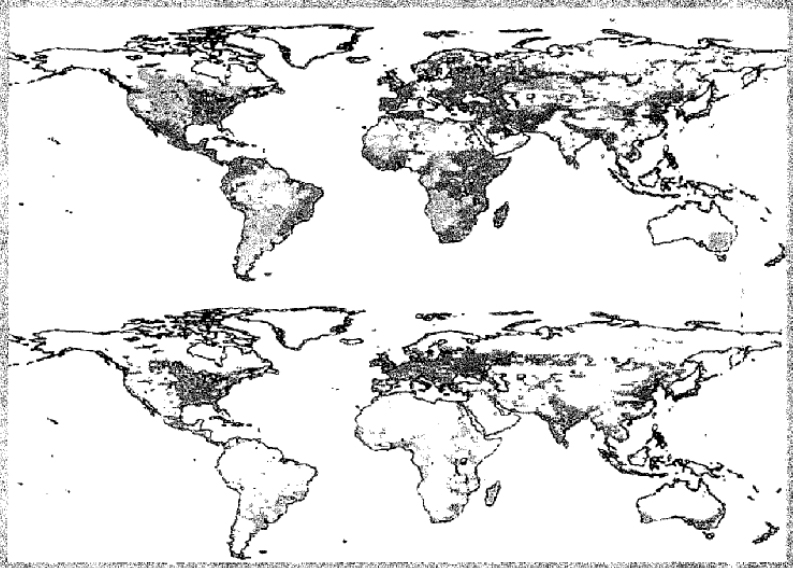
SAMPLING THEORIES AND METHODS

Sampling is taking small portion to make a complete inference about the whole — population.

Census is the complete enumeration of the entire population under study in order to obtain some relevant information about the population.

SAMPLING THEORY AND METHODS



DR NDIDIAMAKA OZOFOR

Published 2008 by
GOD'S WILL PRINTS ENTER.
#12 Udoji Street Ogui New Layout, Enugu.
Phone: 08065818580, 08073505717.

ISBN: 978-2473-71-5

SAMPLING THEORIES

AND

METHODS

STA 311

NDIDIAMAKA MIKEOZOFOR

(B.Sc., Med., Ph.d.)

(Otome-oha Research Series)

SAMPLING THEORY AND METHODS

STA 311

Sampling is taking small portion to make a complete inference about the whole – population.

Census is the complete enumeration of the entire population under study in order to obtain some relevant information about the population.

(1) CENSUS AND SAMPLE SURVEY

Sample survey deals with methods of selecting and observing a part of the population in order to make inferences about the whole population.
Sampling is found in many diverse fields e.g. Demography, industry etc.

ADVANTAGES OF SAMPLING

(1) It saves money
(2) Sampling saves labour
(3) Saves time
(4) It permits ever all high level of accuracy than a complete enumeration because of a higher quality of field staff.

TYPES OF SAMPLING

It can be classified into two: - random sampling or probability sampling and non random sampling.

RANDOM SAMPLING

This is the case where every element in the population has a non-zero probability of being selected in the sample.

Probability samples help us to make numerical statement concerning measures or variability.

NON – RANDOM SAMPLING

Elements are not selected with any known probability. Example of non-random sampling

(1) **Haphazard Sampling:** It includes sampling of volunteered subjects. Conclusion is drawn from which even item that comes to hand.

This method lacks representativeness of the population study

(2) **Purposive and judgment Sampling:** This is used by experts to pick typical or representative specimen's unit proportion e.g. picking a typical city or village to represent an urban or rural population.

(3) **Capture – tag – recapture:** It is suitable for sampling mobile population e.g. insects, fish in the pond etc.

QUOTA SAMPLING

It is a form of purposive sampling widely used in opinion market and similar survey.

Here sampling is continued until a specified quota is obtained from which to build a sample roughly proportional to the population.

DEFINITION OF TERMS

Elementary unit or unit: It is an element or a group of elements living or non-living for which information is sought. The nature of an element is determined by the survey objectives e.g. a person living in a city, a household, a school, an animal etc.

Population or Universe: This is the collection of all units of a specified type has a particular time or specified period is called population.

Population is defined in terms of

(1) Content
(2) Unit
(3) Extent
(4) Time

e.g. We might be interested in the school loss days. We may desire to specify the population.
(2) affecting secondary, elementary school in Nsukka town in 1988.

A population is said to be finite or infinite according to the number of unit in it is finite or infinite. The survey (sampled) population actually achieved may defer some what from some desired target population.

The chief difference frequency arises from non-response and non coverage.

SAMPLING UNITS

It contains the elements and they are used for selecting elements into the sample. In element sampling, each sampling unit contains only one element but in cluster sampling, any sampling unit called cluster may contain several elements.

SAMPLING

One or more units selected from a population according to some specified procedure is said to constitute sample. Thus, a sample in a part or fraction of the population.

LIST OR FRAME

When the elements of a population has been numbered or otherwise identified, we called that population together with its identification system list or a frame.

Example: A France for school children consists of school zones containing schools, their classes and finally children "the frame consists of previously available descriptions of the material in the form of maps, lists, directories etc from which sample units may be constructed and a set of limit is selected. The specification of the frame should define the geographical scope of the survey and the categories of the material covered, also the date and source of the frame" (UN 1950).

A frame is perfect if every element appears on the list separately once and only once and nothing else appears on the list.

Types of sampling procedure (schemes) The three basic methods of selecting a sample of n units from N units in the population are (a) simple random sampling with replacement or without replacement

(b) systematic sampling (sys)

(c) Probability proportional to size (pps)

(i) with replacement ppswr.

(ii) without replacement ppswor.

Types of Sampling Design

Sample Design: consider a population with 6 units 1,2,3,4,5,6. Suppose we are interested in selecting 2 units.

Size n = 2

1, 2 ..... S_1, 2, 3, 3, 4, 4, 5, 1, 2

1, 3 ..... S_2, 2, 4, 3, 5, 4, 6, 1, 3

1, 4 ..... S_3, 2, 5, 3, 6, 5, 6, 1, 4

1, 5 ..... S_4, 2, 6 1, 6

1, 6 ..... S_5,

Suppose $P(S_1)$

$P(S_2)$

$P(S_1)$

We define sample design as all possible samples of given size together with the probability of selecting them P(s).

Types of Design

(1) Unistage Design
(2) Stratified Sampling Design
(3) Cluster Sampling Design
(4) Multistage Design
(5) Multiphase Design

Basic Symbols

Population values:

N = no of units (elements) in the population

Y_1 = value of the y variable for the 11^{th} population

$Y = \sum_{i=1}^{N} Yi \equiv$ Population total for the y variable

n = no. of elements in the sample

$\overline{Y} = \frac{Y}{N} \equiv$ the population mean per element of the y value

$$S_y^2 = \frac{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}{N-1} \quad \text{or} \quad \sigma_y^2 = \frac{\sum_{i=1}^{N} (Y_i - Y)^2}{N} \equiv \text{variance}$$

Population elements. It has two defines

Note $S_y^2 = \frac{N\sigma_y^2}{N-1}$ the difference disappears for large N.

Sampling values (statistics)

Y_i = value of Yi for the i^{th} element

$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ simple sample mean per element

$Y = \sum_{i=1}^{n} y_i$

$S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$ = variance of sample elements.

S_y^2 to the unbiased sample estimate of S^2_y.

Procedure for Simple Random Selection (SRS)

Units 1,2,3,4,...n.

Nkechi, Ijeoma, Okey, Joy, Omojo, Chris, Bridget, Uche, Chichi

1 2 3 4 5 6 7 8 9

6	5	8

(6,5,8) without replacement

4	1	2

(4,1,2)

2	8	1

(2,8,1) with replacement.

Simple random sampling is a method of selecting n units out of N such. The unit in the population are numbered from 1 to N.

A series of random numbers between 1 and N is then drawn either by means of a table of random numbers or by placing the numbers 1 in a roll by mixing them.

The units which bear this numbers constitute the sample. At any stage every unselected element has an equal probability of selection but previously selected numbers are disregarded and cannot be reselected. For this reason the sample is regarded as simple random sampling without replacement.

But in sampling with replacement, the selected elements are placed in the selection polls again and may be re-selected on subsequent draws. The sample size n cannot exceed the population size N in sampling without replacement but n can be any size when sampling with replacement.

