Introduction to Sampling Theory

Lecture 27
Varying Probability Sampling

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Varying Probability Scheme Without Replacement:

Let U_i : i^{th} unit,

P_i : Probability of selection of U_i at the first draw, i = 1,2,…N

$$\sum_{i=1}^{N} P_i = 1$$

$P_{i(r)}$: **Probability of selecting U_i at the rth draw**

$$P_{i(1)} = P_i$$

$$P_{i(2)} = P_i \left[\sum_{j=1}^{N} \frac{P_j}{1-P_j} - \frac{P_i}{1-P_i} \right]$$

Varying Probability Scheme Without Replacement:

$P_{i(2)} \neq P_{i(1)}$ **for all *i* unless** $P_i = \dfrac{1}{N}$.

$P_{i(2)}$ **will, in general, be different for each *i* = 1,2,…,*N*.**

So $E\left(\dfrac{y_i}{p_i}\right)$ **will change with successive draws. This makes the varying probability scheme WOR more complex.**

Only $\dfrac{y_1}{Np_1}$ **will provide an unbiased estimator of** $\bar{Y}$**.**

In general, $\dfrac{y_i}{Np_i}(i \neq 1)$ **will not provide an unbiased estimator of** $\bar{Y}$.

Ordered Estimates:

To overcome the difficulty of changing expectation with each draw, associate a new variate with each draw such that its expectation is equal to the population value of the variate under study.

Such estimators take into account the order of the draw.

They are called the ordered estimates. The order of the value obtained at previous draw will affect the unbiasedness of population mean.

We consider the ordered estimators proposed by Des Raj, first for the case of two draws and then generalize the result.

Des Raj Ordered Estimator - Case 1: Case of Two Draws

Let y_1 and y_2 denote the values of units $U_{i(1)}$ and $U_{i(2)}$ drawn at the first and second draws respectively.

Note that any one out of the N units can be the first unit or second unit, so we use the notations $U_{i(1)}$ and $U_{i(2)}$ instead of U_1 and U_2.

Also note that y_1 and y_2 are not the values of the first two units in the population.

Further, let p_1 and p_2 denote the initial probabilities of selection of $U_{i(1)}$ and $U_{i(2)}$, respectively.

Des Raj Ordered Estimator - Case 1: Case of Two Draws

Consider the estimators

$$z_1 = \frac{y_1}{Np_1}$$

$$z_2 = \frac{1}{N}\left[y_1 + \frac{y_2}{p_2/(1-p_1)} \right]$$

$$= \frac{1}{N}\left[y_1 + y_2 \frac{(1-p_1)}{p_2} \right]$$

$$\bar{z} = \frac{z_1 + z_2}{2}.$$

Note that $\frac{p_2}{1-p_1}$ **is the probability** $P(U_{i(2)} \mid U_{i(1)})$.

Estimation of Population Mean:

First we show that $\overline{z}$ is an unbiased estimator of $\overline{Y}$.

$$E(\overline{z}) = \overline{Y}.$$

Note that $\sum_{i=1}^{N} P_i = 1.$

Consider

$$E(z_1) = \frac{1}{N} E\left(\frac{y_1}{p_1}\right)$$

$$\left(\textbf{Note that } \frac{y_1}{p_1} \textbf{ can take any one of out of the } N \textbf{ values } \frac{Y_1}{P_1}, \frac{Y_2}{P_2}, \ldots, \frac{Y_N}{P_N}\right)$$

$$= \frac{1}{N}\left[\frac{Y_1}{P_1} P_1 + \frac{Y_2}{P_2} P_2 + \ldots + \frac{Y_N}{P_N} P_N\right]$$

$$= \overline{Y}$$

Estimation of Population Mean:

$$E(z_2) = \frac{1}{N} E\left[y_1 + y_2 \frac{(1-p_1)}{p_2} \right]$$

$$= \frac{1}{N}\left[E(y_1) + E_1\left\{ E_2\left(y_2 \frac{(1-P_1)}{p_2} \middle| U_{i(1)} \right)\right\}\right] \quad \text{(Using } E(Y) = E_X[E_Y(Y \mid X)]).$$

where E_2 is the conditional expectation after fixing the unit $U_{i(1)}$ selected in the first draw.

