2.4. SAMPLING TECHNIQUES

UNIT-I

BASIC CONCEPTS

Population: The set of all possible observation under study is called population. It is denoted by (N).

Sample: It is the subset of a population that has been collected through data collection. It is denoted by (n).

Statistic: Any function of the sample values is called statistic.

Sample Space: The totality all sample points consistent with the method of sample adapter will be called sample space. For example: $S = \{H, T\}$ When tossing a coin.

Event: A subset of a sample space is called an event.

Estimator: An estimator is a random sample and may take different values from sample to sample.

Parameter: Any population constants are called parameter.

Unbiased estimator: An estimator (t) is set to be unbiased estimator for the parameter (θ), if $E(t) = \theta$ or $E(t) - \theta = 0$. Similarly $E(t) - \theta = B(t)$ is known as biased estimator.

Standard Error: The positive square root of variance is called standard error of the estimator.

Finite Population: A population said to be finite if it has countable number of values. For example: number of students in the M.S.University.

Infinite Population: A population said to be infinite if it has uncountable number of values. For example: Number of stars in the sky.

Sampling Unit: The contribution of population which all the individuals to the sample from the population that cannot be further subdivided for the purpose of sampling at time is called sampling unit. For example, to know that the average income of a family, the head of family is called sampling unit.

Sampling Frame: we need to frame the structure of the survey for adapting any sampling procedure, it is essential to have a list or map to identify each sampling unit by a number, such a list or map is called sampling frame. For example, If list of voters in a particular place.

NEED FOR SAMPLING

The sampling methods have been extensively used for a variety of purposes and in great diversity of situations. In practice it may not be possible to collect the information on all units of a population due to various reasons such as

- ❖ Lack of resources in terms of money, personnel and equipment.
- ❖ The experimentation may be destructive in nature. Eg- finding out the germination percentage of seed material or in evaluating the efficiency of an insecticide. This experimentation is destructive.
- ❖ The data may be was useful if they are not collected within a time limit. The census survey will take longer time as compared to the sample survey. Hence for getting quick results sampling is preferred. Moreover a sample survey will be less costly than complete enumeration.
- ❖ Sampling remains the only way when population contains infinitely many number of units.
- ❖ Greater accuracy.

CENSUS AND SAMPLE SURVEY:

The total count of all units of the population for a certain characteristic is known as complete enumeration or census survey. The money, man power and time required for carrying out complete enumeration is generally large and there are many situations with limited means where the complete enumeration will not be possible.

When only a part of the population is selected denoted as sample and examine it is called sample enumeration or sample survey.

Limitations of Sampling vs. Census:

1) **Less time:** There is considerable saving in time and labor since only a part of the population has to be examined. The sampling results can we obtain more much rapidly and it can analysis such faster since relatively and process.

2) **Reduced cost of the survey:** Sampling usually results in reduction of cost, in terms of money and in terms time. All though the amount of labor and expenses are invalid in collecting information. All generally greater per unit of the sample then complete enumeration the total cost of the sample survey is the expected to be much smaller than that of complete census of since in most of the cases our resources or limited in terms of money and the time with in which the results of the survey should be obtain it is usually imperative to the sampling rather than complete to sampling rather than complete enumerations.

3) **Greater accuracy of results:** The results of the sample are usually much more reliable than those obtained from a complete census due to the following reasons.

a. It is always possible to determine the extent of the sampling errors.

b. Non-sampling errors due to factors such as training of the field workers measuring and recording, observations, location of units incompetents of returns biased due to interviews etc. There are likely to be of a serious nature in complete census than in a sample survey. Non sampling errors can be controlled more effectively by employing more qualified and better trained personal better supervision and better equipment for processing and analysis of relatively limited data. Moreover, it is easier to guard against incomplete and inaccurate returns. There can be a follow up in case of non-response or incomplete response effective control of non-sampling errors in the estimations due to sampling as such sophisticated statistical techniques can be employed to obtain relatively more reliable results.

4) **Greater scope:** Sample survey generally has greater scope as compared with complete census the complete enumeration is impracticable rather inconceivable if the survey requires highly trained personal and more sophisticated equipment for the collection and analysis of the data. Since sample survey saves in time and money it is possible to have a through and intensive enquiry because detailed information can be obtained from a small group of respondents.

5) If the population is too large, for example of trees in a jungle we are left with no way but to resort to sampling.

6) If testing is destructive i.e., if the quality of an article can be determined only of an article in the process of testing as for example:

   i) Testing the quality of milk or chemical salt by analysis.

   ii) Testing the breaking strength of chalks.

   iii) Testing of crackers and explosives.

   iii) Testing the life of an electric tube or bulb etc.

   Complete enumeration is impractical and sampling techniques is the only method to be used in such cases.

7) If the population is hypothetical for example while tossing a coin, the process may continue indefinitely. Sampling method is the only scientific method of estimating parameters of the universe/population.

THE PRINCIPLE STEPS IN A SAMPLE SURVEY:

1) Objectives of the survey:

The first step is to define clear and concrete terms of the objectives of the survey. The sponsors of the survey should take care that these objectives are commensurate with the available resources in terms of money, man power and the time limit required for the availability of the results of the survey.

2) Defining the population to be sampled:

The population that is the aggregate of objects from which sample is chosen should be defined in clear and unambiguous terms.

3) The frame and sampling units:

The population must be capable of division into sampling units for purpose of the sample selection. The sampling units must cover the entire population and must be distinct unambiguous and not overlapping in the sense that every element of the population belongs to one and only one sampling units

In order to cover the population decided upon there should be some list map or other acceptable material called the frame to serve as a guide for the population to be covered.

4) Data to be collected:

The data should be collected keeping in view the objectives of the survey. We should not have the tendency to collect too many data some of which are never subsequently examined and analyzed.

5) The questionnaire or schedule:

Having decided about the type of the data to be collected the next important part of the sample survey is the construction of the questionnaire or schedule of the enquires which requires skill special technique as well as familiarity with the subject matter under study. The question should be clear, brief collaborative non offending courteous in tone unambiguous and to the point, so that not much scope of guessing is the left on the part of the respondent or interviewer suitable and detailed instruction for filling up the questionnaire or schedule should also prepared.

6) Method of collecting information:

i) Interview method:

In this method the investigator goes from house to house and interviews the individuals personally. He asks the questions one by one and fills up the schedule on the basis of the information guess by the individuals.

ii) Mailed questionnaire method:

In this method the questionnaire is mailed to the individuals are required to the individuals who are required to fill up and return it duly completed.

7) Non respondent:

Quite often the data cannot be collected for the sampled units. This incompleteness is called non response which obviously tends to change the results. In such cases of response should be handled with caution in order to draw unbiased and valid conclusions.

8) Selection of proper sampling designs:

The size of the sample (n) the procedure of selection and the estimation of the population parameters along with their margins of uncertainty are some of the important statistical problems that should receive that careful attention.

A number of designs for the selection of a sample are available and a judicious selection will guarantee good and available estimation.

9) Organization of field work:

It is absolutely essential that the person should be thoroughly trained in locating the sample units, recording the measurements the methods of the collection of required data before starting the field work. The success of a survey to a great extent depends upon the reliable field work. It is very necessary to make provision for adequate supervisory staff for inspection after field work.

10) Summary and analysis of the data:

a) *Scrutiny and editing of the data:*

An initial quality check should be carried out by the supervisory staff while the investigators are in the field.

b) Tabulation of the data:

Before carrying out the tabulation of the data we must decide about the procedure for the quality of the data. For the large scale survey, mission tabulation will obliviously be much quicker and economical.

c) Statistical analysis:

Statistical analysis should be made only after the data has been properly scrutinized, edited and tabulated. Different method of estimation may be available for the same data, appropriate formulae should be used to provide final estimates of the required information.

d) Reporting and conclusions:

Finally, the report incorporating detailed statement of the different stages of the survey should be prepared.

11) Information gained for future surveys:

Any complete survey is helpful in providing a note of caution and taking lesson from it for designing future surveys. The information gained from any completed sample in the form of the data regarding means, standard deviation and the nature of the variability of the principle of the measurements tougher sampling.

PRINCIPLE OF SAMPLING SURVEY:

1) Principle of statistical regularity:

This principle as its origin in the mathematical theory of probability. According to the king, the law of statistical regularity lays down moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large principle stress the desirability and the importance of selecting the sample at random so that each every unit in the population has an equal chances of being selecting the sample.

2) Principle of validity:

By the validity of a sample design we mean that it should enable us to obtain valid test and estimates about the parameters of the population. The samples obtain by the techniques of the probability sampling, satisfy this principle.

3) Principle of the optimization:

The principle improves upon obtaining optimum results in terms of efficiency and the cost of the design with the resources at our disposal. The reciprocal of the sampling variance of an estimate provides a measure of its efficiency while a measure of the cost of the design is provided by the total expenses incurred in terms money and man hour.

i) Achieving a given level of efficiency at minimum cost 1 and

ii) Obtaining maximum possible efficiency with given level of cost.

LIMITATION OF SAMPLING:

Advantage:

i) The sampling units are drawn in scientific manner.

ii) Appropriate sampling techniques are used and

iii) The sample size is adequate.

Disadvantage:

i) Proper care should be taken in the planning and execution of the sample, survey otherwise the results obtained might be inaccurate and misleading.

ii) Sampling theory requires services of the trained and qualified person and sophisticated equipment for its planning execution and analysis. In the absence of these results, the sample survey are not trust worthy.

iii) however, if the information required about each and every unit of the universe there is no way but to resort to complete enumeration. If time and money are not important factors or if the universe is not too large, a complete census may be better than any sampling.

TYPES OF SAMPLING:

The technique or method of selection is fundamental importance in the theory of samplings and usually depends upon the nature of the data and the types of the enquiry the procedure a selecting a sample may be broadly classified under the following three heads.

i) Subjective and or judgments sampling.

ii) Probability sampling

iii) Mixed sampling

1) Subjective (or purposive or judgment) sampling:

In this type of sampling, the sample selected is with definite purpose in view and the choice of sampling units depends greatly on the discretion and the judgment of the investigated. These samplings suffer from the drawback of favoritism and nepotism depending upon believes and prejudices of the investigation and thus does not give a representative sample of the population. This sampling method is seldom used and cannot be recommended for general use since it is biased due to element of subjectiveness or the part of the investigator. However, if the investigator is experience and skilled and the sampling is carefully applied, then judgment sample may yield valuable results.

2) Probability sampling:

Probability sampling is a scientific methods of selecting samples according to some law of chance in which each unit in the population has some definite pre assigned probability of being selected in the sample the different types of probability sampling are,

i) Where each unit has equal chances of being selected.

ii) Sampling units have different probability of being selected.

iv) Probability of selection of unit is proportional to the sample size.

3) Mixed sampling:

If the samples are selected partly according to some laws of chance and partly according to fixed sampling rule, they are termed as mixed samples and the techniques of selecting such sample is known as mixed sampling.

The different types of sampling given above have a number of variations. Some of which may be listed below.

i) Simple random sampling

ii) Stratified random sampling
iii) Systemic sampling
iv) Multi stage sampling
v) Quasi random sampling
vi) Area sampling
vii) Simple cluster sampling
viii) Multi stage cluster sampling
ix) Quota sampling.

QUESTIONNAIRE AND SCHEDULE:

QUESTIONNAIRE:

Questionnaire consists of a list questions parading to the enquiry is preferred to have a blank space for answer. This questionnaire is sent respondence who are expected to write the answers in the blank space. A covering letter is also sent along with the questionnaire requesting the respondence to extent their full cooperation by giving the correct replace and retaining the questionnaire duly field in time.

Merits:

a. Questionnaire method is economical.
b. it can we widely used when the area of investing is large.
c. it saves money labor and time.
d. error in the investing isvery small because the information is explain directly to the respondence.

Demerits:

i. In this method there is no direct connection between the investigator and the respondent. Therefore we can sure about the accuracy and reliability of the information.

ii. This method is suitable only for literate people in many countries illiterate people cannot and replay the questionnaire.

iii. There is long delay in receiving questionnaire duly field in time.

iv. People may not give correct answer, thus one is lead to false conclusion.

v. Sometimes the information may not will to gives the return answer.

SCHEDULE:

It is most widely used method of collection of primary data. The number of enumeration are selected and trained, they all provide with standard questionnaire specific training are instructions are given to them for filling of schedule each enumerates will be charge of a certain area the investigator goes to respondents along with the questionnaire and gets answer to the question in schedule and records their answers he explain clearly the adjective and purpose of the enquiry.

Merits:

i. This method is very useful in extensive enquires.

ii. It yield reliable accurate result because the enumerates or educated and trained.

iii. The scope of the enquiry can also be greatly enlarged.

iv. Even if the respondence it liberate this technique can be widely used.

v. As the enumerators personally obtain the information there is less Choice for the non-respondence.

Demerits:

i. This method is an expensive.
ii. This method is time consuming because the enumerators go personally to obtaine the information.
iii. Personal bias of the enumerators may lead to false conclusion.
iv. The quality of the collected data depends upon the personal qualities of the enumerators.

UNIT-II

SAMPLING AND NON-SAMPLING ERRORS:

The errors involved in the collection, processing and analysis data in a survey may be classified as,

i) Sampling error
ii) Non-sampling error

Sampling error:

The error which arises due to only a sample being to estimate the population being used to estimate the population parameter is tripped sample error or sampling fluctuations. The error is inherent and available in any and every sampling scheme. A sample with the smallest sampling error will always be consisted a good reprehensive of the population.

