Ramakrishna Mission Residential College, Narendrapur

Department of Statistics

Notes on

Double Sampling:

To estimate *population mean or population total* ,we often make use of knowledge on an auxiliary variable x at the sampling and/or estimation stage to achieve a gain in precision, e.g., in stratified sampling (stratification on the basis of x), PPS sampling, Ratio and Regression method of estimation etc.

In case of ratio method of estimation, a biased but efficient estimator for $\overline{Y}$

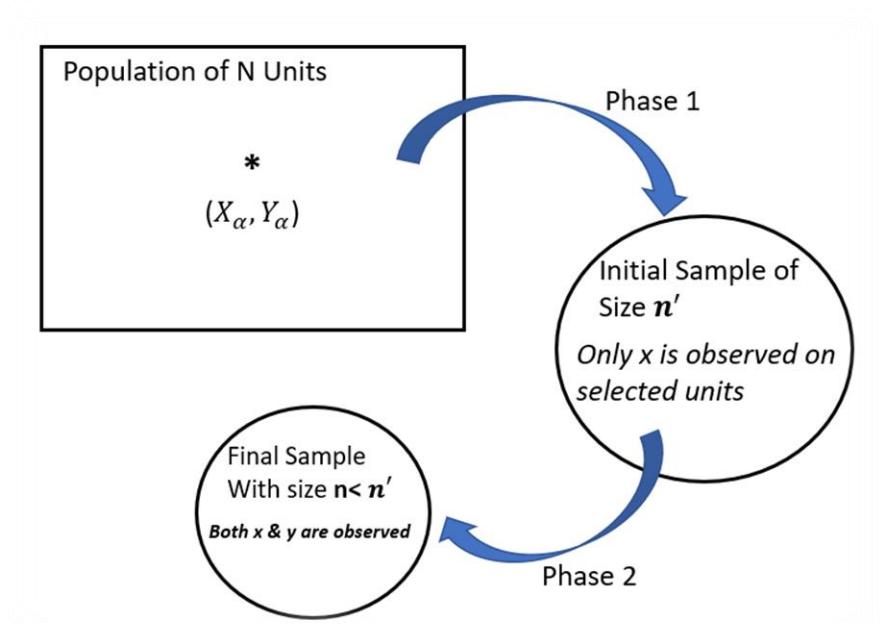
$$\overline{y}_R = \frac{\overline{y}}{\overline{x}}\overline{X} \text{----------(1)}$$

The problem with this estimator is that usually $\overline{X}$ is not known. In that case a double sampling procedure may be used as follows:

i. From the entire population a sample of size n' is drawn using some suitable sampling scheme and the auxiliary variable x is observed for the selected units. (This information is used to estimate $\overline{X}$ in $\overline{y}_R$
ii. At step two, a sub sample of size n is drawn from the initial sample, or an independent sample of size n is drawn from the entire population using some scheme. Both the study variable and the auxiliary variable are observed for the selected units. Usually, n' is much larger than n.

Note: Since in double sampling, a part of the resource is used to collect information on x from the initial sample, the size of the main sample needs to be reduced as compared to the situation where auxiliary information is not used. However, if the cost of collecting information on x is low, the gain due to incorporating auxiliary information more than offsets the loss due to reduction of the size of the main sample.

The Set up and Notations:



The Population: U= $\{U_1, U_2, \dots U_\alpha, \dots, U_N\}$

Size of the population: N (Known)

Study variable: y

Auxiliary variable: x (usually cheap to observe and correlated with y)

The Population values: $y(U_\alpha) = Y_\alpha, x(U_\alpha) = X_\alpha, \ for\ \alpha = 1(1)N$

We can define the parameters on the population values as

Population mean: $\bar{Y} = \sum_1^N Y_\alpha \ , \bar{X} = \sum_1^N X_\alpha$

Population mean square: $S_y^2 = \frac{1}{N-1}\sum_1^N (Y_\alpha - \bar{Y})^2 \ , S_x^2 = \frac{1}{N-1}\sum_1^N (X_\alpha - \bar{X})^2$

Population covariance: $\sigma_{xy} = \frac{1}{N}\sum_1^N ((Y_\alpha - \bar{Y})(X_\alpha - \bar{X}) \quad , \ with \ S_{xy} = \frac{N}{N-1}\sigma_{xy}$

1st Phase: From the entire population a sample of size n' (significantly larger than actual sample size n) is drawn using some suitable sampling scheme and the auxiliary variable x is observed for the selected units.