Estimation of Population Mean:

Since $\frac{y_2}{p_2}$ can take any one of the (N – 1) values (except the value selected in the first draw) $\frac{Y_j}{P_j}$ with probability $\frac{P_j}{1-P_1}$, so

$$E_2\left[y_2 \frac{(1-P_1)}{p_2} \middle| U_{i(1)} \right] = (1-P_1) E_2\left[\frac{y_2}{p_2} \middle| U_{i(1)} \right] = (1-P_1) \sum_j \left[\frac{Y_j}{P_j} . \frac{P_j}{1-P_1} \right]$$

where the summation is taken over all the values of Y except the value y_1 which is selected at the first draw. So

$$E_2\left[y_2 \frac{(1-P_1)}{p_2} \middle| U_{i(1)} \right] = \sum_j Y_j = Y_{tot} - y_1 .$$

Estimation of Population Mean:

Substituting it in $E(z_2)$, we have

$$E(z_2) = \frac{1}{N}\left[E(y_1) + E_1(Y_{tot} - y_1)\right]$$

$$= \frac{1}{N} E(Y_{tot})$$

$$= \frac{Y_{tot}}{N} = \bar{Y}.$$

Thus

$$E(\bar{z}) = \frac{E(z_1) + E(z_2)}{2}$$

$$= \frac{\bar{Y} + \bar{Y}}{2}$$

$$= \bar{Y}.$$

Variance

The variance of $\overline{Z}$ for the case of two draws is given as

$$Var(\overline{z}) = \left(1 - \frac{1}{2}\sum_{i=1}^{N} P_i^2\right)\left[\frac{1}{2N^2}\sum_{i=1}^{N} P_i\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2\right] - \frac{1}{4N^2}\sum_{i=1}^{N} P_i^2\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2.$$

Estimation of $Var(\bar{z})$

$$Var(\bar{z}) = E(\bar{z}^2) - (E(\bar{z}))^2$$

$$= E(\bar{z}^2) - \bar{Y}^2 .$$

Since

$$E(z_1 z_2) = E\left[z_1 E(z_2 \mid U_1)\right]$$

$$= E\left[z_1 \bar{Y}\right]$$

$$= \bar{Y} E(z_1)$$

$$= \bar{Y}^2 .$$

Consider

$$E\left[\bar{z}^2 - z_1 z_2\right] = E(\bar{z}^2) - E(z_1 z_2)$$

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

$$= Var(\bar{z})$$

$\Rightarrow \widehat{Var}(\bar{z}) = \bar{z}^2 - z_1 z_2$ **is an unbiased estimator of** $Var(\bar{z})$.

Alternative form of Estimate of $Var(\bar{z})$:

$$\widehat{Var}(\bar{z}) = \bar{z}^2 - z_1 z_2$$

$$= \left(\frac{z_1 + z_2}{2} \right)^2 - z_1 z_2$$

$$= \frac{(z_1 - z_2)^2}{4}$$

$$= \frac{1}{4} \left[\frac{y_1}{Np_1} - \frac{y_1}{N} - \frac{y_2}{N} \frac{1 - p_1}{p_2} \right]^2$$

$$= \frac{1}{4N^2} \left[(1 - p_1) \frac{y_1}{p_1} - \frac{y_2 (1 - p_1)}{p_2} \right]^2$$

$$= \frac{(1 - p_1)^2}{4N^2} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2 .$$

Case 2: General case

Let $(U_{i(1)}, U_{i(2)}, \ldots, U_{i(r)}, \ldots, U_{i(n)})$ **be the units selected in the order in which they are drawn in *n* draws where** $U_{i(r)}$ **denotes that the *i*th unit is drawn at the *r*th draw.**