This error can be reduced by increasing the size of the sample unfair the degree of sampling error is inversely proportional to the square root of the sample size.

Non sampling error:

Besides sampling error, the sample estimate may be subject to the other errors which group together is term non sampling error. The main sources of non-sampling errors are,

1) Failure to measure some of the units in the selected sample.

2) Observational errors due to defective measurements technique.

3) Errors intituous editing coding and tabulation the results.

4) In practice the census survey results many suffer from non-sampling error. Although these may be free from sampling error. The non-sampling error is likely to increase with increase in sample size while sampling error decrease with increase in sample size.

$$\text{Sampling Error} \propto 1/\sqrt{n}$$

where n- sampling units.

Bias:

The different between estimator (t) and the parameter (θ) is called bias are error. An estimator (t) is said to be unbiased estimator for the parameter θ if $E(t) = \theta$, otherwise biased thus bias is given by $E(t) - \theta = B(t)$.

Mean square error (MSE):

A relative measure of bias is $\frac{B(t)}{\theta}$. The mean of square of the error taken from is called mean square error. Symbolically, $MSE(t) = E(t-\theta)^2$.

The sampling variance of (t) is defined by $V(t) = E(t) - [E(t)]^2$.

In terms of variance, $MSE(t) = V(t) + B^2(t)$

Since t is unbiased, $MSE(t) = V(t)$.

Sources and types of sampling of non-sampling errors:

The non-sampling errors occur at any one are more of this stages of the survey planning field work and tabulation of the survey data. These errors are badly classify errors follows,

Type 1: Non Response Error

Errors resulting from inadequate preparation.

Type 2: Responses Error

Error resulting in the stage of the collection or taking observations.

Type 3: Tabulation Error

Errors resulting from data processing.

Type 1: Non response error

These errors may be assigned mainly,

i) Due to the use of faulty frame of the sampling units,
ii) Biased method of the selection of units.
iii) In adequate schedule,

If the sampling frame is not abraded or old frame is use on account of economic or time saving device it may lead to bias as the targeted population is not enumerated. These a use of such frames may lead either to inclusion of some units not belonging to the population or to omission of units belonging to the population. Such procedure mat brings unknown bias. In some situation apart of sampled units may refuses to respond to the question or may be not at home at the time of interview. It may also lead to this type of error. It can be seen that the method will provide biased estimate. Some of the main sources assigned to these errors may be as follows:

i. Omission on duplication of units due to ambiguous defection of Local units or wrong identify of units and or in accurate and inconsistent o objectives.

ii. In accurate methods of interview or inappropriate schedule.

iii. Difficulty arising arrivers and Sourness on the part of respondents or faulty method of enumeration data collection.

Type 2: Response Error

These errors refer in general to the difference between the individual true value and the corresponding sample value irrespective of the resend of discrepancy. Sometimes there may be interaction between both of them and it may be inflate these errors. The measurements device or technique mat is defective and may cause observational errors may be assigned as under,

a. In adequate supervision and inspection of field staff.

b. Inadequate drained and experience field staffs.

c. Problems involved in data collection and other types of errors on the part of respondence.

Types 3: Tabulation Error

These errors can we assigned number of defective method, number of coding punching, tabulation etc., these method may referred according to the techniques' employed and eccumpents available for the data processing. To these errors bias due to estimation procedure may also include. This bias may be consisted as part of tabulation errors. The main sources to these errors may be assigned as follows.

i. In adequate scrutiny Pasic data.
ii. Errors in data processing abration such coding, punching, listing, verification, etc.
iii. Other errors committed admitted during publication presentation of results.

UNIT-III

SIMPLE RANDOM SAMPLING

A simplest and common most method of sampling is simple random sampling in which the sample drawn unit by unit equal probability of selection for each for unit at each drawn. A simple random sampling is method are selection ‘n’ units out of the population of size ‘N’ by giving equal probability to all units or a sampling procedure in which all possible combination of ‘n’ units that may be form from the population ‘N’ units have the some probability of selection.

SIMPLE RANDOM SAMPLING WITH REPLACEMENT (SRSWOR):

If units is selected and noted then return to the population before the next drawing is made and this procedure repeated ‘n’ times it gives raise to simple random sample of n units this processing is generally known as simple random sampling replacement.

SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT (SRSWOR):

In this procedure is repeated till n distinct units or selected and the reparation are ignored it called simple random sampling without replacement.

Theorem: 3.1

The probability of specified unit is being included in the sample is equal to $\frac{n}{N}$

Proof:

Since, the specified unit can be included in the sample size ‘n’ with ‘n’ mutually exclusive ways that is it can be selected in that the sample at the r^{th} draw $(r = 1,2,\ldots n)$ that is $P(E_r) = \frac{1}{N}$ $\forall$ r=1,2,3…n.

So that the probability that specify units selected in the sample = $\sum_{r=1}^{n} \frac{1}{N}$

$$= \frac{1}{N} + \frac{1}{N} + \frac{1}{N} .... + \frac{1}{N} \text{ (n times)}$$

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

Theorem: 3.2

The probability that specified unit of the population is being selected at any given draw is equal to the probability of its being selected at the first draw.

Proof:

Let E_r be the event that specified unit is selected at the r^{th} draw.

Therefore, $P(E_r)$ = (The probability that specified unit is not selected in any of the previous (r-1) draw) X (The probability that it is selected at the r^{th} with the condition that it is not selected in the previous (r-1) draw) that is,

$$P(E_r) = \frac{N-1}{N}\cdots\frac{N-2}{N-1}\cdots\frac{N-(r-1)}{N-(r-2)}\times\frac{1}{N-(r-1)}$$

$$= \frac{N-(r-1)}{N}\times\frac{1}{N-(r-1)}$$

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

Hence, $P(E_r) = P(E_1)$

Note: The above theorem leads the property of simple random sampling without replacement.

PROCEDURE FOR SIMPLE RANDOM SAMPLING:

Since the theory of sampling is based on the assumption of random sampling, the technique of random sampling is basic significance some of the procedure used for selecting a random sample are follows.

1. Lottery method
2. Random sampling tables method.

1. Lottery method:

This is method where a ticket/chit many be associated with each unit of population. Thus each sampling unit has its identification mark from identification 1 to N. the procedure of selecting an individual is simple. All the ticket/chits are placed in a container, drum or metallic spherical device in which a before each draw. Draw of tickets/chits may be continued until a sample of the required size is obtained.

This procedure of numbering units on tickets/chits and selecting one after reshuffling becomes cabpersons when the population size is large it may be lather difficult to achieve a through shuffling in practice human bias be accruing in this method.

2. Random sampling tables method:

A random number table is and arrangement of digits, in either a linear or rectangular pattern, where each position filled with on those digits. A table of random number is so constructed that all number 0,1,2…9 up a independent of each number some random number tables in common use are,

1. Tippet's random number tables.
2. Fisher and Yates tables.
3. Kendall and smith tables.
4. A Million Random digits

A practical methods of selecting of a random sample is to choose units one by one with the help of the table of random numbers. By considering two digits numbers we can obtain numbers from 00 to 99, all having same frequency. Similarly, three or more digits number may be obtain combining three or more rows or columns of these tables. The simplest way selecting a sample of the required size is by selecting a random number from 1 to N and then taking the unit bearing that numbers. The use random number is, therefore modify some of these modify procedure are

i. Remainder approach
ii. Quotient approach
iii. Independent choice of digits.

Remainder approach:

Let N be the r digits number and let its r digits highest n=multiple be N. A random number K is chosen from 1 to N and the unit with the serial equal to remainder obtain on dividing K by N is selected if the remainder is zero the last units is zero. For example,

i. Let N=123
ii. The highest three digit multiple of 123 is 984.
iii. For selecting unit one random number from 001 to 984 that as selected
iv. Get the random number be selected 287
v. Dividing 287 by 123 is equal to 41.
vi. Hence the unit within the serial number 41 is selected in the sample.

Quotient Approach:

Let N be the r digit number and r digits highest multiple be the N' such that $\frac{N'}{N} = q$. A random k is chosen from 0 to $(N'-1)$. Diving k by q the quotient r is obtained and the unit bearing serial $(r-1)$ is selected in the sample. For example,

i) N=16
ii) The highest two digit multiple of 16 is 96 i.e., N' =96, Hence q=6
iii) Let the two digit random number chosen be 65 which lies between 0 to 95
iv) Diving 65 by 6 is equal to 10
v) Hence the unit bearing number 10-1=9 is selected in the sample.

Independent choice of digits:

This method consists of the selection of two random numbers which are combined to form a one random number. One number is chosen according to the first digit and other according to the remaining digits of the population size. If the number is chosen is zero the last unit is chosen. But if the number is made up is greater than or equal to N, the number is rejected and the operation is repeated.

For Example: Select a random sample of 11 households from a list of 112 households in a village.

(i) By using the 3-digit random numbers given in column 1 to 3, 4 to 6, and so on of the random number table and rejecting numbers greater than 112 (also the number 000), we have for the sample bearing serial numbers 033, 051, 052, 099, 102, 081, 092, 013, 017, 076 and 079.

(ii) In the above procedure, a large number of random number is rejected. Hence, a commonly used device, i.e., remainder approach, is employed to avoid the rejection of such large numbers. The greatest three-digit multiple of 112 is 896. By using three-digit random numbers as above, the sample will comprise of households with serial numbers 086, 033, 049, 097, 051, 052, 066,107, 015, 106 and 020.

(iii) In case the quotient approach is applied, the 3digit multiple of 112 is 869 and 869/112=8. Using the same random numbers and dividing them by 8, we have the sample of households with list numbers 025, 004, 020, 026, 006, 006, 092, 041, 085, 027 and 086 with the replacement method and with list numbers 025, 004, 020, 026, 006, 092, 041, 085, 027, 086 and 042 without the replacement method.

ESTIMATION OF THE POPULATION PARAMETERS:

Notation and Terminology:

Let us consider a finite population of N units and let y be the character under condensation. The capital letters are used to describe the characteristics of the population whereas the small letter refers to the sample letters refers to sample observation. For example:

The N population units may be denoted by $U_1, U_2 \ldots U_N$ and the small n sample units will be denoted by $u_1, u_2 \ldots u_n$.

Let $Y_i \left(i = 1,2 \ldots N\right)$ be the value if the character for the i^{th} unit in the population and the corresponding small letters denoted the value of the character for the units selected in the sample. Then the refine population mean, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$

Sample mean, $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$

Sample mean $\bar{y}_n$ may also be written alternatively as follows,

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} a_i \, y_i$$

where,

$$a_i = \begin{cases} 1 \text{if } i^{th} \text{ unit is included in the sample} \\ 0 \text{if } i^{th} \text{ unit is include in the sample} \end{cases}$$

S^2 = mean s quare for the population .

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}_N\right)^2$$

$$= \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}_N^2\right)$$

Mean square for the sample:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}_n\right)^2$$

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

Population total $= \sum_{i-1}^{n} y_i$

Sample total $= \sum_{i-1}^{n} y_i$

Population variance $\left(\sigma^2\right) = \dfrac{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2}{N}$

Sample variance $\left(s^2\right) = \dfrac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n-1}$

Sample variance of SRSWOR $V\left(\bar{y}\right) = \left(1 - \dfrac{n}{N}\right)\dfrac{s^2}{n}$ *where* $s^2 = \dfrac{N\sigma^2}{N-1}$

Sample variance of SRSWR $\Rightarrow V\left(\bar{y}\right) = \dfrac{\sigma^2}{n}\left(\dfrac{1}{N}\right)$

Estimation of population total $\Rightarrow \hat{y} = \dfrac{N}{n}\sum_{i=1}^{n} y_i = N\bar{y}$

Estimation of population mean $\Rightarrow \hat{y} = \sum_{i=1}^{n}\dfrac{y_i}{n} = \bar{y}$

Theorem: 3.3

In simple random sampling without replacement a sample mean is an unbiased estimator of the population Mean,

ie., $E\left(\bar{y}\right) = \bar{Y}$

Proof:

We know that, $E\left(\bar{y}\right) = E\left(\sum_{i=1}^{n} \frac{y_i}{n}\right)$

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

$$E\left(\bar{y}\right) = \frac{1}{n} \sum_{i=1}^{n} E(y_i) \qquad (1)$$

By definition, $E(y_i) = \sum_{i=1}^{n} y_i p_i$

$$= \sum_{i=1}^{n} y_i \left(\frac{1}{N}\right) \qquad \because p_i = \frac{1}{N}$$

$$E(y_i) = \bar{Y} \qquad (2)$$

From eqn. 1 & 2, we get,

$$E\left(\bar{y}\right) = \frac{1}{n} \sum_{i=1}^{n} \left(\bar{y}\right) = \frac{1}{n} n.\bar{y}$$

$$E\left(\bar{y}\right) = \bar{Y}\ .$$

Theorem: 3.4

In simple random sampling without replacement the sample variance is an unbiased estimated of the population variance.

$$E\left(s^2\right) = S^2$$

Proof:

