**Semester-V Statistics Paper-V**

 **Unit-I**

**Sampling Theory**

**(Design of sample surveys)**

**Introduction:**

 Already we learned some basics of sampling in the previous year. In this year we are going to discuss some other points and sampling designs. Even some basics are presented here.

 There are TWO types of surveys to collect the data.

* Complete Enumeration ( Census )
* Sample Survey

 In the former one, the data are collected from each and every unit of Population ( aggregate of objects under study ).

 In the later one, the data are collected from a portion of population (called as sample).

 In many practical issues, sample survey is advantageous over Census. In particular, in testing and especially in case of destructive it is impossible to conduct Census, and sample survey is only the way.

**Parameters**: The characteristics of the population are called parameters.

 Eg: Population Mean(µ), Population SD(σ), Population Proportion(P),etc

**Statistic**: The function or a measure of Sample observations is called a sample.

 Eg: Sample mean(x̅), Sample SD(s), Sample proportion(p), etc

**Sampling distribution**: The distribution of Statistic is called a sampling distribution.

* The average of the distribution of statistic is called Expectation.
* The standard deviation of the distribution of the statistic is called the **Standard Error(S.E)** of the statistic.

 Eg:

* S.E of x̅ = (σ/$\sqrt{n }$ ) ($\frac{N-n}{N-1}$)
* S.E of p = ($\sqrt{PQ/n }$ ) ($\frac{N-n}{N-1}$)

 Where N=Population Size, n= sample size

 Here $\frac{N-n}{N-1}$ is called finite population correction factor. It can be neglected for large populations.

**Sampling:**

 The process of selecting sample from a population is called sampling. There are different methods of Sampling.

**Methods of Sampling:**

 In this paper, we shall discuss first 3 random sampling methods.

**Principal Steps in a Sample Survey:**

The following are main steps involved in planning and execution of a sample survey.

1. **Objectives of the Survey**: The first and foremost step is to define the objectives of the survey in clear and concrete terms. The sponsoring agency of survey should take care that these objectives are corresponding with the available resources in terms of money, man power and time limit.
2. **Defining the Sampled Population**: The population from which sample is chosen should be defined in clear and unambiguous terms. Some practical difficulties in handling certain segments of population (specially boarder cases) may point to their elimination from the scope of survey. Thus for reasons of practicability or convenience the population to be sampled ( Sampled Population ) may be different than the population for which results are wanted ( Target Population ).
3. **Sampling Units and the Frame**: The units of sampled population are called sampling units. These must cover the entire sampled population and they must be distinct, non-overlapping and unambiguous so that every element of the population belons to only one sampling unit. For Eg, In socio-economic survey for selecting people in a city, the sampling unit might be an individual person, a family or a block in a locality.

 The list of sampling units is called a Frame. It may be in form of a map or any other acceptable material. The frame serves as a guide to the population to be covered. This frame should be accurate and dynamic.

1. **Selection of proper sampling design:** A number of designs (Probability sampling methods and Non-probability sampling methods) for selection of sample are available. A proper design should be selected considering efficiency, cost and time.
2. **Selection of method of collecting information**: A appropriate method should be selected for collecting information keeping accuracy, cost and time. A proper care should be taken in case of non-respondents.

Some important methods are, Interview method and Questionnaire method.

1. **Data to be collecte**d: The data should be collected keeping view the objectives of the survey. A practical method is to chalk out an outline of the tables that the survey should produce. This would help in elimination the collection of irrelevant and too may information and ensure that no essential data are omitted.
2. **Organization of Field work**: It is absolutely essential that the personnel should be well trained in locating sample units, recording the measurements and the methods of collection of required data before starting field work. The success of a survey to a great extent depends upon the reliable field work. It is necessary to inspect the field work after completion by supervisory staff.
3. **The Pre-test**: From practical point of view a small pre-test should be conducted. Here Pre-test means trying out the questionnaire and field methods on a small scale. It always helps to decide upon effective method of asking questions and disclose certain problems, otherwise it will be quite serious on a large scale sample survey.
4. **Summary and Analysis of the data**: The analysis of the data may be classified as,

i) **Scrutiny and editing of data:** An initial quality check should be carried out by the supervisory staff while the investigators are in the field. The scrutiny and editing of completed schedules and questionnaires will help in eliminating the erroneous and inconsistent data.

ii) **Tabulation of data:** Before carrying out the tabulation of data, we must decide the procedure of tabulation of the data which are incomplete due to non-response to certain items in questionnaire. For a large scale sample survey, advanced software tools will help to tabulate the data and these need coding for qualitative variables.

Iii) **Statistical Analysis**: After the data has been properly scrutinized, edited and tabulate, a very careful statistical analysis is to be made. Appropriate formulae should be used to provide unbiased and accurate results.

1. **Reporting and conclusions**: Finally a report incorporating detailed statement of the different stages of the survey should be prepared. In the presentation of the results, it is good practice to report the technical aspect of the design, i.e., the types of the estimators used along with expected error.
2. **Information gained for future surveys:** The information gained from any complicated survey and sample in the form of the data regarding means, SD’s ,costs, time, etc serve as a potential guide for future surveys. It serves as a lesson to the organisers for future surveys in recognising and rectifying the mistakes committed in the execution of the survey.

**Principles of Sampling:**

 The theory of sampling based on the following important principles.

1. **Principle of Statistical Regularity**: The principle has its origin in the theory of probability. According to King “ the law of statistical regularity lays down that a moderately **large numbe**r of items chosen at **random** from a large group are almost sure on the average to process the characteristics of the large group”.

An immediate derivation from the principle of statistical regularity is the **Principle of Inertia of Large Numbers**.

1. **Principle of Validity**: The principle of validity states that the sampling design enable us to obtain valid tests and estimates about the parameters of the population. The samples obtained by the techniques of probability sampling satisfy this principle.
2. **Principle of Optimization**: The principal of optimization states that the results of the design in terms of efficiency and cost with available resources is optimum. The principle of optimization consists in
* Achieving a given level of efficiency at minimum cost
* Obtaining maximum possible efficiency with a given level of cost.

**Scope of Errors (Sampling and Non-sampling Errors)**:

 The errors involved in the collection, processing and analysis of a data can be classified into two heads.

 i) Sampling Errors, and ii) Non-sampling Errors

**i) Sampling Errors:**

 The sampling errors arise due to the fact that only a portion of the population has been used to estimate the parameters and draw inferences about the population. Thus these errors are present in sample survey and absent in census survey.

Sampling errors are due to the following reasons:

1. **Faulty selection of the sample**: Some of bias is introduced by the use of improper sampling technique for the selection of sample. In most of situations, expect few, judgement sampling provides biased results. It can be overcome by using a random sampling.
2. **Substitution**: In general the investigators substitute a convenient unit in place the sampling unit which is difficult to investigate. This leads to some bias since the characteristics may be different from unit to unit.
3. **Faulty demarcation of sampling units:** There will be some bias due to defective demarcation of sampling units. It will be happen in most area surveys dealing with border lines.
4. **Improper choice of statistic:** a constant error due to improper selection of statistic for estimating population parameters.

For example, we know that sampling variance s2 is biased for estimating population variance σ2 where as ns2/(n-1) is unbiased.

**\*\* The sampling errors can be reduced by increasing the sample size, since the S.E is inversely proportional to the square root of sample size.**

**ii) Non-sampling Errors:**

 The non-sampling errors arise at the stages present in both the census and sample surveys. Non-sampling errors can occur at every stage of planning and execution of census and sample survey. These arise from the following factors.

1. **Faulty planning or Definitions**: The planning of the survey consists of definitions various objectives. Here the non-sampling errors can be due to:
2. Errors due to location of units, errors in recording measurements, errors due to ill-designed questionnaire, etc.
3. Data specification being inadequate and inconsistent
4. Lack of trained and qualified personnel.
5. **Response Errors**: These errors resulting from responses may be due to the following reasons.
* May be accidental
* Due to the prestige of informant (it will upgrade or downgrade the response)
* Self interest of informant
* Bias due to interviewer
1. **Non-response Bias**: It occurs if full information is not obtained on all the selected sampling units
2. **Compiling Errors**: The various operations of data such as editing, coding, tabulation, etc are potential sources of error. These can be control through verification, consistency check, etc.
3. **Publication Errors**: The publication errors ( errors committed during presentation and printing of results) due to the two sources – mechanics of publication and failure of survey organisation to point out the limitations of statistics.

 \*\* **The non-sampling errors can reduced by assigning trained, experienced and skilled personnel.**

 \*\* **The data obtained in a complete census, although free from sampling errors, would still be subject to non-sampling errors where as data obtained in a sample survey would be subject to both sampling and non-sampling errors**.

