COURSE CODE: STS 443

COURSE TITLE: SAMPLING TECHNIQUES II

NUMBER OF UNIT: 3 UNITS

COURSE DURATION: THREE HOURS PER WEEK.

COURSE COORDINATOR: DR GODWIN NWANZU AMAHIA B.Sc, M.Sc, ph.D (go.amahia@mail.ui.edu.ng)

LECTURER OFFICE LOCATION: HOD'S STATISTICS OFFICE

COURSE CONTENT:

Difference and regression estimation procedures, Cluster sampling with unequal sizes, Multi – stage and multi – phase sampling, Double sampling, Interpenetrating scheme. Problems of optimal allocation with more than one item. Sources of error in survey.

COURSE REQUIREMENTS:

This is a compulsory course for all statistics students. Students are expected to have a minimum of 75% attendance to be able to write the final examination.

READING LIST:

1.) **Williams G.C Sampling techniques, 3rd edition. John Willey and Sons. New Yolk, 1977.**
2.) **Des R. Sampling theory. Tata Mc Graw – Hill, 1968.**
3.) **Okafor F.C. Sampling survey theory with applications. Afro – Qrbis pub, Nsukka 2002.**
4.) **Kish L. Survey sampling. New york, John Willey, 1965.**
5.) **Chaudhuri A. and S. Hort. Survey Sampling Theory and Methods. Marcel Dekker, New york, 1992.**
6.) **Murthy, M.N. Sampling Theory and Methods. Statistical Publication Society. Calcutta, 1967.**

LECTURE NOTES

LECTURE ONE

USE OF AUXILIARY INFORMATION IN SRS SCHEME

Let us assume that a srs of size n is to be drawn form a finite pop containing N elements. How can we estimate a pop mean μ, a total T_Y, or a ratio R, utilizing sample information on y and an auxiliary variable X?

a. Estimation of the pop ratio R,

$$\hat{R} = r = \sum_{1}^{n} yi \Big/ \sum_{1}^{n} xi = \frac{\bar{y}}{\bar{x}}$$

b. Estimation of the pop mean $\mu = \hat{R}\,\bar{x} = \hat{R}\,\mu_x$

c. Estimation of the pop total $T_y = \hat{R}\,N\,\bar{X} = N\,\hat{R}\,\mu_x = \hat{R}\,X$

Ratio Estimator of the population total $\hat{T}_y$

$$\hat{T}_y = \frac{\bar{y}}{\bar{x}} X, \quad X = \sum_{i=1}^{N} x_i \quad \ldots\ldots\ldots \qquad (2.9)$$

$$\hat{v}\left(\hat{T}_y\right) = X^2 \hat{V}\left(\frac{\bar{y}}{\bar{x}}\right) = N^{2-} \bar{X}^2 \hat{V}\left(\frac{\bar{y}}{\bar{x}}\right) \ldots\ldots\ldots \qquad (2.10)$$

$$= X^2 \left(\frac{N-n}{nN}\right)\left(\frac{1}{N^2 x}\right) \sum_{i-1}^{n} \frac{(y_i - rx_i)^2}{(n-1)} \ldots\ldots \qquad (2.11)$$

Where μ_x is the population mean for the random variable X.

Example 2: In a study to estimate the total sugar content of a trade load of oranges, random sample of μ = 10 oranges was juiced and weighed. The total weight of all the oranges, obtained by first weighing the trade loaded and then unloaded, was found to be 1800pd. Estimate Ty, the total sugar content for the oranges and the standard error of your estimate.

Range	1	2	3	4	5	6	7	8	9	10	
Sugar content: Y	.021	.031	.025	.022	.033	.027	.019	.021	.023	.025	$\sum y_i = 0.246$
Weight of Orange: X	.40	.48	.43	.42	.50	.46	.39	.41	.42	.44	$\sum x_i = 4.35$

Note: N and μ_x are …………..

