SAMPLE SURVEY

BY

TANUJIT CHAKRABORTY

Indian Statistical Institute

Mail : tanujitisi@gmail.com

SAMPLE SURVEY.

Introduction:- Before giving the notion of sampling we will first define population. In a statistical investigation the interest usually lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics relating to individuals belonging to a group. The group of individuals of our interest is called population. The nature of the population can be finite or infinite. If the population is infinite, then complete enumeration is not possible. Now let us explain the notion of finite population.

Finite Population:- Any well-defined set containing a finite number of elements is called a finite population. Ex. plants in a garden, countries in globe, farms in india and so on. Now the population will then consist of certain of these elements; the plants of certain kind in a specified garden, the 3rd world countries in globe, the farms of specific size, etc.

Population Units:- The elements of a finite population will be entities possessing particular characteristics in which an enquirer could be interested and they will be referred to as population units.

Population size, labels, list :- The number of elements in a finite population is called population size, denoted by N. With each unit in a population of size N, a number from 1 through N is assigned. These numbers are called labels of the units and the population together with its identification number system is known as a list; they remain unchanged throughout the study. The values of the population units w.r.t. the characteristic y under study will be denoted by $Y_1, Y_2, \dots, Y_N$. Here, Y_i denotes the value of the unit bearing label i w.r.t. the variable y.

Sample:- A sample is a subset of a population selected, and obtain information concerning the characteristic of the population. In fact, the word 'population' indicates the aggregate from which the sample is chosen. The population to be sampled should coincide with the population about which information is wanted (the target population). Sometimes, for reasons of practicability or convenience, the sampled population is more restricted than the target population. The elements of the population from which we select sample are called sampling units.

Needs for sampling:- The use of sampling in making inference about a population is possibly as old as civilisation itself. When one has to make an inference about a large lot and it is not practicable to examine each individual member of the lot, one invariably takes recourse to sampling; that is to say, one examines only a few members of the lot and on the basis of this sample information one makes decision. [ISS'09]

Sample survey and complete enumeration:- Broadly speaking, information on a population may be collected in two different ways:- (i) Either every unit in the population is enumerated or surveyed (called complete enumeration or census)
(ii) or enumeration is limited to only a part or a sample selected from the population (called sample enumeration or sample survey)

The principal advantages of sampling as compared with complete enumeration are the following:- [CU] (2011)

Reduced cost:- A sample survey will usually be less costly than a complete census because the expense of covering all units would be greater than that of covering only a small fraction.

Greater Speed:- For the same reason, the data can be collected and summarized more quickly from a sample than from a complete count. This is a vital consideration when information is urgently needed.

Greater scope:- In certain types of inquiry highly trained personnel or specialized equipment, limited in availability, must be used to obtain the data. A complete enumeration or census is impracticable. thus surveys that based on sampling have more scope and flexibility regarding the types of information that can be obtained.

Greater Accuracy:- Because personnel of higher quality can be employed and given intensive training and because more careful supervision of the field work and processing of results becomes feasible when the volume of work is reduced, a sample may produce more accurate results than the kind of complete enumeration that can be taken.

Note:- But there is not always a choice of one versus the other. For example, if data are required for every small administrative area in a country, no sample survey of a reasonable size will be able to deliver the desired information; only a complete census can do this.

Distinguish between Design of experiment and sample survey:-

In design of experiment, the enquiries can be answered by conducting an experiment, suitably designed or controlled by the experiment. Thus, if we want to know which five given varities of rice is expected to give the maximum yield, we have to conduct an experiment with a sample of experimental plots, and suitably controlled, and we can then base our conclusions upon the experimental data.

In sample survey technique, the enquiries can be answered by conducting a survey based on samples. Here the individuals to be sampled occur in nature and can't be subjected to any experimental control. Members are sampled as they appear in nature and required informations obtained from them.

Statistic, Sampling distribution and Standard Error:-

statistic is a function of sample values which is itself an observable random variable which does not contain any parameter.

The probability distribution of any statistic is termed as sampling distribution.

The standard deviation of the sampling distribution of a statistic is known as standard error. [CU'2008]

[CU,2008]

Remark on the Utility of standard Error:- Standard Error plays a very important role in the large sample theory and forms the basis of the testing of hypothesis. If T is any statistic, then for large samples,

$$Z = \frac{T - E(T)}{\sqrt{V(T)}} \sim N(0,1)$$

$$\Rightarrow Z = \frac{T - E(T)}{S.E.(T)} \sim N(0,1)$$

Thus if the discrepancy between the observed and expected (hypothetical) value of a statistic is greater than 1.96 times the S.E., the hypothesis is rejected at 5% level of significance. similarly, if $|T - E(T)| \le 1.96 \times S.E.(T)$, the deviation is not regarded significant at 5% level of significance.

The magnitude of standard error gives an index of the precision of the estimate of the parameter. The reciprocal of the standard error is taken as the measure of reliability or precision of the sample.

The principal steps in a sample survey:-

As a preliminary to a discussion of the role that theory plays in a sample survey, it is useful to describe briefly the steps involved in the planning and execution of a survey. surveys vary greatly in their complexity. The principle steps in a survey are grouped under the following heads:

- Objectives of the survey:- A lucid statement of the objectives is most helpful. Without this, it is easy in a complex survey to forget the objectives when engrossed in the details of planning, and to make decisions that are at variance with the objectives.

- Population to be sampled:- The word 'population' is used to denote the aggregate from which the sample is chosen. In sampling a population of farms, rules must be set up to define a farm and these rules must be usable in practice.

  The population to be sampled should coincide with the target population (the population about which information is wanted). sometimes for reasons of practicability or convenience, the sampled population is more restricted than target population.

- Data to be collected:- Only data relevant to the purposes of the survey should be collected. If there are too many questions, then respondents begin to lose interest in answering them. On the other hand, it must be ensured that no important item is missing. A practical procedure is to prepare outlines of the tables that the survey should produce.

- Degree of precision desired:- The result of sample surveys are always subject to some uncertainty because only part of the population has been measured and because of errors of measurement. this uncertainty can be reduced by taking larger samples and by using superior instruments of measurement. Consequently, the specification of the degree of precision wanted in the results is an important step.

- Methods of measurement:- The method of collecting the information (whether by mail or by interview or otherwise) has to be decided, keeping in view the costs involved and the accuracy aimed at. Mail surveys are cost less, but there may be considerable non-response. Interviewers cost more and there are interviewer errors, but without interviews the data collected may be worthless.

- **The Frame & Sampling Units:–** [CU] Before selecting the sample, the population must be divided into parts that are sampling units. These units must cover the whole population and they must not overlap, in the sense that every element in the population belongs to one and only one unit. In sampling the people in a town, the unit might be an individual person, the member of a family, etc.

  The construction of this list of sampling units, called a frame, is often one of the major practical problems. In order to cover the population to be sampled, there should be some list, map or other acceptable materials (called the frame) which serves as a guide to the universe (population) to be covered.

- **Selection of the sample:–** There is now a variety of plans by which the sample may be selected. For each plan that is considered, rough estimates of the size of sample can be made from a knowledge of the degree of precision desired. The relative costs and time involved for each plan are also compared before making a decision.

- **Questionnaire or schedule:–** The questionnaire (to be filled in by respondent) or schedule (to be completed by interviewer) forms a very important part of the sample survey. Having decided upon the data to be collected, the problem of their presentation require considerable skill.

  A schedule contains a list of items on which information is sought, but the exact form of the questions to be asked is not standardised but left to the judgement of the enumerators. A questionaire, on the other hand, is a set of questions that could actually be put to the informants verbatim in a specified order. While either of these may be used in an interview type of enquiry, a mail questionaire type of enquiry necessarily uses the latter.

- **Training of interviewers and their supervision:–** The success of a survey using the interview method depends largely on the ability of the interviewers to get acceptable responses. Their selection and training is very important. Observation by a supervisor during the course of an actual interview is valuable for maintaining standards.

Q. [ISS'07 – 8 Marks]
What do you mean by sampling frame? Why do you need it? Give reasons for preferring the sampling to complete enumeration.

- **Summary & Analysis of Data:-** The analysis of the data may be broadly classified into the following heads:

(a) **Scrutiny of Data:** The first step is to edit the completed questionnaire, in the hope of amending recording errors, or at least of deleting data that are obviously erroneous.

(b) **Tabulation of Data:-** Whether hand tabulation or mechanical tabulation is to be taken recourse to depend upon the quantity of data. For a large-scale survey involving several thousands of individuals, machine-tabulation is expected to be more economical and quicker.

(c) **Statistical Analysis:-** The tables may be further utilised for deriving necessary estimates for population characteristics or for testing hypothesis, if any. Different methods of estimate may be available for the same data.

(d) **Reporting & Conclusions:-** The report should incorporate a detailed statement regarding all the stages of the survey and should present all the statistical information collected. The data should be properly interpreted, the necessary conclusions derived and the right recommendations made. The technical aspects of the design of the survey, e.g., the types of estimators used and their margins of errors.

- **Information gained for future surveys:-** Any completed sample is potentially a guide to improved future sampling, in the data that is it supplies about the means, standard deviations, and nature of the variability of the principal measurements and about the costs involved in getting the data.

Limitations of Sampling:- Sampling theory has its own limitations and problems which may be briefly outlined as follows:

1. Proper care should be taken in the planning and execution of the sample survey, otherwise the results obtained might be inaccurate and misleading.
2. Sampling theory requires the services of trained and qualified personnel and sophisticated equipment for its planning, execution and analysis. In the absence of these, the results of the sample survey are not trustworthy.
3. However, if the information is required about each and every unit of the population, there is no way to resort to complete enumeration. Moreover, if time and money are not important factors or if the population is not too large, a complete census may be better than any sampling method.

Biases and Errors in surveys:- [CU]'10

(a) The error of estimate arises solely from sampling variation that is present when n of the units are measured instead of the complete population of N units. This is called the sampling error. A rough classification of the types of sampling error is as follows:-

(i) Bias due to defective sampling Technique: If a proper random process is not strictly followed, the investigator may allow his desire to obtain a certain result to influence his selection.

(ii) Bias due to faulty demarcation of sampling units: In area surveys, the location of areas by means of a pair of random co-ordinates, though theoretically ensures a random sample, will in practice do so if the field work is done with complete objectivity. In a crop-cutting survey, for instance, there may be an inclination on the part of investigation to include some good plants in the sample, thus resulting in over-estimation.

(iii) Constant bias due to wrong choice of statistic: For example, in estimating the population variance with a sample of independent observations, the sample variance $\frac{1}{n}\sum(x_i-\bar{x})^2$ is biased estimato where as $\frac{1}{n-1}\sum(x_i-\bar{x})^2$ is unbiased.

(b) It is apparent that, apart from sampling error, surveys are subjected to many other kinds of error. These errors are known as non-sampling error. The error that may be present are as follows; [2009]

(i) Non-response error: Failure to measure some of the units in the chosen sample, is known as non-response error. A rough classification of the types of non-response is as follows;

Non-coverage: This is failure to locate or to visit some units in the sample.

Not-at-homes: This type contains persons who reside at hom but are temporarily away from the house.

Unable to answer: The respondent may not have the informati wanted in certain questions or may be unwilling to give it.

The "hard core": The respondents refuse to be interviewer, who are incapacitated constitute this error.

(ii) Measurement errors:- The measuring device may be biased or imprecise, due to this, the measurement error arises.

(iii) Error Introduced in editing, coding and tabulating results: If there is not enough proffessionals in editing or tabulating results, then there will arise some error.

Basic Principles of Sample surveys:- The theory of sampling is based on the following important principles:

1. **Principle of statistical Regularity:-** The law of statistical regularity lays down that a moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristic of the large group. This principle stresses the desirability and importance of selecting the sample at random so that each and every unit in the population has an equal chance of being selected in the sample.

2. **Principle of Validity:-** By validity of a sample design we mean that the sample should be so selected that the results could be interpreted objectively with certain confidence or in terms of probability. In other words, validity of a sample design ensures that valid estimates or tests about the popln. characteristic should be available, for this it is necessary to attach probability to each member of the popln. to be included in the sample.

3. **Principle of Optimisation:-** By the precision of the sample results, we mean how close we can reproduce from a sample the results which would be obtained if we should take a complete count or a census, under the same conditions. The precision is judged by the variance of the estimators concerned. Efficiency of the sample survey is measured by the reciprocal of the sampling variance of the estimator. Cost is measured by expenditure incurred in terms of money or man-hours. The principle of optimisation ensures that a sample strategy to be preferred which gives the highest precision for a given cost of the survey or the minimum cost for a specified level of precision.

Judgement Sampling:- Any type of sampling which depends upon the personal judgement of the sampler himself is called judgement sampling. Here the judgement of the person selecting the sample is significant, for different persons will judge differently. There is no objective method of preferring one judgement to another.

The judgement sampling have two important limitations. One is the difficulty of describing the proper emphasis to the various factors affecting sample design. What is lacking is a theory that will indicate a desirable allocation of resources to such factors of sample design. Some guidance is required for evaluating the various factors entering into the design and contributing to the sampling error, and for selecting the "best" one of a number of alternative designs. The second limitation is the inability to measure the precision of the sample results, and no objective basis is known for measuring the amount of confidence which can be placed in the sample estimates.

Probability Sampling :- Any type of sampling in which each member of the population has a known (non-zero) probability of being selected in the sample is called Probability Sampling or random sampling. With probability sampling it is possible to state an objective basis for choosing from among the alternative methods of sampling and methods of estimation.

With the help of probability theory, we are then in a position to determine the frequency distribution of the estimates derivable from the sampling. With probability designs it will be possible to evaluate the precission of the sample results and compare the precision of different designs and of different modification of the same design; and it gives a measure of amount of confidence which can be placed in the sample estimates.

Sampling design :- A survey population $\hat{U}$ is a set $\{U_1, U_2, \ldots, U_N\}$ of a known number N of units U_j, $j = 1(1)N$, which are identifiable and labeled. With each unit U_j there is associated a value Y_j, the objective of the enquiry is to estimate on the basis of a sample selected from U a specified function of the population values, $\lambda(Y_1, \ldots, Y_N)$. For example, we may wish to estimate the population total $Y = \sum_{i=1}^{N} Y_i$ or the mean $\mu = \sum_{j=1} Y_j / N$. By a sample of size n from U we mean a subset $s = (u_1, u_2, \ldots, u_n)$ of U. Where relevant u_i will be said to be the i^{th} selection in the sample. The number of elements in a sample 's' is denoted by $n(s)$.

We shall denote the class of all samples $\{s\}$ by S. Thus a sample is a point in the sample space S. With each sample s; there is a non-negative number called probability of s and written $P(s)$, such that $\sum_{s \in S} P(s) = 1$.

The probability distribution $\{P(s), s \in S\}$ defined on S, the collection of all possible samples, is known as sampling design.

Problem 1:- Let $U = \{1,2,3\}$. Find the expected values of the estimator $\hat{T}(s) = \frac{N}{n(s)} \sum_{i \in s} y_i$, under the designs defined below

$$P(s) = \begin{cases} \frac{1}{7} & \text{if } s = \{1,3\} \\ \frac{2}{7} & \text{if } s = \{2,3\} \\ \frac{4}{7} & \text{if } s = \{1,2,3\} \end{cases} \quad \text{and } Q(s) = \begin{cases} \frac{1}{3} & \text{if } n(s)=2 \\ 0, & \text{ow} \end{cases}$$

Solution:- Computation of $E_P[\hat{T}(s)]$:

s	$\hat{T}(s)$	$P(s)$	$\hat{T}(s)P(s)$
$\{1,3\}$	$\frac{3}{2}(y_1+y_3)$	$1/7$	$\frac{3}{14}(y_1+y_3)$
$\{2,3\}$	$\frac{3}{2}(y_2+y_3)$	$2/7$	$\frac{3}{7}(y_2+y_3)$
$\{1,2,3\}$	$\frac{3}{3}(y_1+y_2+y_3)$	$4/7$	$\frac{4}{7}(y_1+y_2+y_3)$

$$E_P(\hat{T}(s)) = \frac{11y_1+14y_2+17y_3}{14} = \sum \hat{T}(s)P(s)$$

= Expected value of the estimator $\hat{T}(s)$ under the given design $P(s)$.

Computation of $E_Q(\hat{T}(s))$:-

s	$\hat{T}(s)$	$Q(s)$	$\hat{T}(s)Q(s)$
$\{1,3\}$	$\frac{3}{2}(y_1+y_3)$	$1/3$	$\frac{1}{2}(y_1+y_3)$
$\{2,3\}$	$\frac{3}{2}(y_2+y_3)$	$1/3$	$\frac{1}{2}(y_2+y_3)$
$\{1,2,3\}$	$\frac{3}{2}(y_1+y_2)$	$1/3$	$\frac{1}{2}(y_1+y_2)$

$E_Q(\hat{T}(s)) = y_1+y_2+y_3 = y =$ the population total

Hence $\hat{T}(s)$ is unbiased under sampling design $Q(\cdot)$ but it is not unbiased under the design $P(\cdot)$.

Ex.2. Consider a popln. containing three villages u_1, u_2, u_3 with variate values x_1, x_2 and x_3. A probability sample of two units is selected under the design $P(s) = \begin{cases} \frac{1}{3} & \text{if } n(s) = 2, \\ 0, & \text{ow} \end{cases}$

to estimate the population mean $\mu = \frac{x_1+x_2+x_3}{3}$. The following two estimators are considered:

$$t(s) = \begin{cases} \frac{1}{2}x_1 + \frac{1}{2}x_2 & \text{if } s = \{u_1, u_2\} \\ \frac{1}{2}x_1 + \frac{2}{3}x_3 & \text{if } s = \{u_1, u_3\} \\ \frac{1}{2}x_2 + \frac{1}{3}x_3 & \text{if } s = \{u_2, u_3\} \end{cases}$$

and $$t'(s) = \begin{cases} \frac{1}{2}(x_1+x_2) & \text{if } s = \{u_1, u_2\} \\ \frac{1}{2}(x_1+x_3) & \text{if } s = \{u_1, u_3\} \\ \frac{1}{2}(x_2+x_3) & \text{if } s = \{u_2, u_3\} \end{cases}$$

Show that $Var(t) < Var(t')$ if $x_3(x_3 - 3x_2 + 3x_1) < 0$.

Solution :-

$$E(t(s)) = \frac{1}{3}\left\{\left(\frac{1}{2}x_1 + \frac{1}{2}x_2\right) + \left(\frac{1}{2}x_1 + \frac{2}{3}x_3\right) + \left(\frac{1}{2}x_2 + \frac{1}{3}x_3\right)\right\}$$

$$= \frac{x_1+x_2+x_3}{3}$$

$$= \mu$$

$$E[t^2(s)] = \frac{1}{3}\left\{\left(\frac{1}{2}x_1 + \frac{1}{2}x_2\right)^2 + \left(\frac{x_1}{2} + \frac{2x_3}{3}\right)^2 + \left(\frac{1}{2}x_2 + \frac{1}{3}x_3\right)^2\right\}$$

$$= \frac{1}{3}\left\{\frac{x_1^2}{2} + \frac{x_2^2}{2} + \frac{5}{9}x_3^2 + \frac{1}{2}x_1x_2 + \frac{1}{3}x_2x_3 + \frac{2x_1x_3}{3}\right\}$$

Similarly, $E[t'(s)] = \mu$

and $$E[t'^2(s)] = \frac{1}{3}\left\{\frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 + \frac{1}{2}x_3^2 + \frac{1}{2}x_1x_2 + \frac{1}{2}x_2x_3 + \frac{1}{2}x_3x_1\right\}$$

Now, $Var(t) < Var(t')$

$$\Rightarrow E[t^2] < E[t'^2]$$

$$\Rightarrow x_3(x_3 - 3x_2 + 3x) < 0.$$

Selection of a Simple Random sample :-

Random sample refers to that method of sample selection in which every item has an equal chance of being selected. But the random sample does not depend upon the method of selection only but also on the size and nature of the population. Random sample can be obtained by any of the following methods:

(a) Lottery System:- The simplest method of selecting a random sample is the lottery system, which is illustrated below by means of an example:

Suppose we want to select 'r' candidates out of n. We assign the number 1 to n; one number to each candidate and write these numbers (1 to n) on n slips which are made as homogeneous as possible in shape, size, colour, etc. These slips are then put in a bag and throughly shuffled and then 'r' slips are drawn one by one. The 'r' candidates corresponding to numbers on the slips drawn, will constitute a random sample.