**Advantages of Sampling over Census:**

1. **Less Time**: There is considerable saving in time and labour since only a portion of population has to be examined. At the same time, results can be obtained rapidly and analyzed much faster.
2. **Reduction in cost**: Sampling usually results in reduction in cost in terms of money and man hours. Since in most of cases our resources are limited in terms of money and the time, sampling is more advantageous than census.
3. **Greater Accuracy of results**: The results of a sample survey are usually much more reliable than those obtained from a complete census due to the following reasons:
* It is possible to determine the extent of the sampling error
* Scope of non-sampling errors is less in sampling compared with census
1. **Greater Scope**: The complete census is impracticable if the survey requires a highly trained personnel and more sophisticated equipment for collection and analysis of data. It is possible to have a thorough and intensive enquiry because a more detailed information can be obtained from a small group of respondents.
2. If Population is too large, , if testing is destructive, if population is hypothetical, sampling only the way.

**Simple Random Sampling:**

 If sample is drawn from a population such that each and every unit of the population has equal and independent chance of being included in the sample, then the sampling is called a **Simple Random Sampling** or simply **Random Sampling**. The sample so obtained is called a **Simple Random Sample**

It is TWO types

* Simple Random Sampling Without Replacement (SRSWOR)
* Simple Random Sampling With Replacement (SRSWR)

**SRSWOR**: In SRSWOR, units are drawn from the population one by one by without replacing the selected units in the previous draw in to the population before the next draw giving equal chance.

If Population size is ‘N’ and Sample size is ‘n’,

 the number possible samples by SRSWOR is $\left(\genfrac{}{}{0pt}{}{N}{n}\right)$

 \*\* The probability of each sample is 1/$\left(\genfrac{}{}{0pt}{}{N}{n}\right)$

1. In SRSWOR, the probability of selecting a specified unit in to the sample at any draw is equal to the probability of selecting it at first draw, i.e., 1/N.

Proof:

Let Ei = the event of selecting a specified unit in to the sample at ith draw

=> P(E1) = 1/N, P(E2/E̅1) = 1/(N-1),

 P(E3/(E̅1⋂E̅2) = 1/(N-2), ..... P(Ei/(E̅1⋂..E̅i-1)) = 1/(N-i),....

Now. P(Ei) = P(E̅1⋂E̅2⋂E̅3⋂......E̅i-1⋂Ei)

By multiplication theorem of probability,

 = P(E̅1).P(E̅2/ E̅1).P(E̅3/ (E̅1⋂E̅2))........P(Ei/(E̅1⋂E̅2⋂E̅3⋂......E̅i-1))

 = (1-1/N). (1-1/(N-1)). (1-1/(N-2).........1/(N-i)

 = (N-1)/N. (N-2)/(N-1). (N-3)/(N-2)......1/(N-i)

 = 1/N = P(E1)

1. In SRSWOR, the probability of a specified unit being included in to the sample is n/N.

Proof:

P(a specified unit being included in the sample) = P(E1UE2UE3U.....UEn)

 = P(E1)+P(E2)+....+P(En)

 (Since E1,E2,...En are exclusive)

 = 1/N+1/N+.....+1/N (n times)

 = n/N

**SRSWR**: In SRSWR, units are drawn from the population one by one by replacing the selected units in the previous draw in to the population before the next draw giving equal chance.

If Population size is ‘N’ and Sample size is ‘n’,

 the number possible samples by SRSWOR is Nn

\*\* The probability of each sample is 1/ Nn

\*\* The probability of a specified unit being included in the sample is n/N

**Methods of selecting simple random sample:**

* Lottery Method
* Random Numbers Method or Mechanical Randomization Method

**Lottery Method**: The simplest method of selecting a random sample is the Lottery method. The following are general steps in any lottery system to select ‘n’ units out of ‘N’ units

* Assigning numbers 1,2,....,N to the N units of sampled population.
* Preparing ‘N’ identical (in shape, size, colour, etc) slips/cards one per each number.
* Putting the slips in a bag or box and shuffling thoroughly.
* Now drawing ‘n’ slips from the bag one by one by WOR or WR
* The ‘n’ units corresponding to the drawn ‘n’ slips constitute a Random Sample.

\*\* This is one the most reliable methods of selecting random sample and is independent of the properties of the population.

\*\* This method is quite difficult and time consuming in case of large population. In this case an alternative method is Random Numbers method.

**Random Numbers Method**: This is most practical and inexpensive method for selecting a random sample, which consists in use of **Random Number Tables**. These tables have been constructed that each of the digits 0, 1, 2,...,9 appear with approximately the same frequency and independent of each other.

The method of drawing the random sample consists in the following steps:

* Assigning numbers 1,2,....,N to the N units of sampled population.
* Selecting any page of the Random Number Tables at random and picking the numbers in a row or column or diagonal at random.
* The population units corresponding to the numbers selected in the above step constitute the random sample.

The following are different sets of random number tables commonly used in practice.

* Tippet’s (1927) Random Number Tables – consisting of 10,400 four digit numbers ( 10,440 X 4 = 41,600 digits)
* Fisher and Yates (1938) Tables – 15,000 digits arranged in 7,500 two digit numbers.
* Kendall and Babington Smith (1939) – 1,00,000 digits grouped into 25,000 four digit numbers .
* Rand Corporation (1955) random number tables – 10,00,000 digits grouped into 2,00,000 five digit numbers.

**Theorems:**

**Notations:**

 Population Sample

N = Population Size n = Sample Size

Yi = The measurement of ith unit of population yi = The measurement of ith unit of sample

Y̅N = Population Mean = $\frac{\sum\_{}^{}Yi}{N}$ y̅n = Sample Mean = $\frac{\sum\_{}^{}yi}{n}$

S2 = Population Mean Square s2 = Sample Mean Square

 = $\frac{\sum\_{}^{}(Yi-\overbar{Y})^{2}}{N-1}$ or $\frac{1}{N-1} (\sum\_{}^{}Yi^{2}-N$ $(\overbar{Y})^{ 2}$ ) = $\frac{\sum\_{}^{}(yi-\overbar{y})^{2}}{n-1}$ or $\frac{1}{n-1}( \sum\_{}^{}yi^{2}-n $ $(\overbar{y})^{ 2}$)

σ2 = Population Variance f = n/N = sampling fraction

 = $\frac{\sum\_{}^{}(Yi-\overbar{Y})^{2}}{N}$ or $\frac{1}{N} \sum\_{}^{}Yi^{2}-$ $(\overbar{Y})^{ 2}$

1. **In SRSWOR, the sample mean is an unbiased estimate of the population mean.**

Proof:

 To prove, E(y̅n) = Y̅N

 Consider, E(y̅n) = E($\frac{\sum\_{}^{}yi}{n} $) = $\frac{E(\sum\_{i=1}^{n}yi)}{n}$ ....... (1)

 Let ai = 1, if ith unit is included in the sample ( i= 1,2,... N)

 0, if ith unit is not included in the sample (by WOR)

Now we can write, $\sum\_{i=1}^{n}yi$ = $\sum\_{i=1}^{N}aiYi$ ........(2)

 and we know that ai ~ Bernoulli dist with p=n/N

 => E(ai) = n/N .......(3)

Substituting (2) and (3) in (1),

E(y̅n) = $\frac{E(\sum\_{i=1}^{n}yi)}{n}$ = $\frac{E(\sum\_{i=1}^{N}aiYi)}{n}$ = $\frac{\sum\_{i=1}^{N}E(aiYi)}{n}$ = $\frac{\sum\_{i=1}^{N}E(ai)Yi}{n}$

 = $\frac{\sum\_{i=1}^{N}(\frac{n}{N})Yi)}{n}$ = $\frac{\sum\_{i=1}^{N}Yi)}{N}$ = Y̅N

 Hence, the sample mean is an unbiased estimate of population mean.

Note: The unbiased estimate of Population total (YN) = $\hat{Y}$ = N y̅n

1. **In SRSWOR, the sample mean square is an unbiased estimate of the population mean square**.