Let $(y_1, y_2, .., y_r, \ldots, y_n)$ **and** $(p_1, p_2, \ldots, p_r, \ldots, p_n)$ **be the values of study variable and corresponding initial probabilities of selection, respectively.**

Further, let $P_{i(1)}, P_{i(2)}, \ldots, P_{i(r)}, \ldots, P_{i(n)}$ **be the initial probabilities of** $U_{i(1)}, U_{i(2)}, \ldots, U_{i(r)}, \ldots, U_{i(n)}$, **respectively.**

Case 2: General case

Further Let

$$z_1 = \frac{y_1}{Np_1}$$

$$z_r = \frac{1}{N}\left[y_1 + y_2 + \ldots + y_{r-1} + \frac{y_r}{p_r}(1 - p_1 - \ldots - p_{r-1}) \right] \text{ for } r = 2, 3, \ldots, n.$$

Consider $\bar{z} = \frac{1}{n}\sum_{r=1}^{n} z_r$ **as an estimator of population mean** $\bar{Y}$**.**

We already have shown in case 1 that $E(z_1) = \bar{Y}$.

Case 2: General case

Now we consider $E(z_r), r = 2, 3, \ldots, n.$ **We can write**

$$E(z_r) = \frac{1}{N} E_1 E_2 \left[z_r \middle| U_{i(1)}, U_{i(2)}, \ldots, U_{i(r-1)} \right]$$

where E_2 is the conditional expectation after fixing the units $U_{i(1)}, U_{i(2)}, \ldots, U_{i(r-1)}$ **drawn in the first (*r* - 1) draws.**

Consider

$$E\left[\frac{y_r}{p_r}(1 - p_1 - \ldots - p_{r-1}) \right] = E_1 E_2 \left[\frac{y_r}{p_r}(1 - p_1 - \ldots - p_{r-1}) \middle| U_{i(1)}, U_{i(2)}, \ldots, U_{i(r-1)} \right]$$

$$= E_1 \left[(1 - P_{i(1)} - P_{i(2)} \ldots - P_{i(r-1)}) E_2 \left(\frac{y_r}{p_r} \middle| U_{i(1)}, U_{i(2)}, \ldots, U_{i(r-1)} \right) \right].$$

Case 2: General case

Since conditionally $\frac{y_r}{p_r}$ **can take any one of the** $[N-(r-1)]$

values $\frac{Y_j}{P_j}, j=1,2,\ldots,N$ **with probabilities** $\frac{P_j}{1-P_{i(1)}-P_{i(2)}\ldots-P_{i(r-1)}}$, **so**

$$E\left[\frac{y_r}{p_r}(1-p_1-\ldots-p_{r-1})\right]=E_1\left[(1-P_{i(1)}-P_{i(2)}\ldots-P_{i(r-1)})\sum_{j=1}^{N}{}^{*}\frac{Y_j}{P_j}.\frac{P_j}{(1-P_{i(1)}-P_{i(2)}\ldots-P_{i(r-1)})}\right]$$

$$=E_1\left[\sum_{j=1}^{N}{}^{*}Y_j\right]$$

where $\sum_{j=1}^{N}{}^{*}$ **denotes that the summation is taken over all the values of** y **except the** y

values selected in the first $(r-1)$ **draws like as** $\sum_{j=1(\neq i(1),i(2),\ldots,i(r-1))}^{N}$ **, i.e., except the**

values $y_1, y_2, \ldots, y_{r-1}$ **which are selected in the first** $(r-1)$ **draws.**