Use $\bar{x}$ in place of μ_x and assume an estimator of T_y given by (N-n)/ μ = 1

$$\hat{T}_y = \frac{\sum_{i=1}^{10} yi}{\sum_{i=1}^{10} xi}(Tx) = \frac{0.246}{4.35}(1800) = 101.79\,pd.$$

$$\sum_{i=1}^{10} y^2 i = 0.006224, \quad \sum_{i-1}^{10} x^2 i = 1.9035, \quad \sum_{i=1}^{10} xiyi = 0.10839$$

$$\bar{x} = \frac{4.35}{10} = 0.435, \quad \frac{\bar{y}}{x} = \frac{0.246}{4.35} = 0.05655$$

$$\sum_{i=1}^{n} (y_i - rx_i)^2 = \sum_{i=1}^{n} y_i^r + r^2 \sum_{i=1}^{10} x_i^2 - 2r \sum_{i=1}^{10} x_i y_i$$

= 0.006224 + (0.05655) (1.9035) – 2 (0.05655) (.10839)

= 0.000052285

$$\hat{V}(\hat{T}_y) = (1800)^2 \left(\frac{1}{10}\right)\left[\frac{1}{(0.435)^2}\right]\left(\frac{0.000052285}{9}\right) = 9.94720$$

Se $\left(\hat{T}_y\right) = \sqrt{9.94720}$

= 3.1539

Ratio estimator of a population mean μ_Y

$$\hat{\mu}_y = \frac{\sum_{i=1}^{n} yi}{\sum_{i=1}^{n} x_i}(\mu_x) = r\mu_x \quad \ldots\ldots\ldots\ldots\ldots (2,12)$$

$$\hat{V}\left(\hat{\mu}_y\right) = \mu_x^2 \hat{V}\left(r\right) = \mu_x^2 \left(\frac{N-n}{nN}\right)\left(\frac{1}{\mu_X^2}\right)\sum_{1=i}^{n} \frac{\left(y_i - rx_i\right)^2}{\left(n-1\right)}$$

$$= \left(\frac{N-n}{nN}\right)\sum_{i=1}^{n}\frac{(y_i - rx_i)^2}{(n-1)} \qquad \text{………… (2.13)}$$

Example 3: A company wishes to estimate the average amount of money μ_y paid to employees for medical express during the first three months of the current year. Average quarterly reports are available in the fiscal reports of the previous year. A random sample of 100 employee records is taken from the population of 1000 employees. The sample results are summarized below. Estimate the average amount of money μ_y.

n = 100, N=1000

Total for the current quarter = $\sum_{i=1}^{100} y_i = 1750$

Total for the corresponding quarter of the previous year = $\sum_{i=1}^{100} x_i = 1200$

Population total for the corresponding quarter of the previous year = T_x = 12,500

$\sum_{1}^{n} y_i^2 = 31{,}650$, $\sum_{1}^{n} x_i^2 = 15{,}620$, $\sum_{1}^{n} x_i y_i = 22{,}059.35$

The estimate of μ_x is

$$\hat{\mu}_y = r\,\mu_x$$

Where $\mu_x = \frac{T_x}{N} = \frac{12{,}500}{1000} = 12.50$

Then $\hat{\mu}_y = \frac{\sum_{1}^{n} y_i}{\sum_{1}^{n} x_i}\left(\mu_x\right) = \frac{1750}{1200}\left(12.50\right) = 18.23$

$$\sum_{i=1}^{n}\left(y_i - rx_i\right)^2 = \sum_{1}^{100} y_i^2 + r^2\sum_{1}^{100} x_i^2 - 2r\sum_{1}^{100} x_i y_i$$

$= 31{,}650 + (1.4583)^2 (15{,}620) - (2.9166)(22{,}059.35)$

$= 441.68$

$$\hat{V}(\hat{\mu}_Y) = \left(\frac{N-u}{nN}\right)\sum_{i=1}^{n} \frac{(y_i - rx_i)^2}{(n-1)}$$

$$= \frac{1000-100}{100(1000)}\left(\frac{441.68}{99}\right)$$

$$= 0.0401527$$

$$\text{Se } (\hat{m}_y) = \sqrt{0.0401527}$$

$$\approx 0.20$$

LECTURE TWO

REGRESSION ESTIMATION

We observed that the ratio estimator is appropriate when the relationship between Y and X is linear through the origin. If there is evidence of a linear relationship between the observed Y's and X's, but not necessarily one that would pass through the origin, then this extra information provided by the auxiliary variable x may be taken into account through a regression estimator of the mean μ_y. One must still have a knowledge of μ_x, before the estimator can be employed. The underlying line that shows the relationship between the Y's and X's is referred to as the regression line of Y on X.