This method of selection is quite independent of the properties of population. This is one of the most reliable methods of selecting random sample.

(b) 'Random Numbers' Method:- For large population the lottery system is too labourious and time consuming. The most practical and inexpensive method of selecting a random sample consists in the use of 'Random Number Tables', which have been so constructed that each of the digits 0, 1, 2, ..., 9 appear with approximately the same frequency and independently of each other. The method of drawing the random sample consists of the following steps:

(i) Identify the N units in the population with the numbers from 1 to N,

(ii) Select at random, any page of 'random number tables' and pick up the numbers in any row or column or diagonal at random.

(iii) The population units corresponding to the numbers selected in step (ii) constitute the random sample.

Scanned by CamScanner

Simple Random Sampling (S.R.S.):—

From a population of N units select one by one giving equal probability to all units. Make a note of the unit selected and return it to the population. If this operation is performed n times, we get a simple random sample of n units, selected with replacement (WR). Not returning the unit (or units) selected and selecting a further unit with equal probability from the units that remain in the population, then we get a simple random sample selected without replacement (WOR).

Definition:— If each unit of the population has an equal probability of being selected at each drawing, then the sampling is called simple random sampling.

Theorem:— In SRSWR, the sample space contain N^n samples of size 'n' of the population U. The probability distribution

$$P(s) = \begin{cases} \frac{1}{N^n}, & \text{if } n(s) = n \\ 0, & \text{OW} \end{cases}$$

, is the sampling design of the SRSWR.

In SRSWOR, the sample space contain $\binom{N}{n}$ samples of size 'n' of the population U. The probability distribution

$$P(s) = \begin{cases} \frac{1}{\binom{N}{n}}, & \text{if } n(s) = n \\ 0, & \text{OW} \end{cases}$$

, is the sampling design of the SRSWOR.

Proof:— Let $U = \{U_1, U_2, \ldots, U_N\}$ be a population. Clearly, in SRSWR, any drawing produces U_i, $i = 1(1)N$, has the probability $\frac{1}{N}$ and all draws are independent, since the selected is replaced before the next drawing is made.

$\therefore$ P(selecting a specified sample of n units from a population of N units)

$$= \frac{1}{N} \cdot \frac{1}{N} \cdots\cdots n\text{ times} = \frac{1}{N^n}.$$

Hence, in SRSWR, each of N^n samples has an equal probability $\frac{1}{N^n}$ of being selected.

$\therefore$ The sampling design of SRSWR is

$$P(s) = \begin{cases} \frac{1}{N^n} & \text{if } n(s) = n \\ 0 & \text{OW} \end{cases}$$

In SRSWOR,

Probability of selecting any unit at the first draw $= \frac{1}{N}$

Probability of selecting any unit out of the remaining $(N-1)$ units in the second draw $= \frac{1}{N-1}$,

and so on.

Probability of selecting any unit of the remaining $N-(i-1)$ units at the i^{th} draw $= \frac{1}{N-(i-1)}$, $(i = 3(1)n)$.

Since all the draws are independent, by compound probability theorem, the probability of selecting a sample of size n in a fixed specified order, is

$$\frac{1}{N(N-1)(N-2)\cdots\cdots(N-n+1)}$$

Since this probability is independent of the order of the sample and since there are $n!$ permutations of the sampled units, by addition theorem of probability, the required prob. of obtaining a sample of size n (in any order) is

$$P(s) = \begin{cases} \frac{n!}{N(N-1)\cdots\cdots(N-n+1)} = \frac{1}{{}^N C_n}, & \text{if } n(s) = n \\ 0, & \text{ow} \end{cases}$$

Theorem :- In SRSWOR,

(i) The probability of selecting a specified unit of the population at any given draw is equal to the probability of its being selected at the first draw, is equal to $\frac{1}{N}$.

(ii) The probability of selecting any specified unit in the sample is equal to the probability that a specified unit is included in the sample, is equal to $\frac{n}{N}$. (OR)

$\pi_i = P[\,U_i \text{ in the sample}\,] = \frac{n}{N}$.

(iii) $\pi_{ij} = P[\,U_i, U_j \text{ in the sample}\,] = \frac{n(n-1)}{N(N-1)}$

(iv) The events 'U_i in the sample' and 'U_j in the sample' are not independent.

Proof :- (i) Let E_r be the event that any specified unit is selected at the r^{th} draw.

$\therefore P[E_r] = \Pr[$that the specified unit is not selected in any one of the previous $(r-1)$ draws and then selected at the r^{th} draw$]$

$$= \frac{N-1}{N} \times \frac{N-2}{N-1} \times \cdots\cdots \times \frac{N-r+1}{N+r+2} \times \frac{1}{N-r+1}$$

$$= \frac{1}{N}.$$

Clearly, $P[E_r] = \frac{1}{N} = P[E_1]$ for any r.

This is an important property of SRSWOR.

(ii) Since a specified unit can be included in the sample of size n in n mutually exclusive ways, viz. it can be selected in the sample at the rth draw ($r = 1, 2, \dots, n$) and since

$$P[E_r] = \frac{1}{N}, \; r = 1, 2, \dots, n$$

By addition theorem of probability we get,

The prob. that a specified unit is included in the sample $= \sum_{r=1}^{n} \left(\frac{1}{N}\right) = \frac{n}{N}$.

(OR)

$$\pi_i = P[\,U_i \text{ is in the sample}\,] = \sum_{s \ni i} P(s) = \sum_{s \ni i} \frac{1}{\binom{N}{n}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N},$$

since there are $\binom{N-1}{n-1}$ subsets with i as an element.

(iii) $\pi_{ij} = \sum_{s \ni i,j} P(s) = \sum_{s \ni i,j} \frac{1}{\binom{N}{n}} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$, since there are $\binom{N-2}{n-2}$ subsets with i, j as elements.

(iv) Note that, $\pi_{ij} \neq \pi_i \pi_j$

$\Rightarrow$ Two events are not independent.

Estimation in Simple Random Sampling :—

We assume that to each unit U_i in the population is attached a variate value Y_i for the character y. The population total is $Y = \sum_{i=1}^{N} Y_i$; the mean being $\bar{Y} = \sum_{i=1}^{N} Y_i / N = \mu$. Let the '$n$' units (selected in this order) in the SRS be $u_1, u_2, \dots, u_n$, with variate values $y_1, y_2, \dots, y_n$, respectively.

Theorem:– In SRSWR, the sample mean $\bar{y}$ is unbiased for the popl^n mean and $Var(\bar{y}) = \frac{\sigma_y^2}{n} = \frac{(N-1)}{Nn} s_y^2$, where $N\sigma_y^2 = \sum_{i=1}^{N} (Y_i - \bar{Y})^2 = (N-1) s_y^2$.

Proof:– If y_i, $i = 1(1)n$, is the value of the unit drawn in the ith draw then y_i can take any one of the N values Y_i with probability $\frac{1}{N}$.

$$\therefore E(y_i) = \sum_{j=1}^{N} Y_j P[y_i = Y_j] = \frac{1}{N} \sum_{j=1}^{N} Y_j = \bar{Y}$$

Similarly, $E(y_i^2) = \frac{1}{N} \sum_{j=1}^{N} Y_j^2$

Hence, $V(y_i) = E(y_i^2) - E^2(y_i) = \frac{1}{N} \sum_{j=1}^{N} Y_j^2 - \bar{Y}^2 = \sigma_y^2 = \frac{N-1}{N} s_y^2$

Since draws are independent, $cov(y_i, y_j) = 0$. We get.

$$E(\bar{y}) = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \frac{1}{n} \cdot n\bar{Y} = \bar{Y}.$$

and $Var(\bar{y}) = V\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V(y_i) = \frac{1}{n^2} \cdot n\sigma_y^2 = \frac{\sigma_y^2}{n} = \frac{N-1}{Nn} s_y^2.$

[Proved]

Theorem:- In SRSWOR, the sample mean $\bar{y}$ is an unbiased estimator of the population mean and $Var(\bar{y}) = \frac{N-n}{N-1} \cdot \frac{\sigma_y^2}{n} = \frac{N-n}{Nn} \cdot S_y^2$

Proof:- In SRSWOR, y_i can take any one of the N values $Y_1, Y_2, \ldots, Y_N$ with equal prob. $\frac{1}{N}$ and the product $y_i y_j$ can take the values $Y_r Y_s$, $r \neq s$, with probabilities $\frac{1}{N(N-1)}$.

Hence, we have (i) $E(y_i) = \frac{1}{N} \sum_{i=1}^{N} Y_i = \bar{Y}$

(ii) $E(y_i^2) = \frac{1}{N} \sum_{i=1}^{N} Y_i^2$

(iii) $E(y_i y_j) = \frac{1}{N(N-1)} \sum_{i \neq j} Y_i Y_j$

Therefore, $V(y_i) = \frac{1}{N} \sum_{j=1}^{N} Y_j^2 - \bar{Y}^2 = \frac{1}{N} \sum_{j=1}^{N} (Y_j - \bar{Y})^2 = \sigma_y^2 = \frac{N-1}{N} \cdot S_y^2$

$$Cov(y_i, y_j) = E(y_i y_j) - E(y_i)E(y_j)$$

$$= \frac{1}{N(N-1)} \sum_{i \neq j} Y_i Y_j - \bar{Y}^2$$

$$= \frac{1}{N(N-1)} \left[\left(\sum_{k=1}^{N} Y_k \right)^2 - \sum_{k=1}^{N} Y_k^2 \right] - \bar{Y}^2$$

$$= \frac{1}{N(N-1)} \left[N^2 \bar{Y}^2 - \sum_{k=1}^{N} Y_k^2 \right] - \bar{Y}^2$$

$$= -\frac{1}{N(N-1)} \left[\sum_{k=1}^{N} Y_k^2 - N\bar{Y}^2 \right]$$

$$= -\frac{1}{N-1} \sigma_y^2$$

$$= -\frac{S_y^2}{N}.$$

We know that $Var(\bar{y}) = V\left[\frac{1}{n} \sum_{i=1}^{n} y_i \right]$

$$= \frac{1}{n^2} \left[\sum_{i=1}^{n} V(y_i) + 2 \sum_{i<j} \sum Cov(y_i, y_j) \right]$$

$$= \frac{1}{n^2} \left[\frac{n(N-1)}{N} S_y^2 + 2 \cdot \frac{n(n-1)}{2} \left(-\frac{S_y^2}{N} \right) \right]$$

$$= \frac{N-n}{Nn} S_y^2$$

$$= \frac{N-n}{N-1} \cdot \frac{\sigma_y^2}{n}$$ [Proved]

<u>Remark:</u>— (1) $V(\bar{y}) = \frac{1}{n}\left(1-\frac{n}{N}\right)S_y^2 = (1-f)\frac{S_y^2}{N}$, where $f=\frac{n}{N}$ is the sample fraction, the fraction of the population taken into the sample. For a random sample of size n from an infinite popln, it is well known that $Var(\bar{y}) = \frac{\sigma^2}{n}$. The only change in a random sampling WOR when the popln is finite is the introduction of the factors,

$$1-\frac{n-1}{N-1} = \frac{N-n}{N-1}, \text{ as } Var_{WOR}(\bar{y}) = \left(1-\frac{n-1}{N-1}\right)\frac{\sigma^2}{n}.$$

The quantity $\left(1-\frac{n-1}{N-1}\right)$ is called the <u>finite population correction (f.p.c.)</u>

Also note that, $V_{WOR}(\bar{y}) = \left(1-\frac{n-1}{N-1}\right)V_{WR}(\bar{y}) < V_{WR}(\bar{y})$, <u>the sample mean is more efficient estimator of $\bar{Y}$ in SRSWOR compare to SRSWR.</u>

(2) An unbiased estimator of the popln total $Y = N\bar{Y}$ is given by $\hat{Y} = N\bar{y}$ and

$$Var(\hat{Y}) = N^2 Var(\bar{y}) = \begin{cases} N^2 \cdot \frac{\sigma_y^2}{n}, & \text{in SRSWR.} \\ N^2 \cdot \frac{N-n}{N-1} \cdot \frac{\sigma_y^2}{n}, & \text{in SRSWOR.} \end{cases}$$

[CU'2008]

<u>Estimation of sampling variance or standard error of $\bar{y}$ in SRS:</u>—

In order to obtain an unbiased estimator of $V(\bar{y})$, we prove the theorem;

<u>Theorem:</u>— If $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$, then $E(s_y^2) = \begin{cases} \sigma_y^2, & \text{in SRSWR} \\ S_y^2, & \text{in SRSWOR} \end{cases}$

<u>Proof:</u>— Note that $(n-1)s_y^2 = \sum_{i=1}^{n} y_i^2 - \bar{y}^2 n$. Now $E(y_i^2) = \frac{1}{N}\sum_{i=1}^{N} Y_i^2$

and $E(\bar{y}^2) = V(\bar{y}) + E^2(\bar{y}) = \begin{cases} \frac{N-n}{nN}S_y^2 + \bar{Y}^2, & \text{in SRSWOR} \\ \frac{\sigma_y^2}{n} + \bar{Y}^2, & \text{in SRSWR} \end{cases}$

Hence, in SRSWOR,

$$\begin{aligned}(n-1)E(s_y^2) &= \frac{n}{N}\sum_{i=1}^{N} Y_i^2 - n\left(\frac{N-n}{nN}S_y^2 + \bar{Y}^2\right) \\ &= n\left\{\frac{1}{N}\sum Y_i^2 - \bar{Y}^2\right\} - \frac{N-n}{N}\cdot S_y^2 \\ &= \left\{n\frac{N-1}{N} - \frac{N-n}{N}\right\} S_y^2 \\ &= (n-1)S_y^2.\end{aligned}$$

$$\therefore \boxed{E(s_y^2) = S_y^2}.$$

Similarly, in SRSWR,

$$\begin{aligned}(n-1)E(s_y^2) &= \frac{n}{N}\sum Y_i^2 - n\left(\frac{\sigma_y^2}{n} + \bar{Y}^2\right) \\ &= n\left(\frac{1}{N}\sum Y_i^2 - \bar{Y}^2\right) - \sigma_y^2 \\ &= (n-1)\sigma_y^2.\end{aligned}$$

Hence, $E(s_y^2) = \sigma_y^2$.

Corollary:- An unbiased estimator of $V(\bar{y})$ is given by

$$\hat{V}(\bar{y}) = \begin{cases} \frac{1}{n} s_y^2, & \text{in SRSWR} \\ \frac{1}{n}\left(1-\frac{n}{N}\right) s_y^2, & \text{in SRSWOR.} \end{cases}$$

An unbiased estimator of $V(\hat{Y})$ is

$$\hat{V}(\hat{Y}) = \begin{cases} N^2 \cdot \frac{s_y^2}{n}, & \text{in SRSWR} \\ N^2 \cdot \left(1-\frac{n}{N}\right) \cdot \frac{s_y^2}{n}, & \text{in SRSWOR.} \end{cases}$$

The standard error (S.E.) of $\bar{y}$ is $\sigma_{\bar{y}} = \sqrt{\frac{N-n}{nN}} \cdot S_y$ and we take $\hat{\sigma}_{\bar{y}} = \sqrt{\frac{N-n}{nN}} \cdot s_y$.

Similarly, $\sigma_{\hat{Y}} = N \cdot \sqrt{\frac{N-n}{nN}} \cdot S_y$ and $\hat{\sigma}_{\hat{Y}} = N\sqrt{\frac{N-n}{nN}} \cdot s_y$.

These estimators are slightly biased.

Merits of Simple Random Sampling:-

1. Since the sample units are selected at random giving each unit an equal chance of being selected, the element of subjectivity or personal bias is completely eliminated. As such a simple random sample is more representative of the popln. as compared to the judgement or purposive sampling.
2. The statistician can ascertain the efficiency of the estimates of the parameters by considering the sampling distribution of the statistics.

Limitations of SRS:-

1. The selection of a simple random sample requires an upto-date frame, i.e. a completely catalogued popln. from which samples are to be drawn. Frequently, it is virtually impossible to identify the units in the popln. before the sample is drawn and this restricts the use of SRS technique.
2. Administrative Inconvenience: A simple random sample may result in the selection of the sampling units which are widely spread geographically and in such a case the cost of collecting the data may be much in terms of time and money.
3. At times, a simple random sample might give most non-random looking results. For eg., if we draw a r.s. of size 13 from a pack of cards, we may get all the cards of the same suit. However, the probability of such an outcome is extremely small.
4. For a given precision, SRS usually requires larger sample size as compared to stratified random sampling.

SOLVED EXAMPLES: —

1. A sample of size 4 is to be drawn from a population of size 8. Let Y_i denote the value of study variable for the i^{th} unit, $i=1,\dots,8$. Suppose units 1 and 8 are included in every sample and a simple random sample (without replacement) of size 2 is drawn from units $2,3,\dots,7$. Show that $\hat{Y} = \frac{Y_1+Y_8+6\bar{Y}_2}{8}$ is an unbiased estimator of the population mean, where $\bar{Y}_2$ is the mean of the two units drawn. Obtain an expression for the variance of this estimator. [2008]

Solution:—

$$\hat{Y} = \frac{Y_1+Y_8+6\bar{Y}_2}{8}$$

$$\therefore \hat{Y} = \frac{Y_1+Y_8}{8} + \frac{3}{4}\cdot\frac{1}{2}\sum_{i\in s} Y_i, \text{ as } \bar{Y}_2 \text{ is the mean of the two units drawn.}$$

$$\therefore E(\hat{Y}) = \frac{Y_1+Y_8}{8} + \frac{3}{8}\times E\left[\sum_{i=2}^{7} I_i Y_i\right], \text{ where } I_i = \begin{cases} 1 & \text{if } s\in S \\ 0 & \text{ow} \end{cases}$$

$$= \frac{Y_1+Y_8}{8} + \frac{3}{8}\times\frac{2}{6}\sum_{i=2}^{7} Y_i$$

$$= \frac{Y_1+Y_2+Y_3+Y_4+Y_5+Y_6+Y_7+Y_8}{8} = \bar{Y}.$$

$\therefore$ $\hat{Y}$ is an unbiased estimator of the population mean.

$$V(\hat{Y}) = V\left\{\frac{Y_1+Y_8+6\bar{Y}_2}{8}\right\} = \frac{9}{16}\,Var(\bar{Y}_2).$$

2. Suppose there is a popln $U = (U_1,\dots,U_N)$ with unknown variate values Y_j $(j=1(1)N)$. In order to estimate the popln total Y from a SRSWOR of size n; we use the estimator $t = N\bar{y}$. Now suppose that we have advance information that the y value of U_N is Y_N. Then the estimator based on a SRS of size n taken from the $(N-1)$ units is $t' = Y_N + (N-1)\bar{y}'$. Show that $V(t') < V(t)$.

Solution:— Let μ_N and σ_N^2 be the mean and variance of y in the popln of N units and μ_{N-1}, σ_{N-1}^2 the corresponding quantities in the popln of $(N-1)$ units. Then

$$(N-1)\sigma_{N-1}^2 = \sum_{j=1}^{N-1}\left(Y_j - \mu_N + \mu_N - \mu_{N-1}\right)^2$$

$$= N\sigma_N^2 - \frac{N}{N-1}(Y_N-\mu_N)^2.$$

$$\text{Now, } V(t') = (N-1)^2 V(\bar{y}') = (N-1)^2\cdot\frac{1}{n}\left(1-\frac{n-1}{N-2}\right)\sigma_{N-1}^2$$

$$= (N-1)\frac{1}{n}\left(1-\frac{n-1}{N-2}\right)\left[N\sigma_N^2 - \frac{N}{N-1}(Y_N-\mu_N)^2\right]$$

$$< \frac{N-1}{N-2}\cdot\frac{N-n-1}{n}\cdot N\sigma_N^2.$$

Remark:— These problems show that in actual survey in SRS the sample mean $\bar{y}$ does not have BLUE.