Data: (x'_i, y'_i), i=1(1) n'

Only x values are observed.

So, we can calculate $\bar{x}' = \frac{1}{n'}\sum_{i=1}^{n'} x_i' \ , s_{x'}^{'2} = \frac{1}{n'-1}\sum_{i=1}^{n'} (x_i' - \bar{x}')^2$

Similarly

we can define (cannot be calculated): $\bar{y}', s_{y'}^{'2}, s'_{x'y'} = \frac{1}{(n'-1)}\sum(x'_i - \bar{x}')(y'_i - \bar{y}')$

2nd Phase: At step two, a sub sample of size n is drawn from the initial sample, or an independent sample of size n is drawn from the entire population using some scheme. Both the study variable and the auxiliary variable are observed for the selected units.

Data: (x_i, y_i), i=1(1)n.

In this phase both x and y are observed.

So we observe $\bar{x}, \bar{y}, s_x^{'2}, s_y^{'2}, s'_{xy}$ in this phase.

<u>Case 1</u>: A sub sample of size n is drawn from the initial sample and SRSWOR is used at both the phases:

Double sample ratio estimator for population mean is given by

$$\bar{y}_R{}^d = \frac{\bar{y}}{\bar{x}} . \bar{x}' \text{----------(2)}$$

Let us define $\delta_1 = \frac{\bar{y}-\bar{Y}}{\bar{Y}}, \delta_2 = \frac{\bar{x}-\bar{X}}{\bar{X}}, \delta_3 = \frac{\bar{x}'-\bar{X}}{\bar{X}}$

$$\bar{y}_R{}^d = (1+\delta_1)\bar{Y}\left((1+\delta_2)\bar{X}\right)^{-1}(1+\delta_3)\bar{X}$$

$$= \frac{(1+\delta_1+\delta_3+\delta_1\delta_3)}{(1+\delta_2)}\bar{Y}$$

$$= (1+\delta_1+\delta_3+\delta_1\delta_3)(1-\delta_2+\delta_2{}^2-\cdots)\bar{Y} \text{ [assuming } |\delta_2| < 1]$$

$$= (1+\delta_1+\delta_3+\delta_1\delta_3-\delta_2-\delta_1\delta_2-\delta_2\delta_3-\delta_1\delta_2\delta_3+\delta_2{}^2+\cdots)\bar{Y}$$

Note that

since at the 2nd phase SRSWOR is used to draw <u>final sample of size n</u> from <u>initial sample of size</u> n', using results of SRSWOR we can have

$$E_2(\bar{y}) = \bar{y}', \; V_2(\bar{y}) = \left(\frac{1}{n}-\frac{1}{n'}\right)s_{y'}^{'2}, E_2(s_y^{'2}) = s_{y'}^{'2} \text{ ----------(3)}$$

$$E_2(\bar{x}) = \bar{x}', \; V_2(\bar{x}') = \left(\frac{1}{n}-\frac{1}{n'}\right)s_{x'}^{'2}, \; E_2(s_x^{'2}) = s_{x'}^{'2} \text{ ----------(4)}$$

$$\text{also } cov_2(\bar{x}, \bar{y}) = \left(\frac{1}{n}-\frac{1}{n'}\right)s'_{x'y'} \text{ and } E_2(s'_{xy}) = s'_{x'y'} \text{----------(5)}$$

Again

at the 1st phase SRSWOR is used to draw <u>initial sample of size</u> n' from the <u>population of size N</u> . So using results of SRSWOR we can have

$$E_1(\bar{y}') = \bar{Y}, \; V_1(\bar{y}') = \left(\frac{1}{n'}-\frac{1}{N}\right)S_y{}^2, E_1(s_{y'}^{'2}) = S_y^2 \text{ ----------(6)}$$

$$E_1(\bar{x}') = \bar{X}, \; V_1(\bar{x}') = \left(\frac{1}{n'}-\frac{1}{N}\right)S_x{}^2, \; E_1\left(s_{x'}^{'2}\right) = S_x^2 \text{----------(7)}$$

$$\text{also } cov_1(\bar{x}', \bar{y}') = \left(\frac{1}{n'}-\frac{1}{N}\right)S_{xy} \text{ and } E_1(s'_{x'y'}) = S_{xy} \text{----------(8)}$$

Now, E(δ_1) $= E_1E_2\left(\frac{\overline{y}-\overline{Y}}{\overline{Y}}\right) = \frac{1}{\overline{Y}}\left(E_1\left(\overline{y}'-\overline{Y}\right)\right) = 0.$

Similarly, E(δ_2) = 0.