Regression estimator of a population mean μ_y

$$\hat{\mu}_Y L = \bar{y} + b(\mu_x - \bar{x}) \quad \ldots\ldots\ldots\ldots\ldots (2.14)$$

Where b = $\sum_{i=1}^{n} \frac{(yi - \bar{y})(xi - \bar{x})}{\sum_{i=1}^{n}(xi - \bar{x})^2} = \sum_{i=1}^{n} \frac{xiyi - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$

Estimated variance of $\hat{\mu}yL$ is

$$\hat{V}(\hat{\mu}yL) = \left(\frac{N-n}{nN}\right)\left(\frac{1}{n-2}\right)\left[\sum_{i=1}^{n}(yi-\bar{y})^2 - b^2\sum_{i=1}^{n}(xi-\bar{x})^2\right]$$

Example 4: A mathematics achievement test was given to 486 students prior to their entering a certain college. From these students a simple random sample of n = 10 students was selected and their progress in calculus observed. Final calculus grades were then reported, as given in the table below. It is known that $\mu_x = 52$ for all 486 students taking the achievement test. Estimate μ_Y for this population.

Student	1	2	3	4	5	6	7	8	9	10
Achievement T.S.X	39	43	21	64	57	47	28	75	34	52

Final calculusla. Y	65	78	52	82	92	89	73	98	56	75

$\bar{y}$ =76, $\bar{x}$ =46

$$b = \frac{\sum_{1}^{n} xiyi - m\bar{x}\bar{y}}{\sum x^2 i - n\bar{x}^2} = \frac{36854 = 10(46)(76)}{23{,}634 - 10(46)^2}$$

$$= 0.766$$

$\sum_{i=1}^{n}(yi - \bar{y})^2$ =2056; $\sum_{i=1}^{n}(xi - \bar{x})^2$ = 2474

$\hat{\mu}_Y L = \bar{y}$ - $b(\mu_x - \bar{x})$=76+(0.766)(52-46)=80

$$\hat{V}(\hat{\mu}_Y L) = \left(\frac{N-n}{nN}\right)\left(\frac{1}{n-2}\right)\left[\sum_{i=1}^{n}(yi - \bar{y})^2 - b^2 \sum_{1}^{n}(xi - \bar{x})^2\right]$$

$$= \frac{486-10}{486(10)}\left(\frac{1}{8}\right)[2056 - (0.766)^2(2474)]$$

$\hat{V}(\hat{\mu}yL)$ = 7.397

Se($\hat{\mu}yL$) = 2.71774

$\cong$ 2.7

LECTURE THREE

Difference Estimation

The difference method of estimating a population mean or total is similar to the regression method in that it adjusts the estimation of the parameter under consideration (seq) $\bar{y}$ value up or down by an amount depending on the difference $(\mu x - \bar{x})$. However, the regression coefficient b is not computed. In effect, b to is set equal to unity. The function $(\mu x - \bar{x})$ is called a zero function, its expected value is identically equal to zero.

Difference estimator of a Population Mean μ_y

$$\hat{\mu}_{yD} = \bar{y} + (\mu x - \bar{x}) = (\mu x + \bar{d}) \qquad \text{…. (2.15)}$$

Where $\bar{d} = (\bar{y} - \bar{x})$

Estimated variance of $\hat{\mu}_{yD}$ is

$$\hat{V}(\hat{\mu}_{yD}) = \left(\frac{N-n}{nN}\right) \frac{\sum_{i=1}^{n}(di - \bar{d})^2}{(n-1)}$$

…. (2.16)

Example 5:

Suppose a population contains 180 inventory items with a stated book value of ₦13,320.0. Let xi denote the book value and yi the audit value of the ith items. A simple random sample of n = 10 items yields the results in the table below. Estimate the mean audit value of μ_y by the difference method and estimate the variance of $\hat{\mu}_{yD}$.