Case 2: General case

Thus now we can express

$$E(z_r) = \frac{1}{N} E_1 E_2 \left[y_1 + y_2 + \ldots + y_{r-1} + \frac{y_r}{p_r}(1 - p_1 - \ldots - p_{r-1}) \right]$$

$$= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + \ldots + Y_{i(r-1)} + \sum_{j=1}^{N}{}^{*} Y_j \right]$$

$$= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + \ldots + Y_{i(r-1)} + \sum_{j=1(\neq i(1), i(2), \ldots, i(r-1))}^{N} Y_j \right]$$

$$= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + \ldots + Y_{i(r-1)} + \left\{ Y_{tot} - \left(Y_{i(1)} + Y_{i(2)} + \ldots + Y_{i(r-1)} \right) \right\} \right]$$

$$= \frac{1}{N} E_1 \left[Y_{tot} \right] = \frac{Y_{tot}}{N} = \bar{Y} \qquad \text{for all} \quad r = 1, 2, \ldots, n.$$

Case 2: General case

Then

$$E(\bar{z}) = \frac{1}{n}\sum_{r=1}^{n} E(z_r)$$

$$= \frac{1}{n}\sum_{r=1}^{n} \bar{Y}$$

$$= \bar{Y}.$$

Thus $\bar{z}$ is an unbiased estimator for population mean $\bar{Y}$.

The expression for variance of $\bar{Z}$ in general case is complex but its estimate is simple.

Estimate of Variance:

$$Var(\bar{z}) = E(\bar{z}^2) - \bar{Y}^2$$

Consider for *r < s*, $E(z_r z_s) = E\left[z_r E(z_s \mid U_1, U_2, \ldots, U_{s-1})\right]$

$$= E\left[z_r \bar{Y}\right]$$

$$= \bar{Y} E(z_r)$$

$$= \bar{Y}^2$$

because for $r < s,\ z_r$ **will not contribute**

and similarly for $s < r, z_s$ **will not contribute in the expectation.**

Further, for *s < r*, $E(z_r z_s) = E\left[z_s E(z_r \mid U_1, U_2, \ldots, U_{r-1})\right]$

$$= E\left[z_s \bar{Y}\right]$$

$$= \bar{Y} E(z_s)$$

$$= \bar{Y}^2.$$

Estimate of Variance:

Consider,

$$E\left[\frac{1}{n(n-1)}\sum_{r(\neq s)=1}^{n}\sum_{s=1}^{n} z_r z_s\right] = \frac{1}{n(n-1)}\sum_{r(\neq s)=1}^{n}\sum_{s=1}^{n} E(z_r z_s)$$

$$= \frac{1}{n(n-1)} n(n-1)\overline{Y}^2 = \overline{Y}^2.$$

Substituting $\overline{Y}^2$ **in** $Var(\overline{z})$, **we get**

$$Var(\overline{z}) = E(\overline{z}^2) - \overline{Y}^2 = E(\overline{z}^2) - E\left[\frac{1}{n(n-1)}\sum_{r(\neq s)=1}^{n}\sum_{s=1}^{n} E(z_r z_s)\right]$$

$$\Rightarrow \widehat{Var}(\overline{z}) = \overline{z}^2 - \frac{1}{n(n-1)}\sum_{r(\neq s)=1}^{n}\sum_{s=1}^{n} z_r z_s.$$

Using $\left(\sum_{r=1}^{n} z_r\right)^2 = \sum_{r=1}^{n} z_r^2 + \sum_{r(\neq s)=1}^{n}\sum_{s=1}^{n} z_r z_s \quad \Rightarrow \quad \sum_{r(\neq s)=1}^{n}\sum_{s=1}^{n} z_r z_s = n^2\overline{z}^2 - \sum_{r=1}^{n} z_r^2,$

Estimate of Variance:

the expression of $\widehat{Var}(\bar{z})$ can be further simplified as

$$\widehat{Var}(\bar{z}) = \bar{z}^2 - \frac{1}{n(n-1)}\left[n^2\bar{z}^2 - \sum_{r=1}^{n} z_r^2\right]$$

$$= \frac{1}{n(n-1)}\left[\sum_{r=1}^{n} z_r^2 - n\bar{z}^2\right]$$

$$= \frac{1}{n(n-1)}\sum_{r=1}^{n}(z_r - \bar{z})^2.$$