Session 08

Double or Two-Phase Sampling

Double Sampling for Ratio Estimation

What is double sampling?

Designs in which initially a sample of units is selected for obtaining auxiliary information only, and then a second sample is selected in which the variable of interest is observed in addition to the auxiliary information.

Double sampling is also called two-phase sampling. It is useful in obtaining auxiliary variables for ratio and regression estimation. Double sampling is also useful for finding information to stratified sampling.

Ratio estimation with double sampling

- y_i - variable of interest
- x_i - auxiliary variable
- n' - number of units in the first sample (which includes the second sample)
- n - number of units in the second sample

Only in the second samples, both x_i and y_i values are observed. In the remaining units, (in the first but not the second sample), x_i but not y_i are observed. Note that observing y_i's are expensive whereas observing x_i's are not.

If x_i and y_i are highly linearly correlated and approximately passing through the origin, then the ratio estimate with double sampling may lead to improved estimates. While using the ratio estimate for double sampling, the ratio will be estimated using samples where both (x, y) are observed, i.e., the second sample, whereas τ_x will be estimated by the larger first sample.

The ratio estimator is:

$$\hat{\tau} = r\hat{\tau}_x$$

where

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i},$$

and

$$\hat{\tau}_x = \frac{N}{n'}\sum_{i=1}^{n'} x_i$$

The estimated variance of the ratio estimator is:

$$\hat{Var}(\hat{\tau}_r) = N(N-n')\frac{s^2}{n'} + N^2\frac{n'-n}{n'n(n-1)}\sum_{i=1}^{n}(y_i - rx_i)^2$$

$$\hat{Var}[E(\hat{\tau}_r \mid s_1)] \qquad E[\hat{Var}(\hat{\tau}_r \mid s_1)]$$

Note that s_1 stands for the first sample.

Example for Double Sampling

A forest resource manage is interested in estimating the total number of dead trees in a 400 acre area of heavy infestation. She subdivides the area into 200 plots of equal sizes and uses photo counts to find the number of dead trees in 18 randomly sampled plots. She then randomly samples 8 plots out of these 18 plots and conducts a ground count on these 8 plots.

Estimate the total number of dead trees in the 400 acre area.

Let *x* denote the number of dead trees in the plot by photo count and *y* the number of dead trees by ground count. The data are given as:

Plot	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
x'	5	7	10	6	7	9	3	6	8	11	5	9	12	13	3	20	15	4

Out of these 18 plots, 8 are randomly selected and a ground count is conducted.

Plot	**2**	**3**	**5**	**6**	**12**	**15**	**16**	**17**
x	7	10	7	9	9	3	20	15
y	9	13	10	11	10	4	25	17
y-rx	0.3375	0.6250	1.3375	-0.1375	-1.1375	0.2875	0.2500	-1.5625

Minitab output:

Variable	N	Mean	StDev
x′	18	8.50	4.46
x	8	10.00	5.26
y	8	12.37	6.28

Sum of x′ = 153.00, Sum of x = 80.00,
Sum of y = 99.00.

Sum of squares (uncorrected) of y-rx = 6.1928

For this example:

- $N = 200$
- $n' = 18$
- $n = 8$

Allocation in double sampling for ratio estimation

- c' - the cost of observing an *x*-variable on one unit
- c - the cost of observing an *y* -variable on one unit

The total cost = $c'n' + cn$

For a fixed total cost, the lowest variance of $\hat{\tau}_r$ is obtained by:

$$\frac{n}{n'} = \sqrt{\frac{c'}{c} \times \frac{\sigma_r^2}{\sigma^2 - \sigma_r^2}}$$

where σ_r^2 is the variance of Y about the ratio line. σ^2 is the variance of Y.

Note that here we use s_r^2 to estimate σ_r^2 and s^2 to estimate σ^2. In order for these to be reasonably good estimates, the sample size should not be too small in practical use.