Sample	Audit value, yi	Book value, xi	Di
1	9	10	-1
2	14	12	+2
3	7	8	-1
4	29	26	+3

5	45	47	-2
6	109	112	-3
7	40	36	+4
8	238	240	-2
9	60	59	+1
10	170	167	+3

$\bar{y}$ = 72.1, $\bar{x}$ =71.7, μx =74.0

$\hat{\mu}_{yD} = \mu x + \bar{d}$ =74.0 + (72.1 – 71.7) =74.4

Also $(\frac{1}{ny})\sum_{i=1}^{10}(di - \bar{d})^2$ = $(\frac{1}{n-1})\ (\sum_{i=1}^{10} (d^2 i - n\bar{d}^2)$

= $\frac{58 - 10(0.4)2}{9} = 6.27$

Thus

$\hat{V}(\hat{\mu}_{yD})$ = $(\frac{N-n}{nN})\ \sum_{i=1}^{10}\frac{(di-\bar{d})^2}{(n-1)}$

= $[\frac{180-10}{(10)(180)}]\ (6.27)$

= 0.59

Se $(\hat{\mu}_{yD})$ = 0.768

LECTURE FOUR

Two Stage Cluster Sampling

Introduction

The procedure of first selecting clusters and then choosing a specified number of elements from each selected cluster is known as sub-sampling. It is also called two-stage sampling. The clusters that form the units of sampling at the first stage are called the first stage units or primary sampling units (PSU). The elements or groups of elements within clusters which form the units of sampling at the second stage are called sub-units or second –stage units (SSU).

TWO-STAGE SAMPLING, EQUAL FIRST-STAGE UNITS: Estimation of the Population Mean and Total

We shall assume that the target population has NM_i elements grouped into N first-stage units, each containing M_i second-stage units.

Let:

N = the number of the clusters in the population

n = the number of clusters selected in a simple random sample.

Mi = the number of elements in cluster i.

m_i = the number of elements in a simple random sample from cluster i.

M = $\sum_{I=I}^{N} M_i$ = the number of elements in the population

$\overline{M}$ = $\frac{M}{N}$ = the average cluster size for the population

y_{ij} = the j^{th} observation in the sample from the ith cluster.

$\bar{y}_i$ = $\frac{1}{m_i} \sum_{j=1}^{mi} y_{ij}$ = the sample mean for the ith cluster.

An unbiased estimator of the population mean is given by

$$\hat{\mu} = \left(\frac{N}{M}\right) \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{n} \quad \text{….. (1.1)}$$

The estimate variance of $\hat{\mu}$ is

$$\hat{V}(\hat{\mu}) = (\frac{N-n}{N})(\frac{1}{n\bar{M}^2})S_b^{\,2} + (\frac{1}{nN\bar{M}^2})\sum_{i=1}^{n} M_i^{\,2}(\frac{M_i - M_i}{M_i})\frac{S_{wi}^{\,2}}{m_i} \qquad \text{…. (1.2)}$$

Where

$$S_b^2 \quad = \quad \frac{1}{(n-1)}\sum_{i-1}^{n}(M_i\bar{y}_i - \bar{M}\hat{\mu})^2 \qquad \text{…. (1.3)}$$

And

$$S^2_{wi} \quad = \quad \frac{1}{(mi-1)}\sum_{j-1}^{mi}(y_{ij} - \bar{y}_i)^2, \qquad i = 1, 2, \ldots, n \qquad \text{…. (1.4)}$$

The ratio estimator of the population mean is

$$\hat{\mu}_r \quad = \quad \frac{\sum_{i=1}^{n} M_i\bar{y}_i}{\sum_{i=1}^{n} M_i} \qquad \text{…. (1.5)}$$

Estimated variance of $\hat{\mu}_r$ is

$$\hat{V}(\hat{\mu}_r) = (\frac{N-n}{nN})(\frac{1}{\bar{M}^2})S_b^{\,2} + (\frac{1}{nN\bar{M}^2})\sum_{i=1}^{n} M_i^{\,2}(\frac{M_i - M_i}{M_i})\frac{S_{wi}^{\,2}}{m_i} \qquad \text{…. (1.6)}$$

Where

$$S_b^2 \quad = \quad \sum_{i=1}^{n}\frac{M_i^{\,2}(\bar{y}i - \hat{\mu}r)^2}{(n-1)} \qquad \text{…. (1.7)}$$