3. A SRS of size $n = n_1 + n_2$ with mean $\bar{y}$ is drawn from a finite population, and a simple random subsample of size n_1 is drawn from it with mean $\bar{y}_1$. Show that,

(a) $V(\bar{y}_1 - \bar{y}_2) = S_y^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$, where $\bar{y}_2$ is the mean of the remaining n_2 units in the sample.

(b) $V(\bar{y}_1 - \bar{y}.) = S_y^2\left[\frac{1}{n_1} - \frac{1}{n}\right]$

(c) $Cov(\bar{y}, \bar{y}_1 - \bar{y}) = 0$.

Solution:- Since $\bar{y}_1$ is based on a subsample,

$$V(\bar{y}_1) = E_1 V_2(\bar{y}_1) + V_1 E_2(\bar{y}_1) \quad \text{———} (*)$$

where, E_1 is the unconditional expectation and E_2 is the conditional expectation w.r.t. the subsample.
Similarly, V_1 is the unconditional variance and V_2 is the conditional variance w.r.t. the subsample.

Note that, $E_2(\bar{y}_1) = \bar{y}$ and $V_2(\bar{y}_1) = \frac{n-n_1}{nn_1}\cdot s_y^2$

$$\therefore V_1 E_2(\bar{y}_1) = \frac{N-n}{Nn}\cdot S_y^2, \; E\,V_2(\bar{y}_1) = \frac{n-n_1}{nn_1}\cdot S_y^2$$

From (*), $V(\bar{y}_1) = \left(\frac{1}{n_1} - \frac{1}{N}\right)\cdot S_y^2$.

Again,
$$\begin{aligned}
Cov(\bar{y}, \bar{y}_1) &= E(\bar{y}\bar{y}_1) - E(\bar{y})E(\bar{y}_1) \\
&= E_1E_2[\bar{y}\bar{y}_1] - \bar{Y}E_1E_2[\bar{y}_1] \\
&= E_1[\bar{y}E_2(\bar{y}_1)] - \bar{Y}E_1(\bar{y}) \\
&= E_1[\bar{y}^2] - E_1^2[\bar{y}] \\
&= V(\bar{y}) \\
&= \frac{N-n}{Nn}\cdot S_y^2.
\end{aligned}$$

We know that, $cov(\bar{y}, \bar{y}_1 - \bar{y}) = cov(\bar{y}, \bar{y}_1) - cov(\bar{y}, \bar{y}) = V(\bar{y}) - V(\bar{y}) = 0$. [(c) is proved]

Note
$$\begin{aligned}
V[\bar{y}_1 - \bar{y}] &= V(\bar{y}_1) + V(\bar{y}) - 2cov(\bar{y}, \bar{y}_1) \\
&= V(\bar{y}_1) + V(\bar{y}) - 2V(\bar{y}) \\
&= V(\bar{y}_1) - V(\bar{y}) \\
&= \frac{N-n_1}{Nn_1} S_y^2 - \frac{N-n}{Nn}\cdot S_y^2 = \left[\frac{1}{n_1} - \frac{1}{n}\right] S_y^2
\end{aligned}$$
[(b) is proved]

We know that, $\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1+n_2}$, $\Rightarrow \bar{y}_2 = \frac{n\bar{y} - n_1\bar{y}_1}{n_2}$.

$$\begin{aligned}
\therefore V(\bar{y}_1 - \bar{y}_2) &= V\left(\bar{y}_1 - \frac{n\bar{y} - n_1\bar{y}_1}{n_2}\right) = \frac{1}{n_2^2}V[n(\bar{y}_1 - \bar{y})] \\
&= \frac{n^2}{n_2^2}V[\bar{y}_1 - \bar{y}] \\
&= \frac{n^2}{(n-n_1)^2}\left(\frac{1}{n_1} - \frac{1}{n}\right)S_y^2 \\
&= \left[\frac{1}{n_1} + \frac{1}{n_2}\right]\cdot S_y^2
\end{aligned}$$
[(a) is proved]

4. A simple random sample of size 3 is drawn from a population of size N with replacement. As an estimator of $\bar{Y}$ we take $\bar{y}'$, the unweighted mean over the different units in the sample. Show that the average variance of $\bar{y}'$ is $\frac{(2N-1)(N-1)}{6N^2} \cdot S_y^2$
$= \left(\frac{1}{3} - \frac{1}{6N}\right)\sigma_y^2$.
Show that the probabilities that the sample contains 1, 2 and 3 distinct units are $P_1 = \frac{1}{N^2}$, $P_2 = \frac{3(N-1)}{N^2}$, $P_3 = \frac{(N-1)(N-2)}{N^2}$.
[2007]

Solution:-

P_1 = Prob [one distinct unit in all the three draws]

$$= \frac{N}{N^3}$$

$$= \frac{1}{N^2}.$$

P_2 = Prob [two distinct units in all the three draws]

$$= 3C_2 \cdot \frac{N(N-1)}{N^3} = \frac{3(N-1)}{N^2}.$$

P_3 = Prob [three distinct units in all the three draws]

$$= \frac{N(N-1)(N-2)}{N^3} = \frac{(N-1)(N-2)}{N^2}$$

We know that the variance of the sample mean based on 'n' distinct units (or, a sample drawn in WOR) is

$$\frac{N-n}{Nn} \cdot S_y^2 = V_n.$$

Therefore the average variance of $\bar{y}'$ is $= V_1 P_1 + V_2 P_2 + V_3 P_3$

$$= \left(\frac{N-1}{N} \cdot S_y^2\right) \cdot \frac{1}{N^2} + \left(\frac{N-2}{N.2} S_y^2\right) \cdot \frac{3(N-1)}{N^2} + \left(\frac{N-3}{N.3} \cdot S_y^2\right) \cdot \frac{(N-1)(N-2)}{N^2}$$

$$= \left[(N-1) + \frac{N-2}{2} \cdot 3(N-1) + \frac{N-3}{3}(N-1)(N-2)\right] \cdot \frac{S_y^2}{N^3}$$

$$= \frac{(N-1)}{6N^3} \cdot S_y^2 \left[2N^2 - N\right]$$

$$= \frac{(2N-1)(N-1)}{6N^2} \cdot S_y^2$$

$$= \frac{(2N-1)}{6N} \cdot \sigma_y^2.$$

AH. Ques:- A simple random sample of size 3 is to be taken from a popln. of N units WR. Find the probabilities for the sample to have one, two and three distinct units. Hence show that the sample mean based only distinct units of the sample is unbiased for the popln. mean. Find the average variance of the sample mean. Compare the performance of this estimator with the sample mean based on all the units. (3+4+3+5)

5. From a population of N units sampling with replacement with equal probabilities is continued till the sample contains n distinct units. Denoting by ν the no. of selections made, show that —

(a) $E(\nu) = N\left(\frac{1}{N} + \frac{1}{N-1} + \frac{1}{N-2} + \cdots + \frac{1}{N-n+1}\right)$

(b) $E\left(\frac{1}{\nu}\right) > \{E(\nu)\}^{-1} > \frac{N-n}{n(N-1)}$.

Now two estimators of the popln. mean μ may be formed; one is $\bar{y}_n = \frac{1}{n}\sum y_r$ based the distinct units and the other is $\bar{y}_\nu = \frac{1}{\nu}\sum K_r y_r$, where K_r is the frequency of appearance of the distinct units in the sample. Show that

(c) $\bar{y}_n$ and $\bar{y}_\nu$ are unbiased.

(d) $V(\bar{y}_\nu) > V(\bar{y}_n)$

(e) $V(\bar{y}_\nu) = \sigma_y^2\left(E\left(\frac{1}{\nu}\right)\right)$. Hence obtain (d).

<u>Solution:</u>- (a) Let X_i : the no. of units required after the i^{th} distinct unit have been obtained to obtain the next distinct unit, $i = 1(1)\overline{n-1}$.

When i distinct units have already been obtained, then the Prob. that a new distinct unit will be obtained is $\frac{N-i}{N} = p$, say.

$\therefore \quad P[X_i = k] = pq^{k-1}, \quad k = 1, 2, 3, \ldots$

i.e. $X_i \sim \text{Geo}\left(p = \frac{N-i}{N}\right)$.

Then $E(X_i) = \frac{1}{p} = \frac{N}{N-i}$.

Clearly, $\nu = 1 + X_1 + \cdots + X_{n-1}$,

$$E(\nu) = 1 + E(X_1) + \cdots + E(X_{n-1})$$
$$= 1 + \frac{N}{N-1} + \frac{N}{N-2} + \cdots + \frac{N}{N-n+1}$$
$$= N\left(\frac{1}{N} + \frac{1}{N-1} + \frac{1}{N-2} + \cdots + \frac{1}{N-n+1}\right)$$

(b) $E^2\left(\sqrt{\nu}\cdot\frac{1}{\sqrt{\nu}}\right) \le E(\nu)E\left(\frac{1}{\nu}\right)$, by C-S inequality.

$\Leftrightarrow \quad E\left(\frac{1}{\nu}\right) \geq \frac{1}{E(\nu)}$.

Now, $[E(\nu)]^{-1} = \dfrac{1}{N\left\{\frac{1}{N} + \frac{1}{N-1} + \cdots + \frac{1}{N-n+1}\right\}}$

$> \dfrac{1}{N\left\{\frac{1}{N} + \frac{n-1}{N-n}\right\}}$ since $\frac{1}{N-i+1} < \frac{1}{N-n}$ $\forall\, i = 1(1)\overline{n-1}$.

$\Rightarrow E\left(\frac{1}{\nu}\right) > [E(\nu)]^{-1} > \frac{N-n}{n(N-1)}$.

(c) For a given number n of distinct units, the sample of distinct units is a simple random sample, selected WOR. Hence,

$$E_2[\bar{y}_n | n] = \bar{Y} \text{ and } E[\bar{y}_n] = \bar{Y} \longrightarrow (*)$$

For a given sample, $A_n = (y_1, y_2, \ldots\ldots, y_n)$ of n distinct units, the probability that a specified distinct unit with value y_r will be selected at any selection (there being ν such selection) is $\frac{1}{n}$ and

$$E_2[K_r | A_n] = \frac{\nu}{n}.$$

Hence, $E_2[\bar{y}_\nu | A_n] = \frac{1}{\nu}\sum_{r=1}^{n} y_r . E_r[K_r|A_n] = \frac{1}{\nu}\sum_{r=1}^{n} y_r \cdot \frac{\nu}{n} = \bar{y}_n$

$\therefore\ E(\bar{y}_\nu) = E_1 E_2[\bar{y}_\nu | A_n] = E_1[\bar{y}_n] = \bar{Y}$, from $(*)$.

(d) $V(\bar{y}_\nu) = E_1 V_2(\bar{y}_\nu) + V_1 E_2(\bar{y}_\nu)$

$= E_1 V_2(\bar{y}_\nu) + V_1(\bar{y}_n) \geq V_1(\bar{y}_n)$

(e) $V(\bar{y}_\nu) = E[\bar{y}_\nu - \bar{Y}]^2$

$= E_1 E_2[(\bar{y}_\nu - \bar{Y})^2 | \nu]$

$= E_1\left\{\frac{1}{\nu}\cdot\sigma_y^2\right\}$ [Since, in SRSWR, for a sample of size 'n', $Var(\bar{y}) = \frac{\sigma_y^2}{n}$.]

$= \sigma_y^2 E\left(\frac{1}{\nu}\right)$

Also, we have $E\left(\frac{1}{\nu}\right) > \frac{N-n}{n(N-1)}$,

$\therefore\ V(\bar{y}_\nu) > \frac{N-n}{n(N-1)}\sigma_y^2 = \frac{N-n}{nN}\cdot S_y^2 = V(\bar{y}_n)$

[Proved]

Estimation of Population Proportion / Simple Random sampling for Attributes:—

[CU'2009]

Sometimes we wish to estimate the total number or proportion of units in the popln. that possess some charactoristic or attributes.

Notation: — We suppose that every unit in the popln. can be classified into two categories C and C'.

Number of units in C in	
Population	Sample
A	a

Proportion of units in C in	
Population	Sample
$P = A/N$	$p = a/n$

If we associate with U_i, the ith unit in the population, a variable

$$Y = \begin{cases} 1 & \text{if } U_i \text{ belongs to } C \\ 0 & \text{ow} \end{cases}$$

Clearly, the number of units belonging to C is $\sum_{i=1}^{N} Y_i = Y$ in the popln. and is $\sum_{i=1}^{n} y_i = y$ in the sample of size n.

The proportion of C is $P = \sum_{i=1}^{N} Y_i / N = \frac{Y}{N} = \bar{Y}$ and is $p = \frac{1}{n}\sum_{i=1}^{n} y_i$

$= \frac{y}{n} = \bar{y}$, say.

Note that, $E(p) = E(\bar{y}) = \bar{Y} = P$, in SRS.

Again, $S_y^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2}{N-1} = \frac{NP - NP^2}{N-1}$.

$= \frac{N}{N-1}\cdot PQ$, where $Q = 1-P$.

Similarly, $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{n}{n-1}\cdot pq$.

Now, $Var(p) = Var(\bar{y}) = \frac{N-n}{nN}\cdot S_y^2 = \frac{N-n}{nN}\cdot\frac{N}{N-1}PQ$

$= \frac{N-n}{N-1}\cdot\frac{PQ}{n}$, in SRSWOR

But in SRSWR, $V(p) = Var(\bar{y}) = \frac{\sigma_y^2}{n} = \frac{(N-1)}{nN}\cdot S_y^2 = \frac{N-1}{nN}\cdot\frac{N}{N-1}\cdot PQ$

$= \frac{PQ}{n}$.

Hence p is an unbiased estimator of P, with

$$Var(p) = \begin{cases} \frac{PQ}{n}, & \text{in SRSWR} \\ \frac{N-n}{N-1}\cdot\frac{PQ}{n}, & \text{in SRSWOR.} \end{cases}$$

Theorem:— An unbiased estimator of the $Var(p)$ is given by

$$\hat{V}(p) = v(p) = \begin{cases} \frac{pq}{n-1}, & \text{in SRSWR} \\ \frac{N-n}{(n-1)N}\cdot pq, & \text{in SRSWOR.} \end{cases}$$

[CU'2009]

Proof:— We have, $E(s_y^2) = \begin{cases} \sigma_y^2, & \text{in SRSWR} \\ S_y^2, & \text{in SRSWOR} \end{cases}$

Hence, an unbiased estimator of $V(\bar{y})$ is $\hat{V}(\bar{y}) = \begin{cases} \frac{s_y^2}{n}, & \text{in SRSWR} \\ \frac{N-n}{nN}s_y^2, & \text{in SRSWOR.} \end{cases}$

Now, $s_y^2 = \frac{n}{n-1}\cdot pq$.

$$\therefore \hat{V}(p) = v(p) = \begin{cases} \frac{pq}{n-1}, & \text{in SRSWR} \\ \frac{N-n}{(n-1)N}\cdot pq, & \text{in SRSWOR} \end{cases}$$

Problem:- A simple random sample of size n is drawn without replacement from a population of size N. A r.v. T_i is associated with the i^{th} unit in the popln., such that $T_i = \begin{cases} 1, & \text{if the } i^{th} \text{ unit is selected} \\ 0, & \text{OW} \end{cases}$

Show that ——

(a) $E(T_i) = \frac{n}{N}$, $\quad V(T_i) = \frac{n}{N}\left(1-\frac{n}{N}\right)$, $Cov(T_i,T_j) = \frac{n}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right)$, $1 \le i < j \le N$

(b) sample proportion is an unbiased estimator of the popln. proportion. Also find the variance of the sample proportion.

Solution:- (a) We have ~~$T_i \sim Bin(n, p=\frac{1}{N})\ \forall\ i=1(1)n$~~.

$T_i \sim \text{Bernoulli}\left(p=\frac{n}{N}\right)$

$$E(T_i) = p = \frac{n}{N}$$
$$V(T_i) = p(1-p) = \frac{n}{N}\left(1-\frac{n}{N}\right)$$

$$Cov(T_i,T_j) = E(T_iT_j) - E(T_i)E(T_j)$$
$$= \frac{n(n-1)}{N(N-1)} - \frac{n}{N}\cdot\frac{n}{N}$$
$$= \frac{n}{N}\left[\frac{(n-1)}{(N-1)} - \frac{n}{N}\right]$$

(b) We suppose that every unit in the popln. can be classified into two categories C and C'. Let P be the proportion of units in C in the popln. of size N. Here no. of members in the category C is NP in the population.

Define, $T_i = \begin{cases} 1 & \text{if the } i^{th} \text{ popln. unit belongs to } C \\ 0 & \text{OW} \end{cases}$

Then $\sum_{i=1}^{N} T_i$ is the no. of units belonging to C in the popln. and

$$P = \frac{\sum_{i=1}^{N} T_i}{N} \text{ is the popln. proportion.}$$

Similarly, $t_i = \begin{cases} 1 & \text{if the } i^{th} \text{ sample unit belongs to } C \\ 0 & \text{OW} \end{cases}$

Then $\sum_{i=1}^{n} t_i$ is the no. of units belonging to C in the sample and

$$p = \frac{\sum_{i=1}^{n} t_i}{n} = \text{the sample proportion of } C.$$

$$E(t_i) = P[t_i = 1] = \frac{NP}{N} = P.$$

$\Rightarrow E(p) = \frac{1}{n}\sum_{i=1}^{n} E(t_i) = P$, i.e. sample proportion is an unbiased estimator of the popln. proportion.

Now, $$V(p) = \frac{1}{n^2}\left\{\sum_{i=1}^{n} V(t_i) + 2\sum_{i<j} Cov(t_i,t_j)\right\}$$
$$= \frac{1}{n^2}\left\{n\cdot\frac{NP}{N}\left(1-\frac{NP}{N}\right) + n(n-1)\cdot\frac{NP}{N}\left(\frac{NP-1}{N-1} - \frac{NP}{N}\right)\right\}$$
$$= \frac{PQ}{n}\left[1 - \frac{n-1}{N-1}\right].$$

☒ Determination of sample size in SRS:— In planning any sample survey the problem that a statistician is faced with is to determine the size of the sample so that the unknown population parameters may be estimated with a specified degree of precision. The statement of precision desired may be made by giving the amount of error that we are willing to tolerate in the sample estimates. The precision can be specified in several ways:

(1) Sample size for obtaining estimate with specified coefficient of variation (c.v.):— [Useful for Practical]

We want to find n in SRSWOR so that

$$\therefore \; C.V.(\bar{y}_n) = \frac{S.E.(\bar{y}_n)}{E(\bar{y}_n)} = \frac{S_y\sqrt{\frac{1}{n}-\frac{1}{N}}}{\bar{Y}} = C_0, \text{ say}.$$

$$\Rightarrow \frac{S_y^2 \cdot \frac{1}{n}\left(1-\frac{n}{N}\right)}{\bar{Y}^2} = C_0^2$$

[If N is large, so that f.p.c. $\left(\frac{n}{N}\right)$ can be neglected, then we have $\frac{S_y^2}{n\bar{Y}^2} = C_0^2$.

$$\Rightarrow n = \frac{S_y^2}{C_0^2\bar{Y}^2} = \left[\frac{\text{C.V. of } Y \text{ in the popln.}}{C_0}\right]^2$$

$= n_0$, say.]

For any N, we have

$$\frac{1}{n} - \frac{1}{N} = \left[\frac{\bar{Y}C_0}{S_y}\right]^2 = \frac{1}{n_0}, \text{ say}.$$

$$\Rightarrow \frac{1}{n} = \frac{1}{n_0}\left(1+\frac{n_0}{N}\right)$$

i.e. $$n = \frac{n_0}{1+\frac{n_0}{N}}.$$

Hence, for C.V. $= C_0$, we have

$$n = \begin{cases} \dfrac{n_0}{1+\frac{n_0}{N}} \\ n_0, & \text{if } N \text{ is sufficiently large } \& \; n_0 = \left(\frac{S_y}{C_0\bar{Y}}\right)^2. \end{cases}$$

(2) Sample-size for Given Margin of error (d) in estimate of P and confidence coefficient $(1-\alpha)$.