In sampling without replacement we have

$$\begin{bmatrix} N \\ n \end{bmatrix} = {}_{N}C_{n} = \frac{N!}{(N-n)!n!}$$

The probability of selecting n samples is

$$\frac{1}{{}_{N}C_{n}}$$

Probability that any pairs or any two units appear in the sample $= \frac{{}_{N-2}C_{n-2}}{{}_{N}C_{n}} = \frac{n(n-1)}{N(N-1)}$

Possible pairs in sampling without replacement $= {}_{N-2}C_{n-2}$

The probability that any unit will be selected in the sample = in the sample $= \frac{{}_{N-1}C_{n-1}}{{}_{N}C_{n}} = \frac{n}{N}$

Prob $(\mu i) = \frac{n}{N}$

Prob $(\mu i \mu j) = \frac{n(n-1)}{N(N-1)}$ for srs wor

Mean & Variance In Simple Random Sampling:

Assuming we have sample of n units of which information is sought, we have $Y_1, Y_2, \ldots Y_n$ from this sample of n units. Our interest is to estimate mean.

$$\bar{Y} = 1/n \sum_{i=1}^{n} y_i$$

$E(\bar{Y})$ expected value of the sampling mean

$$= \sum \left(1/n \sum_{i=1}^{n} y_i \right) = 1/n \sum_{i=1}^{n} E(y_i)$$

$$\therefore E[\bar{Y}] = 1/n \sum_{i=1}^{n} \sum_{i=1}^{N} y_i \, p(y_i) = 1/n \sum_{i=1}^{n} \sum_{i=1}^{N} \frac{y_i}{N}$$

$$= 1/n \sum_{i=1}^{n} \bar{Y}$$

$$= \frac{n \cdot \bar{Y}}{n} = \bar{Y}$$

Similarly

$$E(\bar{Y}) = \sum 1/n \sum_{i=1}^{n} y_i = 1/n \sum_{i=1}^{n} n/N \, y_i$$

$$= 1/N \sum_{i=1}^{n} y_i = \bar{Y};$$

Sample mean in an unbiased of the population mean and is applied in SNR and SWOR.

Consider the population of 3 units 1,2,3; Assuming we selected 2 each. $\bar{Y} = 2$

	$\bar{Y}$	prob. of (S)	$E(\bar{Y}) = \bar{Y}$ prob.
1,2	1.5	⅓	1.5 x ⅓
1,3	2	⅓	2 x ⅓
2,3	2.5	⅓	2.5 x ⅓
	6		$= 6/3 = 2$

Mean of the possible sample will give you population mean and that is why we say that sample mean is an unbiased of the population mean. This applies both to sampling with replacement and sampling without replacement.

A sample is a random variable because it varies from sample to sample.

VARIANCE

$$\therefore V(\overline{Y}) = E\ [\overline{Y} - E(\overline{Y})]^2$$

$$= E\ \left[1/n \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} E(y_i)\right]^2$$

$$= 1/n = E\left[\sum_{i=1}^{n} y_i - \sum E\ (y_i)\right]^2$$

divide both sides by n^2

$$n^2V(\overline{Y}) = E\left[\sum_{i=1}^{n} y_i - n\overline{y}\right]^2 = \left[E \sum (y_i - \overline{Y})\right]^2$$

you should expand

Note that $\left[\sum_{i=1}^{n} ai\right]^2 = \sum_{i=1}^{n} ai_2 + 2\sum\sum_{i\ j} (y_i - \overline{y})\ (YK - \overline{y})$

In sampling with replacement only the first term remains. The 2nd term covariance will not relate but we are sampling without replacement, the co-variance is related.

$$\therefore n2V(\overline{Y}) = E\sum_{i=1}^{n} (y_i - \overline{Y})^2$$

$$= \sum_{i=1}^{n} e(y_i - \overline{Y})^2$$

$$= \sum_{i=1}^{n} \sigma^2$$

$V(\bar{y}) = n\sigma^2 \Rightarrow V(\bar{y}) = \frac{\sigma}{n} = \frac{(N-1)S^2}{N^n}$;

for sampling with replacement.

obtain the variance of the sample mean without placement

$$nV(\bar{y}) = E\left\{\sum_{i=1}^{n}(y-\bar{y})^2 + 2\sum\sum_{i>k}^{m} (y_k-\bar{y})(y_k-\bar{y})\right\}$$

$$= \sum^{N} \frac{n}{N}(y_i-\bar{y}) + 2\frac{n(n-1)}{N(N-1)}\sum_{i>k}^{N}\sum^{N}(y_i-\bar{y})(y-\bar{y})$$

$$n^2V(\bar{y}) = n\sigma^2 + \frac{n(n-1)}{N(N-1)}\left\{\sum_{i=1}^{n}(y_i-\bar{y})\right\}^2$$

$$= \sum (y_i-\bar{y})^2\Big] \quad \sigma^2$$

$$= n\sigma^2 - \frac{n(b-1)}{N(N-1)} N\sigma^2 = \sigma^2 n - \left(\frac{n(n-1)N}{N(N-1)}\right)$$

Note $(\sum ai)^2 = \sum ai^2 + 2\sum\sum_{i>j} ai$

$(\sum ai)^2 - \sum ai^2 = 28$

$$n\sigma^2\left(1-\frac{n-1}{N-1}\right)$$

$$= n^2V(\bar{y}) = n\sigma^2\left(1-\frac{n-1}{N-1}\right) \Rightarrow V(\bar{y}) = \frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$

$$= \frac{N-n}{N}\frac{Sy2}{n}$$

$$V(\bar{y}) = \frac{N-n}{N-1}\frac{\sigma^2}{n} = \frac{N-n}{N}\frac{S^2}{n} = \left(1-\frac{n}{N}\right)\frac{S^2}{n} = \frac{1-f}{n}S^2$$

$F = \frac{n}{} =$ sampling factor

$1 - f$ = finite population correction (fpc)

The square root of sample variance is called standard error.

$$\hat{S}^2 = \sum_{i=1}^{n} \frac{(Yi - \overline{y})^2}{n-1}$$

$\hat{S}^2$ is an unbiased estimate of S^2

$$\hat{S}^2 = \sum_{i=1}^{n} \frac{(y_i - \overline{y})^2}{n-1}$$

$$(n-1)\,\hat{S}^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum y_i^2 - n\overline{y}^2$$

$$(n-1)\,E(\hat{S}^2) = E\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right) = E\sum_{i=1}^{n} y_i^2 = E\sum_{i=1}^{n} y_i^2 - nE(\overline{y})^2$$

$$= \sum_{i=1}^{N} \frac{n}{N} y_i^2 - n\left[V(\overline{y}) + \overline{y}^2\right]$$

$$V(\overline{y}) = E(\overline{y}^2) - [E(\overline{y})]$$
$$(\overline{y}) + [E(\overline{y})]^2 = E(y^2)$$

$$(n-1)\,E(\hat{S}^2) = \frac{n}{N}\sum_{i=1}^{N} y_i^2 - n\left(\frac{N-n}{N}\,\frac{S^2}{n} + \overline{y}^2\right)$$

$$= \frac{n}{N}\sum_{i=1}^{N} y_i^2 - n\overline{y}^2 - \frac{(N-n)}{N}\,s^2$$

$$= n\,\frac{1}{N}\sum_{i=1}^{n} y_i^2 - S^2 + \frac{nS^2}{N}$$

$$= n\sigma^2 - S^2 + \frac{nS^2}{N}$$

$$(n-1)\,E(\hat{S}^2) = n\,\frac{N-1}{N}\,S^2 - S^2 + n\,\frac{S^2}{N}$$

$$= (n-1)\, S^2$$

$$E(\hat{S}^2) = S^2$$

SUMMARY

$\bar{y}$ = sample mean

$$V(\bar{y}) = \frac{1-f}{n} S^2$$

$$V(\bar{y}) = \frac{1-f}{n} S^2$$

Proportions

We now want to estimate from a single random sample the population proportion of element belonging to a defined class or possessing a defined attribute.

A proportion is the mean of a dichotomous variable where members of a class receive value y = 1 and non-members the value y = 0. we sometimes call this a binomial variable.

Denote by N_o and n_o the number of units belong to the defined class A in the population and in sample of size n respectively.

Then population proportion $P = N_o/N$

$$\Rightarrow NP = Y = \sum_{i=1}^{n} y_i^2$$

$$np = y = \sum_{i=1}^{n} y_i = n_o = \sum_{i=1}^{n} y^2_i \quad \text{where } p = n_o/n \text{ the sample.}$$

$$E(\hat{P}) = P$$

$$V(\hat{P}) = \frac{1}{N}(1-f)\frac{NP(1-P)}{N-1}$$

$$\text{Since } S^2 = \frac{\sum_{C=1}^{N} y_i^2 - \frac{(\sum_{i-1}^{N} y_i)^2}{N}}{N-1} = \frac{NP-\frac{(NP)^2}{N}}{N-1}$$

$$= \frac{NP - NP^2}{N-1} = \frac{NP(1-P)}{N-1}$$

The unbiased estimate of S^2

$$\hat{S}^2 = \frac{n\hat{P}(1-\hat{P})}{n-1}$$

The sample estimate of $V(\hat{P})$ is

$$\hat{V}(\hat{P}) = \frac{1-f}{n} \cdot \frac{n\hat{P}(1-\hat{P})}{n-1} = \frac{1-f}{n-1}\hat{P}(1-\hat{P})$$

for w.r $\quad n-1 = n$

$$\hat{V}(\hat{P}) = \frac{1-f}{n}\hat{P}(1-\hat{P})$$

$$\hat{P} = \frac{n_o}{n}$$

$$N_o = N\hat{P}$$

$$V(\hat{N}_o) = N^2V(\hat{P})$$

The estimate of $V(\hat{N}_o)$

$$\hat{V}(\hat{N}_o) = N^2\hat{V}(\hat{P})$$

Example: class N = 8

Take a sample size n = 3.