We know that, Sample Variance,

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

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i^{\,2} - n\,\bar{y}_n^{\,2}\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i^{\,2} - n\left(\frac{\sum y_i}{n}\right)^2\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i^{\,2} - \frac{n}{n^2}\left(\sum_{i=1}^{n} y_i\right)^2\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left[y_i^{\,2} - \frac{1}{n}\left(\sum_{i=1}^{n} y_i^{\,2} + \sum_{i\neq j=1}^{n} y_i y_j\right)\right]$$

$$= \left[\frac{1}{n-1}\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} y_i y_j\right]$$

$$= \sum_{i=1}^{n} y_i^2\left(\frac{1}{n-1} - \frac{1}{n(n-1)}\right) - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} y_i y_j$$

$$= \sum_{i=1}^{n} y_i^{\,2}\left[\frac{n-1}{n(n-1)}\right] - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} y_i y_j$$

$$\Rightarrow s^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} y_i y_j$$

Taking expectation on both sides,

$$E\left(s^2\right) = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} y_i y_j\right]$$

$$E\left(s^2\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(y_i^2\right) - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} E\left(y_i y_j\right) \tag{1}$$

Consider, $E\left(\sum_{i=1}^{n} y_i^2\right) = E\left(a_i y_i^2\right)$

$$= y_i^2 E\left(a_i\right)$$

$$= y_i^2\left(0\times p\left(a_i = 1\right) + 1\times p\left(a_i = 1\right)\right)$$

$$= y_i^2\begin{bmatrix}\left(o\times i^{th}\ unit\ of\ the\ sample\ \ is\ not\ included\ in\ the\ sample\ size\ n\right)+\\ (1\times i^{th}\ unt\ of\ the\ sample\ is\ included\ in\ the\ sample\ size\ n)\end{bmatrix}$$

$$= y_i^2\left[0\times\left(1-\frac{1}{n}\right) + \left(\frac{n}{N}\right)\right]$$

$$E(y_i)^2 = \frac{n}{N}y_i^2 \tag{2}$$

Consider, $E\left(\sum_{i=1}^{n} y_i y_j\right) = E\left(a_i a_j y_i y_j\right)$

$$= \sum_{i \neq j=1}^{n} y_i y_j E(a_i a_j)$$

$$= \sum_{i \neq j=1}^{n} y_i y_j [0 \times P(a_i = 1, a_j = 1) + 1 \times P(a_i = 1, a_j = 1)]$$

$$= \sum_{i \neq j=1}^{n} y_i y_j [P(a_i = 1 \cap a_j = 1)]$$

$$= \sum_{i \neq j=1}^{n} y_i y_j [P(a_i = 1) P(a_j = 1)]$$

$$= \sum_{i \neq j=1}^{n} y_i y_j \begin{bmatrix} P\left(i^{th} \text{ unit of the sample include in the sample}\right) \\ \times P\begin{pmatrix} j^{th} \text{ unit of the sample} \\ \text{include} \end{pmatrix} \end{bmatrix}$$

$$= \sum_{i \neq j=1}^{n} y_i y_j \left[\frac{n}{N} \times \frac{n-1}{N-1} \right] \qquad (3)$$

Substitute eqn. 2 and 3 in eqn. 1, then

$$E(s^2) = \frac{1}{n}\left(\frac{n}{N} \sum_{i=1}^{n} y_i^2 \right) - \frac{1}{n(n-1)} \frac{n(n-1)}{N(N-1)} \sum_{i \neq j=1}^{n} y_i y_j$$

$$= \frac{1}{N} \sum_{i=1}^{n} y_i^2 - \frac{1}{N(N-1)} \sum_{i \neq j=1}^{n} y_i y_j$$

$$= \frac{1}{N-1} \left[\sum_{i=1}^{n} y_i^2 - N \bar{Y}_N^2 \right]$$

$$= S^2$$

Therefore , $E(s^2) = S^2$.

Theorem: 3.5

In simple random sampling without replacement the variance of the sample mean is given by,

$$\operatorname{var}\left(\bar{y}_n\right) = \left(\frac{1}{n} - \frac{1}{N} \right) S^2$$

$$= \left(\frac{N-n}{nN} \right) S^2$$

Proof:

$$\operatorname{var}\left(\bar{y}_n\right) = E\left(\bar{y}_n^2\right) - \left[E\left(\bar{y}_n\right) \right]$$

$$= E\left(\bar{y}^2\right) - \bar{Y}_N^2 \qquad \because E\left(\bar{y}_n\right) = \bar{Y}_N \qquad (1)$$

Consider, $E\left(\bar{y}_n^{\,2}\right) \quad = E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2$

$$= \left(\frac{1}{n}\right)^2 E\left(\sum_{i=1}^{n} y_i\right)^2$$

$$= \left(\frac{1}{n}\right)^2 E\left(\sum_{i=1}^{n} y_i\right)^2 + E\left[\sum_{i\neq j=1}^{n} y_i y_j\right]$$

$$= \left(\frac{1}{n}\right)^2 \left[E\left(\sum_{i=1}^{n} y_i^{\,2}\right) + E\left(\sum_{i\neq j=1}^{n} y_i y_j\right)\right] \qquad (2)$$

Consider,

$$E\left(\sum_{i=1}^{n} y_i^2\right) \quad = E\left(\sum_{i=1}^{n} a_i y_i^2\right)$$

$$= \sum_{i=1}^{n} y_i^2 E(a_i)$$

$$= \sum_{i=1}^{n} y_i^2 \left(\frac{n}{N}\right) \qquad \because E(a_i) = \frac{n}{N} \qquad (3)$$

But,

$$\sum_{i=1}^{N}\left(y_i - \bar{Y}_N\right)^2 = \sum_{i=1}^{n} y_i^{\,2} - N\bar{Y}_N^{\,2}$$

$$\Rightarrow \sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N}\left(y_i - \bar{Y}_N\right)^2 + N\bar{Y}_N^{\,2}$$

$$\sum_{i=1}^{N} y_i^2 = (N-1)S^2 + N Y_N^2 + N\bar{Y}_N^2 \qquad \because S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}_N^{\,2}\right) \qquad (4)$$

Substitute eqn. 4 in eqn. 3, we get

$$E\left(\sum_{i=1}^{n} y_i^{\,2}\right)=\left[(N-1)S^2+N\bar{Y}_N^2\right]\left(\frac{n}{N}\right) \qquad \ldots\ldots\ldots(5)$$

$$E\left(\sum_{i\neq j=1}^{n} y_i y_j\right) \quad = E\left(\sum_{i\neq j=1}^{n} a_i a_j\, y_i y_j\right)$$

$$=\sum_{i\neq j=1}^{n} y_i y_j\, E(a_i a_j)$$

$$=\sum_{i\neq j=1}^{n} y_i y_j \left(\frac{n(n-1)}{N(N-1)}\right)$$

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

$$=\left(\frac{n(n-1)}{N(N-1)}\right)\left[(N\bar{Y}_N)^2-\left[(N-1)S^2+N\bar{Y}_N^{\,2}\right]\right]$$

$$=\left(\frac{n(n-1)}{N(N-1)}\right)\left[N^2\bar{Y}_N^{\,2}-(N-1)S^2-N\bar{Y}_N^{\,2}\right]$$

$$=\left(\frac{n(n-1)}{N(N-1)}\right)\left[N\bar{Y}_N^{\,2}(N-1)-(N-1)-(N-1)S^2\right]$$

$$E\left(\sum_{i\neq j=1}^{n} y_i y_j\right) \qquad =\left(\frac{n(n-1)(N-1)}{N(N-1)}\right)\left[N\bar{Y}_N^2-S^2\right]$$

$$=\left(\frac{n(n-1)}{N}\right)\left[N\bar{Y}_N^{\,2}-S^2\right] \qquad (6)$$

Substitute eqn. 5 & 6 in eqn. 2,

$$E(\bar{y}_n)^2=\frac{1}{n^2}\left(\frac{n}{N}\left[(N-1)S^2+N\bar{Y}_N^2\right]\right)+\frac{n(n-1)}{N}\left[N\bar{Y}_N^{\,2}-S^2\right]$$

$$=\frac{n(N-1)}{Nn}S^2+\frac{Nn}{n^2N}\bar{Y}_N^{\,2}+\frac{n(n-1)N}{n^2N}-\frac{n(n-1)}{Nn^2}S^2$$

$$=\frac{(N-1)}{Nn}S^2+\frac{1}{n}\bar{Y}_N^2+\frac{(n-1)}{n}\bar{Y}_N^2-\frac{(n-1)}{Nn}S^2$$

$$=\frac{N}{Nn}S^2-\frac{1}{Nn}S^2+\frac{1}{n}\bar{Y}_N^2+\frac{n}{N}\bar{Y}_N^2-\frac{1}{n}\bar{Y}_N^2-\left[\frac{n}{Nn}S^2-\frac{1}{Nn}S^2\right]$$

$$=\frac{S^2}{n}-\frac{1}{Nn}S^2+\frac{1}{n}\bar{Y}_N^2+\bar{Y}_N^2-\frac{1}{n}\bar{Y}_N^2-\frac{1}{N}S^2+\frac{1}{Nn}S^2$$

$$=\frac{S^2}{n}+\bar{Y}_N^2-\frac{S^2}{N}$$

$$=\left(\frac{1}{n}-\frac{1}{N}\right)S^2+\bar{Y}_N^2 \qquad (7)$$

Substitute eqn. 7 in eqn. in 1, $\Rightarrow$

$$\operatorname{var}(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2 + \bar{Y}_N^2 - \bar{Y}_N^2$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right)S$$

$$\operatorname{var}(\bar{y}_n) = \left(\frac{N-n}{Nn}\right)S^2$$

Merits of Simple random sampling:

i. since the sample units are selected at the random going each units at equal chances of being selected the subjectively or personal bias is completely eliminated as such a simple random sampling is more representative population of a judgment is compared to purposive sampling.

ii. the statisticians are as certain the efficiency of the estimates of the parameters by considering the sampling distribution of the statistics or estimates i.e, if the sample size 'n' increase, $\bar{y}_n$ an estimates of $\bar{Y}_N$ becomes more efficient.

Drawbacks:

i) The selection of simple random sampling requires up to data frame, i.e a completely catalogued population from which samples are to be drawn. Frequently it is the virtually impossible to identify the units in the population before the sample is drawn and this restricts the use of sample random sampling.

ii) Administrative inconvenience: A simple random sample may results in the selection of the sampling units are widely spread geographically and in such case that the cost of collecting data may be much in terms of time and money.

iii) At times a simple random sampling might most non-random looking results. For example, If we draw a random sample of size 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.

iv) For a given precision the simple random sampling usually requires a larger sample size as compared to stratified sampling.

PROPLEM 1: Draw a random sample (without replacement) of size 15 from a population of size 500.

Solution:

- ❖ Identify the 500 units in the population with the numbers from any of the random number series.
- ❖ Starting at random with any number on that page and moving row-wise, column-wise or diagonally, we select one by one the three digited numbers, discarded the numbers over 500, until 15 numbers below 500 are obtained.
- ❖ Finally, the units in the population, corresponding to these 15 numbers will constitute our random sample without replacement.
- ❖ The 15 random samples are
  193, 226, 234, 182, 164, 157, 497, 264, 350, 361, 337, 357, 020, 374, 394.

PROPLEM 2: The following data refer to the Kapas yield of 96 plants.

82	102	88	93	97	38	103	92
102	62	63	72	64	68	59	69
73	65	46	79	87	84	29	52
28	36	46	79	87	84	29	52
56	66	42	37	35	97	32	35
89	99	54	72	26	67	18	27
60	72	33	42	52	82	14	22
57	73	63	61	63	92	40	58
62	61	43	25	42	36	17	30
75	87	47	56	76	36	35	44
56	51	111	73	93	58	49	89
50	80	54	55	91	12	82	76

Select a sample of 25 plants by using simple random sampling method. Also calculate the 25 samples and verify whether the sample mean is equal to the population mean.

Solution:

Kapas Yield of 96 plants
(ie.) Population, N=96

$$\text{Population Mean}(\overline{X}_N) = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{\sum_{i=1}^{96} X_i}{96}$$

$$= \frac{82+102+88+\ldots+12+82+76}{96} = \frac{5798}{96} = 60.40$$

82 1	102 13	88 25	93 37	97 49	38 61	103 73	92 85
102 2	62 14	63 26	72 38	64 50	68 62	59 74	69 86
73 3	65 15	46 27	79 39	87 51	84 63	29 75	52 87
28 4	36 16	46 28	79 40	87 52	84 64	29 76	52 88
56 5	66 17	42 29	37 41	35 53	97 65	32 77	35 89
89 6	99 18	54 30	72 42	26 54	67 66	18 78	27 90
60 7	72 19	33 31	42 43	52 55	82 67	14 79	22 91
57 8	73 20	63 32	61 44	63 56	92 68	40 80	58 92
62 9	61 21	43 33	25 45	42 57	36 69	17 81	30 93
75 10	87 22	47 34	56 46	76 58	36 70	35 82	44 94
56 11	51 23	111 35	73 47	93 59	58 71	49 83	89 95
50 12	80 24	54 36	55 48	91 60	12 72	82 84	76 96

S.No.	**Random Number**	**Rapsi Yield**	**S.No.**	**Random Number**	**Rapsi Yield**
1.	19	72	14.	34	47
2.	16	36	15.	76	29
3.	88	52	16.	93	30
4.	02	102	17.	61	38
5.	47	73	18.	09	62
6.	73	103	19.	06	89
7.	46	56	20.	84	82
8.	15	65	21.	25	88
9.	14	62	22.	29	42

10.	04	28	23.	94	44
11.	65	97	24.	48	55
12.	23	51	25.	55	52
13.	89	35			

$$\text{Sample Mean}(\bar{x}_n) = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{25} x_i}{25}$$

$$= \frac{72+36+52+...+55+52}{25} = \frac{1490}{25} = 59.6$$

Population Mean($\overline{X}$) = 60.4 and Sample Mean($\bar{x}$)=59.6

Therefore, Population Mean $\cong$ Sample Mean.