Some margin of error 'd' in the estimated proportion p has been agreed on and there is a small risk α that we are willing to incur that the actual error is $\geq d$; i.e.

$$P\left[\,|p-P| \geq d\,\right] = \alpha.$$

Simple random sampling is assumed, then p is taken as normally distributed with $\sigma_p^2 = V(p) = \frac{N-n}{N}\cdot\frac{PQ}{(n-1)}$ and $\mu_p = P$ for large n.

Hence $\frac{p-P}{\sigma_p} \overset{a}{\sim} N(0,1)$ as $n\to\infty$. and, we have

$$\alpha = P\left[|p-P|/\sigma_p \geq d/\sigma_p\right] = 2\left\{1-\Phi\left(\frac{d}{\sigma_p}\right)\right\} \Rightarrow \frac{d}{\sigma_p} = \tau_{\alpha/2}, \text{ the}$$

upper α-point of $N(0,1)$.

Scanned by CamScanner

Therefore, $d^2 = \tau^2_{\alpha/2} \cdot \frac{N-n}{N-1} \cdot \frac{PQ}{n} \Rightarrow n = \dfrac{\tau^2_{\alpha/2} \cdot \frac{PQ}{d^2}}{1 + \frac{1}{N}\left(t^2 \frac{PQ}{d^2} - 1\right)}$ ---(*)

If N is large, 1st approximation is

$$n_0 = \tau^2_{\alpha/2} \cdot \frac{pq}{d^2} = \frac{pq}{V}$$, where $V = \frac{pq}{n_0}$ is the desired variance of the sample proportion. If $\frac{n_0}{N}$ is negligible then n_0 is a satisfactory approximation to n of (*). If not, it is apparent that a better approximation is

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \cong \frac{n_0}{1 + \frac{n_0}{N}}.$$

(3) The formula for n with continuous data:-

Most commonly, we wish to control the relative error 'r' in the estimated popln. total or mean with a SRSWOR having mean $\bar{y}$, we want

$$P\left[\left|\frac{\bar{y} - \bar{Y}}{\bar{Y}}\right| \geq r\right] = P\left[\left|\frac{N\bar{y} - N\bar{Y}}{N\bar{Y}}\right| \geq r\right] = P\left[|\bar{y} - \bar{Y}| \geq r\bar{Y}\right] = \alpha$$, where α is a small probability.

We assume that, $\bar{y} \overset{a}{\sim} N(\bar{Y}, \sigma^2_{\bar{y}})$, where $\sigma^2_{\bar{y}} = \frac{N-n}{N} \cdot \frac{S_y^2}{n}$, for large n.

Now, $\alpha = P\left[|\bar{y} - \bar{Y}| \geq r\bar{Y}\right] = P\left[\frac{|\bar{y} - \bar{Y}|}{\sigma_{\bar{y}}} \geq r.\bar{Y}/\sigma_{\bar{y}}\right]$

$$\Rightarrow \frac{r\bar{Y}}{\sigma_{\bar{y}}} = \tau_\alpha \Rightarrow (r\bar{Y})^2 = \tau_\alpha^2 \cdot \frac{N-n}{Nn} \cdot S_y^2$$

$$\Rightarrow n = \left(\frac{\tau_\alpha S_y}{r\bar{Y}}\right)^2 \Big/ \left[1 + \frac{1}{N}\left(\frac{\tau_\alpha S_y}{r\bar{Y}}\right)^2\right]$$

Population C.V. $= \frac{S}{\bar{Y}}$.

The 1st approximation is taken as $n_0 = \left(\frac{\tau_\alpha S_y}{r\bar{Y}}\right)^2 = \frac{1}{C}\left(\frac{S_y}{\bar{Y}}\right)^2$ if N is large. If $\frac{n_0}{N}$ is appreciable we compute n as

$$n = \frac{n_0}{1 + n_0/N}.$$

Remark:- In any sampling design, the basic purpose is to obtain a sample which is a proper representative of the popln.. In SRS, the sample is selected randomly from the popln. (entire). An observed sample may be obtained from a particular part of other popln. then it may not be a good representative of the popln. If the popln. units are more or less homogeneous w.r.t. the study variable, then SRS produces samples which are good representative of the popln.

If the popln is not homogeneous or heterogeneous then SRS is not a proper sampling design.

Note:- Use SRS method if popln. is homogeneous.
Use Stratified sampling method if popln. is not homogeneous.

Randomized Response Techniques : Warner's Model :— [CU'2009]

A situation likely to lead either to refusals to answer or to evasive answers occurs when a question in a survey is sensitive or highly personal (e.g. does the respondent regularly engage shoplifting or use drugs?)

Consider first the estimation of a binomial proportion— the proportion π_A of respondents who belong to a certain class A or have committed a certain act. By ingenious use of a randomizing device, Warner (1965) showed that it is possible to estimate this proportion without the respondent revealing his or her personal status w.r.t. this question.

The randomizing device, such a box with red and white balls, selects one of the two statements or questions, each requiring a "yes" or "no" response, to be presented to the respondent. The interviewer does not know which question any respondent has answered, but does know the relative probability P and (1−P) with which the two statements are presented. The success of the method depends, of course, on the respondent's being convinced that by participating he or she will not be revealing personal status with regard to the sensitive issue.

In Warner's original proposal the two statements are: "I am a member of class A" presented with probability P, "I am not a member of class A", presented with prob. (1−P).

With a r.s. of n respondents the interviewer records a binomial estimate $\hat{\phi} = m/n$ of probability ϕ of 'yes' answers. If the questions are answered truthfully, the relation between ϕ and π_A in the popln. is

$$\phi = P\pi_A + (1-P)(1-\pi_A) = (2P-1)\pi_A + (1-P).$$

With known P, this relation suggests the estimate

$$\hat{\pi}_{AW} = \frac{\hat{\phi} - (1-P)}{(2P-1)},$$

if $P \neq 1/2$, this estimate turns out to be the MLE of π_A.

The estimate is unbiased, with variance $V(\hat{\pi}_{AW}) = \dfrac{\phi(1-\phi)}{n(2P-1)^2}$, since $m \sim Bin(n, \phi)$.

Writing in the form $1-\phi = (2P-1)(1-\pi_A) + (1-P)$, we find easily, $$V(\hat{\pi}_{AW}) = \frac{\pi_A(1-\pi_A)}{n} + \frac{P(1-P)}{n(2P-1)^2}.$$

Ques:- (2010) Suppose you want to estimate the proportion of people in a popln. who are drug addict. Assuming that a sampled person may not give a correct reply to a direct question, Discuss an alternative procedure to answer your question.

Ratio and Regression Estimators : Use of auxilliary information :-

(A) **Ratio Estimator :-** Frequently we come across situations in which the ratio of y to another character x is believed to be less variable than the y's themselves. In that case it would be better to estimate R, the ratio of y to x in the population, from the sample and then multiply it by the known total of x to estimate the total for y. This procedure is called ratio estimation.

Frequently we wish to estimate a ratio rather than a total or mean, for example, it is desired to estimate the total agricultural area in a region containing N communes. There are very big communes and very small communes and this makes the character y vary tremendously over the region. But the ratio of agricultural area and the popln. size of the commune, which is the per capita agricultural area, would be less variable.

Let Y and X be the total agricultural area and the total popln. in the region. Then the per-capita agricultural area in the region is $R = \frac{Y}{X}$. If a simple random sample of n communes gives $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i$ as the total for y and x, respectively. It is natural to estimate R by $\hat{R} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i = \frac{\bar{y}}{\bar{x}}$ and the total of y (i.e. Y) is estimated by $\hat{Y}_R = \hat{R} \cdot X = \frac{\bar{y}}{\bar{x}} \cdot X$, where X is known total of x. It should be noted that the two problems are different, though they are connected. For estimating Y we could have used information on any character x; this information need not to be recent, but must be known for the entire population. On the other hand, information on a sample basis is required for y as well as for x (the denominator of the ratio) if the purpose is to eliminate the ratio $R = \frac{Y}{X}$ in population.

Since the theory is same in the either case, most of the subsequent results will relate to the problem of estimating a ratio.

▪ **Bias of the Ratio Estimator :-** The following theorem gives the exact bias associated with $\hat{R}$.

• **Theorem :-** In simple random sampling, bias of the ratio estimator $\hat{R} = \frac{\bar{y}}{\bar{x}}$ is given by $B(\hat{R}) = -\frac{Cov(\hat{R}, \bar{x})}{E(\bar{x})}$.

Proof :- As $Cov\left(\frac{\bar{y}}{\bar{x}}, \bar{x}\right) = E(\bar{y}) - E(\bar{y}/\bar{x})E(\bar{x})$,

$$\Rightarrow \bar{X} . E\left(\frac{\bar{y}}{\bar{x}}\right) = \bar{Y} - Cov\left(\frac{\bar{y}}{\bar{x}}, \bar{x}\right)$$

$$\Rightarrow E(\hat{R}) = R - \frac{1}{\bar{X}} Cov\left(\frac{\bar{y}}{\bar{x}}, \bar{x}\right) = R - \frac{Cov(\hat{R}, \bar{x})}{E(\bar{x})}$$

$$\therefore B(\hat{R}) = E(\hat{R}) - R = -\frac{Cov(\hat{R}, \bar{x})}{E(\bar{x})}.$$

Q [ISS EXAM '10 '10 Marks] What are ratio and regression estimates? How would you obtain bias of these estimators?

Corollary:- Denoting the standard deviation of $\hat{R}$ by $\sigma(\hat{R})$, we have

$$B(\hat{R}) = -\frac{1}{\bar{X}} \cdot \sigma(\hat{R})\,\sigma(\bar{x}) \cdot \rho(\hat{R}, \bar{x})$$

$$\text{or, } \frac{B(\hat{R})}{\sigma(\hat{R})} = -\rho(\hat{R}, \bar{x}) \cdot \frac{\sigma(\bar{x})}{\bar{X}} = -\rho(\hat{R}, \bar{x}) \cdot C.V(\bar{x})$$

Hence, $\frac{|B(\hat{R})|}{\sigma(\hat{R})} \leq C.V.(\bar{x})$, since $|\rho(\hat{R}, \bar{x})| \leq 1$,

where, C.V. stands for the coefficient of variation. the same bound applies, of course, to the bias in $\hat{Y}_R$ and $\hat{\bar{Y}}_R$.

Remark:- (1) $\hat{R}$ is consistent for R in the sense that $\hat{R} = R$ when the sample size is N,

(2) The bias associated with $\hat{Y}_R = \hat{R}X$ is $XB(\hat{R})$.

(3) $\hat{R}$ is unbiased if $\rho(\hat{R}, \bar{x}) = 0$.

Theorem:- The approximate bias and mean square error (MSE) of the ratio estimator $\hat{R}$ are $B(\hat{R}) \simeq \frac{\left(\frac{1}{n} - \frac{1}{N}\right)}{\bar{X}^2}(RS_x^2 - \rho S_y S_x)$

and $MSE(\hat{R}) = \frac{\left(\frac{1}{n} - \frac{1}{N}\right)}{\bar{X}^2}(S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y)$ [CU]

Proof:- Define, $e_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}}$ and $e_1 = \frac{\bar{x} - \bar{X}}{\bar{X}}$.

It may be noted that (i) $E(e_0) = E\left(\frac{\bar{y} - \bar{Y}}{\bar{Y}}\right) = 0$ (ii) $E(e_1) = 0$

(iii) $E(e_0^2) = E\left(\frac{\bar{y} - \bar{Y}}{\bar{Y}}\right)^2 = \frac{V(\bar{y})}{\bar{Y}^2}$

(iv) $E(e_1^2) = \frac{Var(\bar{x})}{\bar{X}^2}$

(v) $E(e_0 e_1) = E\left\{\frac{(\bar{x} - \bar{X})(\bar{y} - \bar{Y})}{\bar{X}\bar{Y}}\right\} = \frac{Cov(\bar{x}, \bar{y})}{\bar{X}\bar{Y}}$

Assume that the sample size is large enough so that $|e_0| < 1$ and $|e_1| < 1 \Leftrightarrow 0 < \bar{x} < 2\bar{X}$, $0 < \bar{y} < 2\bar{Y}$.

Since $\bar{y} = \bar{Y}(1 + e_0)$, $\bar{x} = \bar{X}(1 + e_1)$, the estimator $\hat{R} = \frac{\bar{y}}{\bar{x}}$ can be written as

$$\hat{R} = \frac{\bar{Y}(1 + e_0)}{\bar{X}(1 + e_1)} = R(1 + e_0)(1 + e_1)^{-1}$$

$$= R\{1 + e_0 - e_1 + e_1^2 - e_0^2 e_1 + \ldots\ldots\}$$

Hence, $E(\hat{R}) - R = B(\hat{R}) \simeq R\{E(e_1^2) - E(e_0 e_1)\}$

$$= R\left\{\frac{V(\bar{x})}{\bar{X}^2} - \frac{Cov(\bar{x}, \bar{y})}{\bar{X}\bar{Y}}\right\}$$

$$= \frac{\left(\frac{1}{n} - \frac{1}{N}\right)}{\bar{X}^2}(RS_x^2 - \rho S_x S_y)$$

[In SRSWOR, $V(\bar{x}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_x^2$, $V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2$ and $Cov(\bar{x}, \bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{xy} = \left(\frac{1}{n} - \frac{1}{N}\right) \rho S_x S_y$]

Again, $MSE(\hat{R}) = E(\hat{R}-R)^2 \simeq R^2 E[e_0^2 + e_1^2 - 2e_0e_1]$, ignoring terms of degree greater than two.

Therefore $MSE(\hat{R}) \simeq R^2 \left\{ \frac{V(\bar{x})}{\bar{X}^2} + \frac{V(\bar{y})}{\bar{Y}^2} - \frac{2Cov(\bar{x},\bar{y})}{\bar{X}\bar{Y}} \right\}$

$$\simeq \frac{\left(\frac{1}{n}-\frac{1}{N}\right)}{\bar{X}^2} \left\{ R^2 S_x^2 + S_y^2 - 2\rho S_x S_y R \right\}$$

Remark:- (1) $B(\hat{Y}_R) = B(N\bar{X}\hat{R}) = N\left(\frac{N-n}{Nn}\right)\frac{1}{\bar{X}}\{RS_x^2 - \rho S_x S_y\}$

and $MSE(\hat{Y}_R) = N^2\left(\frac{N-n}{Nn}\right)\{R^2S_x^2 + S_y^2 - 2R\rho S_x S_y\}$

(2) The quantity $\frac{\text{Bias}}{\text{S.E.}}$, which is the same for $\hat{R}, \hat{Y}_R, \hat{\bar{Y}}_R$, may be expressed as $\frac{\text{Bias}}{\text{S.E.}} = C.V.(\bar{x}) \frac{(RS_x - \rho S_y)}{\sqrt{\{R^2S_x^2 + S_y^2 - 2R\rho S_x S_y\}}}$

where, $C.V.(\bar{x}) = \frac{\sqrt{V(\bar{x})}}{\bar{X}} = \frac{\left(\frac{1}{n}-\frac{1}{N}\right)^{1/2} \cdot S_x}{\bar{X}}$.

The following theorem gives the condition under which the ratio estimator will be more efficient than the conventional expansion estimator or estimator based on the mean per unit $(\bar{y})$.

Theorem:- The ratio estimator $\hat{Y}_R = \frac{\bar{y}}{\bar{x}}X = \hat{R}X$ is more efficient than the expansion estimator $\hat{Y}$, with simple random sample, if $\rho > \frac{1}{2}\cdot\frac{CV(x)}{CV(y)} = \frac{1}{2}\left(\frac{S_x}{\bar{X}}\right)\Big/\left(\frac{S_y}{\bar{Y}}\right)$.

[CU] [ISS'2012]

Proof:- $V(\hat{Y}) > MSE(\hat{Y}_R)$, under SRS.

$\Rightarrow N^2 \cdot \frac{N-n}{nN} S_y^2 > N^2 \cdot \frac{N-n}{Nn}\{S_x^2R^2 + S_y^2 - 2R\rho S_x S_y\}$, approximately.

$\Rightarrow S_y^2 > \{S_y^2 + R^2S_x^2 - 2R\rho S_x S_y\}$

$\Rightarrow \rho > \frac{1}{2}\cdot R \cdot \frac{S_x}{S_y} = \frac{1}{2}\left(\frac{S_x}{\bar{X}}\right)\Big/\left(\frac{S_y}{\bar{Y}}\right) = \frac{1}{2}\cdot\frac{CV(x)}{CV(y)}$, if $R > 0$

Hence the proof. [The theorem also holds for large samples]

Estimated MSE under Simple Random Sampling:—

Note that $\sum_{i=1}^{N}[Y_i - RX_i]^2 = \sum_{i=1}^{N}[Y_i - \bar{Y} + \bar{Y} - RX_i]^2$

$$= \sum_{i=1}^{N}[Y_i - \bar{Y} + R\bar{X} - RX_i]^2, \text{ since } R = \frac{\bar{Y}}{\bar{X}}.$$

$$= \sum_{i=1}^{N}[Y_i - \bar{Y}]^2 + R^2\sum_{i=1}^{N}[X_i - \bar{X}]^2 - 2R\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

$\Rightarrow \frac{1}{N-1}\sum_{i=1}^{N}[Y_i - RX_i]^2 = S_y^2 + R^2S_x^2 - 2R\rho S_x S_y$.

Then, we have $MSE(\hat{R}) = \left(\frac{1}{n}-\frac{1}{N}\right)\cdot\frac{1}{\bar{X}^2}\cdot\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - RX_i)^2$.

Therefore a reasonable estimator for the MSE of the ratio estimate is $v(\hat{R}) = \left(\frac{1}{n}-\frac{1}{N}\right)\cdot\frac{1}{\bar{x}^2}\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2/(n-1)$, where $\hat{R} = \bar{y}/\bar{x}$.

∴ The estimator is biased.

[ISS EXAM'12] Show that the ratio estimator is better than the one based on SRSWOR if $\rho > 1/2$ when $C_x = C_Y$. [10 marks]

Now, $\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\hat{R}x_i)^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n}y_i^2+\hat{R}^2\sum_{i=1}^{n}x_i^2-2\hat{R}\sum_{i=1}^{n}x_iy_i\right\}$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left\{y_i-\bar{y}-\hat{R}(x_i-\bar{x})\right\}^2$$

$$= \left\{s_y^2+\hat{R}^2 s_x^2-2\hat{R}s_{xy}\right\}, \text{ where}$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

Hence, $v(\hat{R}) = \left(\frac{1}{n}-\frac{1}{N}\right)\cdot\frac{1}{\bar{x}^2}\left\{s_y^2+\hat{R}^2 s_x^2-2\hat{R}s_{xy}\right\}$

Since $\hat{Y}_R = \hat{R}.\bar{X}N$, $v(\hat{Y}_R) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)\left\{s_y^2+\hat{R}^2 s_x^2-2\hat{R}s_{xy}\right\}$

Unbiased Ratio-type estimator:

Theorem:- In SRS, an unbiased estimator of $R=\bar{Y}/\bar{X}$ is given by

$$\hat{R} = \bar{r} + \frac{(N-1)n}{N(n-1)}\cdot\frac{\bar{y}-\bar{r}\bar{x}}{\bar{X}}, \text{ where } \bar{r} = \frac{1}{n}\sum_{i=1}^{n}r_i = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i}$$

Proof:- $\frac{1}{N}\sum_{i=1}^{N}R_i(X_i-\bar{X})$, where $R_i = \frac{Y_i}{X_i}$, $i=1(1)N$.

$$= \frac{1}{N}\sum_{i=1}^{N}(Y_i-\bar{X}R_i) = \bar{Y}-\bar{X}\cdot\frac{1}{N}\sum_{i=1}^{N}R_i = \bar{Y}-\bar{X}.E(r_i)$$

But in SRS, $E(\bar{r}) = E(r_i)$.