We want to estimate their average age.

Ages 21, 26, 35, 22, 31, 16, 23, 30 $\overline{Y} = \mu = 25.5$

n = 3

30
21
31 $\quad \frac{81}{3}$ $\quad \overline{X} = 27.8$
82

$$y = \tfrac{1}{3}\sum_{i=1}^{n} yi = 27.3 \qquad yi = 30;21,31$$

Total Age $= N\overline{y} = 218.4$

$$f = \frac{3}{8}$$

$$\hat{S}^2 = \sum_{i=1}^{n} \frac{(yi - 27.3)}{3-1} = \tfrac{1}{2}\left[\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right]$$

$\hat{V}(\bar{y}) = \frac{1-f}{n} \hat{S}^2$ Variance estimate of the sample mean

Exercises:

1. 20 trees were selected by SRSWOR from 176 trees in a forest plantation. The length of the trees in metres are given below:

12	10	9	8	4	6	7	3
5	9	9	8	7	1	3	8
10	10	10	6				

(a) Obtain the estimate of the mean length of the trees.
(b) Obtain the variance of your estimate.
(c) Find the total length of all the trees in the plantation and give its variance.

2(a) Estimate the proportion of trees taller than 9 meters.
(b) Calculate the variance of the sample proportion.

DETERMINATION OF THE SAMPLE SIZE

$P - \{/\bar{y} - \bar{y}/ \geq d\} = \alpha$: where d is the margin of error in the estimation of $\bar{y}$ and α is a small risk which we are willing to incur that the actual error is large than d.

Note $\Pr\{\hat{\theta} - d \leq K\} = 1-\alpha$ k - $Z_{\alpha/2}\sigma$

$\sigma_{\bar{y}} = \sqrt{\frac{N-n}{N}\frac{S^2}{n}}$ standard error

Hence $d = \equiv \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{n}}$ where Z in the normal devise for a given probability that the error will exceed the desired margin

$d^2 = \frac{(ZS)^2}{n} \frac{N-n}{N}$

$= \frac{(ZS)^2}{n} - \frac{(ZS)^2}{N}$ $\frac{N-n}{N} = 1 - \frac{n}{N}$

$$d^2 + \frac{(\not{Z}S)^2}{N} = \frac{(\not{Z}S)^2}{n} \Rightarrow n = \frac{(\not{Z}S)^2}{d^2+\frac{(\not{Z}S)^2}{N}} \quad \text{divide } \frac{2}{\lambda^2}$$

$$n = \frac{(\not{Z}S)^2}{1 + \frac{(\not{Z}S)2}{a} \times \frac{1}{N}}$$

If N is large the first approximation

$$no = \left[\frac{\not{Z}S}{d}\right]^2 = \frac{S^2}{V} \text{ where } V = \frac{d^2}{\not{Z}^2}$$

V is the desired variance of the sample mean.

If $\underline{n_o}$ is negligible, n_o is a satisfactory approximation to n; otherwise.

$$n = \frac{n_o}{1+\frac{n_o}{N}}$$

If the population total Y is to be estimate with margin of error d, take as a 1st approach.

$$n_o = \frac{(NS)^2}{V}$$

Note = $\hat{Y} = N\hat{y}$

SYSTEMATIC SAMPLING

A more convenient method of sampling selection when the units are suitably numbered. It consists of taking every kth unit after a random start

N = 21 we want an interval

n = 3 $\underline{N} = k$
n

$\underline{21} = 7$
3 N = nk

1,2,...k

1,2,...n first person is r, 2nd r+k.....

PROCEDURE

Suppose the population contains N units and that N is equal to nk (N = nk) where n is the sample size and k is an integer.

A number is taking at random from the number s1 to k, if the number is r then the sample contains the n units with serial numbers i, i+k, i+2k,.... i+(n-1)k. Thus the sample consists of the 1st unit selected at random and every kth unit thereafter. It is therefore called a systematic sampling. And the procedure of selection is known as the systematic sampling.

K is known as the sampling interval. Assuring you have N = 100,000 and to select n = 1000, To use SRS is difficult because we may not remember the numbers we have selected using the table of random numbers so it is better to use (sys).

Also when n is so large and N population units unknown, it is very difficult to select the required samples using SRS. But we can conveniently use systematic sampling and continue our selection until we reach the required sample.

ADVANTAGE

(1) It is easy to apply and faster to samples than SRS.
(2) It is easier to check the application of intervals than of random selection.
(3) Systematic sample is evenly distributed over the whole population and easily yield a proportionate sample.

Example: A systematic sample over and alphabetical list of names we yield about the same proportion of for each data.

It is likely to be more precious than the simple random samplings (SRS).

DISADVANTAGES

(1) Unbiased estimate of the variance can be obtained from single systematic sample at least 2 input systematic samples are needed.

(2) Poor arrangement of units may proof a very inefficient sample (e.g.) when the is a periodic variation in the population. The sampling interval falls in line with it.

ESTIMATION IN SYSTEMATIC SAMPLING

$$N = nk$$

$$\bar{y}_{sy} = \frac{1}{N}\sum_{i=1}^{n} y_i = K\sum_{i=1}^{n} y_i = \frac{y}{N} \qquad \therefore n = \frac{N}{k}$$

If $N = nk$, $\bar{y}$ is an unbiased estimate of the population mean $\bar{y}$.

VARIANCE OF THE SYSTEMATIC MEMO

The interval k divides the population into k large sampling units or clusters each of which contains n of the original units thus, choosing a randomly located systematic sample means in effect selecting with probability 1/k one group or cluster of units from the following k clusters forming the entire population.

CLUSTER

e.g. compositing of Clusters

1	2	...	i	...	k
y_1	y_2	...	y_i	...	y_k
y_{k+1}	y_{k+2}	...	y_{k+i}	...	y_{2k}
.					
.					
.					
.					
$y_{(n-1)k+1}$	$y_{(n-1)k+2}$		$y_{(n-1)k+i}$		y_{nk}
$\bar{y}_1$	$\bar{y}_2$		$\bar{y}_i$		$\bar{y}_k$

$$V(\bar{y}) = \frac{1}{K} \sum_{i=1}^{k} (\bar{y}i - \bar{Y})^2$$

The random start from 1 to k imparts to each limit the selection is a simple random sample of one cluster unit from a population of k cluster units.

MEAN

$$E(\bar{y}_{sy}) = \frac{1}{k} \sum_{i=1}^{k} \bar{y}i = \frac{1}{nk} \sum_{i=1}^{k} \sum y_{ij} = \frac{\bar{Y}}{N} = Y$$

y_{ij} denotes the jth member of ith systematic sample; j = 1,2,...; i=1,2,...

$$V(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{K(n-1)}{N} S^2_{wsy} = \text{population variance}$$

Where $S^2_{wsy} = \frac{1}{i=1} \sum_{i=1}^{k} \sum_{i=1}^{k} (y_{ij} - \bar{y}i)^2$

In the variance among units that lie within the same systematic sample. K(n-) completed by the usual rules of ANOVA, since each of the k samples constitute n-1 degree of freedom.

Proof:

Sum of square

$$(N-1)S^2 = \sum_{j=1}^{k}\sum_{j=1}^{n}(\overline{y}_{ij}-\overline{Y})^2 \quad \text{add } -\overline{y}_i + \overline{y}_i$$

$$= \sum_{i}\sum_{j}\left[(\overline{y}_i - \overline{Y}) + (\overline{y}_{ij} - \overline{y}_i)\right]^2,$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}\left[(\overline{y}_i - \overline{Y}) + (\overline{y}_{ij} - \overline{y}_i)\right]^2 + 2(\overline{y}_{ii} - \overline{Y})(\overline{y}_{ij} - \overline{Y})$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}(\overline{y}_{ii} - \overline{Y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{y}_i)^2 + 2\sum_{i=1}^{k}\sum_{i=1}(\overline{y}_{ij} - \overline{Y})(y_{ij} - \overline{y}_i)$$

Since $\sum_{j=1}^{n}(y_{ij} - \overline{y}_{ii}) = \sum_{j=1}^{n} y_{ij} - n\overline{y}_i = n\overline{y}_i - n\overline{y}_i = 0$

$$= n\sum_{i=1}^{k}(\overline{y}_{ii} - \overline{Y})^2 + \sum_{j=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{y}_i)^2$$

But by definition

$$V(\overline{y}_{sy}) = \frac{1}{K}\sum_{j=1}^{k}(\overline{y}_{ii} - \overline{Y})^2 \text{ and } S^2_{wsy} = \frac{\sum_{j=1}^{k}\sum_{j=1}^{n}(y_{ij} - y_i)^2}{K(n-1)}$$

$$\therefore (N-1)S^2 = \frac{nk}{K}\sum_{j=1}^{k}(\overline{y}_{ii} - \overline{Y})^2 + K(N-1)S^2_{wsy}$$

$$\therefore \frac{(N-1)}{N}S^2 = V(\overline{y}_{sy}) + K\frac{(n-1)}{N}S^2_{wsy} \qquad \text{Note ink} = N$$

$$V(\overline{y}_{sy}) = \frac{N-1}{N}S^2 - K\frac{(n-1)}{N}S^2_{wsy}$$

The usual practice is to assume that the population is random and use simple random sampling estimate of variance for the system sampling.