REMARKS:

(i) Sampling fraction:

The ratio between population and sample is called sampling fraction i.e, $f = \frac{n}{N}$. The factor $(1-f)$ in called the finite population correction (FPC). If the population size N is very large or of N is small compared with N then, $f = \frac{n}{N} \to 0$ and consequently FPC i.e $1-f \to 1$

(ii) Sampling variance, $\operatorname{var}(\bar{y}_n) = \left(1-\frac{n}{N}\right)\frac{s^2}{n} = (1-f)\frac{s^2}{n}$

Standard error of the sampling distribution of $\bar{y}_n$ is given by,

$$\text{S.E } (\bar{y}_n) = \sqrt{\frac{N-n}{N}}$$

<u>SIMPLE RANDOM SAMPLING BY ATTRIBUTES:</u>

An attributes is a qualitative characteristics which cannot be measured quantitative. For example, honestly, beauty intelligence etc., in such a situation the information may not be possible to classify to measure but it may be possible to classify the whole population into various classes with respect to a attribute. We consider the simplest the cause where the population of deviated to classes only say ‘A’ and ‘A’ with respect to an attribute hence any sampling unit in the population may be place in class A and A' respectively. In this study of attributes we may interested in estimating the total number of proportion of,

i. Defective items in a large consignment such items.
ii. The literates’ in a down.

iii. The educated unemployed persons.

Notations and Terminology :

For population:

Let us suppose that a population with n units $U_1, U_2 .....U_n$ is classified into two mutually disjoint and exhaustive classes A and A' such that A + A' =N

Then, proportion of units processing the given attributes is $\frac{A}{N}$,

i.e $P = \frac{A}{N} \Rightarrow A = PN$

Q=the proportion of units which do not possess the given attributes,

$$\Rightarrow Q = \frac{A'}{N} = 1 - P$$

Where, P & Q represents population of success and failure respectively in the population.

With the i^{th} sampling units let us associated a variate $y_i = (i = 1,2,\ldots N)$ defined as follows

$$y_i = \begin{cases} 1 \, if \; it \, belong \; to \, class \; A \\ 0 \, if \; it \, belong \; to \, class \; A' \end{cases}$$

Thus $\sum_{i=1}^{N} y_i = A$ the number of units in the population processing the given attribute.

For sample:

Let us consider a simple random sampling without replacement of size n from the population N if 'a' is the simple possessing the given attribute. Then proportion of sampled units possessing the given attribute is $\frac{a}{n}$.

i.e, $p = \frac{a}{n} \Rightarrow a = np$

proportion of sampled units which do not possess given attribute is $\frac{a}{n}$

i.e $q = \frac{a'}{n}$

$a' = nq$

$p + q = 1, \quad q = 1 - p$.

With the i^{th} sampled units. Let us associated a variate $y_i = (i = 1,2,\ldots N)$ defined as follows,

$$y_i = \begin{cases} 1 \, if \; i^{th} \; sampled \; unit \; posses \; the \; given \, attribute \\ 0 \, if \; i^{th} \; sampled \; unit \; is \, does \; not \; posses \, \sin g \; the \; given \, attributes \end{cases}$$

Thus, $\sum y_i = a$ the number of units possessing the given attributes

Population mean:

w.k.t $\bar{x} = \dfrac{\sum_{i=1}^{N} x_i}{N}$

the total number of units in the population possessing the attributes is ‘A’

i.e ., $\sum_{i=1}^{N} x_i = A$

$$\bar{X} = \frac{A}{N} \Rightarrow \frac{NP}{N} = P$$

$$\bar{X} = P$$

Sample mean:

w.k.t $\bar{x} = \dfrac{\sum_{i=1}^{N} x_i}{N}$

the total number of sampling units possessing the attribute is ‘a’

$$\sum_{i=1}^{n} x_i = a$$

i.e, $\bar{x} = \dfrac{a}{n} \Rightarrow \dfrac{np}{n} = p$

$$\Rightarrow \bar{x} = p$$

also $\sum y_i^2 = a = np$

For population mean square:

W.K.T,

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}_N\right)^2$$

$$S^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i - N\bar{Y}_N^2\right)$$

$$= \frac{1}{N-1}\left(NP - NP^2\right) \qquad \because \sum_{i=1}^{N} Y_i = A$$

$$= \frac{NP}{N-1}(1-p)$$

$$S^2 = \frac{NPQ}{N-1}$$

For sample mean square:

W.K.T,

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

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

$$= \frac{1}{n-1}\left(np - np^2\right) \qquad \because \sum_{i=1}^{N} y_i = a$$

$$= \frac{np}{n-1}(1-p)$$

$$s^2 = \frac{npq}{n-1}$$

Theorem: 3.6

Sample proportion ‘p’ is an unbiased estimate of the population proportion ‘p’.

i.e $E(p) = P$

Proof:

Lemma:

In simple random sampling without replacement a sample mean is an unbiased estimate of the population mean,

i.e., $E(\bar{y}_n) = \bar{Y}_N$

Proof of the Lemma:

Consider, $E(\bar{y}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} a_i a_j\right)$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i E(a_i) \qquad (1)$$

Hence, $E(a_i) = 1$

$$= 1.P\{i^{th} \text{ unit is included in the sample of size } n\}$$
$$+ 0.P\{i^{th} \text{ unit not included in the sample of size } n\}$$

$$E(a_i) = 1.\frac{n}{N} + 0.\left(1 - \frac{n}{N}\right) \Rightarrow \frac{n}{N} \qquad (2)$$

Substitute equation 2 in equation 1,

$$E(\bar{y}_n) = \frac{1}{n}\sum_{i=1}^{N} y_i\left(\frac{n}{N}\right)$$

$$= \sum_{i=1}^{N} \frac{y_i}{N} \Rightarrow \bar{Y}_N \tag{3}$$

Hence proof lemma is $E(\bar{y}_n) = \bar{Y}_N$

Proof of the main theorem:

W.k.t $\sum_{i=1}^{n} y_i = a$ & $\sum_{i=1}^{N} Y_i = A$

$a = np$ & $A = NP$

Substitute the above information in equation 3

$$E(\bar{y}_n) = \bar{Y}_N \quad \Rightarrow E(p) = P$$

Remarks:

$$\begin{aligned} E(NP) &= NP \\ &= N\,E(P) \\ &= NP \end{aligned}$$

Theorem: 3.7

In simple random sampling without replacement,

$$\text{var}(p) = \frac{N-n}{N-1} \cdot \frac{PQ}{n}$$

Proof:

w.k.t, $\text{var}(p) = \text{var}(\bar{y})$

$$\begin{aligned} &= \frac{N-n}{nN} S^2 && \because \bar{y} = p \\ &= \frac{N-n}{nN} \cdot \frac{NPQ}{N-1} && \because S^2 = \frac{NPQ}{N-1} \\ &= \frac{N-n}{n} \cdot \frac{PQ}{N-1} \end{aligned}$$

$$\text{var}(p) = \frac{N-n}{N-1} \cdot \frac{PQ}{n}$$

CONFIDENCE LIMITS:

After having the estimate of an unknown parameter it becomes necessary measure the reliability of these estimates and the construct some confidence limits with the given degree of confidence if we assume that the estimated $\bar{y}_n$ is normally.

Distributed about the Population Mean $\bar{Y}$ lower and upper confidence limits for the population mean are given,

$$\hat{Y}_L = \bar{y} - t_{(\alpha,n-1)} s \left(\frac{(1-f)}{n} \right)^{\frac{1}{2}}$$

And

$$\hat{Y}_U = \bar{y} + t_{(\alpha,n-1)} s \left(\frac{(1-f)}{n} \right)^{\frac{1}{2}}$$

where $t_{(\alpha,n-1)}$ stands for the value of students 't' distribution with (n-1) degrees of freedom at α level of significance. Similarly the confidence limits for the population,

$$\bar{Y}_L = N\bar{y} - \frac{t_{(\alpha,n-1)} s \sqrt{1-f}}{\sqrt{n}}$$

And,

$$\bar{Y}_U = N\bar{y} + \frac{t_{(\alpha,n-1)} s \sqrt{1-f}}{\sqrt{n}}$$

DETERMINATION OF SAMPLE SIZE:

In sampling analysis the most ticklish question is: What should be the size of the sample or how large or small should be '*n*'. If the sample size ('*n*') is too small, it may not serve to achieve the objectives and if it is too large, we may incur huge cost and waste resources. As a general rule, one can say that the sample must be of an optimum size i.e., it should neither be excessively large nor too small. Technically, the sample size should be large enough to give a confidence interval of desired width and as such the size of the sample must be chosen by some logical process before sample is taken from the universe. Size of the sample should be determined by a researcher keeping in view the following points:

- *Nature of universe:* Universe may be either homogenous or heterogeneous in nature. If the items of the universe are homogenous, a small sample can serve the purpose. But if the items are heterogeneous, a large sample would be required. Technically, this can be termed as the dispersion factor.

- *Number of classes proposed:* If many class-groups (groups and sub-groups) are to be formed, a large sample would be required because a small sample might not be able to give a reasonable number of items in each class-group.

- *Nature of study:* If items are to be intensively and continuously studied, the sample should be small. For a general survey the size of the sample should be large, but a small sample is considered appropriate in technical surveys.

- *Type of sampling:* Sampling technique plays an important part in determining the size of the sample. A small random sample is apt to be much superior to a larger but badly selected sample.

- *Standard of accuracy and acceptable confidence level:* If the standard of accuracy or the level of precision is to be kept high, we shall require relatively larger sample. For doubling the accuracy for a fixed significance level, the sample size has to be increased fourfold.

- *Availability of finance:* In practice, size of the sample depends upon the amount of money available for the study purposes. This factor should be kept in view while determining the size of sample for large samples result in increasing the cost of sampling estimates.

- *Other considerations:* Nature of units, size of the population, size of questionnaire, availability of trained investigators, the conditions under which the sample is being conducted, the time available for completion of the study are a few other considerations to which a researcher must pay attention while selecting the size of the sample.

There are two alternative approaches for determining the size of the sample. The first approach is "to specify the precision of estimation desired and then to determine the sample size necessary to insure it" and the second approach "uses Bayesian statistics to weigh the cost of additional information against the expected value of the additional information." The first approach is capable of giving a mathematical solution, and as such is a frequently used technique of determining '*n*'. The limitation of this technique is that it does not analyse the cost of gathering information *vis-a-vis* the expected value of information. The second approach is theoretically optimal, but it is seldom used because of the difficulty involved in measuring the value of information. Hence, we shall mainly concentrate here on the first approach.

DETERMINATION OF SAMPLE SIZE THROUGH THE APPROACH BASED ON PRECISION RATE AND CONFIDENCE LEVEL

To begin with, it can be stated that whenever a sample study is made, there arises some sampling error which can be controlled by selecting a sample of adequate size. Researcher will have to specify the precision that he wants in respect of his estimates concerning the population parameters. For instance, a researcher may like to estimate the mean of the universe within ± 3 of the true mean with 95 per cent confidence. In this case we will say that the desired precision is ±3, i.e., if the sample mean is Rs 100, the true value of the mean will be no less than Rs 97 and no more than Rs 103. In other words, all this means that the acceptable error, *e*, is equal to 3. Keeping this in view, we can now explain the determination of sample size so that specified precision is ensured.
(a)

The confidence interval for the population, μ is given by $\overline{X} \pm z \dfrac{\sigma_p}{\sqrt{n}}$

where $\overline{X}$ = sample mean

z = the value of the standard variate at a given confidence level (to be read from the table giving the areas under normal curve as shown in appendix) and it is 1.96 for a 95% confidence level;

n = size of the sample

σ_p = standard deviation of the population (to be estimated from past experience or on the basis of a trial sample). Suppose, we have σ_p = 4.8 for purpose.

If the difference between m and X or the acceptable error is to be kept within ±3 of the sample mean with 95% confidence, then we can express the acceptable error, 'e' as equal to

$$e = z\frac{\sigma_p}{\sqrt{n}}$$

$$\Rightarrow 3 = 1.96\frac{4.8}{\sqrt{n}}$$

Hence, $n = \frac{(1.96)^2 (4.8)^2}{(3)^2} = 9.834 \cong 10$

In a general way, if we want to estimate μ in a population with standard deviation σ_p with an error no greater than "e" by calculating a confidence interval with confidence corresponding to z, the necessary sample size, n equals as under $n = \frac{z^2\sigma^2}{e^2}$.