E(δ_3) = $E_1\left(\frac{\overline{x}'-\overline{X}}{\overline{X}}\right) = 0.$

Thus, up to first order of approximation, $\overline{y}_R{}^d$ is unbiased for $\overline{Y}$.

Now, E($\delta_1\delta_2$) = E($\frac{(\overline{y}-\overline{Y})(\overline{x}-\overline{X})}{\overline{X}\overline{Y}}$)

$= \frac{1}{\overline{X}\overline{Y}} . E_1[E_2\{(\overline{x}-\overline{x}'+\overline{x}'-\overline{X})(\overline{y}-\overline{y}'+\overline{y}'-\overline{Y})\}]$

$= \frac{1}{\overline{X}\overline{Y}} . E_1[E_2(\overline{x}-\overline{x}')(\overline{y}-\overline{y}') + (\overline{x}'-\overline{X})E_2(\overline{y}-\overline{y}') + (\overline{y}'-\overline{Y})E_2(\overline{x}-\overline{x}') + (\overline{x}'-\overline{X})(\overline{y}'-\overline{Y})]$

$= \frac{1}{\overline{X}\overline{Y}} E_1[cov_2(\overline{x},\overline{y}) + (\overline{x}'-\overline{X})(\overline{y}'-\overline{Y})]$

[since with respect to E_2(when we are at 2nd phase ,i.e., initial sample is treated as population and final sample as sample) , $(\overline{x}'-\overline{X})$ and $(\overline{y}'-\overline{Y})$ are constants and $E_2(\overline{y}-\overline{y}') = E_2(\overline{x}-\overline{x}') = 0$]

$= \frac{1}{\overline{X}\overline{Y}} E_1[\left(\frac{1}{n}-\frac{1}{n'}\right)s'_{x'y'} + (\overline{x}'-\overline{X})(\overline{y}'-\overline{Y})]$ *[by* (5)*]*

$= \frac{1}{\overline{X}\overline{Y}} [\left(\frac{1}{n}-\frac{1}{n'}\right)E_1(s'_{x'y'}) + cov_1(\overline{x}',\overline{y}')]$

$= \frac{1}{\overline{X}\overline{Y}}[\left(\frac{1}{n}-\frac{1}{n'}\right)S_{xy} + (\frac{1}{n'}-\frac{1}{N})S_{xy}]$ *[by* (8)*]*

$= \frac{1}{\overline{X}\overline{Y}} [(\frac{1}{n}-\frac{1}{N})S_{xy}]$----------(9)

E($\delta_1\delta_3$) = $\frac{E_1E_2(\overline{y}-\overline{Y})(\overline{x}'-\overline{X})}{\overline{X}\overline{Y}} = \frac{1}{\overline{X}\overline{Y}}[E_1(\overline{x}'-\overline{X})E_2(\overline{y}-\overline{Y})]$

[with respect to E_2 , $(\overline{x}'-\overline{X})$ is constant]

$= \frac{1}{\overline{X}\overline{Y}}[E_1\{(\overline{x}'-\overline{X})(\overline{y}'-\overline{Y})\}]$ *[$E_2(\bar{y}) = \bar{y}'$ by* (3)*]*

$=cov_1(\overline{x}',\overline{y}') = (\frac{1}{n'}-\frac{1}{N})\frac{S_{xy}}{\overline{X}\overline{Y}}$ *[by* (8)*]*----------(10)

E($\delta_2\delta_3$) = $\frac{1}{\overline{X}^2}\left[E_1\left\{(\overline{x}'-\overline{X})\left(E_2(\overline{x}-\overline{X})\right)\right\}\right] = \frac{1}{\overline{X}^2}\left[E_1(\overline{x}'-\overline{X})^2\right]$ *[$E_2(\bar{x}) = \bar{x}'$ by* (4)*]*