Hence, bias in $\bar{r} = E(\bar{r})-R = -\frac{1}{\bar{X}N}\sum_{i=1}^{N}R_i(X_i-\bar{X})$ —— (*)

Again, an unbiased estimator of

$$\frac{1}{N-1}\sum_{i=1}^{N}R_i(X_i-\bar{X}) \text{ is } \frac{1}{n-1}\sum_{i=1}^{n}r_i(x_i-\bar{x}) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-r_i\bar{x})$$

$$= \frac{n}{n-1}(\bar{y}-\bar{r}\bar{x}).$$

From (*), bias in $\bar{r} = E(\bar{r})-R = -\frac{(N-1)n}{N(n-1)}E\left\{\frac{\bar{y}-\bar{r}\bar{x}}{\bar{X}}\right\}$

$$\Rightarrow E\left\{\bar{r}+\frac{(N-1)n}{N(n-1)}\cdot\frac{\bar{y}-\bar{r}\bar{x}}{\bar{X}}\right\} = R.$$

Hence, $\hat{R}_* = \bar{r}+\frac{(N-1)n}{N(n-1)}\cdot\left(\frac{\bar{y}-\bar{r}\bar{x}}{\bar{X}}\right)$ is an unbiased estimator of $R = \frac{\bar{Y}}{\bar{X}}$.

Remark:-

(i) The corresponding UE of the popln. total Y is $\hat{Y}_R = \hat{R}_*X$

$$= \bar{r}X + \frac{(N-1)n}{n-1}(\bar{y}-\bar{r}\bar{x}).$$

(ii) An unbiased estimator of the popln mean $\bar{Y}$ is $\hat{\bar{Y}}_{R*} = \hat{R}_*\bar{X}$

$$= \bar{r}.\bar{X} + \frac{(N-1)n}{N(n-1)}(\bar{y}-\bar{r}\bar{x}).$$

(B). Regression Estimator :— [CU'2010]

Like the ratio estimator, the linear regression estimate is designed to increase precision by the use of an auxiliary variate x_i that is correlated with y_i. The ratio estimator is at its best when the relation between y and x is a straight line through the origin, that is, $y - Kx = 0 \Leftrightarrow y/x = K$. When the relation between y_i and x_i is examined, it may be found that although the relation is (approximately) linear, the line does not go through the origin. This suggests an estimator based on the linear regression of y on x rather than on the ratio of the variables.

We suppose that y_i and x_i are each obtained for every unit in the sample and that the popl^n mean $\bar{X}$ of the x_i is known. The linear regression estimator of $\bar{Y}$, the popl^n mean of y_i, is

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}),$$

where, b is an estimator of the change in y when x is increased by unit. The rationale behind this estimator is that if $\bar{x}$ is below average, we should expect $\bar{y}$ also to be below average by an amount $b(\bar{X} - \bar{x})$ because of the regression of y on x. For an estimator of the popl^n Y, we take $\hat{Y}_{lr} = N\bar{y}_{lr}$.

Suppose that we can take a rapid estimate x_i of some characteristic for every unit and can also, by some more costly method, determine the correct value y_i of the characteristic for a simple random sample of the units. For an example, an eye estimate of the volume of timber was made on each of a popl^n of $\frac{1}{10}$-acre plots, and the actual timber volume was measured for a simple random sample of the plots. The regression estimate

$$\bar{y} + b(\bar{X} - \bar{x})$$

adjusts the sample mean of the actual measurements by the regression of the actual measurements on the rapid estimates.

By a suitable choice of b, the regression estimate includes as particular cases both the mean per unit and the ratio estimate. Obviously if 'b' is taken as zero, then $\bar{y}_{lr} = \bar{y}$.

If $b = \frac{\bar{y}}{\bar{x}}$, $\bar{y}_{lr} = \bar{y} + \frac{\bar{y}}{\bar{x}}(\bar{X} - \bar{x}) = \frac{\bar{y}}{\bar{x}} \cdot \bar{X} = \hat{Y}_R$

Regression Estimator when b is computed from the sample :—

Let $y = \bar{Y} + B(x - \bar{X})$ be the popl^n regression line of y on x, where

$$B = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N} (X_i - \bar{X})^2}$$

is the popl^n regression coefficient.

Here 'b' must be the least squares estimate of B, that is,

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Comparison with the Ratio and the Mean per unit, in large sample :—

Theorem:- Under simple random sampling, with large sample, $V(\bar{y}) > MSE(\bar{y}_{lr})$ and $MSE(\bar{y}_R) > MSE[\bar{y}_{lr}]$. [CU'08]

Proof:- $MSE(\bar{y}_{lr})$ or $V(\bar{y}_{lr}) \simeq \frac{N-n}{Nn} S_y^2 (1-\rho^2)$, [regression]

$MSE(\bar{y}_R)$ or $V(\bar{y}_R) \simeq \frac{N-n}{Nn}(S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y)$, [Ratio]

$V(\bar{y}) = \frac{N-n}{Nn} S_y^2$ [Mean per unit]

since $|\rho| < 1$, $(1-\rho^2) < 1. \Rightarrow V(\bar{y}) > V(\bar{y}_{lr})$ or $MSE(\bar{y}_{lr})$

Now, $MSE(\bar{y}_R) - MSE(\bar{y}_{lr}) = \frac{N-n}{Nn}\{S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y - S_y^2 + S_y^2\rho^2\}$

$$= \frac{N-n}{Nn}\{R S_x - S_y \rho\}^2 > 0.$$

The regression estimator is more precise than the ratio estimator.

☐ **In large samples, when is MSE of regression estimator equal to that of the ratio estimator?** [CU'08]

Sol. When $B = R$

$\Leftrightarrow y = kx$, i.e., the relation between y and x is straight line through the origin.

The approximate MSE of the regression estimator under SRSWOR :—

Theorem:- If b is the least square estimate of B and $\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x})$, then in SRSWOR of size n, with large n,

$Var(\bar{y}_{lr}) / MSE(\bar{y}_{lr}) \simeq \frac{1-f}{n} S_y^2 (1-\rho^2)$, where $\rho = \frac{S_{yx}}{S_x S_y}$ is the population correlation between y and x. [CU'2010]

Proof:- The sampling error of $\bar{y}_{lr}$ arises from the quantity

$$\bar{y}_{lr} - \bar{Y} = \bar{y} - \bar{Y} + b(\bar{X} - \bar{x}).$$

As an approximation, replace $\bar{y}_{lr}$ by $\bar{y}_{lr}^* = \bar{y} + B(\bar{X} - \bar{x})$, where B is the population linear regression coefficient of y on x.

The error committed in this approximation is $(B-b)(\bar{X}-\bar{x})$.
Note that $(b-B)=O\left(\frac{1}{\sqrt{n}}\right)$ and $(\bar{x}-\bar{X})=O\left(\frac{1}{\sqrt{n}}\right)$, hence $(B-b)(\bar{X}-\bar{x})$ is of order $\frac{1}{n}$ in SRS.
Again, $V(\bar{y}^*_{lr})$ is of order $\frac{1}{n}$, since it is the variance of the sample mean of the variate $(y-Bx)$.

Hence, $E(\bar{y}_{lr}-\bar{Y})^2 = E\left\{\bar{y}^*_{lr}-\bar{Y}-(B-b)(\bar{X}-\bar{x})\right\}^2$

$$= V(\bar{y}^*_{lr}) + E\left[(b-B)^2(\bar{X}-\bar{x})^2\right] + 2E\left[(\bar{y}^*_{lr}-\bar{Y})(b-B)(\bar{X}-\bar{x})\right]$$

Now, $E\left[(b-B)^2(\bar{x}-\bar{X})^2\right] \le \left\{E(b-B)^4 E(\bar{x}-\bar{X})^4\right\}^{1/2}$, which is of order $1/n^2$. Similarly,

$$E\left[(b-B)^2(\bar{y}_{lr}-\bar{Y})(\bar{X}-\bar{x})\right] \le \left\{E(b-B)^2\right\}^{1/2}\left\{E(\bar{y}^*_{lr}-\bar{Y})^4 E(\bar{X}-\bar{x})^4\right\}^{1/4}$$

which is of order $1/n^{3/2}$.

Thus the large sample variance of the regression estimator $\bar{y}_{lr}$ is (MSE)

$$V(\bar{y}_{lr}) \simeq V(\bar{y}^*_{lr}) = Var\left(\bar{y}+B(\bar{X}-\bar{x})\right)$$
$$= Var(\bar{y}-B\bar{x})$$

let $e = y-Bx$
Then $e_i = y_i - Bx_i \ \forall\ i=1(1)n$

$$\therefore V(\bar{y}^*_{lr}) \simeq V(\bar{e}) = \left(\frac{1}{n}-\frac{1}{N}\right)\cdot S_e^2, \text{ under SRSWOR.}$$
$$= \left(\frac{1}{n}-\frac{1}{N}\right) S_y^2(1-\rho^2)$$
$$= \frac{1-f}{n}\cdot S_y^2(1-\rho^2).$$

Sample estimate of the MSE or Variance:- Note that $V(\bar{y}_{lr}) \simeq \frac{1-f}{n} S_e^2$, where, $S_e^2 = S_y^2(1-\rho^2)$.
Note that, an unbiased estimator of $S_e^2 = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2$ is

$$s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(e_i-\bar{e})^2.$$

Now, $e_i-\bar{e} = y_i-\bar{y}-B(x_i-\bar{x}) = \{y_i-\bar{y}-b(x_i-\bar{x})\}+(b-B)(x_i-\bar{x})$.
The 2nd term on the right, of order $\frac{1}{\sqrt{n}}$, may be neglected in relation to the 1st term, which of order unity.
Hence, in large sample $s_e^2 \simeq \frac{1}{(n-1)}\sum_{i=1}^{n}\{y_i-\bar{y}-b(x_i-\bar{x})\}^2$ is an estimate of S_e^2. The estimator $\frac{1}{(n-2)}\sum_{i=1}^{n}\{y_i-\bar{y}-b(x_i-\bar{x})\}^2$ is suggested since it is used in regression theory.

Q. [ISS EXAM '10] (10 marks) Write a critical note on the method of Double Sampling.

Double Sampling :— [CU]

The ratio and regression estimators assume the advance information about an auxiliary variable x. However there are some situations where the population mean or total of the auxiliary variable will not be known in advance. When such information is lacking, it is sometimes relatively cheap to take a large preliminary sample in which x alone is measured. A sample of size n' is selected initially by using a suitable sampling design and the popln mean $\bar{X}$ is estimated and then a sample of size n is selected to estimate the popln means of the study variable (y) and auxiliary variable (x); the second phase sample can be either a subsample of the first phase sample or it can be directly drawn from the given popln. this technique is known as <u>double sampling or two-phase sampling</u>. Two phase sampling is recommended only when the cost of conducting first phase survey is more economical when compared to that of the second phase.

<u>Ratio Estimators</u>:— In some application of double sampling, the auxiliary variable x has been used to make a ratio estimator of $\bar{Y}$. In the first (large) sample of size n', we measure only x; in the second, a random subsample of size $n = \nu n' = \frac{n'}{k}$ where the fraction 'ν' is chosen in advance, we measure both x and y. If the first sample is used to obtain $\bar{x}'$ as an estimator of $\bar{X}$ in a ratio estimate of $\bar{Y}$, the estimator of $\bar{Y}$ is

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}} . \bar{x}' = \hat{R} . \bar{x}'$$

To find the approximate variance, write

$$\bar{y}_R - \bar{Y} = \frac{\bar{y}}{\bar{x}} . \bar{x}' - \bar{Y} = \left(\frac{\bar{y}}{\bar{x}} \bar{X} - \bar{Y}\right) + \left(\frac{\bar{y}}{\bar{x}}(\bar{x}' - \bar{X})\right)$$

$$= \frac{\bar{X}}{\bar{x}}(\bar{y} - R\bar{x}) + \frac{\bar{y}}{\bar{x}}(\bar{x}' - \bar{X}).$$

The first component is the error of the ordinary ratio estimator. We replace $\bar{X}/\bar{x}$ by unity in this term. We replace the factor $\bar{y}/\bar{x}$ in the second component by the popln ratio $R = \bar{Y}/\bar{X}$. Thus

$$\bar{y}_R - \bar{Y} \simeq (\bar{y} - R\bar{x}) + R(\bar{x}' - \bar{X})$$

If the second sample is a random subsample of the first,

$$E_2(\bar{y}_R - \bar{Y}) \simeq \bar{y}' - \bar{Y} \; ; \; V_2(\bar{y}_R - \bar{Y}) \simeq \left(\frac{1}{n} - \frac{1}{n'}\right) s_d'^2,$$

where $s_d'^2$ is the variance within 2nd sample of the variate $d = (y - Rx)$.

Averaging over repeated random selections of the 1^{st} sample,

$$V(\bar{y}_R) = V_1 E_2(\bar{y}_R) + E_1 V_2(\bar{y}_R)$$

$$\simeq \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)\left(S_y^2 - 2\rho R S_x S_y + R^2 S_x^2\right).$$

since $E_2(s_d'^2) = S_d^2 = S_y^2 - 2\rho R S_x S_y + R^2 S_x^2$.

Seperating the term $1/n$, $1/n'$, we get,

[CU'09] $$V(\bar{y}_R) \simeq \frac{S_y^2 - 2\rho R S_x S_y + R^2 S_x^2}{n} + \frac{2R S_{xx} - R^2 S_x^2}{n'} - \frac{S_y^2}{N}.$$

Regression Estimators:– In some applications of double sampling the auxiliary variate x has been used to make a regression estimator $\bar{Y}$.

First sample size: n': measure only x

second sample size: $n = vn'$, v is given: measure both x and y.

The estimator of $\bar{Y}$ is $\bar{y}_{lr} = \bar{y} + b(\bar{x}' - \bar{x})$, where $\bar{x}', \bar{x}$ are the means of x in the 1^{st} and 2^{nd} samples and b is the least square regression coefficient of y on x, computed from the 2^{nd} sample.

☒ $MSE(\bar{y}_{lr})$ or $V(\bar{y}_{lr}) \simeq \frac{S_y^2(1-\rho^2)}{n} + \frac{\rho^2 S_y^2}{n'} - \frac{S_y^2}{N}$, assuming $1/n$ and $1/n'$ are negligible:–

Proof:– In finding the sampling error of $\bar{y}_{lr}$ in SRS, we showed that if b in $\bar{y}_{lr}$ is replaced by the finite popln region coefficient $B = \rho \frac{S_y}{S_x}$, the error in the approximation is of order $1/\sqrt{n}$ relative to that in $\bar{y}_{lr}$, we therefore examine the variance or MSE of the approximation, $\bar{y}_{lr}^* = \bar{y} + B(\bar{x}' - \bar{x})$.

Let $u_i = y_i - Bx_i$. Since the 2^{nd} sample is drawn at random from the (large) first sample, $E_2(\bar{y}_{lr}^*) = \bar{y}'$;

$V_2(\bar{y}_{lr}^*) = \left(\frac{1}{n} - \frac{1}{n'}\right) s_u'^2$, where $s_u'^2$ is the variance in the first phase sample. Then $V(\bar{y}_{lr}) \simeq V(\bar{y}_{lr}^*) = V_1 E_2(\bar{y}_{lr}^*) + E_1 V_2(\bar{y}_{lr}^*)$

$$= V(\bar{y}') + E_1\left\{\left(\frac{1}{n} - \frac{1}{n'}\right) s_u'^2\right\}$$

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

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

since $E(s_u^2) = S_u^2 = (1-\rho^2) S_y^2$.

Hence, $V(\bar{y}_{lr}) \simeq \frac{S_y^2(1-\rho^2)}{n} + \frac{\rho^2 S_y^2}{n'} - \frac{S_y^2}{N}$; (Proved)

Estimated Variance (or MSE) in Double Sampling for regression: (39)

If the terms in $\frac{1}{n}$ are negligible, $V(\bar{y}_{lr})$ is given by [CU'09]

$$V(\bar{y}_{lr}) \simeq \frac{S_y^2(1-\rho^2)}{n} + \frac{\rho^2 S_y^2}{n'} - \frac{S_y^2}{N}.$$

With a linear regression model, the quantity,

$$s_{y.x}^2 = \frac{1}{n-2}\left\{\sum_{i=1}^{n}(y_i-\bar{y})^2 - b^2\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}$$ is an UE of $S_y^2(1-\rho^2)$.

Since $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$ is an UE of S_y^2, it follows that

$(s_y^2 - s_{y.x}^2)$ is an UE of $\rho^2 S_y^2$.

Thus an estimator of $V(\bar{y}_{lr})$ or $MSE(\bar{y}_{lr})$ is

$$v(\bar{y}_{lr}) = \frac{s_{y.x}^2}{n} + \frac{s_y^2 - s_{y.x}^2}{n'} - \frac{s_y^2}{N}.$$

If the 2nd phase sample size is small and terms in $\frac{1}{n}$ are not negligible relative to 1, an estimate of variance suggested for SRS is

$$v(\bar{y}_{lr}) = s_{y.x}^2\left\{\frac{1}{n} + \frac{(\bar{x}'-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right\} + \frac{s_y^2 - s_{y.x}^2}{n'} - \frac{s_y^2}{N}.$$

Optimum allocation and comparison with single sampling :-

When $\frac{1}{n}$ is negligible, we have

$$V + \frac{S_y^2}{N} = \frac{S_y^2(1-\rho^2)}{n} + \frac{\rho^2 S_y^2}{n'} ;\ C = cn + c'n'.$$

By C-S inequality, the product VC is minimized when

$$\frac{cn^2}{S_y^2(1-\rho^2)} = \frac{c'n'^2}{\rho^2 S_y^2}$$

$$\Leftrightarrow \frac{n}{n'} = \left\{\frac{c'}{c}\frac{(1-\rho^2)}{\rho^2}\right\}^{1/2}.$$

Substitution in VC gives

$$(VC)_{min} = S_y^2\left(\sqrt{c(1-\rho^2)} + \sqrt{c'\rho^2}\right) - \frac{C S_y^2}{N}.$$

Thus for a specified cost C,

$$V_{min} = \frac{S_y^2\left(\sqrt{c(1-\rho^2)} + \sqrt{c'\rho^2}\right)^2}{C} - \frac{S_y^2}{N}.$$

If all resources are devoted instead to a single sample with no regression adjustment, this sample has size C/c and the variance of its mean is $V(\bar{y}) = \frac{c S_y^2}{C} - \frac{S_y^2}{N}$.

Hence, optimum use of double sampling gives a smaller variance if $c > \left\{\sqrt{c(1-\rho^2)} + \sqrt{c'\rho^2}\right\}^2$.

Bias in the Linear Regression Estimator :—

Introduce the variate $e_i = Y_i - \bar{Y} - B(X_i - \bar{X})$.
The properties of e_i are: $\sum_{i=1}^{N} e_i = 0$ and

$$\sum_{i=1}^{N} e_i(X_i - \bar{X}) = \sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X}) - B\sum_{i=1}^{N}(X_i - \bar{X})^2 = 0 \text{, by defn. of } B.$$

Now, $b = \sum_{i=1}^{n} y_i(x_i - \bar{x}) \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2$

$$= \left\{\sum_{i=1}^{n}\left[e_i + \bar{Y} + B(x_i - \bar{X})\right](x_i - \bar{x})\right\} \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= B + \left\{\sum_{i=1}^{n} e_i(x_i - \bar{x}) \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2\right\}$$

We have, $E(\bar{y}_{lr}) = \bar{Y} - E\{b(\bar{x} - \bar{X})\}$.

Thus one expression for bias, is

$$-E\{b(\bar{x} - \bar{X})\} = -\operatorname{cov}(b, \bar{x}).$$

Now, $-\operatorname{cov}(b, \bar{x}) = -\operatorname{cov}\left(B + \dfrac{\sum_{i=1}^{n} e_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \bar{x}\right)$

$$= -\frac{E(\bar{u} - \bar{U})(\bar{x} - \bar{X})}{S_x^2}$$

, where, $u_i = e_i(x_i - \bar{X})$ and $U = 0$.

$$= -\left(\frac{1}{n} - \frac{1}{N}\right)\cdot\frac{E(u_i - \bar{U})(x_i - \bar{X})}{S_x^2}$$

$$= -\frac{1-f}{n}\cdot\frac{E\{e_i(x_i - \bar{X})^2\}}{S_x^2},$$

which is the bias turns out to be in the estimator $\bar{y}_{lr}$.