All this is applicable when the population happens to be infinite. But in case of finite population, the above stated formula for determining sample size will become

$$n = \frac{z^2.N.\sigma^*{}_p}{(N-1)e^2 + z^2\sigma_p^2}$$

*In case of infinite population, the confidence interval for μ is given by

$$\bar{X} \pm z\frac{\sigma_p}{\sqrt{n}}\sqrt{\frac{(N-n)}{(N-1)}}$$

where $\sqrt{\frac{(N-n)}{(N-1)}}$ is the finite population multiplier and all other mean thing as stated above. If the precision is taken as equal "e" then we have,

$$e = z.\frac{\sigma_p}{\sqrt{n}}.\sqrt{\frac{(N-n)}{(N-1)}}$$

$$\Rightarrow e^2 = z^2.\frac{\sigma_p^2}{n}.\frac{N-n}{N-1}$$

$$\Rightarrow e^2(N-1) = \frac{z^2.\sigma_p^2.N}{n} - \frac{z^2.\sigma_p^2.n}{n}$$

$$\Rightarrow e^2(N-1) + z^2.\sigma_p^2 = \frac{z^2.\sigma_p^2.N}{n}$$

$$\Rightarrow n = \frac{z^2.\sigma_p^2.N}{e^2(N-1) + z^2.\sigma_p^2}$$

where N= Size of the population
n= Size of the sample
e= Acceptable error
σ_p = Standard deviation of the population
z = Standard Normal variate at given confidence level.

UNIT-IV

STRATIFIED RANDOM SAMPLING:

Stratification means division into layers auxiliary information related to the character under study may be used to divided population into various groups such that,

i. Units within each groups or us homogenous as possible.
ii. The group means as widely difference as possible.

Thus a population consisting of 'N' sampling units is divided into 'k' relatively homogenous mutually disjoint subgroups terms as strata of sizes $N_1, N_2 \cdots N_k$ such that $N = \sum_{i=1}^{k} N_i$ if a sample random sampling (generally without replacement) of sizes $n_i = (i = 1,2..k)$ is drawn from the stratum respectively such that $n = \sum_{i=1}^{k} n_i$.

This sample is termed as stratified sample of size 'n' and the technique of drawing such a sample is called stratified random sampling.

$n = \sum_{i=1}^{k} n_i$ = total sample size from all strata.

$$\bar{y}_n = population\, mean = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} y_{ij}$$

$$= \frac{1}{N} \sum_{i=1}^{k} N_i \bar{Y}_{N_i} = \sum_{i=1}^{k} W_i \bar{Y}_{N_i} \text{ s}$$

$\bar{y}_{n_i}$ = mean of sample selected from i^{th} stratum

$s_i^2 = sample\ mean\ square\ of\ the\ i^{th}\ stratum$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{ni} \right)^2, i = 1,2..k$$

Consider the following two estimate of population mean which are,

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{y}_{ni}$$

$$\bar{y}_{st} = \frac{1}{N} \sum_{i=1}^{k} N_i y_{ni} = \sum_{i=1}^{k} w_i \bar{y}_{ni}$$

Theorem: 4.1

If in every stratum the sample estimator $\bar{y}_i$ is unbiased and samples are drawn independently in difference strata then $\bar{y}_{st}$ is an unbiased estimator of the population mean and its sampling variance is given by.

$$V(\bar{y}_{st}) = \sum_{i}^{k} w_i^2 v(\bar{y}_i)$$

Proof:

w.k.t $\qquad \bar{y}_{st} = \sum_{i=1}^{k} \frac{N_i \bar{y}_i}{N}$

taking expectation on both sides,

$$E(\bar{y}_{st}) = E\left[\sum_{i=1}^{k} \frac{N_i \bar{y}_i}{N}\right]$$

$$= \sum_{i=1}^{k} \frac{N_i}{N} E(\bar{y}_i)$$

$$= \sum_{i=1}^{k} W_i E(\bar{y}_i)$$

$$E(\bar{y}_{st}) = \bar{Y}_{st}$$

Hence, Sample mean is unbiased estimator of the population for given stratified random sampling.

To prove:

$$V(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2 V(\bar{y}_i)$$

$$= \sum_{i=1}^{k} \frac{N_i^2}{N^2} V(\bar{y}_i)$$

$$= \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 V(\bar{y}_i)$$

$$\text{var}(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2 V(\bar{y}_i)$$

Theorem: 4.2

With stratified random sampling without replacement an unbiased estimator of the variance $\bar{y}_{st}$ is given by,

$$v(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2 \frac{s_i^2}{n_i} - \sum_{i=1}^{k} W_i \frac{S_i^2}{N_i}$$

Proof:

Since the sample in each stratum is simple random sampling without replacement then,

$$\text{var}(\bar{y}_i) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_i^2$$

$$v(\bar{y}_{st}) = v\left(\sum_{i=1}^{k} \frac{N_i \bar{y}_i}{N}\right)$$

$$= \sum_{i=1}^{k} \frac{N_i^2}{N^2} v(\bar{y}_i)$$

$$= \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2$$

$$= \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2 \frac{S_i^2}{n_i} - \sum_{i=1}^{k} \frac{N_i^2}{N^2} - \frac{S_i^2}{N_i}$$

$$= \sum_{i=1}^{k} W_i^2 \frac{S_i^2}{n_i} - \sum_{i=1}^{k} \frac{N_i}{N} \frac{S_i^2}{N} \qquad \because W_i = \left(\frac{N_i}{N}\right)$$

$$= \sum_{i=1}^{k} W_i^2 \frac{S_i^2}{n_i} - \sum_{i=1}^{k} W_i \frac{S_i^2}{N}$$

Allocation of sampling size in difference strata:

1. In stratified sampling the allocation of the sample to different strata is done by the consideration of three factors such as,
   i. The total number of units in the stratum.
   ii. The variability with in this stratum and
   iii. The cost in taking observation per sampling unit in the stratum.

2. A good allocation is one were maximum precision is obtained with minimum resources.

3. There are four method of allocation of sample size is to different strata in a stratified sampling procedure these are,
   i. Equal allocation
   ii. Proportional allocation
   iii. Neyman's allocation
   iv. Optimum allocation

Equal Allocation:

This is a situation of considerable practical interest for reasons of administrative or field work convenience. In this method a total sample size 'n' is divided equally among all the strata, i.e., for the i^{th} stratum

$$n_i = \frac{n}{k}$$

Proportional Allocation:

This allocation proposed by Bowely (1926) the procedure of allocation is very common in practice because of its simplicity when no other information except 'N' the total of number of units in the i^{th} stratum is available a allocation of a given sample of size 'n' to different strata it's done in proportion to their sizes. i.e., in the i^{th} stratum

$$n_i = \frac{nN_i}{N}$$

Neyman's Allocation:

This allocation of the total sample size to strata of is called minimum variance allocation and is due to Neyman (1934). In this allocation is assume that the sampling cost per unit among difference strata in the same and the size of the its fixed. Sample size is are allocated by $n_i = \frac{nW_iS_i}{\sum_i^k W_iS_i}$

w.k.t $\quad w_i = \frac{N_i}{N}$

$$n_i = \frac{nN_iS_i}{\sum_i^k N_iS_i}$$

The minimum variance of Neyman's allocation with fixed 'n' is obtained by

$$V_{\min}(\bar{y}_{st}) = \frac{\left(\sum_i^k w_i s_i\right)^2}{n} - \frac{\sum_i^k W_i S_i^2}{N}$$

Optimum Allocation:

In this method of allocation a sample sizes 'n_i' in the respective strata are determine with the view to minimize variance for $\bar{y}_{st}$ for a specified cost of containing the sample survey or to minimize the cost for a specified value of variance of ($\bar{y}_{st}$) . The simplest cost function in stratified sampling that we can take is , $c = a + \sum_i^k n_i c_i$, where overhead cost a is constant and c_i is the average cost of surveying one units in the i^{th} stratum which may depend the nature and size of the units in the stratum.

Then the allocation of a given sample of size 'n' to different and given cost function $c = a + \sum_i^k n_i c_i$,

$$n_i = n\frac{\left(\frac{W_i S_i}{\sqrt{c_i}}\right)}{\sum_i^k \left(\frac{w_i s_i}{\sqrt{c_i}}\right)}$$

Theorem: 4.3 (*OPTIMUM ALLOCATIONS)*

The stratified random sampling with a given cost function, $c = c_0 + \sum_i^k n_i c_i$ the variance of estimate $\bar{y}_{st}$ is minimum of $n_i = \frac{N_i S_i}{\sqrt{c_i}}$ then

$$n_i = n\frac{\left(\frac{N_i S_i}{\sqrt{c_i}}\right)}{\sum_i^k \left(\frac{n_i s_i}{\sqrt{c_i}}\right)}$$

Proof:

w.k.t the variance of the stratified random sampling,

$$v(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^k N_i (N_i - n_i)\frac{s_i^2}{n_i} \qquad ....(1)$$

The cost function, $c = c_0 + \sum_i^k n_i c_i \qquad ....(2)$

To determine the optimum value of 'n_i'consider the function,

$$\phi = v(\bar{y}_{st\ wor}) + \lambda c \qquad ....(3)$$

$$\phi = \frac{1}{N^2}\sum_{i=1}^k N_i (N_i - n_i)\frac{s_i^2}{n_i} + \lambda c$$

$$\phi = \frac{1}{N^2}\sum_{i=1}^k \frac{N_i^2 S_i^2}{n_i} - \frac{1}{N^2}\sum_{i=1}^k N_i S_i^2 + \lambda c_0 + \lambda \sum_{i=1}^k n_i c_i$$

Differentiate, w.k.t n_i and equal to 0

i.e $\frac{\partial \phi}{\partial n_i} = 0$

$$\frac{\partial \phi}{\partial n_i} = 0 \Rightarrow -\frac{1}{N^2}\sum_{i=1}^k \frac{N_i^2 S_i^2}{n_i} - 0 - 0 + 0 + \lambda \sum_{i=1}^k c_i = 0$$

$$\Rightarrow -\frac{1}{n^2}\sum_{i=1}^k \frac{N_i^2 S_i^2}{n_i} + \lambda \sum_{i=1}^k c_i = 0$$

$$\Rightarrow \sum_{i=1}^k \left[-\frac{N_i^2 S_i^2}{N^2 n_i^2} + \lambda c_i \right] = 0$$

$$\Rightarrow -\frac{N_i^2 S_i^2}{N^2 n_i^2} + \lambda c_i = 0$$

$$\Rightarrow \lambda c_i = \frac{N_i^2 S_i^2}{N^2 n_i^2}$$

$$\Rightarrow n_i^2 = \frac{N_i^2 S_i^2}{N^2 \lambda c_i}$$

$$\Rightarrow n_i = \frac{N_i S_i^2}{N\sqrt{\lambda c_i}}$$

Here λ is unknown and we have the value of λ is fixed. The sum overall all stratum,

i.e., $n = \sum_{i=1}^{k} n_i$

$$\sum_{i=1}^{k} \frac{N_i S_i}{N\sqrt{\lambda c_i}}$$

$$n = \frac{1}{\sqrt{\lambda N}} \sum_{i=1}^{k} \frac{N_i S_i}{\sqrt{c_i}} \qquad \text{......(5)}$$

Dividing equation 4 by equation 5, we get,

$$\frac{n_i}{n} = \frac{N_i S_i / N\sqrt{\lambda c_i}}{\frac{1}{\sqrt{\lambda N}} \sum_{i=1}^{k} \frac{N_i S_i}{\sqrt{c_i}}}$$

$$\frac{n_i}{n} = \frac{\left(N_i S_i / \sqrt{\lambda c_i}\right)\left(\frac{1}{\sqrt{\lambda N}}\right)}{\frac{1}{\sqrt{\lambda N}} \sum_{i=1}^{k} \frac{N_i S_i}{\sqrt{c_i}}}$$

$$n_i = n \frac{\left(N_i S_i / \sqrt{\lambda c_i}\right)}{\sum_{i=1}^{k} \frac{N_i S_i}{\sqrt{c_i}}}$$

Theorem: 4.4 (Neyman;s Allocation)

In stratified random sampling the v($\bar{y}_{st}$) is minimum for a fixed total size of the sample 'n'.

$$n_i = \frac{nW_i S_i}{\sum_i^k W_i S_i} = \frac{nN_i S_i}{\sum_i^k N_i S_i} \qquad where, W_i = \frac{N_i}{N}$$

Proof:

Here the total sample size 'n' is fixed then we have to point 'n_i' such that $\text{var}\left(\bar{y}_{st}\right)$ is minimum subject to $c = \sum_i^k n_i \Rightarrow n$

The cost function is written as $c = \sum_{i}^{k} n_i - n$

w.k.t,

The variance of stratified random sampling

$$\text{var}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i (N_i - n_i)\frac{s_i^2}{n_i} \qquad \text{.......(2)}$$

Consider, the function,

$$\phi = v(\bar{y}_{st\ wor}) + \lambda c$$

$$\phi = \frac{1}{N^2}\sum_{i=1}^{k} N_i (N_i - n_i)\frac{s_i^2}{n_i} + \lambda\left(\sum_{i=1}^{k} n_i - n\right)$$

$$\phi = \frac{1}{N^2}\sum_{i=1}^{k}\frac{N_i S_i^{\ 2}}{n_i} + \frac{1}{N^2}\sum_{i=1}^{k}\frac{N_i S_i^{\ 2}}{} + \lambda\sum_{i=1}^{k} n_i - \lambda n$$

Differentiate w.r. to n_i and equate to 0.

i.e $\frac{\partial\phi}{\partial n_i} = 0$

$$\Rightarrow -\frac{1}{N^2}\sum_{i=1}^{k}\frac{N_i S_i^2}{n_i^2} + 0 + \lambda\sum_{i=1}^{k}(1) + 0 = 0$$

$$\Rightarrow -\frac{1}{N^2}\sum_{i=1}^{k}\frac{N_i S_i^2}{n_i^2} + \lambda\sum_{i=1}^{k}(1) = 0$$

$$\Rightarrow \sum_{i=1}^{k}\left[\frac{N_i S_i^2}{n_i^2 N^2} + \lambda\right] = 0$$

$$\Rightarrow \frac{N_i S_i^2}{n_i^2 N^2} = \lambda$$

$$\Rightarrow n_i^2 = \frac{N_i S_i^2}{\lambda N^2}$$

Taking square root on both sides,

$$\Rightarrow n_i = \frac{N_i S_i}{\sqrt{\lambda} N} \qquad \text{........(3)}$$

Here λ is fixed the total overall stratum.

$$\Rightarrow n = \sum_{i=1}^{k} n_i$$

$$\Rightarrow n = \sum_{i=1}^{k}\frac{N_i S_i}{N\sqrt{\lambda}} \qquad \text{........(4)}$$

Dividing equation 3 in by equation 4, we get

$$\frac{n_i}{n} = \frac{N_i S_i / N\sqrt{\lambda}}{\sum_{i=1}^{k} N_i S_i / N\sqrt{\lambda}} = \frac{N_i S_i}{\sum_{i=1}^{k} N_i S_i}$$

$$n_i = \frac{nN_iS_i}{\sum_{i=1}^{k} N_iS_i}$$

COMPARISON OF STRATIFIED RANDOM SAMPLING WITH SIMPLE RANDOM SAMPLING WITHOUT STRATIFICATION:

The comparative study of simple random sampling without stratification and stratified random sampling under different systems of allocation such as proportion allocation and Neyman's optimum allocation.