Proof:

 To prove, E(s2) = S2

 Consider, E(s2) = E($\frac{1}{n-1}( \sum\_{}^{}yi^{2}-n $ $(\overbar{y})^{ 2}$))

 = $\frac{1}{n-1}E( \sum\_{}^{}yi^{2}-n $ $(\frac{\sum\_{}^{}yi}{n})^{ 2}$)

 = $\frac{1}{n-1}E( \sum\_{}^{}yi^{2}- $ $\frac{(\sum\_{}^{}yi)}{n}^{ 2}$)

 = $\frac{1}{n-1}E( \sum\_{}^{}yi^{2}- $ $\frac{\sum\_{}^{}yi^{2}+2\sum\_{}^{}\sum\_{}^{}yiyj}{n}^{ }$)

 = $\frac{1}{n-1}E( (n-1)/n\sum\_{}^{}yi^{2}- $ $\frac{2\sum\_{}^{}\sum\_{}^{}yiyj}{n}^{ }$)

 E(s2) = $\frac{1}{n}E( \sum\_{}^{}yi^{2})- $ $\frac{E(2\sum\_{}^{}\sum\_{}^{}yiyj)}{n(n-1)}^{ }$ ........ (1)

Let ai = 1, if ith unit is included in the sample

 0, if ith unit is not included in the sample ( i= 1,2,... N)

and

 Let aiaj = 1, if both ith and jth units are included in the sample

 0, otherwise ( i‡j= 1,2,... N)

Now we can write, $\sum\_{}^{}yi^{2}$= $\sum\_{}^{}aiYi^{2}$ and

 $\sum\_{}^{}\sum\_{}^{}yiyj$ = $\sum\_{}^{}\sum\_{}^{}aiajYiYj$ ........ (2)

 and we know that ai ~ Bernoulli dist with p=n/N

 => E(ai) = n/N and E(aiaj) = n(n-1)/N(N-1).......(3)

Substituting (2) and (3) in (1),

E(s2) = $\frac{1}{n}E( \sum\_{}^{}yi^{2})- $ $\frac{E(2\sum\_{}^{}\sum\_{}^{}yiyj)}{n(n-1)}^{ }$

 = $\frac{1}{n}E( \sum\_{}^{}aiYi^{2})- $ $\frac{E(2\sum\_{}^{}\sum\_{}^{}aiajYiYj)}{n(n-1)}^{ }$

 = $\frac{1}{n} \sum\_{}^{}E(ai)Yi^{2}- $ $\frac{2\sum\_{}^{}\sum\_{}^{}E(aiaj)YiYj}{n(n-1)}^{ }$

 = $\frac{1}{n} \sum\_{}^{}\left(\frac{n}{N}\right)Yi^{2}- $ $\frac{2\sum\_{}^{}\sum\_{}^{}\left(\frac{n(n-1)}{N(N-1)}\right)YiYj}{n(n-1)}^{ }$

 = $\frac{1}{N} \sum\_{}^{}Yi^{2}- $ $\frac{2\sum\_{}^{}\sum\_{}^{}YiYj}{N(N-1)}^{ }$

 = $\frac{1}{N} \sum\_{}^{}Yi^{2}- $ $\frac{(\sum\_{}^{}Yi)^{2}-\sum\_{}^{}Yi^{2}}{N(N-1)}^{ }$

 = $\frac{1}{N-1} \sum\_{}^{}Yi^{2}- $ $\frac{(\sum\_{}^{}Yi)^{2}}{N(N-1)}^{ }$

 = $\frac{1}{N-1}( \sum\_{}^{}Yi^{2}- $ $\frac{(\sum\_{}^{}Yi)}{N}^{ 2}$)

 = $\frac{1}{N-1}( \sum\_{}^{}yi^{2}-N $ $(\frac{\sum\_{}^{}Yi}{N})^{ 2}$)

 = $\frac{1}{N-1}( \sum\_{}^{}yi^{2}-N $ $(Y̅)^{ 2}$)

 E(s2) = S2

Hence, s2 is an unbiased estimate of S2

1. **In SRSWOR, the variance of the estimate of the population mean is,**

 **V(y̅n) =** $\frac{N-n}{nN} $ **S2 or (**$\frac{1}{n}-\frac{1}{N}) $ **S2 or (1-f ) S2/n**

Proof:

Consider, V(y̅) = E(y̅2) – (E(y̅))2

 = E($\frac{\sum\_{}^{}yi}{n})$2 - Y̅N2

 = $E$ $\frac{(\sum\_{}^{}yi)}{n^{2}}^{ 2}$ - Y̅N2

 = $\frac{1}{n^{2}}E$ $(\sum\_{}^{}yi^{2}+2\sum\_{}^{}\sum\_{}^{}yiyj$) - Y̅N2 ........ (1)

Let ai = 1, if ith unit is included in the sample

 0, if ith unit is not included in the sample ( i= 1,2,... N)

and

 Let aiaj = 1, if both ith and jth units are included in the sample

 0, otherwise ( i‡j= 1,2,... N)

Now we can write, $\sum\_{}^{}yi^{2}$= $\sum\_{}^{}aiYi^{2}$ and

 $\sum\_{}^{}\sum\_{}^{}yiyj$ = $\sum\_{}^{}\sum\_{}^{}aiajYiYj$ ........ (2)

 and we know that ai ~ Bernoulli dist with p=n/N

 => E(ai) = n/N and E(aiaj) = n(n-1)/N(N-1).......(3)

Substituting (2) and (3) in (1),

V(y̅) = $\frac{1}{n^{2}}E$ $(\sum\_{}^{}yi^{2}+2\sum\_{}^{}\sum\_{}^{}yiyj$) - Y̅N2

 = $\frac{1}{n^{2}}E$ $(\sum\_{}^{}aiYi^{2}+2\sum\_{}^{}\sum\_{}^{}aiajYiYj$) - Y̅2

 = $\frac{1}{n^{2}}(\sum\_{}^{}E(ai)Yi^{2}+2\sum\_{}^{}\sum\_{}^{}E(aiaj)YiYj$) - Y̅2

 = $\frac{1}{n^{2}}(\sum\_{}^{}(\frac{n}{N})Yi^{2}+2\sum\_{}^{}\sum\_{}^{}\frac{n(n-1)}{N(N-1)}YiYj$) - Y̅2

 = $\frac{1}{nN} \sum\_{}^{}Yi^{2}+ $ $\frac{(n-1)2\sum\_{}^{}\sum\_{}^{}YiYj}{nN(N-1)}^{ }$ - Y̅2

 = $\frac{1}{nN} \sum\_{}^{}Yi^{2}+ $ $\frac{\left(n-1\right)((\sum\_{}^{}Yi)^{2}-\sum\_{}^{}Yi^{2})}{nN(N-1)}^{ }$ - Y̅2

 = $\frac{N-n}{nN(N-1)} \sum\_{}^{}Yi^{2}+ $ $\frac{(n-1)(NY̅)^{2}}{nN(N-1)}^{ }$ - Y̅2

 = $\frac{N-n}{nN(N-1)} \sum\_{}^{}Yi^{2}+ $ $\frac{(n-1)NY̅^{2}}{n(N-1)}^{ }$ - Y̅2

 = $\frac{N-n}{nN(N-1)} \sum\_{}^{}Yi^{2}+ $ $(\frac{(n-1)N}{n(N-1)}^{ }$ -1) Y̅2

 = $\frac{N-n}{nN(N-1)} \sum\_{}^{}Yi^{2}+ $ $(\frac{n-N}{n(N-1)}^{ }$) Y̅2

 = $\frac{N-n}{nN(N-1)} [\sum\_{}^{}Yi^{2}$ - NY̅2 ]

V(y̅n) = $\frac{N-n}{nN} $ S2

Note: The variance of the estimate of Population total = V($\hat{Y}$) = V(Ny̅n) = N2 V(y̅n)

1. In SRSWR, the sample mean is an unbiased estimate of the population mean.

Proof:

To prove, E(y̅n) = Y̅N

 Consider, E(y̅n) = E($\frac{\sum\_{}^{}yi}{n} $) = $\frac{E(\sum\_{i=1}^{n}yi)}{n}$

 = $\frac{\sum\_{i=1}^{n}E(yi)}{n}$ ..... (1)

Since each y1, y2, y3,....,yn are drawn by WR from population with mean Y̅ and variance σ2 , y1,y2,...,yn are independent and E(yi)= Y̅ and V(yi) = σ2

 From (1), E(y̅n) = $\frac{\sum\_{i=1}^{n}E(yi)}{n}$ = $\frac{\sum\_{i=1}^{n}Y̅}{n}$ = n Y̅/n = Y̅

 Hence, the sample mean is an unbiased estimate of population mean.

1. In SRSWR, the variance of the estimate of the population mean is,

 V(y̅n) = $\frac{N-1}{nN} $ S2 or σ2/n

Proof:

Since each y1, y2, y3,....,yn are drawn by WR from population with mean Y̅ and variance σ2 , y1,y2,...,yn are independent and E(yi)= Y̅ and V(yi) = σ2

 V(y̅n) = V($\frac{\sum\_{}^{}yi}{n}$) = $\frac{\sum\_{}^{}V(yi)}{n^{2}}$ = $\frac{\sum\_{}^{}σ^{2}}{n^{2}}$ = n σ2/n2

 = σ2/n

 = $\frac{N-1}{nN} $ S2 ( since N σ2 = (N-1) S2)

1. SRSWOR is more efficient than SRSWR. i.e., V(y̅)WOR ≤ V(y̅)WR

 Proof:

V(y̅)WOR = $\frac{N-n}{nN} $ S2

V(y̅)WR = $\frac{N-1}{nN} $ S2

 Clearly , N-n ≤ N-1

 => $\frac{N-n}{nN} $ S2 ≤ $\frac{N-1}{nN} $ S2 ( since all coefficients are non-negative)

 => V(y̅)WOR ≤ V(y̅)WR

 Hence SRSWOR is efficient than SRSWR.