Or using t independent samples i different random starts then $\text{var}(\bar{y}_{sy})$.

$$= \frac{\sum_{j=1}^{t} (\bar{y} - \bar{y})^2}{t(t-1)} \qquad -\bar{y} = \frac{1}{t} \sum_{j=1}^{t} y_i$$

example: There are 169 industrial establishments employing 20 or more persons in town of Lagos in Nigeria. This following are the employment figures based on a 1-in-5 systematic sample. y_i, $i = 1 \ldots\ldots 34$.

35,	88,	35,	36,	156,	25,	24,	237	
80,	568,	22,	139,	163,	37,	37,	27	
25,	26,	38,	24,	62,	331,	28,	31	
81,	121,	49,	23,	34,	23,	22,	53,	50
50.								

Solution:

$N = 169$

$K = 5$

$n = 34$

To estimate the mean of employees in the 169 establishments

$$\bar{y} = \frac{1}{n} \sum y_i \qquad = \frac{2680}{34} = 78.82$$

≈ 78 persons per establishment

$\hat{Y} = N \times \bar{y} = 169 \times 78.82 = 13320.$

The estimate of the total number of people employed in each establishment.

$(\bar{y})$ = $\frac{1-f}{N}\hat{S}^2$

= $\frac{1}{5}$, $1-f = \frac{4}{5}$

= $\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 9387.73$

$(\bar{y}_{sy})$ = $\frac{4}{5} \times \frac{1}{34} \times 9387.73 = 220.87$

Example 2: 10 independent systematic samples each being a 1-in-50 samples are selected from the industrial establishment employing 20 or more persons in the town of Lagos.

The t_i (total sample, establishment) in the sample establishments;

Sample	1	2	3	4	5	6	7	8	9	10
t_i	18,	169,	679,	141,	141,	216,	154,	123,	234,	141

$\sum_{i=1}^{10} t_i = 2376, \quad \sum t_i^2 = 925{,}986$

$\bar{t} = 237.6$

$\sum(t_i - \bar{t})^2 = \frac{\sum t_i^2 - 10\bar{t}^2}{9} = \frac{925{,}986 - 10 \times (237.6)}{9}$

$= \frac{361448}{9}$

The variance estimate = $\frac{361448}{10 \times 9} = \frac{\sum(t_1 - t_2)}{t(t-1)}$

An estimate of total employment is $50 \times 237.6 = 11880$.
Its variance estimate is $(50)^2 \times \frac{361448}{10 \times 9}$

PROBLEMS WITH INTERVALS

If population size N is not an integral of K interval $N \neq nk$.

Remedy:

1. Permit the sample size to either n or n+1. Choose k such that nk<N,(n+1)k.

then the random start to determine whether the sample size is n or n+1.

Assume having added enough blanks to make it exactly nk+k long. The probability of selection is 1/k.

Alternatively, keep the sample size constant at n by omitting one element at random if 1/n were selected, through the procedure is not equal probability of selection method.

2. Eliminate with equal probability enough units to reduce the listing to exactly nk before election in the interval k. the probability of selection in 1/k.

3. Consider the list to be circular. Choose a random start from 1 – N (between 1 to N). Now add the interval k until exactly n elements are chosen.

Going to the end of the list, and then continue to the beginning. Any convenient interval k will remit in an equal probability of selection of n elements with the probability of n/N.

4. using fractional intervals is simple in a decimal fraction.

Example: To select a sample of n = 100u from a population N = 920.

The interval $k = \frac{N}{n} = \frac{920}{100} = 9.2$

is applied. Because the interval = 9.280 x by 10 is a fraction

Select a random start from 1 to 92 then add the interval 92 successively until n = 100 is obtained.

Cluster

Example: Suppose a human population is divided into N = 160 area segments each containing 4 households. A simple random sample of size n = 20 segment is selected and information collected on the number of persons in the sample household, the date obtained are given below:

Segment	# of persons by household				Segment	#of households			
1	6	5	2	4	11	4	7	4	9
2	3	5	5	6	12	6	7	3	6
3	4	2	13	5	13	4	7	7	6
4	6	4	2	7	14	5	4	1	6
5	2	10	3	4	15	3	4	4	8
6	10	5	2	5	16	3	3	3	5
7	3	3	2	5	17	7	5	5	5
8	5	6	1	8	18	4	6	6	4
9	3	6	4	4	19	2	4	4	3
10	5	4	3	3	20	5	4	5	9

Question: 1.

a. Estimate the total number of persons in the population.

b. Estimate the average number of persons per household

c. Obtain the standard errors of your estimates.

2. The 3510 farms in a village are allocated to 90 clusters, the number of farms; in different clusters is not sample of 15 clusters selected and the number of cattle (the variety) determined the sample data are given below:

	No. of farms	Total No. of Cattle (y).
1	35	418
2	25	402
3	48	362
4	30	394
5	70	515
6	55	910
7	66	600
8	18	316
9	30	288
10	32	350
11	64	784
12	24	290
13	48	795
14	40	478
15	82	906

(a) Estimate the average number of Cattle per farm, using the unbiased estimate as well as the ration to size estimate.
(b) Estimate the total number of cattle in the village using the unbiased estimate as well as the ratio to size estimate.
(c) Obtain the variance of your estimate in each case.

USE OF AUXILIARY INFORMATION

(i) RATIO: In a multipurpose large scale sample survey in addition to estimating the means, totals and proportions, we may wish to estimate the ratio of two different characters.

For example you may be interested in average house hold income, average household size, ratio of income to expenditure, yield per hecter, ratio of farmer's income to non-farmers income, male-female ratio in school, in labour force etc.

In estimating the population ratio

$R = \frac{\overline{Y}}{X} = \frac{\overline{Y}}{X}$ of two characters Y and X

X, what is done is to first obtain an unbiased sample estimate of the mean (total) of y and x and then take the ratio of the two estimates the ratio so formed is called a **ratio estimate.** Suppose we are interested in estimating the population ratio of say yam production, y, to cassava production, x, in a given community. Using a SRSWOR of size n farms from a total of N farms in such a community, an estimate of the average yam production $\overline{y}$ and cassava production $\overline{x}$ are then used to form a ratio given by $\hat{R} = \overline{y}/\overline{x}$. The ration estimator $\hat{R}$ is biased for R, it becomes necessary to derive the expression of the bias. For this derivate we proceed as follows:

$$\mu = \overline{X} + \overline{x} - \overline{X} = \overline{X}\ 1 + \left(\frac{\overline{x} - \overline{X}}{\overline{x}}\right)$$

$$= \overline{X}\,(1+\sigma_{\overline{x}})$$

$$\boxed{\text{Bias } (\hat{R}) = E(\hat{R}) - R}$$

$$\check{R} - R = E(\hat{R}-R)$$

$$= \frac{\overline{y}}{\overline{x}} - R$$

$$= \frac{\overline{y} - R\overline{x}}{x}$$

Hence

$$\hat{R} - R = \frac{\overline{y} - R\overline{x}}{\overline{X}(1+\sigma_x)} = \frac{y - Rx}{\overline{X}} \quad (1+\sigma_x)^{-1}$$

$/\sigma x/<1$

$$(1+\sigma x)^{-1}$$

$$(\hat{R} - R) - \frac{\bar{y} - R\bar{x}}{} \quad (1 - \sigma\bar{x} + \sigma^{2}\bar{x} + \sigma^{-3}\bar{x} + \ldots)$$