$= \frac{V_1(\overline{x}')}{\overline{X}^2} = (\frac{1}{n'} - \frac{1}{N})\frac{{S_x}^2}{\overline{X}^2}$ *[by (7)]*----------(11)

$E({\delta_1}^2) = \frac{1}{\overline{Y}^2}E_1\left(E_2(\overline{y} - \overline{Y})^2\right) = \frac{1}{\overline{Y}^2}E_1\left(E_2(\overline{y} - \overline{y}' + \overline{y}' - \overline{Y})^2\right)$

$= \frac{1}{\overline{Y}^2}E_1[V_2(\overline{y}) + (\overline{y}' - \overline{Y})^2 + 2(\overline{y}' - \overline{Y})E_2(\overline{y} - \overline{y}')]$ *[expanding and taking E_2]*

[reasons are analogous to those explained earlier]

$= \frac{1}{\overline{Y}^2}E_1[\left(\frac{1}{n} - \frac{1}{n'}\right){s_{y'}}^2 + (\overline{y}' - \overline{Y})^2]$ *[by (3)]*

$= \frac{1}{\overline{Y}^2}[\left(\frac{1}{n} - \frac{1}{n'}\right){s_y}^2 + V_1(\overline{y}')]$ *[by (6)]*

$= \frac{1}{\overline{Y}^2}[\left(\frac{1}{n} - \frac{1}{n'}\right){s_y}^2 + (\frac{1}{n'} - \frac{1}{N}){s_y}^2]$ *[by (6)]*

$= \frac{1}{\overline{Y}^2}\left(\frac{1}{n} - \frac{1}{N}\right){s_y}^2$ ----------(12)

[Alternatively, $E({\delta_1}^2) = \frac{1}{\overline{Y}^2}E(\overline{y} - \overline{Y})^2 = \frac{1}{\overline{Y}^2}[E_1V_2(\overline{y}) + V_2E_1(\overline{y})]$**]**

Similarly, $E({\delta_2}^2) = \frac{1}{\overline{X}^2}(\frac{1}{n} - \frac{1}{N}){s_x}^2$----------(13)

And $E({\delta_3}^2) = \frac{1}{\overline{X}^2}E_1(\overline{x}' - \overline{X})^2 = \frac{1}{\overline{X}^2}V_1(\overline{x}') = \frac{1}{\overline{X}^2}(\frac{1}{n'} - \frac{1}{N}){s_x}^2$----------(14)

$MSE({\overline{y}_R}^d) = E({\overline{y}_R}^d - \overline{Y})^2 \cong E(\delta_1 + \delta_3 - \delta_2)^2\overline{Y}^2$

[Neglecting the terms involving δ_1, δ_2 and δ_3 with power more than 2]

$= \overline{Y}^2[E({\delta_1}^2 + {\delta_2}^2 + {\delta_3}^2 + 2\delta_1\delta_3 - 2\delta_1\delta_2 - 2\delta_2\delta_3)]$

$= \overline{Y}^2[\left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{{S_y}^2}{\overline{Y}^2} + \frac{{s_x}^2}{\overline{X}^2} - \frac{2S_{xy}}{\overline{XY}}\right) + (\frac{1}{n'} - \frac{1}{N})(\frac{{S_x}^2}{\overline{X}^2} + \frac{2S_{xy}}{\overline{XY}} - \frac{2{s_x}^2}{\overline{X}^2})\ ]$

$= (\frac{1}{n} - \frac{1}{N})\ ({S_y}^2 + {S_x}^2R^2 - 2S_{xy}R) + (\frac{1}{n'} - \frac{1}{N})\ (2RS_{xy} - R^2{S_x}^2)$ ----------(15)

[using (9) to (14)]

A biased but useful estimator of MSE may be obtained by replacing ${S_y}^2, {S_x}^2, S_{xy}, R$ *by* ${s_{y'}}^2$, ${s_{x'}}^2, s_{xy}'$ *and* $\frac{\overline{y}}{\overline{x}}$.

<u>Case 2</u>: The main sample is drawn from the entire population, from independent of the initial sample.