**Merits and Limitations of Simple Random Sampling:**

**Merits:**

1. Since the sample is selected at random giving each unit an equal chance of being selected, personal bias is completely eliminated.
2. Simple Random Sample is more presentative of the population compared with non-random sample
3. This is very simple and efficient compared to other random sampling methods when the population is small and homogeneous.
4. As sample size ‘n’ increases, it provides more efficient estimates for the population parameters.

**Limitations:**

1. The selection simple random sample requires a complete up-to-date frame. Practically it may not possible to identity the all units before sample is drawn.
2. It requires more time and money for selecting the units that spread widely.
3. It usually requires large sample size for efficient estimates compared to the stratified random sampling.
4. It is inefficient if the population is heterogeneous.

**Problem:**

Suppose that a population consists of 6 units with measurements 2, 4, 6, 8, 10 and 12. Write all the possible samples of size 2 by without replacement from the population and verify

1. The sample mean is unbiased estimate of the population
2. The sample mean square is unbiased estimate if the population mean square

Also calculate the sampling variance of the estimate, sample mean and verify

1. It with the formula of variance
2. SRSWOR is efficient than SRSWR and find the gain in efficiency.

**Solution:**

Given, N = 6

 Yi : 2, 4, 6, 8, 10, 12

 => Population mean, Y̅ = $\frac{\sum\_{}^{}Yi}{N}$ = $\frac{2+4+6+8+10+12}{6}$ = 7

 and Population mean square, S2 = $\frac{1}{N-1} (\sum\_{}^{}Yi^{2}-N$ $(\overbar{Y})^{ 2}$ )

 = $\frac{1}{6-1} $( 22+42+62+82+102+122 – 6(7)2)

 = $\frac{1}{5}$ (70) = 14

also given n = 2

 => the no. of possible samples of size 2 from 6 units by WOR = 6c2 = 15

The list of samples and verification of the bits (a), (b) and (i), (ii) are given below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample No. | Sample values (n=2)(y1, y2) | Sample meany̅= (y1+y2)/2 | (y̅)2 | Sample mean squares2 = $\frac{1}{n-1}( \sum\_{}^{}yi^{2}-n $ $(\overbar{y})^{ 2}$)s2 = y12+y22 – 2(y̅)2 |
| 1 | 2, 4 | 3 | 9 | 2 |
| 2 | 2, 6 | 4 | 16 | 8 |
| 3 | 2, 8 | 5 | 25 | 18 |
| 4 | 2, 10 | 6 | 36 | 32 |
| 5 | 2, 12 | 7 | 49 | 50 |
| 6 | 4, 6 | 5 | 25 | 2 |
| 7 | 4, 8 | 6 | 36 | 8 |
| 8 | 4, 10 | 7 | 49 | 18 |
| 9 | 4, 12 | 8 | 64 | 32 |
| 10 | 6, 8 | 7 | 49 | 2 |
| 11 | 6, 10 | 8 | 64 | 8 |
| 12 | 6, 12 | 9 | 81 | 18 |
| 13 | 8, 10 | 9 | 81 | 2 |
| 14 | 8, 12 | 10 | 100 | 8 |
| 15 | 10. 12 | 11 | 121 | 2 |
| **Total** | **105** | **805** | **210** |

a) Sampling mean, E(y̅) = $\frac{\sum\_{}^{}y̅}{15}$ = 105/15 = 7 = Y̅

 Therefore, sample mean is an unbiased estimate of population mean

b) E(s2) = $\frac{\sum\_{}^{}s2}{15}$ = 210/15 = 14 = S2

Therefore, sample mean square is an unbiased estimate of population mean square

 and Sampling variance, V(y̅) = $\frac{\sum\_{}^{}y̅2}{15}$ - Y̅2 = 805/15 – (7)2 = 4.6667

i) by formula, V(y̅)WOR = $\frac{N-n}{nN} $ S2 = (6-2)x14/(2x6) = 56/12 = 4.6667

 Therefore, sampling variance agrees with formula

ii) V(y̅)WR = $\frac{N-1}{nN} $ S2 = (6-1)x14/(2x6) = 70/12 = 5.8333

 V(y̅)WOR = 4.6667 < V(y̅)WR = 5.8333

Therefore, SRSWOR is efficient than SRSWR

 The Efficiency = 5.8333/4.6667 = 1.25

The gain in efficiency = 25 %

i.e., for the given figures, SRSWOR is 25% more efficient than SRSWOR

**Self Assessment Questions:**

 **Multiple Choice Questions (MCQ):**

1. If information is collected from each and every unit of population then the survey is

called ( )

a) sample survey b) census survey c) pilot survey d) none

1. In sample survey the information is collected from ( )

a) every unit of the population b) few selected units of the population

c) both d) none

1. Sampling frame is a list of ( )

a) Voters b) Random numbers c) Sampling units of populationd) None

1. Which of the following method is used to collect data ( )

a) Interview method b) Questionnaire method c) schedule methodd) all

1. Which of the following is not a principal step of sample survey ( )

a) preparation of frame b) summary and analysis of data

c) validity of sampling design d) information gained for future survey

1. Which of the following is the principle of sample survey ( )
a) principle of statistical regularity b) principle of validity

c) principle of optimization d) all

1. Sampling errors occurred in ( )

a) sample survey b) census survey c) both d) none

1. Non-sampling errors occurred in ( )

a) sample survey b) census survey c) both d) none

1. The reason of sampling error is ( )

a) improper selection of sample b) improper selection of statistics for estimation

c) faulty identification of sampling units d) all

1. Which of the following are the factors of non-sampling errors ( )

a) respondents b) non-respondents c) compiling of data d) all

1. Sampling errors can be reduced by ( )

a) increasing sample size b) decreasing sample size

c) controlling response errors d) controlling errors in publishing

1. Non-Sampling errors can be reduced by ( )

a) increasing sample size b) decreasing sample size

c) controlling response errors d) assigning qualified, trained and experienced personal

1. Which of the following is advantage of sampling over census ( )

a) less time and cost b) accuracy of results

c) greater scope d) all

1. Which of the following is not non-probability sampling ( )

a) purposive sampling b) quota sampling

c) area sampling d) snowball sampling

1. Which of the following is a probability sampling ( )

a) random sampling b) systematic sampling

c) cluster sampling d) all

1. If the sample is selected by giving equal importance to each and every unit of

the population then the method is known as ( )

a) simple random sampling b) systematic sampling

c) judgement sampling d) quota sampling

1. The no. of possible samples of size 3 that can be drawn from a population of size 10 by

SRSWOR is ( )

a) 10 x 3 b) 10C3 c) 103 d) 310

1. The no. of possible samples of size 3 that can be drawn from a population of size 10 by

SRSWR is ( )

a) 10 x 3 b) 10C3 c) 103 d) 310

1. In SRSWOR, the probability of selecting a specified unit at ith (i=1,2,..n) draw is equal to ( )

a) 1/N b) n/N c) 1/(N-i+1) d) N/n

1. In SRSWOR, the probability of including a specified unit in to the sample is ( )

a) 1/N b) n/N c) 1/(N-1) d) N/n

1. Which of the following method is used to draw a random sample ( )

a) lottery method b) random numbers method

c) roulette wheel d) All of these

1. If N=population size , n=sample size , y̅=sample mean, Y̅=population mean, then by simple

random sampling the estimate of population total $\hat{Y}$ is ( )

a) ny̅ b) Ny̅ c) y̅ d) Y̅

1. If S2 is population mean square, in SRSWOR, the variance of sample mean is ( )

a) V(y̅)= $\frac{N-n}{Nn}$ S2 b) V(y̅)= $\frac{n-N}{Nn}$ S2 c) V(y̅)= $\frac{N-n}{(N-1)n}$ S2 d) V(y̅)= $\frac{n-N}{(N-1)n}$ S2

1. If S2 is population mean square, in SRSWR, the variance of sample mean is ( )

a) V(y̅)= $\frac{N-n}{Nn}$ S2 b) V(y̅)= $\frac{n-N}{Nn}$ S2 c) V(y̅)= $\frac{N-1}{Nn}$ S2 d) V(y̅)= $\frac{N}{(N-1)n}$ S2

1. Sampling fraction is ( )

a) N/n b) n/N c) 1/N d) 1/n

1. Which of the following is always true ( )

a) V(y̅WOR) = V(y̅WR) b) V(y̅WOR)≥ V(y̅WR)

c) V(y̅WOR)≤ V(y̅WR) d) None

**Descriptive Questions:**

 **SHORT QUESTIONS:**

1. Explain the principles of sample survey.
2. Write the difference between sampling versus census.
3. Explain the types of sampling.