If we now stop at degree one in the expansion of $(1 + \sigma\bar{x}$

$$(\hat{R}\text{-}R) = \frac{\bar{y}\text{-}R\bar{x}}{\bar{X}} (1 - \sigma\bar{x}) = \frac{\bar{y} - R\bar{x}}{\bar{X}} - \frac{(\bar{y}\text{-}R\bar{x})}{\bar{X}} \sigma\bar{x}$$

$$B(\hat{R}) = E(\hat{R}\text{-}R) = 0 - \frac{E(\bar{y}\text{-}R\bar{x})\sigma\bar{x}}{\bar{X}}$$

$$E\frac{(\bar{y}\text{-}R\bar{x})}{\bar{X}} = \frac{E(\bar{y}\text{-}R\bar{x})}{\bar{X}} = \frac{E(\bar{y}) - R\,E(\bar{x})}{\bar{X}} = \frac{\bar{y}\text{-}x}{\bar{X}}$$

$$= \frac{\bar{Y}\text{-}\bar{Y}}{\bar{X}} = 0$$

$$= \frac{E[\bar{y}\sigma\bar{x} - R\bar{x}\sigma\bar{x}}{\bar{X}} = \frac{E\,\bar{y}\{(\bar{x} - \bar{x})\} - E\{R\bar{x}\}}{\bar{X}^{2}}$$

$$E\bar{y}(\bar{x} - \bar{x}) = E(\bar{y} - \bar{y})(\bar{x} - \bar{x}) = \frac{1 - f}{n} S^{2}_{xy}$$

$$E\bar{x}(\bar{x} - \bar{x}) = E(\bar{x} - \bar{x})^{2} = \frac{1 - f}{n} S^{2}_{xy}$$

So the first order approximate to the bias of $\hat{R}$ is

$$B(\hat{R}) = -\frac{1\text{-}f}{X^{2}n}[S_{xy} - R\,S^{2}_{xy}] = \frac{1 - f}{nX^{2}}(R\,S^{2}_{x} - S_{xy})$$

$$S_{xy} = \sum_{i=1}^{n} \frac{(y_i - Y)(x_i - X)}{N\text{-}1} = \rho S_x S_y$$

Writing in terms of the coefficient of variation

$$B(\hat{R}) = \frac{1-f}{n} R \{C^2_x - C_x C_y\}$$

$$C^2x = \frac{S^2x}{\overline{X}^2}$$

ρ is the correction coefficient between x and correction coefficient between x and .

Mean Square Error Of $\hat{R}$

$$MSE = E(\hat{\theta} - \theta)^2 = E[\hat{\theta} - E(\hat{\theta}) - \theta]^2$$

$$= E[\{\hat{\theta} - E(\hat{\theta})\}^2 + \{E(\hat{\theta}) - \theta\}^2 + 2\{\hat{\theta} - E(\theta)\}\{E(\hat{\theta})$$

$$= E\{\hat{\theta} - E(\hat{\theta})\}^2 + E\{E(\hat{\theta}) - \theta\}^2 + 2[\{\hat{\theta} - E(\hat{\theta})\}\{E(\hat{\theta}) - \theta$$
$$\{= \overline{y} - \overline{y}\}$$

$$= E\{\hat{\theta} - E(\hat{\theta})\}2 + E\{E(\hat{\theta}) - \theta\}^2 + 0$$

$$MSE(\hat{\theta}) = V(\hat{\theta} - [B(\hat{\theta})]^2$$

But y is until

For MSE $(\overline{y}) = V(\overline{y}) + [B(\overline{y})]^2$ $\{B[\overline{y}]\}^2 = 0$

Since the ratio estimator is biased we now derive its MSE.

$$M(\hat{R}) = E(\hat{R} - R)^2$$

$$(\hat{R} - R)^2 = \frac{(\overline{y} - R\overline{x})^2}{\overline{X}^2} (1 + \sigma\overline{x} + \sigma\overline{x}^2 - \ldots)$$

Taking only the leading term in the above result the first order approximate to the MSE of $\hat{R}$ which is equal to its variance is

$$E(\hat{R}-R)^2 = E\frac{(\bar{y}-R\bar{x})^2}{\bar{X}^2} = \frac{V(\bar{y}-R\bar{x})}{\bar{X}^2} \quad \Big| \quad \text{Note: } E(\bar{y}-Rx)$$

$$V(\hat{R}) = V(\bar{y}) - R^2\frac{V(\bar{X})}{\bar{X}^2} - 2R\,\text{Corr}(\bar{x},\bar{y})$$

$$= \frac{1-f}{\bar{X}^2 n}\left[S^2_y + R^2S^2_x - 2RS_{xy}\right] \quad \Big| \quad \text{Note } V(\bar{y})$$

It can also be written as

$$= \frac{1-f}{\bar{X}^2 n}\sum_{i=1}^{N}\frac{(y_i - Rx_i)^2}{N-1}$$

The sample estimate of the $V(\hat{R})$ is given by

$$\hat{V}(\hat{R}) = \frac{1-f}{X^2 n}\sum_{i=1}^{n}\frac{(y_i - \hat{R}x_i)^2}{n-1}\frac{1-f}{X^2 n}\left[S^2_y + R^2\hat{S}^2_x - 2R\hat{S}_{xy}\right]$$

If the population mean $\bar{X}$ is not known, the sample estimate x could be used provided that n is large.

Ratio Mean

$$\hat{R}\bar{x} = \frac{\bar{y}}{\bar{x}}\bar{x} = \bar{y}r$$

$$\bar{y}r = \hat{R}\bar{x}$$

$$V(\bar{y}r) = \bar{X}^2V(\hat{R}) = \frac{1-f}{n}\left[S_y^2 + R^2S_x^2 - 2RS_{xy}\right] = \frac{1-f}{}\sum_{i=1}^{N}\frac{(y_i-Rx_i)^2}{N-1}$$

$(\overline{yr}) = \overline{X}\, B(\hat{R})$

Sample estimate of the variance of the $\overline{yr}$ is

$V(\overline{yr}) \leq V(\overline{y})$ iff

$(\overline{yr}) \leq V(\overline{y})$ iff

$[Sy^2 + R_2Sx_2 - 2RS_{xy}] \leq \frac{1-f}{n} S_2y$

$R^2S^2_x - 2RS_{xy} \leq 0$

$R^2S^2_x \leq 2RS_{xy}$

$\left\{\begin{matrix} R^2S^2_x \leq 2RPS_xS_y \\ R^2S^2_x \leq 2RPS_y \end{matrix}\right\}$

$\frac{1}{2} R^2S^2_x \leq RS_{xy}$

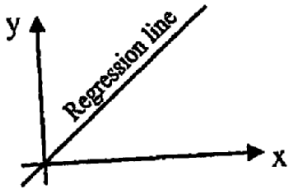
$\frac{1}{2} R^2 \leq R\beta$

$\frac{1}{2} R \leq \beta$

$\frac{1}{2} R \leq \beta$

$R = \frac{S^2_y}{S^2_x}$

We use the ratio mean when the regression lier passes through the origin.



Regression Estimation:

As stated above, ratio method of estimation is used when regression line of y of x is linear and passes through the origin. However, it is not in all occasion that the regression has to pass through the origin.

Some times the line has to pass through the x-axis.

Under this condition, the regression method of estimation is used to obtain the estimation of the population mean and character y.

Suppose it is derived to estimate the current population total Y of a give state in Nigeria from the 1273 census the total population X of the state as well as that of the individual towns and village x_i.

From that state, simple random sample of tons and villages in selected and both the current population and 1973 population of each selected town or village are obtained.

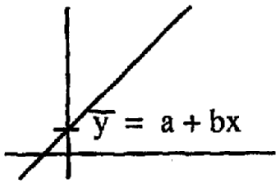
The estimate of the current population of the state can be obtained by using the estimator

$$\hat{Y}_{lr} = \hat{Y} - K(\hat{X} - X)$$
$$= N\{(\overline{y} - K(\overline{x} - \overline{x})\}$$

Villages population x_i $i = \ldots\ldots 10$ 1973
Current population y_i $i = \ldots\ldots 10$ 1989.

Mean of 1973 $\overline{x} = \frac{1}{10} \sum_{i=1}^{10} xi$

Mean of 1989 $\overline{y} = \frac{1}{10} \sum_{i=1}^{10} yi$



The optimum value of k is the regression coefficient β.

So the linear regression mean (total)

$$\overline{y}_{lr} = \overline{y} - \beta(\overline{x} - \overline{x})$$

oth is known, value of β

$= \bar{y} - bo\ (\bar{x} - \bar{x})$ (different estimator since the population is known)

Practice the population regression coefficient β is not own. A solution to the problem is to estimate β from the mple at hand, by

$$\hat{\beta} = \frac{\hat{S}_{xy}}{\hat{\sigma}^2_x}$$

Using the estimated value of $\hat{\beta}$, the regression estimator of the population becomes

$$\bar{y}_{lr} = \bar{y} - \beta(\bar{x} - \bar{X})$$

$$E(\bar{y}_{lr}) = E[\bar{y} - \beta(\bar{x} - \bar{x})]$$
$$= \bar{y} - \beta(\bar{X} - \bar{x}) = \bar{y}$$

Bias and MSE of the Regression Estimator

The exact variance of $\bar{y}_{lr}$, $(\bar{y} - \hat{\beta}(\bar{x} - \bar{x})$, cannot be obtained since it involved the product of two random variables $\hat{\beta}$ and $\bar{x}$.