PROPORTIONAL ALLOCATION vs. SIMPLE RANDOM SAMPLING :

The variance of estimate of population mean in simple random sampling is given by,

$$\text{var}(\bar{y}_{st})_{srss} = \left(\frac{1}{n} - \frac{1}{N}\right)S^2$$

where, $S^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{k}\left(Y_{ij} - \bar{Y}_N\right)^2$

The variance of the estimate of the population mean in stratified random sampling with proportional allocation is given by,

$$\begin{aligned}
\text{var}(\bar{y}_{st})_{prop} &= \frac{1}{N^2}\sum_{i=1}^{k} N_i\left(N_i - n_i\right)\left(\frac{S_i^2}{n_i}\right) \\
&= \sum_{i=1}^{k}\frac{N_i}{N^2}\left(\frac{N_i - n_i}{n_i}\right)S_i^2 \\
&= \sum_{i=1}^{k}\frac{p_i}{N^2}\left(\frac{N_i}{n_i} - 1\right)S_i^2 \qquad \because\ p_i = \frac{N_i}{N} \\
&= \sum_{i=1}^{k}\frac{p_i}{N}\left(\frac{N}{n} - 1\right)S_i^2 \\
&= \frac{1}{N}\left(\frac{N}{n} - 1\right)\sum_{i=1}^{k} p_iS_i^2 \\
&= \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_iS_i^2 \qquad ....(2)
\end{aligned}$$

In order to compare equation 1 and 2 first express S^2 terms of S_i^2

$$\text{Let } S^2 = \frac{1}{N-1}\sum_{i=1}^{k} N_i\left(Y_{ij} - \bar{Y}_N\right)^2 \qquad ....(3)$$

Adding and subtracting $\bar{Y}_{Ni}$ in equation 3 we get,

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{N_i}\left(Y_{ij} - \bar{Y}_{Ni} + \bar{Y}_{Ni} - \bar{Y}_N\right)^2$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{N_i}\left[\left(Y_{ij} - \bar{Y}_{Ni}\right) + \left(\bar{Y}_{Ni} - \bar{Y}_N\right)\right]^2$$

$$S^2(N-1)=\sum_{i=1}^{k}\left[\sum_{j=1}^{N_i}\left(Y_{ij}-\overline{Y}_{ni}\right)^2+\sum_{i=1}^{k}\sum_{j=}^{N_i}\left(\overline{Y}_{Ni}-\overline{Y}_N\right)^2+2\sum_{i=1}^{k}\left(\overline{Y}_{Ni}-\overline{Y}_N\right)\sum_{j=1}^{N_i}\left(Y_{ij}-\overline{Y}_N\right)\right]$$

Since the algebraic sum of the deviations from mean is zero.

i.e., $\sum_{j=1}^{N_i}\left(Y_{ij}-\overline{Y}_{Ni}\right)=0$

Assume that N_i and N are sufficiently large so that take $N_i-1\cong N_i$ and $N-1\cong N$ then,

$$NS^2\cong\sum_{i=1}^{k}N_iS_i^2+\sum_{i=1}^{k}N_i\left(\overline{Y}_{Ni}-\overline{Y}_N\right)^2$$

$$S^2\cong\sum_{i=1}^{k}\frac{N_i}{N}S_i^2+\sum_{i=1}^{k}\frac{N_i}{N}\left(\overline{Y}_{Ni}-\overline{Y}_N\right)^2$$

$$S^2\cong\sum_{i=1}^{k}p_iS_i^2+\sum_{i=1}^{k}p_i\left(\overline{Y}_{Ni}-\overline{Y}_N\right)^2 \qquad \ldots\ldots\ldots(4)$$

Substituting eqn. 4 in eqn. 1, we get,

$$\operatorname{var}(\overline{y}_n)_{srs}\cong\left(\frac{1}{n}-\frac{1}{N}\right)\left[\sum_{i=1}^{k}p_iS_i^2+\sum_{i=1}^{k}p_i\left(\overline{Y}_{Ni}-\overline{Y}_N\right)^2\right]$$

$$\operatorname{var}(\overline{y}_n)_{srs}\cong\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{k}p_iS_i^2+\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{k}p_i\left(\overline{Y}_{Ni}-\overline{Y}_N\right)^2$$

$$\operatorname{var}(\overline{y}_n)_{srs}\cong\operatorname{var}(\overline{y}_{st})_{prop}+\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{k}p_i\left(\overline{Y}_{Ni}-\overline{Y}_N\right)^2$$

Since the finite population correlation ignored,

$$\operatorname{var}(\overline{y}_n)_{srs}\geq\operatorname{var}(\overline{y}_{st})_{prop}$$

Therefore the difference in the stratum means greater is the gain in precision stratified random sampling with proportional allocation over unstratified simple random sampling.

NEYMAN'S ALLOCATION vs. PROPORTIONAL ALLOCATION:

The variance of population mean in stratified random sampling with proportional allocation is,

$$\operatorname{var}(\overline{y}_{st})_{prop}=\left(\frac{1}{n}-\frac{1}{N}\right)$$

To compute the estimate of the variance of population mean in stratified random sampling with Neyman's optimum allocation

$$\operatorname{var}(\overline{y}_{st})_{prop}=\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{k}p_iS_i^2 \qquad \ldots\ldots\ldots(1)$$

$$\operatorname{var}(\overline{y}_{st})_{ney}=\frac{1}{N^2}\sum_{i=1}^{k}N_i\left(\frac{N_i}{n_i}-1\right)S_i^2 \qquad \ldots\ldots(2)$$

Substitute the optimum allocation sample size is,

$n_i = \dfrac{nN_iS_i}{\sum_{i=1}^{k} N_iS_i}$ in equation 2 we get,

$$\text{var}(\bar{y}_{st})_{ney} = \frac{1}{N^2}\sum_{i=1}^{k} N_i\left(\frac{N_i\sum_{i=1}^{k} N_iS_i}{nN_iS_i} - 1\right)S_i^2$$

$$= \frac{1}{N^2}\sum_{i=1}^{k}\left(\frac{N_{.i}^2 S_i}{nN_iS_i} - N_i\right)S_i^2$$

$$= \frac{1}{N^2}\sum_{i=1}^{k}\left(\frac{N_i^2}{n} - N_i\right)S_i^2$$

$$= \frac{1}{n}\sum_{i=1}^{k}\left(\frac{N_i^2S_i^2}{N^2} - \frac{N_iS_i^2}{N^2}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{k}\left(p_i^2S_i^2 - \frac{p_iS_i^2}{N}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{k} p_iS_i^{\;\;2} - \frac{1}{n}\sum_{i=1}^{k}(p_iS_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{k} p_i\left(S_i^2 - p_iS_i^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{k} p_i\left[S_i^2 - p_iS_i^2 + p_iS_i^2 - p_iS_i^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{k} p_i\left[S_i^2\, 2p_iS_i^2 + p_iS_i^2\right] \qquad \because \bar{S}_w = \sum_{i=1}^{k} p_iS_i$$

$$= \frac{1}{n}\sum_{i=1}^{k} p_i\left[S_i - \bar{S}_w\right] \qquad ....(4)$$

The RHS of equation is non-negative,

$$\therefore \text{var}(\bar{y}_{st})_{prop} - \text{var}(\bar{y}_{st})_{ney} \geq 0$$
$$\text{var}(\bar{y}_{st})_{prop} \geq \text{var}(\bar{y}_{st})_{ney}$$

NEYMAN'S OPTIMUM ALLOCATION VS SIMPLE RANDOM SAMPLING:

The variance of estimate of the population mean in stratified random sampling with neyman's optimum allocation.

$$\text{var}(\bar{y}_{st})_{ney} = \frac{1}{n}\sum_{i=1}^{k}(p_iS_i)^2 - \frac{1}{N}\sum_{i=1}^{k}\left(p_iS_i^2\right) \qquad ....(1)$$

The variance of estimate of population mean in sample random sampling without stratification,

$$\text{var}(\bar{y}_n)_{srs} = \left(\frac{1}{n} - \frac{1}{N}\right)\left(\sum_{i=1}^{k} p_i S_i^2 + \sum_{i=1}^{k} p_i \left(\bar{Y}_{Ni} - \bar{Y}_N\right)^2\right) \qquad ......(2)$$

Substitute eqn. 2 in by equation, we get

$$\text{var}(\bar{y}_{sn})_{srs} - \text{var}(\bar{y}_{st})_{ney} = \left(\frac{1}{n} - \frac{1}{N}\right)\left(\sum_{i=1}^{k} p_i S_i^2 + \sum_{i=1}^{k} p_i \left(\bar{Y}_{Ni} - \bar{Y}_N\right)^2\right)$$
$$- \left(\frac{1}{n}\sum_{i=1}^{k}(p_i S_i)^2 - \frac{1}{N}\sum_{i=1}^{k}\left(p_i S_i^2\right)\right)$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_i S_i^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_i \left(\bar{Y}_{Ni} - \bar{Y}_N\right)^2 - \frac{1}{n}\sum_{i=1}^{k}(p_i S_i)^2 + \frac{1}{N}\sum_{i=1}^{k}\left(p_i S_i^2\right)$$
$$= \frac{1}{n}\sum_{i=1}^{k} p_i S_i^2 - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_i \left(\bar{Y}_{Ni} - \bar{Y}_N\right)^2 - \frac{1}{n}\sum_{i=1}^{k}(p_i S_i)^2 + \frac{1}{N}\sum_{i=1}^{k}\left(p_i S_i^2\right)$$
$$= \frac{1}{n}\left[\sum_{i=1}^{k} p_i S_i^2 - \sum_{i=1}^{k} p_i S_i^2\right] + \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_i \left(\bar{Y}_{Ni} - \bar{Y}_N\right)^2$$
$$= \frac{1}{n}\sum_{i=1}^{k} p_i \left(S_i - \bar{S}_w\right)^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_i \left(\bar{Y}_{Ni} - \bar{Y}_N\right)^2 \qquad ....(3)$$

In equation 3 the RHS is non- negative then,

$$\text{var}(\bar{y}_n)_{srs} - \text{var}(\bar{y}_{st})_{ney} \geq 0$$
$$\text{var}(\bar{y}_n)_{srs} \geq \text{var}(\bar{y}_{st})_{ney}$$

Theorem: 4.5

In finite population correction is ignore such that

$$\text{var}_{srs} \geq \text{var}_{prop} \geq \text{var}_{ney}$$

w.k.t, the variance of the estimate of population mean in simple random sampling without stratification with,

w.k.t, $$\text{var}_{srs} = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} \qquad ....(1)$$

and the variance of the estimate of population mean in stratified random sampling without replacement.