**ESSAY QUESTIONS:**

1. Explain the principal steps involved in conducting a large scale sample survey.
2. Explain sampling errors and non sampling errors.
3. Explain SRSWR and SRSWOR.
4. In SRSWOR, Show that E(s2)=S2
5. In SRSWOR, Derive the formula of the variance of the estimate of the population mean.

**Unit-II**

**Sampling Theory (continued...)**

**Stratified Random Sampling:**

**Introduction:**

 In SRS, clear that the variance of the estimate of the population mean is

* Inversely proportional to the sample size ‘n’
* Directly proportional to the variability of the population S2

 If the population is heterogeneous, we need some technique to increase the efficiency without increasing sample size. One such technique is Stratified Random Sampling.

**Stratification**: The division of heterogeneous population in to a no. of homogeneous sub groups based on some criteria is called stratification. These sub groups are called **Strata** or **Blocks** and each sub group is called a **Stratum** or a **Block**.

The criterion used in stratification is called a **stratifying factor**. Some important stratifying factors are, Age, Gender, Area, Income, Education, etc.

**Definition:**

 If a heterogeneous population of size ‘N’ be divided into ‘k’ mutually exclusive strata of sizes N1, N2, ..., Nk respectively such that ∑Ni = N and a sample of size ‘n’ is selected from the population by drawing k simple random samples of sizes n1, n2, ..., nk such that ∑ni = n respectively from the strata, then the sample is called a **Stratified Random Sample** and the technique is called a **Stratified Random Sampling.**

**Notations:**

**Population** **Sample**

N = Population Size n = Sample Size

k= No. of strata

Ni = Size of ith Stratum ni = Size of the sample from ith stratum

Yij = The measurement of jth unit of ith strata yij = The measurement of jth unit of

 (i=1,2,...,k, j= 1,2, ...., Ni) ith sample ( i=1,2,...,k, j=1,2,...ni)

Y̅Ni = ith Stratum Mean = $\frac{\sum\_{}^{}Yij}{Ni}$ y̅ni = ith Sample Mean = $\frac{\sum\_{}^{}yij}{ni}$

Y̅N = Population Mean = $\frac{\sum\_{}^{}∑Yij}{N}$ y̅n = Sample Mean = $\frac{\sum\_{}^{}∑yij}{n}$

 Or $\frac{\sum\_{}^{}NiY̅i}{N}$ or $\sum\_{}^{}wiY̅i$ or $\frac{\sum\_{}^{}niy̅i}{n}$

**wi** = Ni/N = weight of ith stratum y̅st = estimate of population mean

**fi** = ni/Ni = Sampling fraction of ith stratum (or) Stratified random sample mean

 = $\frac{\sum\_{}^{}Niy̅i }{N}$ or $\sum\_{}^{}wiy̅i $

Si2  = ith Stratum Mean Square si2 = ith Sample Mean Square

 = $\frac{\sum\_{}^{}(Yij-\overbar{Yi})^{2}}{Ni-1}$ or $\frac{1}{Ni-1} (\sum\_{}^{}Yij^{2}-N$ $(\overbar{Yi})^{ 2}$) = $\frac{\sum\_{}^{}(yij-\overbar{yi})^{2}}{ni-1}$or $\frac{1}{ni-1}( \sum\_{}^{}yij^{2}-n $ $(\overbar{yi})^{ 2}$)

S2  = Population Mean Square

 = $\frac{\sum\_{}^{}∑(Yij-\overbar{Y})^{2}}{N-1}$ or $\frac{1}{N-1} (\sum\_{}^{}∑Yij^{2}-N$ $(\overbar{Y})^{ 2}$)

**Unbiased estimate of population and its variance:**

* The stratified random sample, y̅st =$\sum\_{}^{}wiy̅i $ is an unbiased estimate of Population mean y̅N

Proof: E(y̅st)= E( $\sum\_{}^{}wiy̅i $) = $\sum\_{}^{}wiE(\overbar{y}i)$ = $\sum\_{}^{}wiY̅i$ = y̅N

 Since samples from strata are drawn by SRSWOR

* The variance of the estimate is, V(y̅st) = $\sum\_{}^{}wi^{2}(\frac{1}{ni}-\frac{1}{Ni})Si^{2} $

**Allocation of Sample sizes in to strata:**

The sample sizes can be allocated in to different strata by the following two ways.

1. Proportional Allocation 2. Optimum Allocation

**Proportional Allocation:**

An allocation is said to be proportional allocation if the sampling fraction is constant for each stratum.

i.e., n1/N1 = n2/N2 = ...... = ni/Ni = ...... = nk/Nk = ∑ni/∑Ni = n/N (= C, a constant)

 or ni = C Ni or ni α Ni, for i = 1,2,....,k

 or ni/Ni = n/N

 or **ni =** $\frac{n}{N}$ **Ni, i=1, 2, ..., k**

This formula is used to allocate sample sizes in each stratum.

\*\* Drawback: The proportional allocation neglects the stratum variability.

For example, If N=1000, N1=200, N2=300 and N3 =500 and if n=200, then by proportional allocation, n1 = (200/1000)x200 = 40

 n2 = (200/1000)x300 = 60

 n3 = (200/1000)x500 = 100

**Result: Variance of the estimate of population mean under Proportional allocation:**

 Under proportional allocation, **ni/Ni = n/N ......(1)**

 Therefore, V(y̅st) = $\sum\_{}^{}wi^{2}(\frac{1}{ni}-\frac{1}{Ni})Si^{2}$

 **=** $\sum\_{}^{}(\frac{Ni}{N})^{2}(\frac{1}{ni}-\frac{1}{Ni})Si^{2}$

 **=** $\sum\_{}^{}(Ni/N^{2}) (\frac{Ni}{ni}-1)Si^{2}$

 **=** $\sum\_{}^{}(Ni/N^{2}) (\frac{N}{n}-1)Si^{2}$ **( from (1))**

 **=** $\sum\_{}^{}\left(\frac{Ni}{N}\right)(\frac{1}{n}-\frac{1}{N})Si^{2}$

 **V(y̅st)prop =** $\left(\frac{1}{n}-\frac{1}{N}\right)\sum\_{}^{}wiSi^{2}$

**Optimum Allocation:**

An allocation is said to be an optimum allocation, if the sample sizes are allotted to each stratum such that

1. variance is minimum for fixed total sample size

or

1. variance is minimum for fixed total cost

or

1. cost is minimum for a fixed variance.

**Cost Function**: Practically, it is clear that the cost of investigating a unit is different from stratum to stratum. Thus total cost is defined as a cost function as given below.

 The cost function, **C = C0 +** $\sum\_{}^{}nici$

 Where, **C0** = over head cost, ci = cost per unit in ith stratum

**Theorem**: The variance v(y̅st) is minimum for a fixed sample size ‘n’, if ni α NiSi

and hence show that ni = $\frac{nNiSi}{\sum\_{}^{}NiSi}$

**Proof**:

 V(y̅st) = $\sum\_{}^{}wi^{2}(\frac{1}{ni}-\frac{1}{Ni})Si^{2}$ **......(1)**

To minimize (1) subject to the condition ∑ni = n (fixed).

But it is difficult to minimize (1) conditionally, so using Lagrange multipliers technique it is equivalently minimizing,

 Z = V(y̅st) + λ (∑ni - n) unconditionally in the variations of ni.

 => Z = $\sum\_{}^{}wi^{2}(\frac{1}{ni}-\frac{1}{Ni})Si^{2}$+ λ (∑ni - n) .........(2)

 To minimize (2), we have,

 $\frac{∂Z}{∂ni}$ = 0 => $\frac{∂\sum\_{}^{}wi^{2}(\frac{1}{ni}-\frac{1}{Ni})Si^{2} + λ (∑ni - n)}{∂ni}$ = 0

 => wi2$(\frac{-1}{ni^{2}})Si^{2}$+ λ (1-0) = 0

 => ni2 = wi2Si2/ λ

 = > ni = wiSi / $\sqrt{λ}$

 => ni = NiSi/N$\sqrt{λ}$ ..... (3)

 => ni α NiSi Hence theorem

 Formula for sample sizes:

 Summing (3) on both sides over i=1,2..,k ( for calculating λ )

 ∑ni =∑NiSi/N$\sqrt{λ}$ => N$\sqrt{λ}$ = ∑NiSi /n

 Replacing N$\sqrt{λ}$ value in (3)

 ni = NiSi/[∑NiSi/n]

 => **ni =** $\frac{nNiSi}{\sum\_{}^{}NiSi}$ **for i = 1,2,...,k**

 This is the formula to allocate the sample sizes in strata by optimum allocation under criterion (i). This formula is called Neyman’s formula and allocation is called Neyman’s optimum allocation.