In that case, $E(\delta_1\delta_3)=0$

$E(\delta_2\delta_3)=0$

$E(\delta_1\delta_2) = = \frac{cov(\overline{x},\overline{y})}{\overline{XY}} = (\frac{1}{n} - \frac{1}{N})\frac{S_{xy}}{\overline{XY}}$

$E(\delta_1{}^2) = \frac{{s_y}^2}{\overline{Y}^2}(\frac{1}{n} - \frac{1}{N})$

$E(\delta_2{}^2) = \frac{{s_x}^2}{\overline{X}^2}(\frac{1}{n} - \frac{1}{N})$

$E(\delta_3{}^2) = = \frac{1}{\overline{X}^2}(\frac{1}{n'} - \frac{1}{N}){s_x}^2$

[using the results of SRSWOR]

$MSE({\overline{y}_R}^d) = \left(\frac{1}{n} - \frac{1}{N}\right)\left[{S_y}^2 + R^2{s_x}^2 - 2RS_{xy}\right] + (\frac{1}{n'} - \frac{1}{N})R^2{S_x}^2$----------(16)

<u>Double Sampling Regression Estimator:</u>

Usual regression estimator is given by $\overline{y}_{Reg}{}' = \overline{y} - b(\overline{x} - \overline{X})$.

When $\overline{X}$ is not known, it can be replaced by $\overline{x}'$, the sample mean obtained from the initial sample of size n', say.

Thus, we have the double sampling regression estimator as $\overline{y}_{Reg}{}^d = \overline{y} - b(\overline{x} - \overline{x}')$

Let $\delta_1 = \frac{\overline{y}-\overline{Y}}{\overline{Y}}, \delta_2 = \frac{\overline{x}-\overline{X}}{\overline{X}}, \delta_3 = \frac{\overline{x}'-\overline{X}}{\overline{X}}, \delta_4 = \frac{b-\beta}{\beta}$

Then, $\overline{y}_{Reg}{}^d = \overline{Y}(1 + \delta_1) + (1 + \delta_4)\beta\overline{X}(\delta_3 - \delta_2)$

$MSE({\overline{y}_{Reg}}^d) = E\left[\overline{(y}_{Reg}{}^d - \overline{Y})^2\right] = E[\{\delta_1\overline{Y} + \beta\overline{X}(\delta_3 - \delta_2 + \delta_3\delta_4 - \delta_2\delta_4)\}^2]$

$= \overline{Y}^2E\left({\delta_1}^2\right) + \beta^2\overline{X}^2E(\delta_3 - \delta_2 + \delta_3\delta_4 - \delta_2\delta_4)^2 + 2\beta\overline{XY}E(\delta_1\delta_3 - \delta_2\delta_1 + \cdots)$

$\cong \overline{Y}^2E\left({\delta_1}^2\right) + \beta^2\overline{X}^2E\left({\delta_3}^2 + {\delta_2}^2 - 2\delta_2\delta_3\right) + 2\beta\overline{XY}E(\delta_1\delta_3 - \delta_2\delta_1)$

[Neglecting the terms in $\delta_1, \delta_2 and\ \delta_3$ in power greater than 2]

$= \left(\frac{1}{n} - \frac{1}{N}\right){S_y}^2 + \beta^2\left(\frac{1}{n'} - \frac{1}{N}\right){S_x}^2 + \beta^2\left(\frac{1}{n} - \frac{1}{N}\right){S_x}^2 - 2\beta^2\left(\frac{1}{n'} - \frac{1}{N}\right){S_x}^2 + 2\beta\left(\frac{1}{n'} - \frac{1}{N}\right)S_{xy} - 2\beta(\frac{1}{n} - \frac{1}{N})S_{xy}$

$= \left(\frac{1}{n} - \frac{1}{N}\right)[{S_y}^2 + \beta^2{S_x}^2 - 2\beta S_{xy}] + (\frac{1}{n'} - \frac{1}{N})(2\beta S_{xy} - \beta^2{S_x}^2)$

$= \left(\frac{1}{n} - \frac{1}{N}\right)(1 - \rho^2){S_y}^2 + (\frac{1}{n'} - \frac{1}{N})\rho^2{S_y}^2$

Note that if the main sample is drawn independently of the initial sample, we get the same result.