To understand the above result, we can see that if the study is very large scale, for example, if $n' = 1000$, then we will select n as 75. The proportion is small since 0.885 is small compared to $(6.28)^2$.

Double Sampling for Stratification

In some sampling situations, units can be assigned to strata only after the sample is selected.

The method of post-stratification is useful only if the relative proportion of each stratum in the population $W_h = \frac{N_h}{N}$ is known for each stratum *h*. If these proportions are not known, double sampling may be used, with an initial (large) sample used to classify the units into strata and then a stratified sample selected from the initial sample.

The two steps of double sampling for stratification:

Step 1: n' initial simple random samples are selected from a population of N units. These units are classified into strata, with n'_h observed to be in stratum h. The population proportion $W_h = \frac{N_h}{N}$ is estimated by the sample proportion: $w_h = \frac{n'_h}{n'}$, $h = 1, \ldots, L$.

Step 2: A second sample is then selected by stratified random sampling from the first sample. These units are classified into strata, with n_h units selected from the n'_h sample units in stratum h. Measurement of y_{hi} is recorded for each unit in the second sample.

We denote the sample mean in stratum h in the second sample as: $\bar{y}_h = \sum_{i=1}^{n_h} y_{hi} \Big/ n_h$

An estimate for population mean is thus: $\bar{y}_d = \sum_{h=1}^{L} w_h \bar{y}_h$

Note that $\bar{y}_d$ is unbiased.

The decomposition of variances of two phase sampling is:

Let s_1 denote the first-phase sample, then

$$Var(\bar{y}_d) = Var\left[E(\bar{y}_d \mid s_1)\right] + E\left[Var(\bar{y}_d \mid s_1)\right]$$

Thus, the variance of the estimate for population mean is:

$$Var(\bar{y}_d) = \frac{N-n'}{N} \times \frac{\sigma^2}{n'} + E\sum_{h=1}^{L}\left[\left(\frac{n'_h}{n'}\right)\left(\frac{n'_h - n_h}{n'_h}\right)\frac{\sigma^2_{h(s_1)}}{n_h}\right]$$

where σ^2 is the overall population variance and $\sigma^2_{h(s_1)}$ is the population variance within stratum h for the particular first–phase sample s_1.

An unbiased estimate for the variance of the estimate is:

$$\hat{Var}(\bar{y}_d) = \frac{N-n'}{N} \times \frac{1}{n'-1}\sum_{h=1}^{L} w_h(\bar{y}_h - \bar{y}_d)^2 + \frac{N-1}{N}\sum_{h=1}^{L}\left[\left(\frac{n'_h - 1}{n'-1} - \frac{n_h - 1}{N-1}\right)\frac{w_h s_h^2}{n_h}\right]$$

where s_h^2 is the stratum sample variance from the second sample.

Example for Double Sampling for Stratification

Average number of pairs of shoes owned by students living in a College town neighborhood.

A shoe store wants to estimate the average number of pairs of shoes owned by the students who live in a certain college town neighborhood. They think that a stratified sample based on gender is a good approach to take but do not know the makeup of the gender in that neighborhood. They also do not know the gender of the respondent until after contacting them. So, they use double sampling by first contacting 160 randomly selected students in that neighborhood and ask them about their gender. It turns out that 64 are males and 96 are females. They then randomly sample 8 males and 12 females, provide them a $10.00 incentive for going home to count the number of pairs of shoes and report them.

The data are given in the table below:

Male	5	6	9	5	9	7	5	8				
Female	17	19	13	16	8	11	15	19	12	13	33	20

Variable	N	Mean	StDev
male	8	6.750	1.753
female	12	16.33	6.37

Solution:

To estimate the average pairs of shoes, they use a double sampling:

Step 1: 160 students are randomly sampled to find out their gender. Result: 64 male, 96 female

Step 2: stratify by gender and randomly sample 8 males and 12 females.

male: $n'_1 = 64$, female: $n'_2 = 96$

$n' = n'_1 + n'_2 = 160$