And

$$S_{wi}^2 \quad = \quad \sum_{j=1}^{mi}\frac{(yi_j - \bar{y}_i)^2}{(m_i - 1)}, \quad i = 1, 2, .., n \qquad \text{…. (1.8)}$$

An estimator of the population total in given by

$$\hat{Y} \quad = \quad M\hat{\mu} = (\frac{N}{n})\sum_{i=1}^{n} M_i\bar{y}_i \qquad \text{…. (1.9)}$$

The estimated variance is

$$\hat{V}(\hat{Y}) \quad = \quad M^2\hat{V}(\hat{Y})$$

$$= \quad (\frac{N-n}{N})(\frac{N^2}{n})\, S^2{}_b + \frac{N}{n}\sum_{i=1}^{n} M_i^{\,2}(\frac{M_i - M_i}{M_i})\frac{S^2{}_{wi}}{m_i} \qquad \text{…. (1.10)}$$

Where S_b^2 and S_{wi}^2 are given by equations (1.7) and (1.8) respectively.

Example 1:

A nursery man wants to estimate the average height (in inches) of 1200 seedlings in a field that is sub-divided into 50 plots that vary in size. A two-stage cluster sample design produced the following data.

Plot	**Number of seedlings M_i**	**Number of seedlings sampled m_i**	**Height of seedlings y_{ij}**
1	63	6	5, 2, 4, 3, 1, 5
2	57	8	4, 2, 7, 2, 7, 2
3.	30	3	3, 2, 5
4.	23	2	4, 4,
Total	173	17	

(i) Estimate the average height of seedlings in the field and the standard error of the estimate

(ii) Construct a 95 per cent confidence interval on the population mean

Solution

Plot	**Number of seedlings M_i**	**Number of seedlings sampled m_i**	$M_i \bar{y}_i$	S_{wi}^2	$M_i^2 (\frac{1}{m_i} - \frac{1}{M_i}) S^2_{wi}$
1	63		210.00	2.67	1,706.575
2	57		228.00	6.00	2,907.00
3.	30		100.00	2.34	631.80
4.	23		92.00	0.00	-
Total	173		630.00	-	5,245.375

(i) The average height of seedlings in the field is given by

$$\hat{\bar{Y}} = ((\frac{N}{M}) \sum_{i-1}^{n} \frac{M_i \bar{y}_i}{n}$$

$= \quad (\frac{50}{1200})(\frac{630}{4}) = 6.5625$

$\hat{=} \quad 6.6$

The estimated variable of $\hat{\hat{Y}}$ is given by

$$\hat{V}(\hat{\hat{Y}}) = (\frac{N-n}{N})(\frac{S_b^{\ 2}}{n\overline{M}^2}) + (\frac{1}{nN\overline{M}^2})\sum_{i-1}^{n} M_i^{\ 2}(\frac{M_i - m_i}{M_i})\frac{S_{wi}^{\ 2}}{m_i}$$

$= \quad 2.0395 + 0.045532769$

$\hat{=} \quad 2.09$

The standard error is given by

$\text{Se}(\hat{\hat{Y}}) \quad \hat{=} \quad 1.4$

(ii) A 95% confidence interval is given by

$\hat{\hat{Y}} \pm 1.96\sqrt{\hat{V}(\hat{\hat{Y}})}$

Or $\quad 6.5625 \pm 2.744$

i.e. $\quad 6.56 \pm 2.74$

Thus the average height is estimated to be 6.56 inches. The error of estimation should be less than 2.74 inches with a probability of approximately 0.95.

Estimation of a population proportion

Consider the problem of estimating a population proportion P such as the proportion of unemployed in a Local Government Area in a State at a particular time. An estimate of P can be obtained by using $\hat{\mu}$, given in equation (1.1) or $\hat{\mu}_r$ in (1.5) and letting y_{ij}, = 1 or 0 depending on whether or not the j^{th} element in the ith cluster falls into the category of interest. In many problems of practical application, M is usually unknown. Let $\hat{p}i$ denote the proportion of sampled elements from cluster i that fall into the category of interest.