Proof:

w.k.t, $$\text{var}_{srs} = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} \qquad ....(1)$$

and $$v(\bar{y}_{srswor}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i\left(N_i - n_i\right)\frac{S_i^2}{n_i} \qquad ....(2)$$

the population correction is ignored,

$$\therefore \text{var}_{srs} = \frac{S^2}{n} \qquad ....(3)$$

w.k.t, the allocation of sample size in stratified random sampling with proportion allocation is,

$$n_i = \frac{nN_i}{N}$$

Substitute n_i in equation 2 we get,

$$\text{var}(\bar{y}_{st})_{prop} = \frac{1}{N^2}\sum_{i=1}^{k} N_i\left(N_i - \frac{nN_i}{N}\right)\left(\frac{NS_i^2}{nN_i}\right)$$

$$\text{var}(\bar{y}_{st})_{prop} = \frac{1}{N}\sum_{i=1}^{k} N_i\left(1 - \frac{n}{N}\right)\left(\frac{S_i^2}{n}\right)$$

$$\text{var}(\bar{y}_{st})_{prop} = \frac{1}{N^2}\sum_{i=1}^{k} \frac{N_i S_i^2}{n} \qquad ....(4)$$

w.k.t, the allocation of sample size in stratified random sampling with neyman's optimum allocation is,

$$n_i = \frac{nN_iS_i}{\sum_{i}^{k} N_iS_i}$$

Substitute n_i in equation 2 we get,

$$\text{var}(\bar{y}_{st})_{ney} = \frac{1}{N^2}\sum_{i=1}^{k} N_i\left(N_i - \frac{nN_iS_i}{\sum_{i}^{k} N_iS_i}\right)\left(\frac{S_i^2\sum_{i=1}^{k} N_iS_i}{nN_iS_i}\right)$$

$$\text{var}(\bar{y}_{st})_{ney} = \frac{1}{N^2}\sum_{i=1}^{k} (N_iS_i)^2 \frac{\sum_{i}^{k} N_iS_i}{nN_iS_i} - \frac{1}{N^2}\sum_{i=1}^{k} N_i\left(\frac{nN_iS_i}{\sum_{i}^{k} N_iS_i} \times \frac{S_i^2\sum_{i=1}^{k} N_iS_i}{nN_iS_i}\right)$$

$$\text{var}(\bar{y}_{st})_{ney} = \frac{1}{N^2}\sum_{i=1}^{k} \frac{(N_iS_i)^3}{nN_iS_i} - \frac{1}{N^2}\sum_{i=1}^{k} N_iS_i^{\,2}$$

$$\text{var}(\bar{y}_{st})_{ney} = \frac{1}{N^2}\sum_{i=1}^{k} N_iS_i^{\,2}\left[\frac{N_i^2S_i}{nN_iS_i} - 1\right]$$

$$\text{var}(\bar{y}_{st})_{ney} = \frac{1}{nN^2}\sum_{i=1}^{k} N_i^{\,2}S_i^{\,2}\left[1 - \frac{n}{N_i}\right]$$

$$\text{var}(\bar{y}_{st})_{ney} = \frac{1}{nN^2}\left(\sum_{i=1}^{k} N_iS_i\right)^2 \qquad ....(5) \because f.p.c\ ignored$$

From equation 3, 4 and 5

$$\text{var}_{srs} \geq \text{var}_{prop} \geq \text{var}_{ney}$$

UNIT-V

SYSTEMATIC RANDOM SAMPLING

A sampling technique in which only the first unit is selected with the help of random numbers and the rest get selected automatically according to some pre-designed pattern is known as systematic random sampling. Systematic random sampling is also referred to briefly as systematic sampling. Suppose N units of the population are numbered from 1 to N in some order. Let N=nk, where n is the sample size and k is an integer, and a random number less than or equal to k be selected and every k^{th} unit thereafter. The resultant sample is called every k^{th} systematic sample and such a procedure termed linear systematic sampling. If N $\neq$ nk, and every k^{th} unit be included in a circular manner till the whole list is exhausted it will be called circular systematic sampling.

K possible systematic sample together with their means are given in the following table. Thus k rows of the table given the k systematic sample. The column of the above table are also sometimes referred to as k strata. From the above table the N units in the population occurs once one of the k sample and thus has an equal chance of being included in the sample. Since the probability of selecting the i^{th} sample $(i = 1,2,3..k)$ as the systematic is $\frac{1}{k}$.

We get,

$$E(\bar{y}_i) = \frac{1}{k}\sum_{i=1}^{k}\bar{y}_i = \bar{y}_{\bullet\bullet}.$$

Thus if N=nk the sample mean provide an unbiased estimate of the population mean.

The variance of systematic sampling is,

$$v(\bar{y}_{sys}) = \frac{1}{k}\sum_{i=1}^{k}\left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2$$

Theorem 5.1:

If N=nk the systematic sample mean $\bar{y}_{sys}$ is an unbiased estimate of the population mean $\bar{y}$

Proof:

w.k.t
$$E(\bar{y}_{sy}) = \frac{1}{k}\sum_{i=1}^{k}\bar{y}_i$$
$$= \frac{1}{k}\sum_{i=1}^{k}\left(\sum_{j=1}^{n}\frac{y_{ij}}{n}\right)$$
$$= \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}y_{ij}$$
$$= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}y_{ij} \qquad \because N = nk$$
$$= \bar{y}$$
$$E(\bar{y}_{sy}) = \bar{y}$$

Theorem 5.2:

In systematic sampling the variance of the sample is given by,

$$v(\bar{y}_{sys}) = (N-1)\frac{S^2}{N} - (n-1)\frac{S^2{}_{wsy}}{n}$$

Proof:

The population variance of the systematic sample is

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}\right)^2$$

Add & subtract by $\bar{y}_{i\bullet}$,

$$(N-1)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \bar{y}_{i\bullet} + \bar{y}_{i\bullet} - \bar{Y}\right)^2$$

$$(N-1)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \bar{y}_{i\bullet}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(\bar{y}_{i\bullet} - \bar{Y}\right)^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \bar{y}_{i\bullet}\right)\left(\bar{y}_{i\bullet} - \bar{Y}\right)$$

$$(N-1)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \bar{y}_{i\bullet}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(\bar{y}_{i\bullet} - \bar{Y}\right)^2$$

$$(N-1)S^2 = k(n-1)S^2{}_{wsy} + \sum_{i=1}^{k}\sum_{j=1}^{k}\left(\bar{y}_{i\bullet} - \bar{Y}\right)^2 \qquad \ldots\ldots\ldots(1)$$

w.k.t, $$v(\bar{y}) = \frac{1}{N}\sum_{i=1}^{k}(y_i - \bar{y})^2 \Rightarrow N(v(\bar{y})) = \sum_{i=1}^{k}(y_i - \bar{y})^2$$

now consider,

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

$$= nk\, v(\bar{y}_{i\bullet}) \qquad \ldots\ldots\ldots\ldots(2)$$

Substitute equation 2 in equation 1 we get,

$$(N-1)S^2 = k(n-1)S^2{}_{wsy} + nk\, v(\bar{y}_{i\bullet})$$

Divided on both sides by nk

$$\frac{(N-1)S^2}{nk} = \frac{k(n-1)}{nk}S^2{}_{wsy} + \frac{nk}{nk}v(\bar{y}_{i\bullet})$$

$$\frac{(N-1)}{nk}S^2 = \frac{k(n-1)}{nk}S^2{}_{wsy} + v(\bar{y}_{i\bullet})$$

$$v(\bar{y}_{i\bullet}) = \frac{(N-1)}{N}S^2 - \frac{(n-1)}{n}S^2{}_{wsy}$$

Hence, $$v(\bar{y}_{sys}) = \frac{(N-1)}{N}S^2 - \frac{(n-1)}{n}S^2{}_{wsy}$$

Theorem 5.3: Systematic sampling vs. simple random sampling.

The systematic sample is more precise then a simple random sample without replacement if the mean square within the systematic sample is larger than the population mean square. In other words systematic sampling will yield better results only if the units within the same sample are heterogeneous.

Proof:

w.k.t the variance of simple random sample is without replacement is given by,

$$v(\bar{y}_{srs}) = \left(\frac{N-n}{N}\right)\frac{S^2}{n}$$

Also we know that the variance of systematic sample is given by,

$$v(\bar{y}_{sys}) = (N-1)\frac{S^2}{N} - (n-1)\frac{S^2{}_{wsy}}{n}$$

For claiming the above statement the assumption in given by,

$$v(\bar{y}_{srs}) > v(\bar{y}_{sys})$$

$$\left(\frac{N-n}{N}\right)\frac{S^2}{n} > (N-1)\frac{S^2}{N} - (n-1)\frac{S^2{}_{wsy}}{n}$$

$$(n-1)\frac{S^2{}_{wsy}}{n} > (N-1)\frac{S^2}{N} - \left(\frac{N-n}{N}\right)\frac{S^2}{n}$$

$$(n-1)\frac{S^2{}_{wsy}}{n} > S^2\left(\frac{(N-1)}{N} - \frac{N-n}{Nn}\right)$$

$$(n-1)\frac{S^2{}_{wsy}}{n} > S^2\left(\frac{n(N-1)-(n-n)}{Nn}\right)$$

$$(n-1)\frac{S^2{}_{wsy}}{n} > S^2\left(\frac{nN-n-N+n}{Nn}\right)$$

$$(n-1)\frac{S^2{}_{wsy}}{n} > S^2\left(\frac{nN-N}{Nn}\right)$$

$$(n-1)\frac{S^2{}_{wsy}}{n} > S^2\left(\frac{N(n-1)}{Nn}\right)$$

$$S^2{}_{wsy} > S^2$$

Theorem 5.4:

$$\mathrm{var}(\bar{y}_{srs}) = \frac{nk-1}{nk}.\frac{S^2}{n}\{1+(n-1)\rho\}$$

Where ρ is the intra class correlation between the units of the same systematic sampling and is given by,

$$\rho = \frac{\sum_{i=1}^{k}\sum_{j\neq j'=1}^{n}(y_{ij}-\bar{y}_{\bullet\bullet})(y'_{ij}-\bar{y}_{\bullet\bullet})}{nk(n-1)\sigma^2}$$

$$\rho = \frac{\sum_{i=1}^{k}\sum_{j\neq j'=1}^{n}\left(y_{ij}-\bar{y}_{\bullet\bullet}\right)\left(y'_{ij}-\bar{y}_{\bullet\bullet}\right)}{(n-1)(n-1)s^2}$$

Since , $N\sigma^2 = (N-1)S^2$
$nk\sigma^2 = (nk-1)S^2$

Proof:

w.k.t,

$$v(\bar{y}_{srs}) = \frac{1}{k}\sum_{i=1}^{k}\left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2$$

$$v(\bar{y}_{srs}) = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{n}\sum_{j=1}^{n}y_{ij} - \frac{1}{n}\sum_{j=1}^{n}\bar{y}_{\bullet j}\right)^2$$

$$v(\bar{y}_{srs}) = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{n}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\bullet j}\right)\right)^2$$

$$v(\bar{y}_{srs}) = \frac{1}{n^2k}\sum_{i=1}^{k}\left(\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\bullet j}\right)\right)^2$$

$$n^2kv(\bar{y}_{srs}) = \sum_{i=1}^{k}\left(\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\bullet j}\right)\right)^2$$

$$n^2kv(\bar{y}_{srs}) = \sum_{i=1}^{k}\left(\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\bullet j}\right)\right) + \sum_{j\neq j=1}^{n}\left(y_{ij} - \bar{y}_{\bullet j}\right)\left(y'_{ij} - \bar{y}_{\bullet j}\right)$$

$$v(\bar{y}_{srs}) = (nk-1)S^2 + (n-1)(nk-1)S^2\rho$$

$$v(\bar{y}_{srs}) = (nk-1)S^2\left[1+(n-1)\rho\right]$$

$$v(\bar{y}_{srs}) = \frac{(nk-1)}{nk}.\frac{S^2}{n}\left[1+(n-1)\rho\right]$$

Remarks:

If $v(\bar{y}_{srs}) \geq 0$

$$\Rightarrow \frac{(nk-1)}{nk}.\frac{S^2}{n}\left[1+(n-1)\rho\right] \geq 0$$

$$\Rightarrow \left[1+(n-1)\rho\right] \geq 0$$

$$\Rightarrow \left[(n-1)\rho\right] \geq -1$$

$$\Rightarrow \rho \geq \frac{-1}{n-1}$$

Thus the minimum value of $\rho\, is \frac{-1}{n-1}$ when $v(\bar{y}_{srs}) = 0$

Theorem 5.5: Systemic sampling vs. Simple Random Sampling without Replacement.

The relative efficiency of the estimate of the population mean is systematic sampling over simple random sampling without replacement is given by,

$$E = \frac{\operatorname{var}(\bar{y}_{srswor})}{\operatorname{var}(\bar{y}_{sys})}$$

$$= \frac{\frac{N-n}{Nn}S^2}{\left[\frac{(nk-1)S^2}{n^2k}\{1+(n-1)\rho\}\right]}$$

$$= \frac{n(k-1)}{(nk-1)\{1+(n-1)\rho\}}$$

Obviously this depend on the value of ρ

If $\quad E > 1 \quad \Rightarrow \frac{n(k-1)}{n(k-1)\{1+(n-1)\rho\}} > 1$

$$\Rightarrow nk - n > nk - 1 + (nk-1)(n-1)\rho$$

$$\Rightarrow nk - n - nk + 1 > (nk-1)(n-1)\rho$$

$$\Rightarrow -(n-1) > (nk-1)(n-1)\rho$$

$$\Rightarrow -1 > (nk-1)\rho$$

$$\Rightarrow \frac{-1}{nk-1} > \rho$$

Thus systematic sampling would be more efficiency as compared with simple random sampling without replacement

If $\rho < \frac{1}{nk-1}$

On other can simple random sampling without replacement with the supervise to the systematic sampling $\rho > \frac{1}{nk-1}$

Theorem 5.6: Systematic sampling vs. Stratified Random Sampling

Let us now record the population of N=nk units to be divided into ‘n’ strata corresponding to the N columns and suppose that one units is drawn randomly from in each stratum thus stratified random sampling of size n the, mean of the j^{th} stratum, $\bar{y}_{\bullet j} = \frac{1}{k}\sum_{i=1}^{k} y_{ij}$

Population mean, $\bar{y}_{\bullet\bullet} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \frac{1}{n}\sum_{j=1}^{n}\bar{y}_{\bullet j}$

Stratum mean square, $S_j^2 = \frac{1}{N_j - 1}\sum_{i=1}^{k}\left(y_{ij} - \bar{y}_{\bullet j}\right)^2$

$$= \frac{1}{k-1}\sum_{i=1}^{k}\left(y_{ij} - \bar{y}_{\bullet j}\right)^2$$

$S^2{}_{wst}$ =pooled mean square between units within strata $= \dfrac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}$

ρ_{wst} in the correlation co-efficient between deviations from stratum means of pair of items that are in the same systematic sample.