——— * ——— * ——— * ——— * ——— * ———

Stratified Random Sampling :-

In stratified sampling the popln. of N units is first divided into subpopulations of $N_1, N_2, \ldots, N_L$ units, respectively. These sub-poplns are non-overlaping and together they comprise the whole set of the popln., so that $N_1 + N_2 + \cdots + N_L = N$. The subpopulations are called strata. To obtain the full benefit from stratification, the values of the N_h must be known. When the strata have been determined, a sample is drawn from each, the drawing being made independently in different strata. The sample sizes within the strata are denoted by $n_1, n_2, \ldots, n_L$, respectively. If a simple random sample is taken in each stratum, the entire procedure is described as stratified random sampling.

Notations :-
N : Population size.,
L : Number of strata in the population.
N_h : Number of units in the stratum h, $h = 1, 2, \ldots, L$.
Y_{hj} : the value of the jth unit in the stratum h, $j = 1(1)N_h$; $h = 1(1)L$.
n_h : sample size corresponding to the stratum h, $h = 1(1)L$
Y_h : Stratum total of the stratum h, $h = 1(1)L$.
$\bar{Y}_h$: Stratum mean of the stratum h, $h = 1(1)L$.
y_{hj} : the value of the jth sampled unit in the stratum h, $j = 1(1)n_h$, $h = 1(1)L$.
$\bar{y}_h$: stratum sample mean of the stratum h, $h = 1(1)L$.

$$S_h^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} \left[Y_{hj} - \bar{Y}_h \right]^2$$ is the true variance of the stratum 'h'.

and $$s_h^2 = \frac{1}{n_h - 1} \sum_{j=1}^{n_h} \left[y_{hj} - \bar{y}_h \right]^2$$ is the sample variance of the stratum 'h'.

Further $W_h = \frac{N_h}{N}$ is the stratum weight and

$f_h = \frac{n_h}{N_h}$ is the sampling fraction in the stratum 'h'.

[ISS EXAM '10]

Q. [10 marks]

Explain the concept of stratification in stratified Random sampling. What is proportional and optimum allocation in stratified simple Random sampling?

With usual notations show that:

$$V(\bar{y}_{st})_{opt} < V(\bar{y}_{st})_{prop} < V(\bar{y})_{ran}.$$

(1) Properties of the estimators:— For the popln mean per unit, the estimator used in stratified sampling is $\bar{y}_{st}$, where $\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h = \sum_{h=1}^{L} W_h \bar{y}_h$.

The estimator $\bar{y}_{st}$ is not in general the same as the sample mean. The sample mean $\bar{y} = \sum_{h=1}^{L} n_h \bar{y}_h$

(2) The estimated variance of $\bar{y}_{st}$:—

- Theorem:- With stratified random sampling, an unbiased estimator of the variance of $\bar{y}_{st}$ is

$$v(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h (N_h - n_h)\frac{s_h^2}{n_h}.$$

Proof:- If a simple random sample is taken within each strata, an unbiased estimator of S_h^2 is $s_h^2 = \frac{1}{(n_h-1)}\sum_{h=1}^{L}(y_{hj}-\bar{y}_h)^2$

Note that,
$$E\left(v(\bar{y}_{st})\right) = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h(N_h-n_h)}{n_h}.E\left(s_h^2\right)$$

$$= \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h\left(N_h-n_h\right)}{n_h}.S_h^2$$

$$= V\left(\bar{y}_{st}\right).$$

Alternative form for computing purposes is

$$v(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h . s_h^2}{N}.$$

~~Allocation of sample size to strata:~~

Principal Advantages of Stratified Random Sampling:—

1. More Representative: In an unstratified random sample some strata may be over-represented, others may be under-represented while some may be excluded altogether. Stratified sampling ensures any desired representation in the sample of the various strata in the population.

2. Greater Accuracy: Stratified random sampling provides estimates with increased precision. Moreover, stratified random sampling enables us to obtain the results of known precision for each of the stratum.

3. Administrative Convenience: As compared with SRSample, the stratified samples would be more concentrated geographically. Accordingly, the time and money involved in collecting the data and interviewing the individuals may be considerably reduced and the supervision of the field work could be allotted with greater ease and convenience.

Allocation of Sample size:-

Allocation of n_h's to various strata is called proportional if the sample fraction is constant for each stratum, i.e.

$$\frac{n_1}{N_1}=\frac{n_2}{N_2}=\cdots=\frac{n_L}{N_L}=\frac{\sum_{h=1}^{L} n_h}{\sum_{h=1}^{L} N_h}=\frac{n}{N}=c \text{ (constant)}$$

Proportional Allocation

$$\Rightarrow \frac{n_L}{N_L}=c=\frac{n}{N} \Rightarrow n_h \propto N_h \quad (h=1,2,\ldots,L)$$

This, in proportional ~~[illegible]~~ allocation each stratum is presented according to its size.

In proportional allocation, var $(\bar{y}_{st})$ is given by:

$$V_{prop}(\bar{y}_{st})=\sum_{h=1}^{L}\frac{N_h(N_h-n_h)}{N^2 n_h}\cdot S_h^2$$

$$=\sum_{h=1}^{L}\frac{N_h}{N}\left(\frac{N_h}{n_h}-1\right)\cdot\frac{S_h^2}{N}$$

$$=\sum_{h=1}^{L}\frac{N_h}{N}\cdot S_h^2\cdot\frac{1}{N}\left(\frac{N}{n}-1\right) \quad ;\left[\text{as } \frac{N_h}{n_h}=\frac{N}{n}\ \forall h.\right]$$

$$=\frac{N-n}{Nn}\sum_{h=1}^{L}\frac{N_h}{N}\cdot S_h^2$$

Theorem:- If in every stratum the estimator $\bar{y}_h$ is unbiased, then $\bar{y}_{st}$ is an unbiased estimator of the popln. mean $\bar{Y}$.

Proof:-

$$E[\bar{y}_{st}]=E\left[\sum_{h=1}^{L}W_h\bar{y}_h\right]=\sum_{h=1}^{L}W_h E(\bar{y}_h)$$

$$=\sum_{h=1}^{L}W_h\bar{Y}_h=\bar{Y},$$

since the estimators are unbiased in the individual strata.

and here $\bar{Y}=\sum_{h=1}^{L}\sum_{j=1}^{N_h}Y_{hj}/N=\sum_{h=1}^{L}\frac{N_h\bar{Y}_h}{N}=\sum_{h=1}^{L}W_h\cdot\bar{Y}_h$.

Theorem:- If the samples are independently drawn from the different strata, $V(\bar{y}_{st})=\sum_{h=1}^{L}W_h^2 V(\bar{y}_h)$, where $V(\bar{y}_h)$ is the variance of an unbiased estimator $\bar{y}_h$ in the stratum h.

Proof:- Since samples are drawn independently from different strata, so $\text{Cov}(\bar{y}_h,\bar{y}_k)=0$, $h\neq k$.

Therefore,

$$V(\bar{y}_{st})=V\left(\sum_{h=1}^{L}W_h\cdot\bar{y}_h\right)$$

$$=\sum_{h=1}^{L}W_h^2V(\bar{y}_h)+2\sum\sum_{h<k}\text{Cov}(\bar{y}_h,\bar{y}_k)$$

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

Optimum Allocation:- The proportional allocations described above do not take into account any factor other than strata sizes. They completely ignore the internal structure of strata like within stratum variability etc, and hence is desirable to consider an allocation scheme which takes into account these aspects. A guiding principle in the determination of the n_i's is to choose them as to:

(a) Minimize the variance of the estimator for
(i) fixed sample size 'n' and (ii) fixed cost.
(b) Minimize the total cost for fixed variance.

Since minimum variance or minimum total cost is an optimal property, the allocation of n_h's to the strata in accordance with the above principles is known as optimum allocation. Thus, in optimum allocation n_h's are to be obtained such that

(i) $Var(\bar{y}_{st})$ is minimum for fixed n.
(ii) $Var(\bar{y}_{st})$ is minimum for fixed total cost C (say).
(iii) Total cost C is minimum for fixed value of $Var(\bar{y}_{st}) = V_0$ (say)

Cost Function:- In any sample survey, the value of information on the experimental units must always be balanced against the cost of obtaining it. In stratified sampling it may cost more to obtain information about a sample in one stratum than in another. For example, interviewing people in rural areas is going to be more costly because of travel expenses than interviewing people in urban areas. Thus, in its simplest form the cost function C in stratified sampling may be given by the linear model:

$$C = a + \sum_{h=1}^{L} c_h n_h$$

where 'a' is the overhead cost and C_h is the cost per unit in the h^{th} stratum.

Theorem 1:- $Var(\bar{y}_{st})$ is minimum for fixed total size of the sample (n) if $n_h \propto N_h S_h$.

Solution:- Here the problem is to minimize:

$$Var(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h},$$

subject to the given condition $\sum_{h=1}^{L} n_h = n$ (fixed).

This is equivalent to minimizing:

$$\phi = Var(\bar{y}_{st}) + \lambda \left(\sum_{h=1}^{L} n_h - n \right)$$

$$= \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h} + \lambda \left(\sum_{h=1}^{L} n_h - n \right)$$

where, λ is the Lagrange's multiplier.

$$\therefore \frac{\partial \phi}{\partial n_h} = -\frac{N_h^2 S_h^2}{N^2 n_h^2} + \lambda = 0$$

$$\Rightarrow n_h = \frac{N_h S_h}{N\sqrt{\lambda}} \quad \cdots\cdots (*)$$

Also, $\frac{\partial^2 \phi}{\partial n_h^2} = \frac{2N_h^2 S_h^2}{N^2 n_h^3} > 0$

From $(*)$

$$\therefore \sum_{h=1}^{L} n_h = \frac{1}{\sqrt{\lambda}.N} \sum_{h=1}^{L} N_h S_h$$

$$\Rightarrow n = \frac{1}{N\sqrt{\lambda}} \sum_{h=1}^{L} N_h S_h$$

$$\Rightarrow \sqrt{\lambda} = \frac{1}{Nn} \sum_{h=1}^{L} N_h S_h$$

Substituting in $(*)$, we have, $n_h = \frac{n N_h S_h}{\sum_{h=1}^{L} N_h S_h}$

$\therefore \boxed{n_h \propto N_h S_h}$, for a fixed total sample size.

This is known as Neyman's formula for optimum allocation. This suggests that greater the value of $N_h S_h$ for a given stratum, greater is the number of sampling units to be selected from the stratum in order to obtain the most precise estimate of the popln. mean.

Theorem 2:- In stratified Random sampling with given cost function of the form: $C = a + \sum_{h=1}^{L} c_h n_h$, then $Var(\bar{y}_{st})$ is minimum if

$$n_h \propto \frac{N_h S_h}{\sqrt{C_h}}.$$

Proof:-

$$C = a + \sum_{h=1}^{L} c_h n_h$$

We have to minimize $Var(\bar{y}_{st})$ subject to the condition: $\sum_{h=1}^{L} c_h n_h = C - a$

Equivalently, we have to minimise:

$$\psi = Var(\bar{y}_{st}) + \lambda\left(\sum_{h=1}^{L} c_h n_h - C + a\right)$$

$$= \frac{1}{N^2}\sum_{h=1}^{L} N_h (N_h - n_h)\frac{S_h^2}{n_h} + \lambda\left[\sum_{h=1}^{L} c_h n_h - C + a\right]$$

where λ being Lagrange's multiplier.

$$\frac{\partial \psi}{\partial n_h} = -\frac{N_h^2 S_h^2}{N^2 n_h^2} + \lambda C_h = 0 \Rightarrow n_h = \frac{N_h S_h}{N\sqrt{C_h}\sqrt{\lambda}}$$

$$\therefore \sum_h n_h = \frac{\sum_h [N_h S_h/\sqrt{C_h}]}{N\sqrt{\lambda}} = n$$

$$\Rightarrow \sqrt{\lambda} = \frac{\sum_{h=1}^{L} [N_h S_h/\sqrt{C_h}]}{nN}.$$

Substituting it, we get $n_h = \dfrac{n N_h S_h / \sqrt{C_h}}{\sum_{h=1}^{L} [N_h S_h / \sqrt{C_h}]}$

Thus, in optimum allocation for a fixed cost,

$$n_h \propto \frac{N_h S_h}{\sqrt{C_h}} \Rightarrow n_h = \lambda . \frac{N_h S_h}{\sqrt{C_h}} \ \forall h.$$

Result:-> Find the optimum allocation when cost (C) is fixed. [CU]

Sol. $C = a + \sum_{h=1}^{L} C_h n_h$, we get

$$C - a = \lambda \sum_{h=1}^{L} C_h . \frac{N_h S_h}{\sqrt{C_h}}$$

$$\Leftrightarrow \lambda = \frac{C-a}{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}}$$

Then the optimum allocation is

$$n_h = \frac{(C-a)}{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}} . \left\{ \frac{N_h S_h}{\sqrt{C_h}} \right\}$$

Remark:- This leads to the conclusion:

A larger sample would be required from a stratum if —

(i) Stratum size (N_h) is large,

(ii) Stratum variability (S_h) is large,

(iii) Sampling cost per unit is low in the stratum.

Relative Precision between Stratified Random and simple random Sampling

Theorem:- In connection with stratified random sampling, show that

$$V_{ran} \geq V_{prop} \geq V_{opt},$$

where the symbols have their usual significance. [CU]

Proof:- $V_{ran} = (1-f)\frac{S^2}{n}$, $f = \frac{n}{N}$ and $V_{prop} = \frac{(1-f)}{n} \sum_h W_h S_h^2$

also, $V_{opt} = \frac{1}{n}\left(\sum_h W_h S_h\right)^2 - \frac{1}{N}\sum_h (W_h S_h^2)$.

Note that, $(N-1)S^2 = \sum_{h=1}^{L}\sum_{j=1}^{N_h} (Y_{hj} - \bar{Y})^2 = \sum_h \sum_j (Y_{hj} - \bar{Y}_h)^2 + \sum_h (\bar{Y}_h - \bar{Y})^2 . N_h$

$$= \sum_h (N_h - 1) S_h^2 + \sum_h N_h (\bar{Y}_h - \bar{Y})^2$$

If the terms $\frac{1}{N_h}$ are negligible and hence in $\frac{1}{N}$, then we have

$$S^2 = \sum W_h S_h^2 + \sum W_h (\bar{Y}_h - \bar{Y})^2$$

Hence $V_{ran} = (1-f)\frac{S^2}{n} = \frac{(1-f)}{n} \sum W_h S_h^2 + \frac{(1-f)}{n} \sum W_h (\bar{Y}_h - \bar{Y})^2$

$$= V_{prop} + \frac{1-f}{n} \sum W_h (\bar{Y}_h - \bar{Y})^2 \geq V_{prop}.$$

Again $V_{prop} - V_{opt} = \frac{1}{n}\left[\sum_h W_h S_h^2 - \left(\sum W_h S_h\right)^2\right] = \frac{1}{n}\sum W_h (S_h - \bar{S})^2$.

where $\bar{S} = \sum W_h S_h$.

Hence, $\boxed{V_{ran} \geq V_{prop} \geq V_{opt}}$

Efficiency of Stratified Random Sampling over Simple Random Sampling

The efficiency (E) of stratified random sampling over simple random sampling depends on the method of allocation of the sample size to various strata and is defined as:

$$E = \frac{1/[V(\bar{y}_{st})]}{1/[V(\bar{y}_n)_R]} = \frac{V(\bar{y}_n)_R}{V(\bar{y}_{st})}.$$

Gain in efficiency due to stratification $= E-1 = \frac{V_{ran} - V_{yst}}{V_{yst}}$.

Percentage gain in efficiency due to stratification $= 100 \times (E-1)$.

Estimation from a sample of the "Gain due to stratification" :-

Theorem:- Given the results of a stratified random sampling, an unbiased estimator of V_{ran}, the variance of the mean of a simple random sample from the same popln. is

$$v_{ran} = \frac{N-n}{n(N-1)}\left[\frac{1}{N}\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{j=1}^{n_h} y_{hj}^2 - \bar{y}_{st}^2 + v(\bar{y}_{st})\right]$$

Proof:-

$$V_{ran} = \frac{N-n}{nN}S^2 = \frac{N-n}{n(N-1)}\left[\frac{1}{N}\sum_h\sum_j Y_{hj}^2 - \bar{Y}^2\right]$$

Note that
$$E[y_{hj}^2] = \sum_{k=1}^{N_h} Y_{hk}^2 \cdot \frac{1}{N_h}, \quad j = 1(1)n_h, \forall\, h = 1(1)L.$$
$$= \frac{1}{N_h}\sum_{k=1}^{N_h} Y_{hk}^2$$

and
$$\frac{1}{N}E\left(\sum_{h=1}^{N_h}\frac{N_h}{n_h}\sum_{j=1}^{n_h} y_{hj}^2\right) = \frac{1}{N}\sum_{h=1}^{L}\frac{N_h}{n_h}\cdot n_h \cdot \frac{1}{N_h}\sum_{j=1}^{N_h} Y_{hj}^2$$
$$= \frac{1}{N}\sum_h\sum_j Y_{hj}^2$$

Also, since $v(\bar{y}_{st})$ and $\bar{y}_{st}$ are unbiased estimator of $V(\bar{y}_{st})$ and $\bar{Y}$, respectively.

$$E[v(\bar{y}_{st})] = V(\bar{y}_{st}) = E(\bar{y}_{st}^2) - \bar{Y}^2$$
$$\Rightarrow E[\bar{y}_{st}^2 - v(\bar{y}_{st})] = \bar{Y}^2.$$
$\Rightarrow \bar{y}_{st}^2 - v(\bar{y}_{st})$ is an unbiased estimator of $\bar{Y}^2$.

Hence, $$V_{ran} = \frac{N-n}{n(N-1)}\left\{\frac{1}{N}\sum_h\sum_j Y_{hj}^2 - \bar{Y}^2\right\}$$

~~estimator of V_{ran}~~

i.e. $$v_{ran} = \frac{N-n}{n(N-1)}\left\{\frac{1}{N}\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{j=1}^{n_h} y_{hj}^2 - \bar{y}_{st}^2 + v(\bar{y}_{st})\right\}$$
is an unbiased estimator of V_{ran}.