Optimum choice of n and n'

For MSE of both ratio and regression estimator in case of double sampling, we may write

MSE = $\frac{V_1}{n} + \frac{V_2}{n'}$ with proper choice of $V_1 \& V_2$.

A simple cost function in case of double sampling may be given as,

$C_0 = c_1 n' + c_2 n$ where c_1 is average cost of surveying one unit in the initial sample and $c_2$is that for the final sample.

If we want optimum choice of n and n' by minimizing MSE when cost is fixed at c_0, say.

Then the objective function can be taken as follows

$\varphi = \frac{V_1}{n} + \frac{V_2}{n'} + \boldsymbol{\lambda}(c_1 n' + c_2 n - c_0)$

$$\frac{\partial}{\partial n}\varphi = 0 \Rightarrow \frac{V_1}{n^2} = \lambda c_2 \Rightarrow n = \sqrt{\frac{V_1}{\lambda c_2}}$$

Similarly, $n' = \sqrt{\frac{V_2}{\lambda c_2}}$

Also, $c_0 = c_1 n' + c_2 n = \frac{\sqrt{c_1 V_2}}{\sqrt{\lambda}} + \frac{\sqrt{c_2 V_1}}{\sqrt{\lambda}}$

$\Rightarrow \frac{1}{\sqrt{\lambda}} = \frac{c_0}{\sqrt{c_1 V_2} + \sqrt{c_2 V_1}}$

$$\therefore\ n_{opt} = \frac{C_0}{\sqrt{C_1 V_2} + \sqrt{C_2 V_1}} \sqrt{\frac{V_1}{C_2}}$$

$$\therefore\ n'_{opt} = \frac{C_0}{\sqrt{C_1 V_2} + \sqrt{C_2 V_1}} \sqrt{\frac{V_2}{C_1}}$$

$$= n_{opt} \sqrt{\frac{V_2}{V_1} \mathrm{x} \frac{C_2}{C_1}}$$

Comparison of Performance of Mean per unit Estimator in Direct Sampling with that of Regression Estimator in case of Double Sampling

In case of double sampling, a sample of size n', is chosen to observe x, and then a sub sample of size n is drawn from initial sample, and y is observed. If $c'$and c respectively be the average cost of surveying x and y for one unit. Then for the cost of survey in case of double sampling is $c'n' + cn$. Usually n' is much larger than n. Suppose we draw a sample of size n_0. *T*hen we must have $n_0 = \frac{C'n'}{C} + n$ ----------(17)

so that the cost of survey remains same for both the schemes.

Now MSE for double sampling regression estimator is

$\left(\frac{1}{n} - \frac{1}{N}\right)(1-\rho^2){S_y}^2 + (\frac{1}{n'} - \frac{1}{N})\rho^2{S_y}^2 = \left(\frac{1}{n} - \frac{1}{N}\right){S_y}^2 - \left(\frac{1}{n} - \frac{1}{n'}\right)\rho^2{S_y}^2$

And variance of mean per unit estimator in case of direct sampling is

$$(\frac{1}{n_0} - \frac{1}{N}){S_y}^2$$

Hence, double sampling regression estimator will be beneficial if,

$$\left(\frac{1}{n} - \frac{1}{N}\right){S_y}^2 - \left(\frac{1}{n} - \frac{1}{n'}\right)\rho^2{S_y}^2 < (\frac{1}{n_0} - \frac{1}{N}){S_y}^2$$

$\Rightarrow \rho^2(\frac{1}{n'} - \frac{1}{n}) < (\frac{1}{n_0} - \frac{1}{n})$

$\Rightarrow \left(\frac{1}{n} - \frac{1}{n'}\right)\rho^2 > (\frac{1}{n} - \frac{1}{n_0})$

$\Rightarrow \rho^2 > \frac{\frac{1}{n} - \frac{1}{n_0}}{\frac{1}{n} - \frac{1}{n'}} = \left(\frac{n_0 - n}{n_0}\right)\frac{n'}{n' - n} = \frac{\frac{C'n'}{C}}{\frac{C'n'}{C} + n} \cdot \frac{1}{1 - \frac{n}{n'}} = \frac{1}{(1 + \frac{cn}{C'n'})(1 - \frac{n}{n'})}$