An estimator of the population proportion p is given by

$$\hat{p} = \frac{\sum_{i=1}^{n} Mi\hat{p}_i}{\sum_{i=1}^{n} M_i} \quad \text{.......(1.11)}$$

The estimated variance of $\hat{p}$ is

$$\hat{V}(\hat{p}) \quad = \quad (\frac{N-n}{N})(\frac{1}{n\bar{M}^2})S_p{}^2 + (\frac{1}{nN\bar{M}^2})\sum_{i-1}^{n} M_i{}^2(\frac{M_i = m_i}{M_i})(\frac{\hat{p}_i \hat{r}i}{m_i - 1}) \quad \text{…. (1.12)}$$

$$\text{Where } S_p{}^2 \quad = \quad (\frac{1}{n-1})\sum_{i-1}^{n} M_i{}^2(\hat{p}_i - \hat{p}^2)^2 \quad \text{….. (1.13)}$$

And $\hat{q}i$ = $(1\text{-}p_i)$

Example 2:

In an urban household survey, a Local Government Area (LGA) consists of 26 Enumeration Areas (EAs) from which a random sample of 4 EAs was selected. Within each selected EA, a probability sample of one in five households was selected. Information on households headed by women was collected as shown below.

(i) Calculate the proportion of households headed by woman and its standard error.

(ii) What is the social significance of the result in (i)?

Solution

EA NO	Household H/H		H/H Headed by woman	$\hat{p}_i$	$M_i\hat{p}_i$	$M_i{}^2(\hat{p}_i - \hat{p})^2$	$(M_i\hat{p}_i - \frac{i}{n}\sum_1^n M_i\hat{p}_i)^2$	$M_i{}^2(\frac{1}{m_i} - \frac{1}{M_i})\frac{\hat{p}_i\hat{q}_i}{m_i - 1}$
	M_i	m_i						
1	70	14	5	0.357	24.99	48.3164	1.3514	69.216
2	104	21	5	0.238	24.752	4.1976	1.9614	78.2559
3	116	23	8	0.348	40.368	109.7214	202.0804	111.2954
4	116	24	3	0.1	14.5	236.9506	135.7808	50.7298

				25	0			
Total	406	-	-		104.61	399.186	341.174	309.4971

(i) The estimate of the proportion headed by women is

$\hat{p} \quad = \quad \sum_{i=1}^{n} M_i \hat{p}_i \,/\, \sum_{i=1}^{n} Mi$

$= \quad (104.61)/406$

$= \quad 0.2577$

$\triangleq \quad 25.8\%$

$$\hat{V}(\hat{p}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{n\bar{M}^2}\right)S^2 p + \left(\frac{1}{nN\bar{M}^2}\right)\sum_{i=1}^{n} M^2 i\left(\frac{Mi-mi}{Mi}\right)\left(\frac{\hat{p}i\hat{q}i}{mi-1}\right)$$

$= \quad 0.003021$

Se($\hat{p}$) $\triangleq$ 5.5%

(ii) $\hat{p}$ = 25.8% provides a useful measure of women autonomy and parity with men.

LECTURE FIVE

DOUBLE SAMPLING: $\bar{X}$ not known

The classical regression type estimator of $\bar{Y}$ assumes knowledge of the population mean $\bar{X}$. However, $\bar{X}$ is often unknown. Assume a large random sample of size n_1 drawn to estimate $\bar{X}$, while a subsample of size n is drawn from n_1 to observe the characteristics Y under study.

Since $\bar{x}'$ based on n_1 units is an unbiased estimate of $\bar{X}$, a regression type estimator appropriate to this situation is

$\bar{y}_d = \bar{y} + \hat{\beta}(\bar{x}' - \bar{x})$………(1.1)

Clearly, $\bar{y}_d$ is a biased estimate of $\bar{Y}$

$E(\bar{y}_d) = E_1E_2(\bar{y}_d)$……..(1.2)

Where the subscripts 1, 2 denote varieties on the first and second phases of sampling

$E_2(\bar{y}_d) = \bar{y}' + \beta(\bar{x}' - \bar{x}')$

$= \bar{y}'$ (replacing $\hat{\beta}$ by β)……. .(1.3)

$E_1(\bar{y}'_d) = \bar{Y}$……...(1.4)

$V(\bar{y}_d) = V_1E_2(\bar{y}_d) + E_1V_2(\bar{y}_d)$……..(1.5)