**Result: Variance of v(y̅st) under Neyman’s allocation:**

By Neyman’s optimum allocation, ni = $\frac{nNiSi}{\sum\_{}^{}NiSi}$ => ni/Ni = $\frac{nSi}{\sum\_{}^{}NiSi}$ ......(a)

Therefore, V(y̅st) = $\sum\_{}^{}wi^{2}(\frac{1}{ni}-\frac{1}{Ni})Si^{2}$

 **=** $\sum\_{}^{}(\frac{Ni}{N})^{2}(\frac{1}{ni}-\frac{1}{Ni})Si^{2}$

 **=** $\sum\_{}^{}(Ni/N^{2}) (\frac{Ni}{ni}-1)Si^{2}$

 **=** $\sum\_{}^{}(Ni/N^{2}) (\frac{\sum\_{}^{}NiSi}{nSi}-1)Si^{2}$ **(from (a))**

 **=** $\sum\_{}^{}(Ni/N^{2}) \frac{(\sum\_{}^{}NiSi)}{n}Si-\sum\_{}^{}(Ni/N^{2})Si^{2}$

 = $\frac{(\sum\_{}^{}NiSi)^{2}}{nN^{2}}-\frac{\sum\_{}^{}NiSi^{2}}{N^{2}}$

 **V(y̅st)N =** $\frac{(\sum\_{}^{}wiSi)^{2}}{n}-\frac{\sum\_{}^{}wiSi^{2}}{N}$

**Theorem:**

Show that **V(y̅st)opt ≤ V(y̅st)prop ≤ V(y̅n)R**

(Or ) Show that Stratified Random Sampling with Neyman’s optimum allocation is more efficient than proportional allocation and simple random sampling.

**Proof:**

 V(y̅st)opt = $\frac{(\sum\_{}^{}wiSi)^{2}}{n}-\frac{\sum\_{}^{}wiSi^{2}}{N}$ ...... (1)

 V(y̅st)prop = $\left(\frac{1}{n}-\frac{1}{N}\right)\sum\_{}^{}wiSi^{2}$ ......(2)

 V(y̅n)R = $\left(\frac{1}{n}-\frac{1}{N}\right)$S2 ......(3)

At first, let us prove V(y̅st)opt ≤ V(y̅st)prop :

 Consider, V(y̅st)prop - V(y̅st)opt

 = $\left(\frac{1}{n}-\frac{1}{N}\right)\sum\_{}^{}wiSi^{2}$ - [$\frac{(\sum\_{}^{}wiSi)^{2}}{n}-\frac{\sum\_{}^{}wiSi^{2}}{N}$] from (1) and (2)

 = $\left(\frac{1}{n}\right)[\sum\_{}^{}wiSi^{2}$ - $(\sum\_{}^{}wiSi)^{2}$]

 = $\left(\frac{1}{n}\right)$ [ var(S)] ≥ 0

 => V(y̅st)prop ≥ V(y̅st)opt

 => **V(y̅st)opt ≤ V(y̅st)prop** ---- (4)

Now, let us prove V(y̅st)prop ≤ V(y̅n)R:

Consider, S2 = $\frac{\sum\_{}^{}∑(Yij-\overbar{Y})^{2}}{N-1}$

Let us re-write, (N-1) S2 = $\sum\_{}^{}∑((Yij-\overbar{Yi})+(\overbar{Yi}-\overbar{Y}))^{2}$

 = $\sum\_{}^{}∑(Yij-\overbar{Yi})^{2}$ + $\sum\_{}^{}∑(\overbar{Yi}-\overbar{Y})^{2}$+$2\sum\_{}^{}\sum\_{}^{}(Yij-\overbar{Yi})(\overbar{Yi}-\overbar{Y})$

 = $∑(Ni-1)Si^{2}$ + $∑Ni(\overbar{Yi}-\overbar{Y})^{2}$ + 0 (by property of AM)

 => (N-1) S2 = $∑(Ni-1)Si^{2}$ + $∑Ni(\overbar{Yi}-\overbar{Y})^{2}$

Let us assume, N-1~N and Ni-1~Ni

 => N S2 = $∑NiSi^{2}$ + $∑Ni(\overbar{Yi}-\overbar{Y})^{2}$

 => S2 = $∑wiSi^{2}$ + $∑wi(\overbar{Yi}-\overbar{Y})^{2}$

 => S2 ≥ $∑wiSi^{2}$ since last term is always non-negative

Multiplying with $\left(\frac{1}{n}-\frac{1}{N}\right)$ on both sides

 => $\left(\frac{1}{n}-\frac{1}{N}\right)$S2 ≥ $\left(\frac{1}{n}-\frac{1}{N}\right)$ $∑wiSi^{2}$

 => V(y̅n)R ≥ V(y̅st)prop from (2) and (3)

 => **V(y̅st)prop ≤ V(y̅n)R  ...... (5)**

Finally from (4) and (5),

 **V(y̅st)opt ≤ V(y̅st)prop ≤ V(y̅n)R**

Therefore Stratified Random Sampling with Neyman’s optimum allocation is more efficient than proportional allocation and simple random sampling.

**Merits and Limitations of Stratified Random Sampling**:

**Merits:**

1. Stratified random sampling provides **more representative** sample when compared with unstratified random sampling where some strata may over-represented or under-represented or completely excluded.
2. Because of Stratification, Stratified random sampling is **more efficient** than simple random sampling.
3. It provides estimates with **greater accuracy** and also it enables us to obtain the results for each stratum.
4. It is more convenient to collect data from stratified random samples with less time, minimum cost and more concentration comparing with simple random sampling.

**Limitations:**

1. If stratification is faulty, it cannot be compensated with large sample.
2. If there are more stratifying factors, it is difficult to stratify the population.
3. The optimum allocation needs Stratum variability (Si’s) in advance. It will overcome by conducting a pilot survey in each stratum with size n and then estimating Si’s.

**Problem:**

A population of size 800 is divided in to 3 strata. Their sizes, means and S.D’s are given below.

|  |  |
| --- | --- |
|  | **Strata** |
| No. | I | II | III |
| Size(Ni) | 200 | 300 | 300 |
| Mean(Y̅i) | 45 | 60 | 70 |
| S.D (Si) | 6 | 8 | 12 |

A stratified random sample of size 120 is to be drawn from the population. Determine the sizes of samples from the 3 strata in case of

1. Proportional allocation ii) Neyman’s optimum allocation

And also verify that stratified random sampling with optimum allocation is more efficient than proportional allocation and simple random sampling.

**Solution:**

Determination of sample sizes: n = 120

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Stratum No. | Ni | Y̅i | Si | NiSi | i)Proportional allocationni = (n/N) Ni | ii)Neyman’s allocationni = nNiSi/∑NiSi |
| I | 200 | 45 | 6 | 1200 | n1 = (120/800)200 = **30** | n1 = (120x1200)/7200 = **20** |
| II | 300 | 60 | 8 | 2400 | n2 = (120/800)300= **45** | n2 = (120x2400)/7200= **40** |
| II | 300 | 70 | 12 | 3600 | n3 = (120/800)300= **45** | n3 = (120x3600)/7200= **60** |
| **Total** | **N=800** |  |  | **7200** | **n =120** | **n =120** |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Stratum No. | Ni | Y̅i | Si | NiSi | NiSi2 | (Ni-1)Si2 | NiY̅i | Ni(Y̅i-Y̅)2 |
| I | 200 | 45 | 6 | 1200 | 7200 | 7164 | 9000 | 45000 |
| II | 300 | 60 | 8 | 2400 | 19200 | 19136 | 18000 | 0 |
| II | 300 | 70 | 12 | 3600 | 43200 | 43056 | 21000 | 30000 |
| **Total** | **800** |  |  | **7200** | **69600** | **69356** | **48000** | **75000** |

Comparing the efficiencies: Y̅ = ∑NiY̅i /N = 48000/800 = 60.