Hence only the large sample approximates to the MSE of $\bar{y}_{lr}$ will be given.

Adding and subtracting $\beta(\bar{x} - \bar{x})$

$$\bar{y}_{lr} = \bar{y} - \hat{\beta}(\bar{x} - \bar{X}) - \beta(\bar{x} - \bar{X}) + \beta(\bar{x} - \bar{X})$$

$$= \bar{y} - \beta(\bar{x} - \bar{X}) - (\hat{\beta} - \beta)\bar{x} - \bar{X})$$

$$= \bar{y}_{lr} - (\hat{\beta} - \beta)(\bar{x} - \bar{X})$$

$$E(\bar{y}_{lr}) = \bar{Y} - E(\hat{\beta} - \beta)(\bar{x} - \bar{X})$$

$$E(\bar{y}_{lr}) - \bar{Y} = -E(\hat{\beta} - \beta)(\bar{x} - \bar{X})$$

It follows that the bias of $\overline{y}_{lr}$ when β is estimated from the sample at hand is $-E(\hat{\beta}-\beta)(\overline{x}-\overline{X}) = -\text{cov}(\beta, \overline{x})$.

Since $\hat{\beta}$ will tend in probability to β as n increases; the large sample approximation to the m.s.e. of $\overline{y}_{lr}$

$V(\overline{y}_{lr}) = V(\overline{y}_{lr}) = V\{\overline{y}-\beta(\overline{x}-\overline{X})\}$

$\text{cov}(x, y) = \frac{\sum(xi-x)(yi-y)}{n-1}$

$V(\overline{y}_{lr}) = V(\overline{y}) + \beta^2 V(\overline{x}) - 2\beta \text{ cov}(\overline{x}, \overline{y})$

If selection is by S.R.S. WOR, YC variance of y_{lr} becomes

$$V(\overline{y}_{lr}) = \frac{1-f}{n}[S_y^2 + \beta^2 S^2_x - 2\beta S_{xy}]$$

$$V(\overline{y}_{lr}) = \frac{1-f}{n} S_y^2 (1-P^2) \qquad \beta = \frac{S_{xr}}{S^2_x} - \frac{PSNx}{S^2_x} = \frac{PSy}{S_x}$$

Where P is the correlation coefficient. The simple estimate of $V(\overline{y}_{lr})$ for large n is

$$\hat{V}(y_{lr}) = \frac{1-f}{n(n-1)} \sum_{i=1}^{n} \{yi - y - \beta(xi - x)\}^2$$

$$= \frac{1-f}{n}[\hat{S}_y^2 + \hat{\beta}2\hat{S}x2 - 2\hat{\beta}\hat{S}_{xy}]$$

$$= \frac{1-f}{n} \hat{S}_y^2 [1 - \hat{P}^2]$$

Regression estimator with preassigned β. If b_o is the preassigned value of β.

$$\overline{y}_{lr} = \overline{y} - b_o(\overline{x} - \overline{X})$$

$$V(\bar{y}_{lr}) = \frac{1-f}{n}\left[S^2_y + b^2_o S_x^2 - 2b_o S_{xy}\right]$$

The sample estimate is

$$\hat{V}(\bar{y}_{rl}) = \frac{1-f}{n}\left[S^2_y + b^2_o S_x^2 - 2b_o S_{xy}\right]$$

Ex 20 trees were selected by SRS without replacement from 176 such trees in a forest plantation. The length x and timber volume y are given below

x	12	10	9	8	4	6	7	3	5	9	9	8	7
y	762	498	411	36	64	168	174	179	191	394	331	337	248
x	1	3	8	1	1	1	6						
y	43	144	290	26	46	21	187						

(a) The mean length of trees.

(b) The mean volume of trees

(ii) Calculate the ratio of length to volume of the trees.

(iii) Using the length of trees as the auxiliary information obtain:

(a) The ratio mean

(b) The regression mean of the volume of trees ($\bar{X} = 8$)

(iv) Estimate the variance of your estimate in (i – iii).

(a) $\bar{x} = \frac{\sum x_i}{n} = 5.9$ $\quad \sum x_i = 118$, $\sum x^2 = 912$, $n = 20$.

(b)i $\bar{y} = \frac{\sum y_i}{n} = 236.5$ $\quad \sum y_i = 4730$, $\sum y^2_i = 1798500$, $n = 20$

(ii) $\hat{R} = \frac{\bar{x}}{y} = \frac{5.9}{236.5} = \frac{59}{2365} =$

(iii) $a\bar{y}_r = R\bar{x} = \frac{59}{2665} \times 5.9$

$= 0.14718816 = 0.147$

note $\hat{S}_{xy} = \sum_{i=1}^{n} \frac{(xi - \bar{x})(yi - \bar{y})}{n-1} = \frac{\sum xiyi - n\bar{x}\,\bar{y}}{n-1}$

$\hat{V}(\hat{R}) = \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^{n}(yi - \hat{R}xi)^2}{n-1} = \frac{1-f}{n\bar{X}^2} \frac{\sum y_i^2 + \hat{R}\sum x_i^2 - 2r\hat{R}\sum x_i y_i}{n-1}$

STRATIFIED SAMPLING

Introduction: Stratification is one way of using auxiliary information to increase the precision of the estimate of the population character like age.

It should be noted that the precision of an estimate of population characteristics can be improved by either reducing the population variability (S_y^2) or by increasing the sample size.

Suppose we are interested in estimating a total yield of a particular crop in a given village, it may be advisable before selecting the farms to group the farms according to their sizes or any measure of size so that small farms are in a group of their own, moderate size farms are classified into another group while large farms form a separate group.

If a random sample of farms is now drawn from a group a better estimate of the total yield could be obtained without necessarily increasing total sample size than when a random sample of the same is selected directly from a whole population of farm. The yield obtain from each farm varies according to the size of the farm. Consequently, the

variability of the farm yield may be very large. Since there is general the cost or time constraints on the size of the sample we may not get good precisely estimate of the total yield by taking a SR from total farms in the village.

If however, the farms are divided into groups called strata in such a way that farms within each group (strata) are homogenous then the variability is reduced within each stratum thus, a more precise estimate will be obtained by independently sampling each stratum and suitably combining the stratum estimate to obtain the total yield for the entire population of the farms.

This procedure of drawing samples from each stratum after dividing the whole units in the population into homogeneous distinct strata is called **Stratified Sampling.** To achieve the within strata is satisfying variable must be highly related to the character of interest and should noted imported sources of variance. Apart from increase in precision there are other import reasons for stratification.

(i) Because the house hold characteristics in rural areas differ from those in urban areas, the Federal office of statistics in the survey of household usually divides the household into Rural and Urban household and employs different sampling procedures for the selecting of households. For example. The stage sampling may be used in sampling urban household while two stage is used for sampling the Rural household.

Similarly, people living in institution like hotels, barracks, prisons will be placed in one strata and the rest living in other places in another strata.

Different sampling and procedures can then be employed in each strata in obtaining samples.

(ii) Apart from using different sampling techniques strata may be formed because the population within them are also designated as domains of study. Each sub-division being treated as a population in its own ringlet.

For example, in a state agriculture sample survey, estimate could be published for the state as a whole as well as for each local government area in the state. In this case, the local government area which is a population in its own height forms a domain of study for planning and developmental purposes.

(iii) Stratification may be used because of administrative conscience, the FOD has its offices in all the states of he Federation which collected data from each state. Such data can summarize and publish on state basis as well as for the whole country.

ESTIMATION IN STRATIFIED SAMPLING

The procedure on stratified random sampling in estimation of population parameter consists

(1) The population of N units is divided into L homogeneous units respectively N_1, N_2, ... N_L of overlapping units respectively. Such that $N_1 + N_2 + \ldots + N_L = N$.

(2) Samples each of size n_1, n_2, ... n_L are drawn invariably from each stratum such that $n_1 + n_2 + n_s + \ldots + n_L = n$ total sample size.

(3) A separate stratum statistics is calculated in each stratum and then weighted to find combined estimate for the entire population.

N population

L homogenous stratum

N_1	n_1	$\overline{y}_1$
N_2	n_2	y_2
.	.	.
.	.	.
.	.	.
.	.	.
N_L	n_L	$\overline{y}_L$

$$Y = \mu = N_1\overline{y}_1 + N_2\overline{y}_2 + w_s\overline{y}_s + \ldots + N_L\overline{y}_L$$

Mean

$$\mu = \frac{\sum fixi}{\sum fi} = \frac{\sum fixi}{\sum fi} \quad \frac{\sum fixi}{N} \quad \text{for } \sum fi = N$$

From the discussion so far it follows that the use of stratified sampling involved these main cooperation:

(1) The choice of a stratification variable
(2) The choice of the number of strata
(3) The determination of the way in which the population is to be stratified.
(4) The choice of the stratum sample size.
(5) The choice of sampling procedure in each stratum.