$V_1E_2(\bar{y}_d) = V(\bar{y}') = \left(\frac{1}{n_1} - \frac{1}{N}\right)S_y^2$……(1.6)

Now, $\bar{y}_d = \frac{1}{n}\sum_1^n y_i - \frac{\beta}{n}\sum_1^n x_i + \sum_1^1 x_i$

$= \frac{1}{n}\sum_1^n (y_i - \beta x_i)$

$\bar{y}_d = \frac{1}{n}\sum_1^n \mu_i - \frac{\beta}{n}\sum_1^n x_i, \quad \mu_i = y_i - \beta x_i$

Further, regard the large sample as a finite population

$E_1V_2(\bar{y}_d) = E_1V_2\left(\frac{1}{n}\sum_1^n \mu_i\right)$

$= E_1\left(\frac{1}{n} - \frac{1}{n_1}\right)S^2\mu'$…….(1.7)

$\left(\frac{1}{n} - \frac{1}{n_1}\right)S_y^2(1 - p^2)$……(1.8)

Since $S^2\mu'$ is an unbiased estimate of $S_\mu^2 = S_y^2(1 - p^2)$

Substituting (1.6) and (1.8) in (1.5), we have

$V(\bar{y}_d) = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_y^2(1 - p^2)$………(1.9)

$\triangleq \frac{1}{n}S_y^2(1 - p^2) + \frac{1}{n_1}\rho^2S_y^2 - \frac{1}{N}S_y^2$…………(1.10)

Although this result is useful, its usefulness will be greatly increased if the cost of observing both X and Y is also taken into consideration. We should choose that strategy which for a fixed cost C_0 we can estimate $\bar{Y}$ with maximum precision. If c_1 and c_2 are the unit cost of observing Y and X respectively, the total cost of the survey apart from overhead costs may be expressed as

$C = c_1n + c_2n_1$………..(1.11)

Define the Langragian function:

$F(n, n_1, \lambda) = \frac{1}{n}S_y^2(1 - p^2) + \frac{1}{n_1}\rho^2S_y^2 - \frac{1}{N}S_y^2 + \lambda(c_1n + c_2n_1 - c)$…….(1.12)

$\frac{dF}{dn} = 0 \Rightarrow n = \{S_y^2(1-p^2)\}^{1/2}/\sqrt{c_1\lambda}$........(1.13)

$\frac{dF}{dn_1} = 0 \Rightarrow n_1 = (p^2 S_y^2)^{1/2}/\sqrt{c_2\lambda}$........(1.13)

$\frac{n}{n_1} = \left[\frac{(1-p^2)}{p^2}\frac{c_2}{c_1}\right]^{1/2}$.........(1.14)

i.e. $n = n_1\left[\frac{(1-p^2)}{p^2}\frac{c_2}{c_1}\right]^{1/2}$.........(1.15)

substituting for n in (1.11), we have (noting that $c = c_1 n + c_2 n_1$)

$$C = c_1 n_1\left[\frac{(1-p^2)}{p^2}\frac{c_2}{c_1}\right]^{1/2} + c_2 n_1$$

$n_1 = \frac{c}{c_2 + c_1\left[\frac{(1-p^2)}{p^2}\frac{c_2}{c_1}\right]^{1/2}}$..........(1.16)

$n = \frac{c}{c_1 + c_2\left[\frac{(p^2)}{1-p^2}\frac{c_1}{c_2}\right]^{1/2}}$...........(1.17)

Substituting .(1.16) and (1.17) in (1.10) we have

$V_0(\bar{y}_d) = \frac{1}{c}\left[\sqrt{c_1(1-p^2)} + \sqrt{c_2 p^2}\right]^2 S_y^2 - \frac{1}{N}S_y^2$.......(1.18)

The variance of an estimator in single sampling is $V(\bar{y}_s) = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2$ ....(1.19)

When the minimum total cost is C, we have from the cost function $n = \frac{c}{c_1}$. The optimum variance in single sampling becomes

$V_0(\bar{y}_s) = \frac{c_1}{c}S_y^2 - \frac{1}{N}S_y^2$..........(1.20)

Double sampling will be more efficient than single sampling if

$V_0(\bar{y}_d) \le V_0(\bar{y}_s)$ i. e.when

$V_0(\bar{y}_s) - V_0(\bar{y}_d) \ge 0$.........(1.21)

