_i^n E_2 (v_i \bar{y}_i / i) \right]$$

$$= E_1 \left[\frac{1}{n} \sum_i^n v_i \bar{Y}_i \right]$$

$$= \frac{1}{n} \sum_i^n E_1 (v_i \bar{Y}_i) = \bar{Y}$$

The sampling variance of the estimator is given by,

$$V(\bar{y}) = V_1 E_2 (\bar{y}/n) + E_1 V_2 (\bar{y}/n)$$

$$= V_1 \left[\frac{1}{n} \sum_i^n v_i \bar{Y}_i \right] + E_1 \left[\frac{1}{n^2} \sum_i^n \frac{M_i^2}{\bar{M}^2} V(\bar{y}_i/n) \right]$$

$$= (1 - f_1) \frac{S_b^2}{n} + \sum_i^N \frac{M_i^2 (1 - f_{2i})}{n N \bar{M}^2} \frac{S_{wi}^2}{m_i}$$

The units are chosen with equal probability in this method and the contribution made by fsu's to the components of this variance depends upon the variation between the fsu totals. If the units vary considerably in their sizes, this component will be large the second component of variance is also to be large as there is likely to be positive correlation between M_i and S_{wi}^2. Frequently, this component is so large, that this estimator is not preferred.

Theorem 2

Show that the estimator given by relation (2) is biased and its bias is given by

$$B = -\sum_i^N (M_i - \bar{M})\bar{y}_i / N\bar{M} \quad \rightarrow (5)$$

and sampling variance by

$$V(\bar{y}_i) = (1-f_1)\frac{s_b^2}{n} + \frac{1}{nN}\sum_i^N (1-f_{2i})\frac{s_{i0}^2}{m_i} \rightarrow (6)$$

where $s_b^2 = \sum_i^N \frac{(\bar{y}_i - \bar{Y})^2}{N-1}$ and s_{i0}^2 is as usual

Proof

To prove that $\bar{y}_i$ is a biased estimator, we can get

$$E(\bar{y}_i) = E\left(\sum_i^n \frac{\bar{y}_i}{n}\right) = E_1\left[\frac{1}{n}\sum_i^n E_2(\bar{y}_i / i)\right]$$

$$= E_1\left[\frac{1}{n}\sum_i^n \bar{y}_i\right] = \bar{Y}_N = \bar{Y}$$

which shows that $\bar{y}_i$ is a biased estimator.

Its bias can be obtained as

$$B = \bar{Y}_N - \bar{Y} = \sum_i^N \frac{\bar{Y}_i}{N} - \sum_i^N \frac{M_i \bar{Y}_i}{NM}$$

$$= -\frac{1}{N\bar{M}}\left[\sum_i^N M_i \bar{Y}_i - \sum_i^N \bar{M}\bar{Y}_i\right]$$

$$= -\frac{1}{N\bar{M}}\left[\sum_i^N (M_i - \bar{M})\bar{Y}_i\right]$$

The sampling variance of the estimator is given by

$$V(\bar{y}_1) = V_1 E_2 [\bar{y}_1/n) + E_1 V_2 (\bar{y}_1/n)$$

$$= V_1 \left[\frac{1}{n} \sum_i \bar{y}_i\right] + E_1 \left[\frac{1}{n^2} \sum_i V_2 (\bar{y}_i/i)\right]$$

$$= (1-f_1) \frac{S_b^2}{n} + E_1 \left[\frac{1}{n^2} \sum_i (1-f_2) \frac{S_{wi}^2}{m_i}\right]$$

$$= (1-f_1) \frac{S_b^2}{n} + \frac{1}{nN} \sum_i^N (1-f_2) \frac{S_{wi}^2}{m_i}$$

The bias in the estimator $\bar{y}_1$ appears due to the fact that the probabilities of selection of the ssu's vary from one unit to another, in the fsu's, due to their unequal sizes. If the M_i's do not vary considerably and the study variate is not correlated with M_i, the bias may not be large. Here, the MSE of $\bar{y}_1$ will consists of three components: one from the bias, one from variation within fsu's, and one arising from variance between the means of fsu's. The values of m_i's are not specified and a proper choice of m_i can be helpful in controlling these components