Thus,

$$\rho_{wst} = \frac{E(y_{ij} - \bar{y}_{\bullet j})(y'_{ij} - \bar{y}'_{\bullet j})}{E(y_{ij} - \bar{y}_{\bullet j})}$$

$$\rho_{wst} = \frac{1}{k(n-1)} \frac{i = 1\sum^{k}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{\bullet j})(y'_{ij} - \bar{y}'_{\bullet j})}{\frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{\bullet j})}$$

$$\rho_{wst} = \frac{i = 1\sum^{k}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{\bullet j})(y'_{ij} - \bar{y}'_{\bullet j})}{(n-1)n(k-1)S^2_{wst}}$$

Theorem 5.7:

$$\operatorname{var}(\bar{y}_{sys}) = \frac{nk-1}{nk}.S^2_{wst}\{1 + (n-1)\rho_{wst}\}$$

Proof:

$$\operatorname{var}(\bar{y}_{sys}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i - \bar{y}_{\bullet\bullet})^2$$

$$\operatorname{var}(\bar{y}_{sys}) = \frac{1}{k}^2 \sum_{i=1}^{k}\left[\frac{1}{n}\sum_{j=1}^{n} y_{ij} - \frac{1}{n}\sum_{j=1}^{n}\bar{y}_{\bullet j}\right]^2$$

$$\operatorname{var}(\bar{y}_{sys}) = \frac{1}{n^2 k}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet j})\right]^2$$

$$\operatorname{var}(\bar{y}_{sys}) = \frac{1}{n^2 k}\left[\sum_{i=1}^{k}\sum_{j=1}^{k}(y_{ij} - \bar{y}_{ij})^2 + \sum_{i=1}^{k}\sum_{j \neq j=1}^{k}(y_{ij} - \bar{y}_{ij})(y'_{ij} - \bar{y}'_{ij})\right]$$

$$\operatorname{var}(\bar{y}_{sys}) = \frac{1}{n^2 k}\left[n(k-1)S^2{}_{wst} + n(n-1)(k-1)\rho_{wst} S^2_{wst}\right]$$

$$\operatorname{var}(\bar{y}_{sys}) = \frac{n(k-1)S^2{}_{wst}}{n^2 k}\left[1 + n(n-1)\rho_{wst}\right]$$

$$\operatorname{var}(\bar{y}_{sys}) = \frac{(k-1)}{nk}\left[1 + n(n-1)\rho_{wst}\right]$$

Remarks: Systematic sampling vs stratified random sampling $\operatorname{var}(\bar{y}_{st}) = \sum_{j=1}^{n}\left(\frac{1}{n_j} - \frac{1}{N_j}\right)p_i^2 S_i^2$ but $N_j = k \, and \, n_j = 1(j = 1,2,\ldots n) \,\&\, p_j = \frac{N_i}{N} = \frac{k}{nk} = \frac{1}{n}$

Therefore $\operatorname{var}(\bar{y}_{st}) = \sum_{j=1}^{n}\left(1 - \frac{1}{k}\right)\frac{1}{n^2}S_i^2$

$$\operatorname{var}(\bar{y}_{st}) = \left(1 - \frac{1}{k}\right)\frac{1}{n^2}\sum_{j=1}^{n} S_i^2$$

$$\operatorname{var}(\bar{y}_{st}) = \frac{k-1}{n^2 k}\sum_{j=1}^{n}\left[\frac{1}{k}\sum_{i=1}^{k}(y_{ij} - \bar{y}_{\bullet j})^2\right]$$

$$\operatorname{var}(\bar{y}_{st}) = \frac{1}{n^2 k}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet j})^2$$

$$\operatorname{var}(\bar{y}_{st}) = \frac{k-1}{nk} S_{wst}^2$$

The relative efficiency of the estimate of the population mean in systematic sampling stratified random sampling is given by,

$$E' = \frac{\operatorname{var}(\bar{y}_{st})}{\operatorname{var}(\bar{y}_{sys})} = \frac{\frac{k-1}{nk} S_{wst}^2}{\frac{(k-1)}{nk} S_{wst}^2 \{1 + (n-1)\rho\}} = \frac{1}{n + (n-1)\rho_{wst}}$$

Thus the relative efficiency of systematic sampling gives difference upon the values are ρ_{wst} and are nothing can be concluded in general. If $\rho_{wst} \geq 0$ then $E' < 1$ thus in this case stratified random sampling will provided better estimate of $\bar{y}_{\bullet\bullet}$ .if $\rho_{wst} = 0$ then $E' = 1$ and thus in the case both sampling provided estimates of $\bar{y}_{\bullet\bullet}$ with equal precision.

Limitation of systematic sampling merits:

- systematic sampling is operationally more convenience systematic sampling or stratified random sampling
- time and work involve in systematic sampling relatively much less.
- systematic sampling yields a sample which is every spread over the entire population.
- operational convince the job of collecting the systematic sample can be entrescisted to the field works.
- systematic sampling may be more efficient then simple random sampling provided the frame is arranged wholly at random.

Demerits:

- the main disadvantage of systematic sampling is that systematic sample are not in a general random sample since the requirement in merits to is rarely fulfilled.
- if N is not a multiple of n then as the actual sample size is different from that requirement.
- sample mean is not an unbiased estimate of the population mean.
- systematic sampling may yield highly estimates if there are periodic features associated with the sampling interval.

APPENDIX

I. Random Number Table

61 37 37 06 29 96 54 18 15 08 28 21 40 64 85 81 52 42 60 87 50 39 69 24 54 67 10 61 26 96

30 13 85 98 57 19 22 45 50 31 76 16 26 65 79 38 80 95 99 60 97 32 95 38 05 54 94 52 51 13

95 11 75 45 00 72 66 66 97 27 51 83 52 02 27 41 69 72 67 18 95 25 23 85 88 04 74 31 17 63

17 08 99 87 82 89 93 96 38 42 24 99 14 20 23 77 77 30 66 95 28 48 82 62 15 50 31 30 21 99

23 84 05 51 62 85 93 82 75 90 86 68 98 40 87 93 42 37 82 31 63 00 93 03 49 06 64 98 28 65

38 43 10 09 48 90 98 84 88 10 28 22 72 93 97 41 24 94 13 54 51 65 46 06 20 25 44 06 22 49

37 72 55 05 51 43 30 00 72 36 23 91 73 45 94 82 81 77 66 93 40 95 01 73 95 12 86 91 19 04

65 68 39 29 61 76 44 76 56 61 27 79 93 18 02 57 29 82 32 40 93 91 06 52 17 02 56 47 76 09

12 84 69 90 17 89 57 07 48 77 53 06 77 54 00 79 87 88 42 78 70 53 99 02 13 65 08 72 61 09

16 24 64 00 57 44 36 91 22 98 80 87 42 75 50 04 11 82 27 59 47 79 19 59 39 52 23 64 92 64

36 20 51 02 05 11 04 03 90 52 21 55 81 48 53 37 95 67 83 93 46 48 49 60 93 61 67 77 81 75

81 46 54 47 88 08 15 23 49 57 53 41 57 77 89 04 45 11 20 86 35 88 47 04 08 60 57 48 17 14

00 20 67 08 68 75 72 39 08 79 49 36 46 54 14 88 14 61 95 72 11 76 31 18 54 77 30 37 05 13

86 21 64 79 77 11 49 61 37 98 96 40 19 16 18 03 04 61 52 77 88 81 57 77 22 96 52 37 52 35

41 48 87 32 12 63 81 05 52 12 59 82 90 90 06 81 32 77 91 79 82 19 61 77 18 70 57 53 44 71

34 49 18 35 39 09 22 88 33 56 12 72 48 68 49 31 55 05 31 77 74 10 55 72 94 82 56 71 84 46

10 84 03 08 39 13 67 18 94 83 81 98 98 81 93 23 92 23 89 11 49 73 54 16 73 91 29 75 24 31

77 51 26 79 84 46 03 11 91 20 84 50 90 97 51 97 87 87 52 11 33 18 94 82 83 12 58 64 21 39

92 32 80 57 19 14 25 94 83 01 97 61 69 64 90 77 68 65 95 41 54 39 37 94 99 09 54 99 39 48

81 46 13 83 71 79 05 73 26 55 89 34 54 60 60 42 33 84 37 02 68 24 63 62 95 06 07 28 91 64

29 94 28 32 42 17 30 50 47 50 35 85 51 33 19 95 71 93 49 43 64 93 62 67 87 67 16 59 15 51

72 92 92 88 95 75 93 19 20 25 84 58 66 90 60 73 85 43 07 33 40 87 84 96 08 17 33 50 87 76

29 10 09 67 99 76 28 22 89 52 48 20 60 97 32 03 96 75 08 18 44 54 39 24 90 43 61 75 03 62

89 06 72 87 21 81 44 82 79 90 64 48 69 64 78 17 35 49 38 97 24 30 08 07 02 49 17 61 44 57

97 88 35 46 04 24 63 47 94 30 49 25 39 16 92 41 78 47 57 99 20 95 84 01 14 60 87 51 06 80

10 44 83 30 15 86 37 30 61 46 66 09 14 06 22 79 11 15 90 98 06 29 67 73 96 52 57 71 47 90

63 27 80 67 28 37 73 38 55 31 19 63 79 04 32 36 11 55 63 19 82 41 88 02 15 96 66 03 59 73

82 35 14 99 53 05 11 61 98 35 02 38 99 34 43 76 52 45 58 64 91 32 22 46 53 18 09 48 12 83

82 17 50 34 40 38 05 74 35 60 02 95 31 01 96 46 75 69 11 23 93 19 21 08 31 49 30 55 29 56

81 51 30 65 69 42 25 79 82 71 38 61 39 69 04 91 78 28 92 09 14 26 74 94 66 24 87 12 54 99

00 73 77 22 60 57 96 57 77 27 48 54 94 06 42 04 77 89 80 91 23 73 80 95 21 92 49 81 70 70

20 95 66 37 92 09 61 16 24 67 39 52 56 94 88 38 47 09 93 80 36 21 58 84 65 58 55 38 17 76

60 86 49 05 64 03 31 03 01 93 31 80 35 63 08 73 03 66 31 38 53 30 93 11 53 96 83 70 25 16

35 59 54 51 74 86 36 52 67 96 87 64 35 94 28 53 09 28 91 45 56 84 67 28 11 01 19 27 88 70

58 70 27 42 49 32 29 84 81 26 73 84 23 91 19 29 69 37 62 23 18 06 84 63 76 70 75 40 99 90

51 73 94 59 02 17 48 55 50 22 15 61 95 24 85 53 39 97 34 07 54 18 13 85 61 20 13 68 86 18

35 63 45 25 72 77 85 26 55 68 04 66 35 55 93 05 69 26 92 11 34 95 73 34 44 64 07 35 69 87

17 15 15 91 52 25 21 19 40 96 46 48 06 62 94 69 85 35 24 63 74 83 04 14 63 37 39 70 00 37

38 13 97 34 03 11 22 45 29 74 39 54 15 28 71 32 67 94 80 16 47 44 14 84 47 78 63 05 73 94

01 08 75 82 15 68 29 98 57 88 69 90 13 07 95 36 83 81 55 82 51 00 98 67 46 20 95 94 61 63

83 23 27 53 79 60 19 77 82 71 38 91 94 76 24 92 81 24 44 12 17 50 18 08 66 25 22 45 12 52

74 55 87 40 13 43 50 61 30 24 77 31 43 57 65 08 99 34 76 84 88 36 86 21 70 73 32 56 82 33

65 28 70 15 04 38 75 26 29 41 97 77 70 18 16 76 14 62 58 93 97 14 45 62 99 45 39 18 07 36

71 04 46 40 74 99 13 86 22 85 26 49 89 34 01 24 39 57 18 15 20 24 72 11 63 67 17 98 01 97

52 61 30 66 03 41 21 39 83 39 92 04 19 91 71 28 82 27 63 07 79 04 45 09 20 11 33 94 81 17

40 53 21 02 85 92 53 90 99 02 16 84 17 22 57 95 79 01 57 14 22 23 80 11 22 11 36 08 43 24

57 92 36 95 65 66 34 22 64 44 61 48 97 49 70 80 46 40 31 27 57 45 73 11 22 31 62 35 67 08

20 68 16 90 17 64 12 12 18 88 91 85 67 73 97 05 18 16 19 88 08 70 48 78 30 23 22 35 74 82

77 75 14 74 86 48 09 06 64 31 66 99 10 10 09 68 14 45 17 47 63 88 09 92 54 45 51 69 14 78

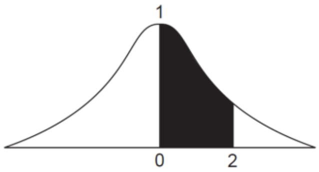
45 15 31 75 71 62 83 08 50 35 94 29 75 02 22 90 56 25 86 97 13 36 72 00 21 26 37 09 90 92

71 27 42 16 71 13 16 03 23 31 27 53 46 00 25 58 08 17 24 81

Note: Numbers are blocked in groups of two digits for convenience only. In using this table you can read numbers of any number of digits in any way you want.

II. Area under Normal Curve

An entry in the table is the proportion under the entire curve which is between z = 0 and a positive value of z. Areas for negative values for z are obtained by symmetry.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2903	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990