Problem :- [CU]

With two strata, a sampler would like to have $n_1 = n_2$ for administrative convenience, instead of using the values given by the Neyman allocation. If V, V_{opt} denote the variance given by the $n_1 = n_2$ and Neyman allocation, respectively, show that fractional increase in variance

$$\frac{V - V_{opt}}{V_{opt}} = \left(\frac{r-1}{r+1}\right)^2, \text{ where } r = \frac{n_1}{n_2} \text{ as given by Neyman allocation.}$$

~~Alternative Form:- The units in a population are allocated to two strata with $N_1/N_2 = d$ and $\sigma_1/\sigma_2 = \lambda$.~~

OR

A popln. is segregated into two strata of sizes N_1, N_2 units. Random samples of sizes n_1 and n_2 are to be drawn with replacement from the two strata to estimate the popln. mean. Suppose $\lambda = \frac{N_1}{N_2}$, and $d = \frac{\sigma_1}{\sigma_2}$ where σ_i^2 is the i^{th} stratum variance, $i = 1, 2$. If V_0 is the variance of the usual unbiased estimator for the best choice of n_1 and n_2 and V_c is the variance for the choice $n_1 = n_2$ then show that

$$\frac{V_c - V_0}{V_0} = \left(\frac{1 - \lambda d}{1 + \lambda d}\right)^2.$$

Solution:- Under equal allocation $n_1 = n_2 = \frac{n}{2}$. We have

$$V(\hat{Y}_{st}) = V = \sum_{h=1}^{2} N_h^2 \cdot \frac{N_h - n_h}{N_h n_h} \cdot S_h^2$$

$$= \sum_{h=1}^{2} N_h^2 \left\{\frac{1}{n_h} - \frac{1}{N_h}\right\} S_h^2$$

$$\simeq \sum_{h=1}^{2} \frac{N_h^2 S_h^2}{n_h} \text{ for large } N_h. \quad \text{———} (*)$$

$$= \frac{2}{n}\left[N_1 S_1^2 + N_2 S_2^2\right], \text{ putting } n_h = \frac{n}{2}$$

Using Neyman allocation, $n_h = \frac{N_h S_h}{N_1 S_1 + N_2 S_2} \cdot n$, $h = 1, 2$ ——— $(**)$

From $(*)$, $V_{opt} = \frac{1}{n}\left[N_1 S_1 + N_2 S_2\right]^2$

By defn. of r, we have $r = \frac{n_1}{n_2} = \frac{N_1 S_1}{N_2 S_2}$, from $(**)$

Then $V = \frac{2}{n} N_2^2 S_2^2 (r^2 + 1)$, $V_{opt} = \frac{N_2 S_2^2}{n}[r+1]^2$

$$\therefore \frac{V - V_{opt}}{V_{opt}} = \frac{\frac{N_2^2 S_2^2}{n}\left\{2(r^2+1) - (r+1)^2\right\}}{\frac{N_2^2 S_2^2}{n}(r+1)^2} = \left(\frac{r-1}{r+1}\right)^2$$

For the alternative form, $r = \frac{n_1}{n_2} = \frac{N_1 S_1}{N_2 S_2} \simeq \frac{N_1 \sigma_1}{N_2 \sigma_2}$, i.e. $r = \lambda d$.

$$\therefore \frac{V_c - V_0}{V_0} = \left(\frac{1 - \lambda d}{1 + \lambda d}\right)^2 \quad [\text{Proved}]$$

- Write a short note on Circular systematic sampling. [CU]

ANS:- Circular Systematic Sampling:— If N is not a multiple of n, i.e., $N \neq nk$, then the sampling interval K can't be uniquely defined. In such a case, take K to be an integer nearest to (N/n). If we select the first unit, say i, randomly between 1 and K, then the systematic samples is:

$$i,\ i+K,\ i+2K,\ i+3K,\ \ldots\ldots,\ i+(n-1)K;\quad 1 \le i \le K.$$

Suppose we want a systematic sample 6 out of 22 units. We have $N=22$ and $n=6$ so that $N/n = 3.67$, we take $K=4$, the integer nearest to 3.67.

Thus, the four systematic samples are:

Sample No.	Random Start	Sample Units	Sample size
1	1	1, 5, 9, 13, 17, 21, ...	n=6
2	2	2, 6, 10, 14, 18, 22, ...	n=6
3	3	3, 7, 11, 15, 19, ...	n=5
4	4	4, 8, 12, 16, 20, ...	n=5

Thus, the sample size is not necessaryily $n(=6)$ but in some cases it is $n-1\ (=5)$.

Moreover, in this case, the sample mean is not an unbiased estimator of the population mean.

The problem of variable sample size when $N \neq nK$ can be overcame by adopting a modified method introduced by Prof. D.B. Lahiri (1952) and known as Circular systematic Sampling (CSS); This ensures a constant sample size.

The procedure consists in selecting the unit 'i' by random start from 1 to N, and thereafter select every K^{th} unit in a circular way, K being an integer nearest to (N/n). The systematic sample is then specified by the units corresponding to the numbers:

$$\left.\begin{array}{l} i+jK, \text{ if } i+jK \le N \\ \text{and } i+jK-N, \text{ if } i+jK > N \end{array}\right\},\ j=0,1,2,\ldots\ldots,(n-1).$$

Using this technique in the above illustration, for the random starts $i=3$ and $i=4$, the corresponding systematic samples of size 6 are given below.

$i=3$; the sample units are: 3, 7, 11, 15, 19, 1 $(=23-22)$

$i=4$; Sample units are: 4, 8, 12, 16, 20, 2 $(=24-22)$;

each with $n=6$, the desired sample size.

For $N \neq nK$, an unbiased estimate of $\overline{Y}_N$ is provided by:

$$\hat{Y}_N = \frac{K}{N}\sum_{j=1}^{n'} y_{ij},$$ where n' is the no. of units that can be expected in the sample.

Estimation of the variance of the estimate:-

$$V(\hat{Y}_{css}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_{ci} - Y)^2.$$

Note that, circular systematic sampling reduces to linear systematic sampling when N/n is an integer; it is thus more general than linear sampling.

Scanned by CamScanner

Comparison between Linear Systematic and Stratified Random sampling

Linear systematic sampling stratifies the popln. into n strata, which consist of the first k units, the second k units, and so on. We might therefore expect the linear systematic sample to be about as precise as the corresponding stratified random sample with one unit per stratum. The difference is that with the systematic sample the units occur at the same relative position in the stratum, whereas with the stratified random sample, the position in the stratum is determined seperately by randomisation within each stratum. The systematic sample is more scattered evenly over the population than that of stratified sample.

X ≡ systematic sample   O ≡ stratified random sample

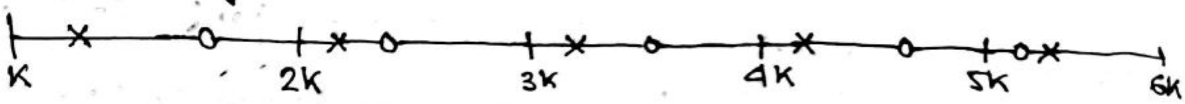

Fig: Systematic and stratified random sampling

The performance of linear systematic sampling in relation to that of stratified or SRS is greatly dependent on the properties of population. For some populations and some values of n, $V(\bar{y}_{st}) = \frac{S^2}{n}\cdot\frac{(N-1)}{N}\{1+(n-1)\rho_{w}\}$ may even increase when a large sample is taken — even a small positive correlation may have a large effect because of the multiplier $(n-1)$.

Thus it is difficult to advice about the situation which systematic sampling is to be recommended — a knowledge of the structure of the population is necessary for its most effective use.

Linear Systematic Sampling

Write a short note on linear systematic sampling.

Ans:- Linear Systematic Sampling :- [CU]

Suppose that the N units in the population are numbered 1 to N in some order. Suppose $N = nK$, where n is the sample size desired and K is an integer. A number is taken at random from the numbers 1 to K (using a table of random numbers). Suppose the random sample is i. Then starting from the i^{th} unit in the popln., every K^{th} unit is selected till a sample size of n is obtained. Then the sample contains n units with serial numbers $i, i+k, i+2k, \ldots\ldots, i+(n-1)k$. Thus the sample consists of the first unit selected at random and every K^{th} unit there after. It is therefore called a systematic sample (with K as the sampling interval) and this procedure of selection is known as systematic sampling or linear systematic sampling.

For example, when $N = 24$, $n = 6$ and $k = 4$, the four possible linear systematic samples are:

Sample Number	Random start	Sampled units
1	1	1, 5, 9, 13, 17, 21
2	2	2, 6, 10, 14, 18, 22
3	3	3, 7, 11, 15, 19, 23
4	4	4, 8, 12, 16, 20, 24

The linear systematic sampling scheme described above can be regarded as dividing the popln of N units into k mutually exclusive and exhaustive groups (clusters) $\{S_1, S_2, \ldots, S_K\}$ of n units each and choosing one of them at random. A linear systematic sample is a simple random sample of one cluster unit from a popln. of k cluster units.

A linear systematic sampling is a mixed sampling, which is partly probabilistic and partly non probabilistic.

[CU] Ques:- Distinguish between linear and circular systematic sampling

[ISS EXAM'11] [8 Marks] Explain the concepts of linear and circular systematic sampling giving suitable illustrations. Further show that for linear systematic sampling, sample mean is an unbiased estimator for population mean.

Unbiased Estimator for the population total and its variance :- [CU]

- Theorem :- An unbiased estimator for the popln. total Y and linear systematic sampling corresponding to the random start r is given by

$$\hat{Y}_{st} = \frac{N}{n}\sum_{j=1}^{n} Y_{r+(j-1)k}, \text{ and variance is given by}$$

$$V(\hat{Y}_{st}) = \frac{1}{k}\sum_{r=1}^{k}(\hat{Y}_r - Y)^2,$$ where $\hat{Y}_r$ is the value of $\hat{Y}_{st}$ corresponding to the random start r.

Proof :- Note that $\hat{Y}_{st}$ can take any one of k values $\hat{Y}_r$, $r = 1(1)k$ with prob. $\frac{1}{k}$.

$$\therefore E(\hat{Y}_{st}) = \sum_{r=1}^{k}\hat{Y}_r \cdot \frac{1}{k} = \frac{1}{k}\sum_{r=1}^{k}\frac{N}{n}\sum_{j=1}^{n} Y_{r+(j-1)k}$$

$$= \frac{N}{nk}\sum_{r=1}^{k}\sum_{j=1}^{n} Y_{r+(j-1)k}$$

$$= \sum_{i=1}^{N} Y_i = Y.$$

Hence $\hat{Y}_{st}$ is unbiased for the population total Y.

Again, $V(\hat{Y}_{st}) = E[\hat{Y}_{st} - E(Y_{st})]^2 = E[\hat{Y}_{st} - Y]^2 = \frac{1}{k}\sum_{r=1}^{k}[\hat{Y}_r - Y]^2$.

An alternative expression for $V(\hat{Y}_{st})$:- [CU]

- Theorem :- In linear systematic sampling interval of k, from a population of size $N = nk$, the variance of $\hat{Y}_{st}$ is given by

$$V(\hat{Y}_{st}) = \frac{N(N-1)}{n} \cdot S^2\{1+(n-1)\rho\}, \text{ where}$$

$\rho = E[(Y_{ij} - \bar{Y})(Y_{ij'} - \bar{Y})] / E[Y_{ij} - \bar{Y}]^2$, $j \neq j'$, is the intra-cluster correlation coefficient.

Proof :- We have $V(Y_{st}) = \frac{1}{k}\sum_{r=1}^{k}(\hat{Y}_r - Y)^2 = \frac{1}{k}\sum_{r=1}^{k}\left\{\sum_{j=1}^{n}(Y_{rj} - \bar{Y})\right\}^2$

$$= \frac{1}{k}\sum_{r}\sum_{j}(Y_{ij} - \bar{Y})^2 + \frac{2}{k}\sum_{i}\sum_{j}\sum_{k>j}(Y_{ij} - \bar{Y})(Y_{ik} - \bar{Y})$$

By definition, $\rho = \frac{2}{kn(n-1)}\sum_{i}\sum_{j}\sum_{k>j}(Y_{ij} - \bar{Y})(Y_{ik} - \bar{Y}) / V(y)$

$$\Rightarrow \sum_{i}\sum_{j}\sum_{k>j}(Y_{ij} - \bar{Y})(Y_{ik} - \bar{Y}) = \frac{kn(n-1)}{2}V(y)\cdot\rho$$

$$= \frac{1}{2}kn(n-1)\rho \cdot \frac{(N-1)}{N} \cdot S_y^2.$$

Hence, $V(\hat{Y}_{st}) = k(N-1)S_y^2 + \frac{2}{k} \cdot \frac{kn(n-1)(N-1)\rho S_y^2}{2N}$

$$= k(N-1)S_y^2 + k(n-1)(N-1)\rho S_y^2 \quad \left[\because \frac{N}{n} = k^{-1}\right]$$

$$= k(N-1)S_y^2\{1+(n-1)\rho\}$$

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

Corollary :- $V(\hat{Y}_{st})$ is systematic sampling be smaller than $V(\hat{Y})$ in SRSWOR if $N(N-1)S_y^2 \cdot \frac{1+\overline{n-1}\rho}{n} < N(N-n)\frac{S_y^2}{n}$

if $\rho < -\frac{1}{N-1}$.

Population with linear trend:-

If the values $Y_1, Y_2, \ldots, Y_N$ of the units with labels $1, 2, \ldots, N$ are modeled by $Y_i = \alpha + \beta i$, $i = 1(1)N$, i.e. the population consists solely a linear trend.

Theorem:- For population possessing linear trend, $V(\hat{Y}_{st}) < V(\hat{Y}_{srs})$ where, $\hat{Y}_{st}$ and $\hat{Y}_{srs}$ are the usual estimators under linear systematic sampling and simple random sampling, respectively. [CU]

Proof:- $V(\hat{Y}_{srs}) = N^2 \cdot \frac{N-n}{nN} \cdot \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$

Let, $Y_i = \alpha + \beta i$, $i = 1(1)N$. Then $\bar{Y} = \alpha + \beta\left(\frac{N+1}{2}\right)$

Now,
$$\sum_{i=1}^{N} (Y_i - \bar{Y})^2 = \sum_{i=1}^{N} \left\{\alpha + \beta i - \alpha - \beta \frac{N+1}{2}\right\}^2$$
$$= \beta^2 \sum_{i=1}^{N} \left(i - \frac{N+1}{2}\right)^2$$
$$= \beta^2 \left\{\sum_{i=1}^{N} i^2 - N\left(\frac{N+1}{2}\right)^2\right\}$$
$$= \beta^2 \cdot \frac{N(N^2-1)}{12}.$$

Now,
$$V(\hat{Y}_{srs}) = \frac{N^2(N-n)}{Nn} \cdot \frac{1}{N-1} \cdot \beta \cdot \frac{N.(N^2-1)}{12}$$
$$= N^2\beta^2 \cdot \frac{(k-1)(nk+1)}{12}, \text{ using } N = nk.$$

Again, $$\sum_{r=1}^{k} \left[\hat{Y}_r - Y\right]^2 = \sum_{r=1}^{k} N^2\beta^2\left[r - \frac{k+1}{2}\right]^2 = \frac{N^2\beta^2 \cdot k(k^2-1)}{12}$$

Therefore, $$V(\hat{Y}_{st}) = \frac{N^2\beta^2(k^2-1)}{12}.$$

Note that, $$\frac{V(\hat{Y}_{st})}{V(\hat{Y}_{srs})} = \frac{k^2-1}{(k-1)(nk+1)} = \frac{k+1}{nk+1} < 1 \ \forall n > 1.$$

Hence, the linear systematic sampling is more precise than SRS in the presence of linear trend.

Interpenetraing sub-sampling technique for unbiased variance estimation in linear systematic sampling :—— [CU]

This technique, particularly useful for the study of correlated errors, was proposed by Mahalanobis (1946). To present it in the simplest terms, a random sample of n units is divided at random into K_1 sub-samples, each subsample containing $m = \frac{n}{K_1}$ units. The field work and processing of the sample are planned so that there is no correlation between the errors of measurement of any two units in different subsamples. For instance, suppose that the correlation with which we have to deal arises solely from biases of the interviewers. If each of K_1 interviewers is assigned to a different subsample and if there is no correlation between errors of measurement for different interviewers, we have an example of the technique:

Consider the mathematical model for errors of measurement :

Let y_{ijr} be the value obtained in the r^{th} repetition of the j^{th} member within the i^{th} subsample (interviewer). Then

$y_{ijr} = \mu_{ij}' + d_{ijr}$, where μ_{ij}' is the true mean of the unit and d_{ijr} is the response deviation on the unit or the fluctuating component of the measurement error.

From the sample results, we can compute an ANOVA table :—

ANOVA Table [On a single unit basis]

Source of variation	d.f.	m.s.
Between interviewers (subsamples)	$K_1 - 1$	$s_b^2 = \frac{m}{K_1-1}\sum_{i=1}^{K_1}(\bar{y}_{ir} - \bar{y}_r)^2$
Within subsamples (interviewers)	$K_1(m-1)$	$s_w^2 = \frac{1}{K_1(m-1)}\sum_{i=1}^{K}\sum_{j=1}^{m}(y_{ijr} - \bar{y}_{ir})^2$

Total $= K_1 m - 1$

Table gives the important results:- $\frac{s_b^2}{n}$ is an unbiased estimator of $\frac{1}{n}E(s_b^2) = V(\bar{y}_r)$. Thus interpenetrating subsamples provide an estimator of $V(\bar{y}_r)$ that takes proper account of both the simple response variance and correlated component.

Note that $\frac{s_b^2}{n} = \frac{m}{n(k_1-1)}\sum_{i=1}^{k_1}(\bar{y}_{in}-\bar{y}_n)^2 = \frac{1}{k_1(k_1-1)}\sum_{i=1}^{k_1}(y_{in}-\bar{y}_n)^2.$

Here m subsamples are interpenetrating in the sense that each is a probability sample over the population.

In linear systematic sampling, $V(\hat{Y}_{LSS}) = V(\hat{Y}_n) = V(N\bar{y}_n)$
$= N^2 V(\bar{y}_n)$.

Hence, by interpenetrating subsamples, an estimator of $V(\hat{Y}_{LSS})$ is

$$v(\hat{Y}_{LSS}) = N^2 \cdot \frac{s_b^2}{n} = N \cdot k \cdot s_b^2$$
$$= N^2 \cdot \frac{1}{k_1(k_1-1)}\sum_{i=1}^{k_1}(\bar{y}_{in}-\bar{y}_n)^2.$$

Problem:- What are the advantages of systematic sampling over simple random sampling. [C.U.]

Answer:- The apparent advantages of systematic sampling over simple random sampling are the following:

(a) It is much easier and quicker to draw a systematic sample and the work may be done by laymen.

(b) Intuitively, systematic sampling seems likely to give more precise estimates than simple random sampling. For example, the method of linear systematic sampling stratifies the popln into n strata of k units each and one unit is selected from each stratum. Moreover, systematic sampling yields a sample which is evenly spread over the entire population. Some of the practical situations where systematic sampling has been found very usual are given below

(i) The selection of every kth strip in forest survey for estimation of timber.

(ii) The selection of every kth village in rural surveys.

Because of its operational convenience, in such situations systematic sampling is better preffered than that of SRS.

Problem:- What are the major disadvantages of Systematic sampling. [CU]

Answer:- 1. The main disadvantage of systematic sampling is that systematic samples are not in general random samples.

2. If N is not a multiple of n, then

(i) the actual sample size is different from that required, and

(ii) sample mean is not an unbiased estimate of the popln. mean.

These disadvantages can be overcome by adopting CSS.

Problem:- In LSS, show that the estimator of the popln. mean is sample mean. Hence find its variance. [CU] [ISS '10]

Sol. If x_{ij} denotes the value of x for the jth unit in the ith group [$i = 1, 2, \ldots, k$ and $j = i+k, i+2k, \ldots\ldots, i+(n-1)k$]

$$E(\bar{x}_{sy}) = \sum_{i=1}^{k} x_{io}/k, \quad \text{if } x_{10}, x_{20}, \ldots, x_{k0} \text{ are the } k \text{ possible estimates.}$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}/nk = \bar{X} \text{ (population mean).}$$

$$Var(\bar{x}_{sy}) = \sum_{i=1}^{k}(x_{io}-\bar{X})^2/k = \frac{k-1}{k} S_c^2 \text{ (say), where } S_c^2 = \frac{1}{k-1}\sum_{i=1}^{k}(x_{io}-\bar{X})^2$$

The variance, however, can't be unbiasedly estimated from a single sample. A way out is interpenetrating samples method.

Cluster Sampling

• Write a short note on cluster sampling.

ANS:- **Cluster Sampling:** — Several references have been made to surveys in which the sampling unit consists of a group or cluster of smaller units that we have called elements or sub-units. There are two main reasons for the widespread application of cluster sampling:

(i) It is found in many surveys that no reliable list of elements in the popln. is available and that it would be prohibitively expensive to construct such a list.

(ii) Even if such a list existed, it would not be economical to base the enquiry on a SRS of persons because this would require interviewers to visit almost every commune in the country and resource do not permit it.

For example, a simple random sample of 600 houses covers a town more evenly than 20 city blocks containing an average of 30 houses a piece. But greater field costs are incurred in locating 600 houses and in travel between them than in locating 20 blocks and visiting all the houses in these blocks. Though, for a given size of sample, a small unit usually gives more precise results than a large unit, but all these considerations point to the need of selecting larger units or clusters, rather than elements directly from the population.