 V(y̅st)opt = $\frac{(\sum\_{}^{}wiSi)^{2}}{n}-\frac{\sum\_{}^{}wiSi^{2}}{N}$

 = $\frac{(\sum\_{}^{}NiSi)^{2}}{nN^{2}}-\frac{\sum\_{}^{}NiSi^{2}}{N^{2}}$

 = 72002/(120x8002) - 69600/8002

 = 0.5662

 V(y̅st)prop = $\left(\frac{1}{n}-\frac{1}{N}\right)\sum\_{}^{}wiSi^{2}$

 = $\left(\frac{1}{n}-\frac{1}{N}\right)\sum\_{}^{}NiSi^{2}$/N

 = (1/120 – 1/800)x69600/800

 = 0.6162

 V(y̅n)R = $\left(\frac{1}{n}-\frac{1}{N}\right)$S2

 = $\left(\frac{1}{n}-\frac{1}{N}\right)$ (1/N-1)[$ ∑(Ni-1)Si^{2}$ + $∑Ni(\overbar{Yi}-\overbar{Y})^{2}$]

 = (1/120 – 1/800)x(1/799)[69356 + 75000]

 = 1.2797

**V(y̅st)opt =0.5662 < V(y̅st)prop = 0.6162 < V(y̅n)R = 1.2797**

**Therefore the stratified random sampling with optimum allocation is efficient than proportional allocation and simple random sampling.**

**The gain in efficiency over proportional allocation is (0.6162/0.5662 - 1)100 = 8.83%**

**The gain in efficiency over SRS is (1.2797/0.5662 - 1)100 = 126%**

**Systematic Sampling ( with N = n.k, k is an integer ):**

**Definition:**

 The systematic sampling consists in selecting only the first unit at random and rest being automatically selected according to some predetermined pattern involving regular spacing of units.

 Suppose that N units are serially numbered from 1 to N in some order and a sample of size n is to be drawn such that N = nk, k is an integer.

 where k = N/n is called sampling interval.

**Now systematic sampling consists in**

* **drawing a random number, say ‘i’ (≤k)**
* **and selecting the unit corresponding to the this number**
* **and every kth unit subsequently.**

Hence a sample of size n will consist of units **i, i+k, i+2k, ...., i+(n-1)k.**

* This sample is called a **systematic sample**.
* ‘**i**’ is called the **random start** that determines whole sample.
* **The no. of possible samples = k** ( one per each i = 1,2..,k)
* The k possible samples are **mutually exclusive** and **exhaustive**.

 i.e., all k samples are not overlapping and cover all units of population.

 i=1 🡪 1, 1+k, 1+2k, ..., 1+(n-1)k

If we arrange these in rows and columns as shown, it seems to be

**Rows -> Systematic samples**

**Columns -> Strata**

**Allocation -> Proportional**

 i=2 🡪 2, 2+k, 2+2k ...., 2+(n-1)k

 ....

 i 🡪 i, i+k, i+2k, ......, i+(n-1)k

 .....

 i=k 🡪 k, 2k, 3k, ........., nk

* So, the probability of each sample = 1/k.

**Other Notations:**

**yij** = the measurement of jth unit in the ith systematic sample.(i=1,2,...,k, j=1,2,...,n)

y̅i. or y̅sys = mean of ith systematic sample = $\frac{\sum\_{j=1}^{n}yij}{n}$

y̅.j = mean of jth stratum = $\frac{\sum\_{i=1}^{k}yij}{k}$

y̅.. or Y̅= over all mean or population mean =$\frac{\sum\_{i=1}^{k}\sum\_{j=1}^{n}yij}{nk}$ or $\frac{\sum\_{i=1}^{k}y̅i. }{k}$ or $\frac{\sum\_{j=1}^{n}y̅.j}{n}$

**S2** = Population mean square = $\frac{1}{nk-1}$ $\sum\_{i=1}^{k}\sum\_{j=1}^{n}(yij-y̅..)^{2}$

**S2wsy** = Mean square among the units which lie within the same systematic sample

 = $\frac{1}{k(n-1)}$ $\sum\_{i=1}^{k}\sum\_{j=1}^{n}(yij-\overbar{y}i.)^{2}$

**Unbiased estimate of population mean and its variance:**

* The systematic sample mean **y̅sys**is an **unbiased estimate** of the population mean Y̅

since, E(y̅sys) = $\sum\_{i=1}^{k}y̅i. P(\overbar{y}i)$ = $\sum\_{i=1}^{k}y̅i. (\frac{1}{k})$ = $\frac{\sum\_{i=1}^{k}y̅i. }{k}$ = Y̅

* The Variance of the estimate,

 V(y̅sys) = $\frac{1}{K}\sum\_{i=1}^{k}(y̅i.-y̅..)^{2}$

**Merits and Limitations of systematic sampling:**

**Merits:**

1. Systematic sampling is operationally more convenient than simple random sampling and stratified random sampling.
2. It consumes less time and work comparing with other methods.
3. It may be more efficient than simple random sampling provided the complete frame is arranged at random.

**Limitations:**

1. In general, systematic samples are not random samples since the merit (3) may not fulfilled.
2. If the frame has a periodic nature with period k, this method may yield highly biased estimates.
3. If k is not an integer, then the actual size of sample may different from that required and sample mean may not be unbiased. In this case, one can use circular systematic sampling.
4. It is not possible to obtain an unbiased estimate of the variance of systematic sample mean. This is important requirement for adopting any sampling method.

**Theorems:**

**1**. Show that V(y̅sys) = $\frac{nk-1}{nk}$ S2 – $\frac{k(n-1)}{nk}$ S2wsy

 and compare systematic sampling with simple random sampling

 **Proof:**

V(y̅sys) = $\frac{1}{K}\sum\_{i=1}^{k}(y̅i.-y̅..)^{2}$

 **S2** = $\frac{1}{nk-1}$ $\sum\_{i=1}^{k}\sum\_{j=1}^{n}(yij-y̅..)^{2}$

 **S2wsy** = Mean square among the units which lie within the same systematic sample

 = $\frac{1}{k(n-1)}$ $\sum\_{i=1}^{k}\sum\_{j=1}^{n}(yij-\overbar{y}i.)^{2}$

 We can write, (nk-1)S2 = $\sum\_{i=1}^{k}\sum\_{j=1}^{n}(yij-y̅..)^{2}$

 = $\sum\_{i=1}^{k}\sum\_{j=1}^{n}(yij-y̅i.+y̅i.-y̅..)^{2}$

 = $\sum\_{i=1}^{k}\sum\_{j=1}^{n}(yij-y̅i.)^{2}$ + $\sum\_{i=1}^{k}\sum\_{j=1}^{n}(y̅i.-y̅..)^{2}$

 = k(n-1) S2wsy  + nk V(y̅sys)

 => **V(y̅sys) =** $\frac{nk-1}{nk}$ **S2 –** $\frac{k(n-1)}{nk}$ **S2wsy**

 **Comparison**:

 Consider, V(y̅)R – V(y̅sys) = $\frac{N-n}{nN}$ **S2  - [**$\frac{nk-1}{nk}$ **S2 –** $\frac{k(n-1)}{nk}$ **S2wsy]**

 **=** $\frac{nk-n}{n(nk)}$ **S2  -** $\frac{nk-1}{nk}$ **S2 +** $\frac{k(n-1)}{nk}$ **S2wsy**

 **=** $\frac{k(1-n)}{nk}$ **S2 +** $\frac{k(n-1)}{nk}$ **S2wsy**

 **=** $\frac{k(n-1)}{nk}$ **[ S2wsy - S2 ]**

* If **S2wsy > S2 , systematic sampling** is more efficient than simple random sampling.
* If **S2wsy < S2 , simple random sampling** is more efficient than systematic sampling.
* If **S2wsy = S2 ,** both systematic sampling and simple random sampling are equally efficient.