MEAN AND ITS VARIANCE IN STRATIFIED SAMPLE

Through out this section we shall deal only with a class of stratified sampling called stratified random sampling, so called because a simple random sample is drawn from the units in each stratum. In other words classes of stratified sampling derive their names from the sampling procedure employed in drawing samples in each stratum. In order to estimate the population mean $\overline{Y}$ a s.r.s of size n_L -

units is drawn without replacement from the population of N_L- units in stratum h. in each stratum, information on character y, is obtained from every unit that appears in the sample. Let y_h be the value obtained from the ith unit in stratum h. the estimate of the population mean is given by

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \text{ (st = stratified)}$$

where $W_h = \frac{Nh}{N}$, the stratum weight

$$y_h = \frac{1}{n_L} \sum_{h=1}^{nn} yhi, \text{ the stratum sample mean}$$

Clearly, the stratified mean $\bar{y}_{st}$ is an unbiased estimate of the population mean since

$$E(\bar{y}_{st}) = E \sum_{h=1}^{L} W_h \bar{y}_h = \sum_{h=1} N_h E(\bar{y}_h)$$

$$= \sum W_L Y_L = \bar{Y}$$

Note $\sum_{h=1}^{L} W_h \bar{Y}_h = \sum_{h=1}^{L} \frac{Nh}{N} \bar{Y}_h = \frac{1}{N} \sum_{h=1}^{L} Y_h$

$$= \frac{Y}{N} = \bar{Y}$$

If interest is on the population total Y we have its estimate as

$$\hat{Y}_{st} = N\bar{y}_{st} = \sum_{h=1}^{L} Nh\bar{y}h = \sum_{h=1}^{L} \hat{Y}_h$$

The variance y_{st} is derived as follows:

$$V = V\sum_{h=1}^{L} W_h\bar{y}_h = \sum_{h=1}^{L} W^2{}_h V(y_h)$$

$$\sum^{L}\sum^{L} W_h W_j \text{ cov } (\bar{y}_h - \bar{y}_i)$$

Since sampling is independent in each stratum, the variance team is zero.

$$\therefore\ V(\bar{y}_{st}) = \sum_{h=1}^{L} N^2{}_L\ (V(\bar{y}_h)$$

and since we are dealing with stratified random sampling

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_{2L}\ \frac{(1-f_h)}{n_h}\ S_L{}^2, \text{ where}$$

$$fn = \frac{n_n}{N_n} \text{ (striatum sampling fraction) and}$$

$$S_h{}^2 = \frac{1}{N_L - 1} \sum_{h=1}^{Nn} (y_{hi} - \bar{Y}_h)^2 \text{ we have to estimate } S^2{}_h \text{ from the}$$

sample estimate of

$V(\bar{y}_{st})$ is given by

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W^2{}_L\ \frac{(1-fn)}{n_L}\ \hat{S}_h{}^2 \text{ where}$$

$\hat{S}^2_h = \frac{1}{n_L-1} \sum_{i=1}^{nh} (y_{hi}-y_h)$ therefore the sample estimate of the $V(\hat{Y}st)$ is

$\hat{V}(\hat{Y}st) = \hat{V}(N\bar{y}st) = NV(\bar{y}st)$

Ratio estimate

$S^2x = \frac{\sum x^2 - n\bar{x}^2}{n-1}$

$S2y = \frac{\sum y^2 - n\bar{y}^2}{n-1}$

Proportion is the Collection of all units of a specified type of a particular point on period of time.
It must be defined in terms of content, unite extent and time.
In element sampling, each sampling unit contains only one element but in cluster sampling, sampling unit contain several elements eg. department.

ESTN of Proportion

In without replacement simple r,s, of size n_n in stratum bean unbiased estimate of the population proportion, P, is given by

$$P_{st} = \sum_{i=1}^{nh} W_h P_h$$

Where P_n is the proportion of units fall in the defined class in the hth stratum. The variance of P_{st} is given by

$$V(P_{st}) = \sum_{h=1}^{L} W_L^2 V(P_n) = \sum_{h=1}^{L} W_h^2 \frac{1-f}{n_L} P_n (1-P_n)$$

$$= \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_L} P_n Q_n \text{ where } Q_n = 1 - P_n.$$

$$\sigma_h^2 = P_h(1 - P_h)$$

$$\frac{H_h \sigma_h^2}{N_h - 1} = \frac{N_h P_L (1 - P_L)}{N_h - 1} S^2_h$$

$V(P_{st})$ can be rewritten as

$$V(P_{st}) = \sum_{h=1}^{L} W_L^2 \frac{N_h - n_L}{N_h - 1} \frac{PnQn}{n_L}$$

$$F_h = \frac{n_n}{N_h}; \quad 1 - f_L = \frac{N_L - n_L}{N_L}$$

The sample estimate of VP_{st} is

$$\hat{V}(P_{st}) = \sum_{h=1}^{L} W_h^2 \frac{NL - nL}{N_h - 1} \frac{P_n}{n_h - 1} \quad \text{where } Z_s = 1 - P_h.$$

For proportional allocation heights of stds

π	=	=	5.172	5.06f 5.11M
S^2	=	=	0.08957	5.10f 5.07f
$V(\overline{x})$	= $\frac{S^2}{n}$	=	0.008957	5.11M 5.10f
$\overline{xf}$	= 5.07	= $\overline{xm}$	5.274	5.06M 5.02f
S^2	= 0.0011	= S^2	0.17441	5.06M
				6.02

Nf = 10 Wf = $\frac{10}{54}$ = 0.185

N = 54

Nm = 44 Nm $\frac{44}{54}$ = 0.815

$\overline{x}st = 0.185 \times 5.07 + 0.815 \times 5.274 = 5.24$

$$\text{variance} = \sum_{h=1}^{L} W_L^2 \frac{S_h^2}{n_L}$$

$S_P^2 = 0.0011$, $S^2_m = 0.17443$

$N_f = n_m = 5 \qquad \frac{Sh2}{5} = 0.00025$

$$V(xm) = \frac{S_h^2}{5} = 0.035$$

$$\sum_{h=1}^{L} N^2_n \frac{Sn^2}{(1-f)\ \frac{S^2}{n}}$$

$$= \sum W^2 h \frac{(1-f)}{n} \frac{S^2}{n} =$$

ALLOCATION OF SAMPLING SIZE TO STRATA

One of the problems of stratified sampling is that of the stratum sample size. In allocating sample size, the variability within strata and the cost per element in each stratum should be taken into consideration.

However, if the variability and cost per element within strata are not known in advance then we may resort to selecting large sample from the stratum that has more number of units, small sample from the sample that has less number of units.

PROPORTION ALLOCATION

In propositional allocation, the stratum sample is selected such that the size of the sample is proportional to the

stratum i.e. n_h x N_h. The constant of proportionality is f = n/N.

$$Nh = N_h \cdot \frac{n}{N} = \frac{nN_h}{N} = NW_L$$

Thus in proportional allocation

$\frac{n}{N} = \frac{n_L}{N_L}$ = f in each stratum or $\frac{n_L}{n} \quad \frac{N_L}{N}$

So it means that if we have $\overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h = \sum_{h=1}^{L} \frac{n_h y_h}{n}$

$$= \frac{1}{n} \sum_{h=1}^{L} \sum_{h=1}^{L} y_{hi} = \overline{y}$$

This proportional allocation results is self weighing sample.

The variance of $\overline{y}_{st}$ which is $\frac{n}{N} = \frac{n_h}{N_h}$

$V_{prop.}(\overline{y}_{st}) = \sum W^2_h \frac{(1-fh)}{n_h} S^2_h$ becomes $n.\frac{N_h}{N} = {}_nW_h$

$$V_{prop}(\overline{y}_{st}) = \frac{1-f}{N} \sum W_h S_h^2$$

For proportion, the variance

$$V_{prop.}(P_{st}) = \frac{1-f}{n} \sum_{h=1}^{L} W_h \frac{N_n P_n Q_n}{N_h - 1}$$

its sample estimate is

$$V_{prop.}(Pst) = \frac{1-f}{N} \sum W_h \frac{n_h P_h q_h}{n_h - 1}$$

The gain made on population all depends on whether the variance within the stratum is with smaller stratum

assuming that the cost of obtaining information from each unit is the same in all stratum.

Advantage

(1) It yields some modest gains in precision
(2) For practical purposes, it s easy and simple to use.
(3) It gives self weighting means.