Two stage pps sampling

Suppose a sample of n fsu's is selected with pps wor. From the selected i^{th} fsu, a selection of m ssu's is made with srswor. If the i^{th} fsu is selected more than once, then a fresh independent drawing of m_i ssu's is being made without replacement from the comple fsu each time. An unbiased estimator of $\bar{Y}$ is given by

$$\hat{\bar{Y}}_{pps} = \frac{1}{n} \sum_{i}^{n} \frac{M_i \bar{y}_i}{p_i} \longrightarrow (1)$$

where $p_i \Rightarrow$ probability of selecting the i^{th} fsu at each draw such that

$$\sum_{i}^{N} p_i = 1 \quad \text{and} \quad \bar{y}_{i.} = \frac{\sum_{i}^{mn} y_{ij}}{m_i}$$

The sampling variance of the estimator is given by

$$V(\hat{\bar{y}}_{pps}) = \frac{1}{n} \sum_{i}^{N} p_i \left(\frac{\bar{Y}_i}{p_i} - \bar{Y} \right)^2 + \frac{1}{n} \sum_{i}^{N} \frac{M_i^2}{p_i} (1 - f_{2i}) \frac{s_{wi}^2}{m_i} \longrightarrow (2)$$

An unbiased estimator of $V(\hat{\bar{y}}_{pps})$ is

$$v(\hat{\bar{y}}_{pps}) = \frac{\sum_{i}^{n} \left(\frac{M_i \bar{y}_i}{p_i} - \hat{\bar{y}} \right)^2}{n(n-1)} \longrightarrow (3)$$

which gives a good procedure of estimation, whatever the method of selection adopted at the second stage, provided the fsu's are selected with replacement. If one is interested in estimating the between and within components of variance, the the between component can be obtained by subtracting the within component from relation (3)

The within fsu variance component can be estimated unbiasedly by

$$V_w(\hat{\bar{Y}}_{pps}) = \sum_i^n \frac{M_i^2 (1-f_{2i})}{n^2 p_i^2} \frac{s_{wi}^2}{m_i}$$

Three stage sampling with equal probability

The procedure of two stage sampling can be carried to a third stage by sampling the ssu's instead of enumerating them completely. For example in crop surveys for estimating the yield average, a village is considered the first stage sampling unit. Within a selected village, only some of the fields growing the crop are selected and taken as the second stage units. When a field is selected, only certain parts (called plots) of it are sampled, which may be termed the third stage units (tsu). Thus the results of three stage sampling can be obtained by extending those of two stage sampling, with further assumptions that each ssu has L third stage units. It is also assumed that the units are selected with equal probability.

Let Y_{ijk} be the value obtained for the k^{th} third stage unit in the j^{th} second stage unit of the i^{th} first stage unit. The relevant population means per element are as follows:

$$\bar{Y}_{ij} = \sum_{k}^{L} \frac{y_{ijk}}{L},$$

$$\bar{Y}_i = \sum_{j}^{M} \sum_{k}^{L} \frac{y_{ijk}}{LM},$$

$$\bar{y} = \sum_{i}^{N} \sum_{j}^{M} \sum_{k}^{L} \frac{y_{ijk}}{LMN}$$

$\bar{y}_{ij}$, $\bar{y}_i$ and $\bar{y}$ will denoted the corresponding values of the sample,

corresponding population variances will be

$$S_b^2 = \sum_{i}^{N} \frac{(\bar{Y}_i - \bar{Y})^2}{(N-1)}$$

$$S_w^2 = \sum_{i}^{N} \sum_{i}^{M} \frac{(\bar{Y}_{ij} - \bar{Y}_i)^2}{N(M-1)}$$

$$S_v^2 = \sum_{i}^{N} \sum_{j}^{M} \sum_{k}^{L} \frac{(y_{ijk} - \bar{Y}_{ij})^2}{NM(L-1)}$$