**2.** If the population consists of linear trend, then show that

 V(y̅st) ≤ V(y̅sys) ≤ V(y̅)R

**Proof**:

 Let the population has the linear trend given by the model,

 **Yi = i** , for i=1,2,...,N ......(1)

 => yij = i + (j-1)k , for i-1,2,...,k and j = 1,2,...,n ......(2)

Then Population mean = y̅.. or Y̅ = $\frac{\sum\_{i=1}^{N}Yi}{N}$ = (1+2+....+N)/N (from (1))

 = (N+1)/2 or (nk+1)/2

 and S2 = $\frac{1}{N-1}$ [ $\sum\_{i=1}^{N}Yi^{2}$- NY̅2 ] = $\frac{1}{N-1}$ [(12+22+...+N2) –N ((N+1)/2)2 ]

 = $\frac{1}{N-1}$ [N(N+1)(2N+1)/6 – N(N+1)2/4]

 = N(N+1)/12 or nk(nk+1)/12 ...... (3)

 Similarly, (from (2) & (3)),

 Stratum variability, Sj2 = K(k+1)/12 for j = 1,2,...,n ...... (4)

And also population variance = σ2 = (N-1)S2/N =(N-1)N(N+1)/12N = $\frac{N^{2}-1}{12}$----(5)

* Now, from (2) & (5),

 V(y̅sys) = $\frac{k^{2}-1}{12}$.......(6)

* V(y̅)R = $\frac{N-n}{nN}$ **S2  =** $\frac{nk-n}{n(nk)}$ **nk(nk+1)/12** (from (3))

 = $\frac{\left(k-1\right)(nk+1)}{12}$...... (7)

* V(y̅st) = $\left(\frac{1}{n}-\frac{1}{N}\right)\sum\_{}^{}(\frac{Nj}{N})Sj^{2}$

 = $\left(\frac{nk-n}{n\left(nk\right)}\right)\sum\_{}^{}\left(\frac{k}{nk}\right)(\frac{k\left(k+1\right)}{12})$(from (4))

 **=** $\left(\frac{n\left(k-1\right)}{n\left(nk\right)}\right)n(\frac{k}{nk})(\frac{k\left(k+1\right)}{12})$

 **=** $\frac{k^{2}-1}{12n}$.......(8)

From (6), (7) and (8),

 V(y̅st) : V(y̅sys) : V(y̅)R = $\frac{k^{2}-1}{12n}$ **:** $\frac{k^{2}-1}{12}$ **:** $\frac{\left(k-1\right)(nk+1)}{12}$

 ~ $\frac{k^{2}-1}{12n}$ **:** $\frac{k^{2}-1}{12}$ **:** $\frac{\left(k-1\right)n(k+1)}{12}$

 = 1/n : 1 : n

 Therefore, **V(y̅st) ≤ V(y̅sys) ≤ V(y̅)R**

**Problem:**

Compare the precision of systematic sampling, simple random sampling and stratified random sampling for the following data in which a population of 40 units exhibits a fairly raising trend with 10 systematic samples and 4 strata.

|  |  |
| --- | --- |
| Strata | Systematic Samples |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| I | 0 | 1 | 1 | 2 | 5 | 4 | 7 | 7 | 8 | 6 |
| II | 6 | 8 | 9 | 10 | 13 | 12 | 15 | 16 | 16 | 17 |
| III | 18 | 19 | 20 | 20 | 24 | 23 | 25 | 28 | 29 | 27 |
| IV | 26 | 30 | 31 | 31 | 33 | 32 | 35 | 37 | 38 | 38 |

**Solution:** Given N =40, k=10 and n=4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Strata (j) | Systematic Samples(i) | **Total** | **Stratum Mean** |  |  |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |  | **y̅.j** | **∑y2ij** | **Sj2** |
| I | 0 | 1 | 1 | 2 | 5 | 4 | 7 | 7 | 8 | 6 | **41** | 4.1 | 245 | 8.5444 |
| II | 6 | 8 | 9 | 10 | 13 | 12 | 15 | 16 | 16 | 17 | **122** | 12.2 | 1620 | 14.6222 |
| III | 18 | 19 | 20 | 20 | 24 | 23 | 25 | 28 | 29 | 27 | **233** | 23.3 | 5569 | 15.5666 |
| IV | 26 | 30 | 31 | 31 | 33 | 32 | 35 | 37 | 38 | 38 | **331** | 33.1 | 11093 | 15.2111 |
| **Total** | **50** | **58** | **61** | **63** | **75** | **71** | **82** | **88** | **91** | **88** | **727** | **72.7** | **18527** | **53.9443** |
| **Sample mean****(y̅i.)** | 12.5 | 14.5 | 15.25 | 15.75 | 18.75 | 17.75 | 20.5 | 22 | 22.75 | 22 | **181.75** |  |  |  |
| **(y̅i.)2** | 156.25 | 210.25 | 232.5625 | 248.0625 | 351.5625 | 315.0625 | 420.25 | 484 | 517.5625 | 484 | **3419.5625** |  |  |  |

y̅.. or Y̅ = $\frac{\sum\_{i=1}^{k}\sum\_{j=1}^{n}yij}{N}$ = 727/40 = 18.175

S2 = $\left(\frac{1}{N-1}\right)[\sum\_{i=1}^{k}\sum\_{j=1}^{n}yij^{2}$ - NY̅2 ]= 1/39 [18527 - 40x18.1752] = 136.2506

V(y̅)R = $\frac{N-n}{nN}$ **S2  =**  $\frac{40-4}{4(40)}$ **(136.2506) =30.6564**

V(y̅sys) = $\frac{1}{K}\sum\_{i=1}^{k}(y̅i.-y̅..)^{2}$ = $\frac{1}{K}\sum\_{i=1}^{k}y̅i.^{2}$ - $(y̅..)^{2}$

 = $\frac{1}{10}(3419.5625)$ - 18.1752 = **11.6256**

V(y̅st) = $\left(\frac{1}{n}-\frac{1}{N}\right)\sum\_{}^{}(\frac{Nj}{N})Sj^{2}$ **=** $\left(\frac{1}{4}-\frac{1}{40}\right)\left(\frac{10}{40}\right)53.9443$

 = **3.0337**

 Therefore,

 V(y̅st) = **3.0337**

V(y̅sys) = **11.6256**

V(y̅)R = **30.6564**

**Hence for this data, Stratified random sampling is more efficient than systematic sampling and simple random sampling.**

**Gain in efficiency over systematic sampling = (11.6256/3.0337 – 1)100 = 283.21%**

**Gain in efficiency over random sampling = (30.6564/3.0337 – 1)100 = 910.53%**

**Self Assessment Questions:**

 **Multiple Choice Questions (MCQ):**

1. Stratified Random Sampling consists ( )

a) classification of population in to different strata

 b)selection of suitable sample size from each stratum

c) both (a) and (b) d) none

1. Which of the following is true in case of Neyman’s optimum allocation ( )

a) ni αNi b) ni αNiSi c) ni αNiSi/√C̅i d) none

1. If N=1000, N1=200, N2=300, N3=500 and n=100, then by proportional allocation, n2 is ( )

a) 20 b) 30 c)50 d) 100

1. Under Neyman’s optimum allocation V(y̅st) is ( )

a) (1/n-1/N) ∑piSi2 b) (1/N-1/n) ∑piSi2

c) 1/n (∑piSi)2-1/N ∑piSi2 d) 1/N (∑piSi)2-1/n ∑piSi2

1. Which of the following is always true ( )

a) V(y̅st)opt ≤ V(y̅st)prop ≤ V(y̅)ran b) V(y̅st)prop ≤ V(y̅st)opt ≤ V(y̅)ran

 c) V(y̅)ran ≤ V(y̅st)prop ≤ V(y̅st)opt d) V(y̅)ran ≤ V(y̅st)opt ≤ V(y̅st)prop

1. If first unit is drawn at random and remaining units are selected with a constant interval then the method is called ( )

a) Simple Random Sampling b) Stratified Random Sampling

c) Systematic Sampling d) Cluster Sampling

1. If N=500, then which of the following value is not possible for selecting sample as size by systematic sampling ( )

a) 25 b) 50 c) 10 d) 15

1. The systematic sampling is more efficient than simple random sampling if ( )

a) S2wsy ≥ S2 b) S2wsy ≤ S2  c) S2wsy = S2 d) S2wst ≤ S2

1. If N units are classified into k rows and n columns(N=nk), then rows are ( )

a) systematic samples b) strata c) stratified random samples d) random samples

1. If N units are classified into k rows and n columns(N=nk), then columns are ( )

a) systematic samples b) strata c) stratified random samples d) random samples

1. If the population is heterogeneous, which of the following method is most appropriate ( )

a) simple random sampling b) judgement method

c) stratified random sampling d) systematic sampling

1. Which of the following sampling methods does not use probability in the selection of sample ( )

a) Purposive sampling b) Judgement sampling

c) Subjective sampling d) All of these

1. A large sample would be required from a stratum if

a) stratum variability is large b) stratum variability is small

c) stratum size is small d) None

1. If the units of the population are serially numbered, then the appropriate sampling method is( )

a) imple random sampling b) Cluster sampling

c) stratified random sampling d) systematic sampling

1. If the population is homogeneous, which of the following method is most appropriate ( )

a) simple random sampling b) judgement method

c) stratified random sampling d) systematic sampling

**Descriptive Questions:**

**SHORT QUESTIONS:**

1. Define stratified random sampling and write its merits
2. Define systematic sampling.
3. Write a short note on cost function.

**ESSAY QUESTIONS:**

1. Explain stratified random sampling with proportional and optimum allocation.
2. Explain Systematic sampling and explain the merits and demerits.
3. Show that V(y̅st)opt ≤ V(y̅st)prop ≤ V(y̅n)R