Using (1.18) and (1.20) this works out as

$p^2(c_1 - c_2) \ge 2\sqrt{c_1 c_2(1-p^2)p^2}$

$p(c_1 - c_2) \ge 2\sqrt{c_1 c_2(1-p^2)}$

$p^2(c_1 - c_2)^2 \ge 4c_1 c_2(1-p^2)$

Or alternatively

$p^2 \ge \frac{4c_1 c_2}{(c_1 + c_2)^2}$...........(1.22)

Extend the results to ratio $\bar{y}_d = \frac{\bar{y}}{\bar{x}}\bar{X}'$

LECTURE SIX

CLUSTER SAMPLING

Introduction

The smallest unit in which a survey population can be subdivided is called an element: a collection of elements is called a cluster.

Definition:

A cluster sample is a simple random sample in which each sampling unit is a collection of elements or cluster.

Cluster sampling is less costly than simple or stratified random sampling if the cost of obtaining a sampling frame that lists all population of elements is very high or if the cost of obtaining observations increases as the distance separating the elements increases.

Cluster Sampling

Consider the following notations:

N= number of clusters in the population

N= number of clusters selected in a simple random sampling

m_i – number of elements in cluster i, $i - 1 \dots N$

$\bar{m} = \frac{1}{n}\sum_{i-1}^{n} m_i$= average cluster size for the sample

M= $\sum_{i=1}^{N} m_i$= number of elements in the population

$\bar{M} = \frac{M}{N} =$ average cluster size for the population

y_i = total of all observations in the ith cluster

Estimate of population mean is

$$\bar{y} = \sum_{i=1}^{n} y_i \Big/ \sum_{i=1}^{n} m_i$$

$$\hat{V}(\bar{y}) = \left(\frac{N-n}{N_n \bar{M}^2}\right)\frac{\sum_{i=1}^{n}(y_i - \bar{y}m_i)^2}{(n-1)} \text{......(1.1)}$$

Estimate of population total

$\hat{Y}=M\bar{y}$

$= M\sum_{i=1}^{n} y_i \,/\, \sum_{1}^{n} m_i$.......(1.2)

$\hat{V}(\hat{Y}) = N^2\left(\frac{N-n}{Nn}\right)\sum_{i=1}^{n}\frac{(y_i - \bar{y}m_i)^2}{(n-1)}$...(1.3)

Or

$N\bar{y}_t = \frac{N}{n}\sum_{i=1}^{n} y_i$......(1.4)

$\hat{V}(N\bar{y}_t) = N^2\left(\frac{N-n}{Nn}\right)\sum_{i=1}^{n}\frac{(y_i - \bar{y}_t)^2}{(n-1)}$......(1.5)

Observe that (1.5) is independent of M.

Estimate of population proportion

Let a_i denote the total number of elements in cluster i that possesses the Characteristics of interest

$\hat{P} = \sum_{i=1}^{n} a_i \,/\, \sum_{i=1}^{n} m_i$.......(1.6)

$\hat{V}(\hat{P}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)\sum_{i=1}^{n}\frac{(a_i - \hat{P}m_i)^2}{(n-1)}$......(1.7)

Example: A simple random sample of 5 blocks from 40 was selected. The objective was to estimate the number of residents aged 65 and above in the population. The result is shown below.

No of residents=m_i	No aged 65 and over a_i	$\hat{p}\, m_i$	$a_i - \hat{p}\, m_i$	$(a_i - \hat{p}\, m_i)^2$
90	15	21.60	-6.60	43.560
32	8	7.08	0.32	0.1024
47	14	11.28	2.72	7.3984
25	9	6.00	3.00	9.0000
16	4	3.84	0.16	0.0256
210	50			60.0864

$$\hat{p} = \frac{\sum_1^5 a_i}{\sum_1^5 m_i} = \frac{50}{210} = 0.24$$

$$\hat{V}(\hat{p}) = \left(\frac{N-n}{N n \bar{M}^2}\right) \sum_{i=1}^{n} \frac{(a_i - \hat{p}\, m_i)^2}{(m-1)}$$

$$\frac{35}{40 \times 5 \times 42^2}\left(\frac{1}{4}\right)(60.0864)$$