Optimum (Neyman) Allocation

The basic principle of optimum allocation is to allocate samples with strata in such a way that large sample is taken from the stratum with large variability aim its units or with large stratum size. As small sample from the stratum with less variability among the units or with smaller stratum size; or to increase sample size in the stratum with low per unit and decrease the sample size in the stratum with high cost per unit. It can be shown that optimum allocation is achieved when the stratum sample size is made proportional the stratum standard deviation and is proportional to the square root of its per unit. Hence the problem of optimum allocation consists in much the sample variance for a given overall cost for a specified sampling variance.

$$N_h \alpha S_h^2$$

$$N_h \alpha \frac{1}{C_h}$$

$$N_h \alpha \frac{S_h^2}{C_h}$$

$$N_h \alpha \frac{S_h}{\sqrt{C_h}}$$

Let us consider the simple linear cost fn $C = C_o + \sum c_h n_h$

Where Co is the overhead cost, C_h is the cost per unit of obtaining the necessary information.

We now wish to obtain the optimum stratum sample size that will minimize the variance of $\overline{y}_{st}$ subject to the given cost constant.

$G = V(\overline{y}_{st}) + \lambda [\sum_{h=1}^{L} C_h n_h + c_o - C]$ where λ is called lagrangier multiplier.

We differentiate G with respect to n_h, equate to zero and solve the resultant equation for n_h.

$$V(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h}$$

$$G = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h^2} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h} + \lambda(\sum_{h=1}^{L} C_h n_h - C + c_o)$$

$$\frac{\sigma G}{\sigma n_h} = -\sum \frac{W_h^2 S_h^2}{n_h^2} + \lambda cn = 0 \quad h = 1,2,\ldots L$$

$$\lambda n^2{}_h C_h = W_h^2 S_h^2 \Rightarrow \lambda n_h^2 = \frac{W^2 S_h^2}{C_h}$$

$$nh\sqrt{\lambda} = \frac{WS_h}{\sqrt{C_h}}$$

$$\sqrt{\lambda} \sum_{h=1}^{L} n_h = \sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{C_h}}$$

$$n\sqrt{\lambda} = \sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{C_h}}$$

$$\sqrt{\lambda} = \frac{1}{n} \sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{C_h}}$$

Substitute for $\sqrt{\lambda}$ in

$$\frac{Nh}{n} \sum \frac{W_h S_h}{\sqrt{C_h}} = \frac{W_h S_h}{\sqrt{C_h}} \Rightarrow n_h = \frac{n W_h S_h / \sqrt{C_h}}{\sum_{h=1}^{L} W_h S_h / \sqrt{C_h}}$$

$$n_h = \frac{n N_h S_h \sqrt{C_h}}{\sum_{h=1}^{L} W_h S_h / \sqrt{C_h}}$$ (1) give total sample size

Note given the total cost

$C = C_o = \sum C_h n_h$

$C - C_o = \sum C_h n_h$

From $n\lambda = \frac{W_h S_h}{\sqrt{C_h}}$

Multiplying both sides by Co and summing over the L strata.

$$\sqrt{\lambda} \sum_{h=1}^{L} C_h n_h = \sum_{h=1}^{L} W_h S_h \sqrt{C_h} \Rightarrow \sqrt{\lambda}(C\text{-}Co) = \sum_{h=1}^{L} W_h S_h \sqrt{C_h}$$

$$\sqrt{\lambda} = \sum_{h=1}^{L} \frac{W_h S_h \sqrt{C_h}}{C\text{-}Co}$$

From $n_h = \frac{(C\text{-}Co)\ W_h S_h / \sqrt{C_h}}{\sum_{h=1}^{L} W_h S_h / \sqrt{C_h}} = \frac{(C\text{-}Co)\ N_h S_h / \sqrt{C_h}}{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}}$

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} \left(\frac{1}{N=} - \frac{1}{N_h}\right) W_h^2 S_h^2 \quad \text{check } \frac{1-f_h}{h} = \frac{1}{n_h} = \frac{2}{\sigma}$$

Substitute $\frac{nN_hS_h}{\sum N_hS_h}$ for n_h in

$$V_{min}(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_hS_h\right) - \frac{1}{N}\sum_{h=1}^{L} WS_h^2$$

One practical probability of optimum allocation is lack of knowledge of
(a) Stratum standard deviation however, this could be dissolved by using the value of S_h stratum standard deviation obtained from (a) guesses (b) past survey (c) Pilot survey.

EQUAL ALLOCATION

The 3rd way of allocating sample to strata is by assigning equal sample sizes to all the strata giving a fixed sample size n irrespective of the size of the given sample, the variability and cost per unit.
Note $n_h = n/l$

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{1-f_h}{n_h} W^2{}_hS_h^{2-} = \sum_{h=1}^{L}\left(\frac{1}{n_h} - \frac{1}{N_h}\right)W_h^2S_h^2$$

$$\frac{L}{n}\sum_{h=1}^{L} W^2{}_hS^2{}_h - \frac{1}{N}\sum_{h=1}^{L} W_hS^2{}_h$$

So the variance of equal allocation

$$V(\bar{y}_{st}) = \frac{L}{n}\sum_{h=1}^{L} W^2{}_hS^2{}_h - \frac{1}{N}\sum_{h=1}^{L} W_hS^2{}_h$$

BIBLIOGRAPHY

Ryans, D.G. (1960); Characteristics of Teachers Nashinton: American council on Education.

Ndinechi G.I. (1990) A guide for Effective Typewriters Instruction in Secondary Schools Business Education Journal Volume II No. 2, Nigeria Association of Business Educators.

Adichie R. (1990): Introduction to College Statistics, Ibadan, Oxford University Press Ltd.

Akudolu, I. R. (1995); Effects of Computer Assisted Language Learnin on Students Achievement and Interest in French. (Unpublished Ph.D thesis).

Ali A. and Ohuche R. O. (1991): Teaching Senior Secondary School Mathematics Creatively. Onitsha: Summer Educational Pub. Ltd.

Harbor – Peters, V. F. (191); Target Task and Formal Methods of Presenting Secondary School Geometric Concepts. Their Effects on Retention, Josic Vol. 1 March.

James R. Flynn (1978): Humanism and Ideology London, Routledge and Kegan Paul Press.

Ohuche R.O. (1986); "Laboratory Approach to Science Teaching", Nigerian Journal of Education. Vol 3.

Onyejemezi, D.A. (1988):The Principles of Educational Technology. Onitsha Summer Educational Publishers.

Sinclair, L. R. (1990): Computer Science a Concise Instruction. Oxfor Heninemann Newness Press.

Fredrick Klemm (1999) A History of Western Technology , London, Press Ltd & Books.

Ozofor, N.M. (2000): Effects of Two Modes of Computer Aided Instructi on Students Achievements and Interest in Statistics probability UNN (Unpublished Ph.D Thesis).

Ernest Nagel (1961): The Structure of Science London Oxford Press.

G. Vygosky (1990) Mathematical Handbook: Moscow, Mins Pub.
Wartofsky, M.W. (1968): Conceptual Foundations of Scientific Thought.
Arthur Danto and Sidney Morgenbesser (1962): Philosophy of Science. Washington DC, Classic Papers Publishers.
John H. Randall, Modern Science (1980) London, Alistar C. Cromob;e Press Ltd.

Huxley T H (1980) Popular Lectures in Science Subjects

Ernest Mach (1990): 20th century delsate in the Philosophy of Science. Boston, Die Mechnik.

Thomas S Kuhn (1970): The Structure of Scientific Revolution , N. Y. Imre Lakatos Press Ltd.
M. Bowden (1991): Science vs Evolution Great Britain, Bath Press, Avon.
Spencer, F. Piltdown: A Scientific Forgery Oxford Press 1990.

Downie, N.M. and Heath, R. W. (1974) Basic Statistical Methods (4th Ed.) Harper and Row Pullshers. N/Y

Edwards, A.L. 9 (1972) Experimental Design in Psychological Research (4th Ed.) Holt, Rinehart and Winston Inc. N/Y.

Gullford, J.P. (1965) Fundamental Statistics in Psychology and Education (4th Ed.) Mc Graw-Hill book company N/Y.

Hays, W.L. 9 (1973) Statistical for the Social Sciences (2nd Ed.) Halt, Rinehart and Winston, Inc. N / Y.

Marascuilo, L.A. (1971) Statistical Methods for Behavioural Science Research. MC Graw – Hill book company. N/Y.

Minimum, E.W. (1978) Statistical Reasoning in Psychology and Education (2nd Ed.) John Wiley and Sons N/Y.

Popham, W. and Sirotnik, K.A. (1973) Educational Statistics: Use and Interpretation (2nd Ed.) Harper and Row Publishers N/Y.

Robinson, P.W. (1981) Fundamentals Experimental Psychology, Prentice hall, Inc. Englewood cliffs.