Theorem

If the n fsu's, mssu's and l ultimate units are chosen by simple random sampling, wor, $\bar{y}$ is an unbiased estimate of $\bar{Y}$ with variance

$$V(\bar{y}) = \frac{(1-f_1)}{n} S_b^2 + \frac{(1-f_2)}{nm} S_w^2 + \frac{(1-f_3)}{nml} S_v^2$$

→ (1)

where $f_1 = n/N$, $f_2 = m/M$, $f_3 = l/L$ are the sampling fractions at three stage respectively. The proof is obvious. The variance given by relation (1) is made up of three components corresponding to the three stages of sampling. The first component is due to the variability of fsu's the second to variation of the ssu and the third to tsu's. If $m = M$, and $l = L$, ie each of the n fsu's were completely enumerated, the variance of the sample mean will be given by first component only, representing the variance of single stage sampling. Similarly if each of the nm selected second stage units were completely enumerated, i.e $l = L$ the variance of the sample mean will be given by the first two terms only, representing the variance of two stage sample design. In $n = N$ or in other words, every fsu in the population is included in the sample, the variance of the sample mean will have the last two terms, ie., it corresponds to a stratified two stage sampling design with fsu's as strata.

Three stage pps sampling

Let a sample of nml units be selected in three stages by adopting pps, wor, at each stage. suppose n fsu's are selected with p_i probabilities of selection for the i^{th} fsu's $(i = 1, \ldots N)$. From each selected fsu, m, ssu's are selected with p_{ij}. probabilities of selection for j^{th} ssu $(j = 1, \ldots M_i)$ and from each selected ssu, l third stage units (tsu's) are selected with p_{ijk} probabilities of selection of the k^{th} tsu of the j^{th} ssu in the fsu $(k = 1, \ldots L_{ij})$. Let y_{ijk} denote the value of the k^{th} tsu in the j^{th} ssu of the i^{th} fsu $(i = 1, \ldots n;$ $j = 1, \ldots m;$ $k = 1, \ldots l)$ in the sample. An estimator of the population total Y can be defined as

$$\hat{Y} = \frac{1}{nml} \sum_{i}^{n} \frac{1}{p_i} \sum_{j}^{m} \frac{1}{p_{ij}} \sum_{k}^{l} \frac{y_{ijk}}{p_{ijk}} \rightarrow (1)$$

It can easily be shown that estimator is unbiased

i.e., $E(\hat{Y}) = E_1 E_2 E_3 (\hat{Y}) = Y$

and the variance of the estimator is obtained by

$$V(\hat{\bar{y}}) = V_1 E_2 E_3 (\hat{\bar{y}}) + E_1 V_2 E_3 (\hat{\bar{y}}) + E_1 E_2 V_3 (\hat{\bar{y}})$$

Thus

$$V(\hat{\bar{y}}) = \frac{1}{n}\left(\sum_i^N \frac{Y_i^2}{P_i} - \bar{Y}^2\right) + \frac{1}{nm}\sum_i^N \frac{1}{P_i}\left(\sum_i^{M_i} \frac{y_{ij}^2}{P_{ij}} - Y_i^2\right)$$

$$+ \frac{1}{nml}\sum_i^N \frac{1}{P_i}\sum_i^{M_i}\frac{1}{P_{ij}}\left(\sum_k^{L_{ij}} \frac{y_{ijk}^2}{P_{ijk}} - \bar{Y}_{ij}^2\right) \rightarrow (2)$$

An unbiased estimator of $V(\hat{\bar{y}})$ is given by

$$v(\hat{\bar{y}}) = \frac{1}{n(n-1)}\left(\sum_i y_{i\cdot}^2 - \frac{\hat{\bar{Y}}^2}{n}\right) \rightarrow (3)$$

$$y_{i\cdot} = \frac{1}{P_i}\sum_j^m \frac{1}{P_{ij}}\left(\sum_k^l \frac{y_{ijk}}{P_{ijk}}\right)$$

It should be noted that, like two stage, the sampling variance function in three stage sampling can also be written as

$$V(\hat{\bar{y}}) = \frac{A_1}{n} + \frac{A_2}{nm} + \frac{A_3}{nml}$$