A simple cluster sampling plan is a sampling plan in which (a) the elementary units of the popln. to be sampled are grouped into clusters, such that each elementary unit is associated with one and only one cluster; (b) a sample is drawn by using the clusters as sampling units and selecting a simple random sample of the clusters, the clusters are referred to as primary sampling units or first stage units (i.e. psu or fsu's)

(A) Single-Stage Cluster Sampling:- No new principles are involved in making estimates when a probability sample of clusters has been taken and each sample cluster is enumerated completely. A problem to be considered is the optimum size of the cluster. This will naturally depend upon the cost of the collecting information from clusters of different size and the resulting variance.

Assume that the popln. contains N clusters $(U_1, \ldots, U_N)$ each containing M elements. The average of y per cluster is $\bar{Y} = \sum_{i=1}^{N} Y_i / N = \sum_{i=1}^{N}\sum_{j=1}^{M} Y_{ij} / N$ and the average per element is $\bar{Y}_e = \sum_i \sum_j Y_{ij}/NM = \bar{Y}/M$; where Y_{ij} be the y value for the j^{th} element within the i^{th} cluster.

Theorem:- In SRSWOR of n clusters, each containing M elements, from a population of N clusters, an unbiased estimate of the popln. total Y is given by $\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i = \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij}$.

and $V(\hat{Y}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\cdot\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2$.

Proof:- Under SRSWOR, $E(\hat{Y}) = N \cdot E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = N.\bar{Y} = Y$.

and
$$
\begin{aligned}
Var(\hat{Y}) &= Var\left(N\cdot\frac{1}{n}\sum_{i=1}^{n} y_i\right) \\
&= N^2\, Var\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2 \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}\left(\sum_{j=1}^{M} Y_{ij}-M\bar{Y}_e\right)^2 \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left\{\sum_i\sum_j (Y_{ij}-\bar{Y}_e)^2 + \sum_i\sum_j\sum_{k(\neq j)}(Y_{ij}-\bar{Y}_e)(Y_{ik}-\bar{Y}_e)\right\} \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\cdot\left\{(NM-1)S_y^2 + (M-1)(NM-1)\rho S_y^2\right\} \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\cdot(MN-1)S_y^2\left\{1+(M-1)\rho\right\}
\end{aligned}
$$

Corollary:- For estimating $\bar{Y}_e$, the average per element, an unbiased estimator is $\bar{y}_e = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$ and the variance is

$$V(\bar{y}_e) = \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i-\bar{Y}_e)^2.$$

Remark:- If, instead of sampling in clusters, a SRSWOR of nM elements is taken directly from the popln., then the estimator is $\hat{Y}' = N\sum_i\sum_j Y_{ij}/n$ and $V(\hat{Y}') = \frac{(NM)^2}{nM}\left(1-\frac{nM}{NM}\right)S_y^2$
$$= \frac{N}{n}\left(1-\frac{n}{N}\right)M S_y^2.$$

We already have, $V(\hat{Y}) \simeq \frac{N^2}{n}\left(1-\frac{n}{N}\right)M S_y^2\left\{1+(M-1)\rho\right\}$

Hence, $\frac{V(\hat{Y})}{V(\hat{Y}')} \cong \left\{1+\overline{M-1}\rho\right\}$

Generally ρ is found to be positive since clusters are usually formed by putting together geographically contiguous farms, stores, families etc. If $\rho<0$, both cost and the variance point to the use of clusters.

Here, $S_y^2 = \frac{1}{(NM-1)}\sum_i\sum_j (Y_{ij}-\bar{Y}_e)^2$

$$\rho = \frac{E(Y_{ij}-\bar{Y}_e)(Y_{ik}-\bar{Y}_e)}{E(Y_{ij}-\bar{Y}_e)^2} = \frac{2\sum_i\sum_{j<k}(Y_{ij}-\bar{Y}_e)(Y_{ik}-\bar{Y}_e)}{(M-1)(NM-1)S_y^2}.$$

(B) Two stage cluster sampling or subsampling :— [CU]

Suppose that a sample of n clusters has been selected from a popln. containing N clusters. If elementary units within a selected cluster give similar result, it seems to be uneconomical to measure them all. A common practice is to select and measure a sample of elementary units from any selected cluster. This technique is called Subsampling, since the cluster selected is not measured completely but is itself sampled or two-stage sampling, because the sample is taken in two steps — first step is to select a sample of cluster (often called primary sampling units (p.s.u) or first stage units (f.s.u.)) and the second is to select a sample of elementary units from each selected clusters (second stage units (s.s.u.)). Here we consider the simplest case in which every cluster contains M elementary units, of which m are chosen from each selected cluster.

Ques:— [CU]

What are the advantages of two-stage sampling? From a population with N first stage units (f.s.u.) each containing M second stage units (s.s.u.) a random sample of n f.s.u.'s is drawn and from each selected f.s.u. a random sample of m s.s.u.'s is drawn. Show that the sample mean per s.s.u. is unbiased for estimating the popln. mean. Derive the variance of the estimator and an unbiased estimator of the variance.

(OR)

~~Solution:~~ [CU].

For a two-stage sampling, where the first-stage units are of equal sizes, obtain an unbiased estimator of the popln. mean. Also obtain an expression for the variance of the estimator. How will you estimate unbiasedly the variance of the estimator from the sample? Assume SRSWOR at each stage.

Solution:—

Principal advantages of two-stage sampling:— It is more flexible than one-stage sampling. It reduces to one-stage sampling when $m=M$, but, unless this is the best choice of m, we have the opportunity of taking some smaller value that appears more efficient. As usual the issue reduces to a balance between statistical precision and cost. When the elementary units (s.s.u.'s) in the same cluster agree very closely, considerations of precision suggest a small value of m. On the other hand, it is sometimes almost as cheap to measure the whole of a cluster as the subsample is, for example, when the cluster is a household and a single respondent can give accurate data about all members of the household.

Two-stage sampling with equal-size p.s.u.'s and subsampling with equal-sized s.s.u.'s:—

Here all p.s.u.'s have the same number M of second-stage units and a constant number 'm' of them are sampled from selected p.s.u.

The following notations are used:

y_{ij} = value obtained for the j^{th} subunit in the i^{th} primary unit

$\bar{y}_i = \sum_{j=1}^{m} y_{ij}/m$ = sample mean per sub-unit.

$\bar{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$

$Y_i = \sum_{j=1}^{M} Y_{ij}$ = total over-all subunits in the i^{th} p.s.u. (or cluster).

$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i - \bar{Y})^2$ = Variance among primary unit means or variance between the p.s.u.s.

$S_w^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij}-\bar{Y}_i)^2/N(M-1)$ = Variance among subunit within primary units.

Theorem:— If n units (p.s.u's) and the m subunits (s.s.u.'s) from each selected p.s.u.'s are selected by SRSWOR, $\hat{Y} = \frac{NM}{n}\sum_{i=1}^{n}\bar{y}_i$ is an unbiased estimate of Y with variance

$$V(\hat{Y}) = \frac{M^2N^2}{n}\left\{\left(1-\frac{n}{N}\right)S_b^2 + \left(1-\frac{m}{M}\right)\frac{S_w^2}{m}\right\}.$$

Proof:— With SRS at both stages,

$$E(\hat{Y}) = E_1E_2(\hat{Y}) = E_1\left[\frac{NM}{n}\sum_{i=1}^{n}E_2(\bar{y}_i)\right] = E_1\left[\frac{NM}{n}\sum_{i=1}^{n}\bar{Y}_i\right]$$
$$= NE_1\left[\frac{1}{n}\sum_{i=1}^{n}M_i\bar{Y}_i\right]$$
$$= NE_1\left[\frac{1}{n}\sum_{i=1}^{n}Y_i\right]$$
$$= N\bar{Y} = Y.$$

Again, we have

$$V(\hat{y}) = V_1(E_2(\hat{Y})) + E_1(V_2(\hat{Y}))$$

Thus, $E_2(\hat{Y}) = N.\frac{1}{n}\sum_{i=1}^{n}Y_i$, $V_1[E_2(\hat{Y})] = M^2N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2$,

$V_2(\hat{Y}) = \frac{N^2}{n^2}\sum_{i=1}^{n}M^2\left(\frac{1}{m}-\frac{1}{M}\right)S_{wi}^2$, where $S_{wi}^2 = \frac{1}{M-1}\sum_{j=1}^{M}(y_{ij}-\bar{Y}_i)^2$

is the variance among subunits for the i^{th} primary unit.

[Here E_2 and V_2 represent the conditional expectation and variance over all selections of sizes of m from the p.s.u.'s which are fixed (like strata); E_1 and V_1 denote similarly the expectation and variance over all possible samples of n p.s.u.'s from the N p.s.u.'s.]

Now,
$$E_1[V_2(\hat{Y})] = \frac{N^2}{n^2}\sum_{i=1}^{n} M^2\left(\frac{1}{m}-\frac{1}{M}\right)E\left(s_{wi}^2\right)$$
$$= \frac{N^2}{n}\cdot M^2\left(\frac{1}{m}-\frac{1}{M}\right)E\left(\frac{1}{n}\sum_{i=1}^{n} s_{wi}^2\right)$$
$$= \frac{N^2}{n}\cdot\frac{M(M-m)}{m}\cdot\left(\frac{\sum_{i=1}^{N} S_{wi}^2}{N}\right)$$
$$= \frac{N^2}{n}\cdot\frac{M^2}{m}\left(1-\frac{m}{M}\right)\cdot S_w^2 .$$

Hence,
$$V(\hat{Y}) = \frac{N^2}{n}\left\{M^2\left(1-\frac{n}{N}\right)S_b^2 + \frac{M^2}{m}\left(1-\frac{m}{M}\right)S_w^2\right\}$$
$$= \frac{(MN)^2}{n}\left\{\left(1-\frac{n}{N}\right)S_b^2 + \frac{1}{m}\left(1-\frac{m}{M}\right)S_w^2\right\}.$$

If $f_1=\frac{n}{N}$, $f_2=\frac{m}{M}$ are sampling fractions in the first and second stages, an alternative form of $V(\hat{Y})$ is
$$V(\hat{Y}) = \frac{N^2}{n}\left\{M^2(1-f_1)S_b^2 + \frac{M^2}{m}(1-f_2)S_w^2\right\}$$
$$= \frac{(MN)^2}{n}\left\{(1-f_1)S_b^2 + \frac{(1-f_2)}{m}S_w^2\right\}.$$

<u>Cor:-</u> $\bar{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$ is an unbiased estimator.

Estimation of Variance:-

<u>Theorem:-</u> Under the conditions of above theorem, an unbiased estimator of $V(\hat{Y})$ is
$$v(\hat{Y}) = M^2N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_b^2 + NM^2\left(\frac{1}{m}-\frac{1}{M}\right)s_w^2$$
i.e.
$$v(\hat{Y}) = M^2N^2\left\{\frac{1-f_1}{n}s_b^2 + \frac{f_1(1-f_2)}{mn}s_w^2\right\},$$
where,
$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{y}_i-\bar{\bar{y}}\right)^2,\quad s_w^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}-\bar{y}_i\right)^2/n(m-1).$$

<u>Proof:-</u>
$$(n-1)s_b^2 = \sum_{i=1}^{n}\bar{y}_i^2 - n\bar{\bar{y}}^2$$

Hence,
$$E[(n-1)s_b^2] = (n-1)E_1[E_2(s_b^2)]$$
$$= E_1\left[\sum_{i=1}^{n}E_2(\bar{y}_i^2) - nE_2(\bar{\bar{y}}^2)\right]$$
$$= E_1\left[\sum_{i=1}^{n}\left\{V_2(\bar{y}_i)+E^2(\bar{y}_i)\right\} - n\left\{V_2(\bar{\bar{y}}) + E_2^2(\bar{\bar{y}})\right\}\right]$$
$$= E_1\left[\sum_{i=1}^{n}\left(\bar{Y}_i-\bar{\bar{Y}}_n\right)^2 + \frac{(n-1)\left(1-\frac{m}{M}\right)}{mn}\sum_{i=1}^{n}S_{wi}^2\right],$$
where $\bar{\bar{Y}}_n = \sum_{i=1}^{n}\bar{Y}_i/n$
$$= (n-1)\left\{S_b^2 + \frac{\left(1-\frac{m}{M}\right)}{mn}\cdot n\sum_{i=1}^{N}S_{wi}^2\right\},$$
taking expectation w.r.t. the first stage simple random sampling.

Hence, $E\left[\left(\frac{1}{n}-\frac{1}{N}\right)s_b^2\right] = \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{\left(1-\frac{n}{N}\right)\left(1-\frac{m}{M}\right)S_W^2}{mn}$

Here, we also have $E(s_W^2) = S_W^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_i)^2$

Therefore,

$$E(v(\hat{Y})) = M^2N^2\left[\left\{\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{\left(1-\frac{n}{N}\right)\left(1-\frac{m}{M}\right)S_W^2}{mn}\right\} + \frac{\frac{n}{N}\cdot\left(1-\frac{m}{M}\right)}{mn}\cdot S_W^2\right]$$

$$= M^2N^2\left\{\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{\left(1-\frac{m}{M}\right)}{mn}S_W^2\right\}$$

$$= M^2N^2\left\{\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)S_W^2\right\}$$

$$= M^2N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + N M^2\left(\frac{1}{m}-\frac{1}{M}\right)S_W^2.$$

Cor:- An unbiased estimator of $V(\bar{\bar{y}})$ is

$$v(\bar{\bar{y}}) = \frac{1-f_1}{n}s_b^2 + \frac{f_1(1-f_2)}{mn}s_\omega^2.$$

Ques:- Assume the cost function $C = c_1 n + c_2 n'$. Find the optimum values of n and n' for which the MSE is minimum subject to a fixed cost. [C.U.]

Solution:-

$$V_n(\bar{y}_{st}) = \frac{1}{n^2}\left(\frac{1}{n'}-\frac{1}{N}\right)S^2$$

For sufficiently large N, $\frac{n'}{N}$ is ignored, we get, $V_n(\bar{y}_{st}) \cong \frac{S^2}{n'n^2}$.

The optimum value of n and n' are obtained minimizing $V_n(\bar{y}_{st})$ subject to the given fixed cost: $C = c_1 n + c_2 n'$.

Using the method of Lagrange's multipliers, we minimise the functions

$$\phi(n,n') = V_n(\bar{y}_{st}) + \lambda(c_1 n + c_2 n' - C)$$

$$= \frac{S^2}{n'n^2} + \lambda(c_1 n + c_2 n' - C);\ \lambda \text{ being Lagrange's multiplier}$$

$$\frac{\partial\phi}{\partial n} = -\frac{2S^2}{n'n^3} + \lambda c_1 = 0 \Rightarrow \lambda = \frac{2S^2}{n'n^3c_1}$$

$$\frac{\partial\phi}{\partial n'} = -\frac{S^2}{n'^2n^2} + \lambda c_2 = 0 \Rightarrow \lambda = \frac{S^2}{n'^2n^2c_2}.$$

$$\therefore \frac{2S^2}{n'n^3c_1} = \frac{S^2}{n'^2n^2c_2} \Rightarrow c_1 = \frac{2n'c_2}{n}.$$

Substituting in $C = c_1 n + c_2 n'$,

$$\Rightarrow C = \frac{2n'c_2 n}{n} + c_2 n' \Rightarrow \boxed{n' = \frac{C}{3c_2}}$$

$$\therefore c_1 = \frac{2n'c_2}{n} \Rightarrow c_1 = \frac{2C\cdot\cancel{c_2}}{3\cancel{c_2}n} \Rightarrow \boxed{n = \frac{2C}{3c_1}}$$ (ANSWER)

CU **Ques:-** Explain the differences between the methods of cluster sampling and stratified sampling. In a given situation, when will you prefer one method over the other?

Solution:-

CU **Ques:-** Consider a population of eight households, say, a, b, c, d, e, f, g, h. Determine the possible samples of size 3 using circular systematic sampling.

Sol. If a sample of size 2 is to be chosen, then $K = N/n$ being 4 in this case, the possible samples in linear systematic sampling will be ae, bf, cg, dh. However, if we like to have a sample of size 3, then the sampling interval N/n is $2\frac{2}{3}$, a fractional number, and we have to go in for circular systematic sampling. Since the integer K nearest to $2\frac{2}{3}$ is 3, the possible systematic samples will be adg, beh, cfa, dgb, ehc, fad, gbe and hcf.

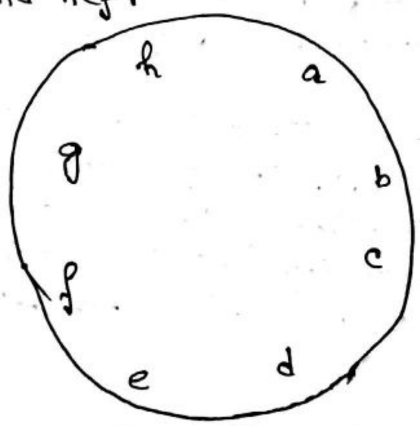

Fig:- The sampling units a, b, ..., h are arranged in a cyclical fashion.

(CU)

Ques:- What are random sampling numbers? State the tests available to test their randomness colaborating any one of them. Using a coin how will you select one unit from a popln of 'N' units with selection probability $1/N$, when the coin is (a) biased (b) unbiased?

Solution:-

Definition of a random sampling number series:-

A random sampling number series is an arrangement, which may be looked upon either as linear or as rectangular, in which each place has been filled in with one of the digits 0, 1, ..., 9. The digit occupying any place is selected at random from these ten digits and independently of the digits occuring in other positions.

Advantages of random sampling numbers:-

If we use random sampling numbers for drawing random samples we need not construct a miniature population. Also, the numbering of the sampling units can be done in any convenient manner.

Secondly, randomisation of the numbers being done once for all. Any part of the series can be used for a random sample of numbers and the problem is simply to interpret these numbers in terms of individuals of the population.

Lastly, a random sampling number series can be used for any enumerable population, so that a series of random numbers has a vide range of application.

Tests applied to random sampling number series:- To examine whether any series is really random, the following tests may be applied.

(a) **Frequency test:-** Here the observed frequencies of the ten digits from 0 to 9 are obtained and tested against the expected frequencies on the basis of the hypothesis that the set of numbers is random, according to which each digit has the prob. $1/10$ to occur in any position of the series. The appropriate statistic is a Pearsonian χ^2 with $df = 9$.

(b) **Serial Test:-** Here the series is considered to be composed of two-digited numbers. The frequencies of all the 100 possible numbers, viz. 00, 01, ..., 99, are obtained and the hypothesis of randomness, according to which each pair has the probability $1/100$, is tested by using the appropriate Pearsonian χ^2 with $df = 99$.

(CU)

Ques:- Write a brief note on the nature and the coverage of work done by NSSO.

Sol. → The National Sample Survey is the biggest set of sample surveys in India being conducted by the Govt. of India. The NSS was initiated in 1950 to conduct sampling require with a view in providing the Govt. and other organisations with socio-economic data which can be used for planning for national development and for research purposes.

The Central statistical organisation (CSO) is responsible for deciding upon the coverage of the survey and the methodology to be used. The major portion of the field work is conducted by National Sample Survey Organisation (NSSO), Government of India. The technical work relating to the NSS, the processing and analysis of data and the final reports preparation has been taken over by the NSSO. The important functions of NSSO are: —

(1) Socio-economic Survey:- Socio-economic survey is the main function of NSSO. It conduct multipurpose sample survey related to land utilization, agricultural production, genetic characteristics and so on. It also conduct some surveys on employment status and these all datas are used by planning commission and other ministries of Govt. of India and other private agencies.

(2) Crop-Estimate Survey:- NSSO extents their help to improve the agricultural statistics by providing standard technique for data collection to both state & central Govt. As a result of this the data collection will be more uniform and comparable and these surveys are related to the crops like oil seeds, vegetables and etc and their estimates.

(3) Industrial Survey:- NSSO conducts anual surveys under the act of collection of statistics 1953, the surveys are related with the fact that employment status, salary and wages, raw materials of industrial production, capital structure of industries.

(4) Price-Statistics Survey:- NSSO collects price statistics regularly on the urban and rural basis seperately.

(a) The price index number conducted for the non-working class of people by using prices of 250 commodities in 59 places.

(b) Price index number conducted for the rural agricultural workers is based on price statistics of 603 villages.