Answer of 2015-16

Q4.Comparison of stratified random sampling with simple random sampling without replacement

We know that

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

and $V(\bar{y}_{st})_p = (\frac{1}{n} - \frac{1}{N})\sum_{i=1}^{k} p_i S_i^2$

Now, to express S^2 in terms of Si^2

We know that

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{K}\sum_{j=1}^{Ni}(y_i - \bar{Y}_N)^2$$

$(N-1)S^2 = \sum_{i=1}^{K}\sum_{j=1}^{Ni}(y_i - \bar{Y}_{Ni} + \bar{Y}_{Ni} - \bar{Y}_N)^2$ after addition and subtraction of $\bar{Y}_{Ni}$

$$= \sum_{i=1}^{K}\sum_{j=1}^{Ni}(y_i - \bar{Y}_{Ni})^2 + \sum_{i=1}^{K}\sum_{j=1}^{Ni}(\bar{Y}_{Ni} - \bar{Y}_N)^2 \text{ + Third term}$$

The third term is zero due to the algebraic property of arithmetic mean.

$$(N-1)S^2 = \sum_{i=1}^{K}(Ni-1)S_i^2 + \sum_{i=1}^{K} Ni(\bar{Y}_{Ni} - \bar{Y}_N)^2 \ldots\ldots\ldots\ldots\ldots(1)$$

We assume that Ni's are large such that

$$\frac{Ni-1}{Ni} = \frac{N-1}{N} \Rightarrow 1$$

Dividing equation (1) by N, we have

$$(\frac{N-1}{N})S^2 = \sum_{i=1}^{K}\frac{(Ni-1)}{N}S_i^2 + \sum_{i=1}^{K}\frac{Ni}{N}(\bar{y}_{Ni} - \bar{Y}_N)^2$$

$$S^2 = \sum_{i=1}^{K}\frac{(Ni-1)}{N}S_i^2 + \sum_{i=1}^{K}\frac{Ni}{N}(\bar{y}_{Ni} - \bar{Y}_N)^2$$

Multiplying both sides by $(\frac{1}{n}-\frac{1}{N})$ we have

$$(\frac{1}{n}-\frac{1}{N})S^2=(\frac{1}{n}-\frac{1}{N})\sum_{i=1}^{K}p_iS_i^2+(\frac{1}{n}-\frac{1}{N})\sum_{i=1}^{K}p_i(\bar{y}_{Ni}-\bar{Y}_{Ni})^2$$

It implies that $V_{random}=V_{prop}$ + positive term

Certainly, there is some gain in precision due to stratification. A gain can be increased by making the stratum means difference among themelves.

If we see the difference between V_{prop} - $V_{Ney=}$ a positive term

Therefore, Neyman allocation is certainly more efficient than that of proportional allocation.

If we combine these two results, then we have V_{random} - $V_{Ney=}$ a positive term

Ultimately, we can say that

$V_{Ney} \leq V_{prop} \leq V_{random}$

Hence proved.

Objective types: 2015-16 (Stat 513)(Answers)

(i) (a) (ii) (a) (iii) (b) (iv) (b) (v) (a) (vi) (b) (vii) (b) (viii) (a) (ix)(c)

Answers of Section B 2013-14

(A):Difference between two stage and two phase sampling:

In the first stage we select several sampling units at a time, for example city blocks—these would be primary sampling units (PSUs). Then we choose a sample within each PSU (e.g. households). These are the ultimate sampling units (or secondary sampling units. For example conducting a survey in a district levels: the PSU's is the blocks and then villages within the block as the second stage units and households within the villages as the third stage units.

Two-phase sampling is typically used when it is very expensive to collect data on the variables of interest, but it is relatively inexpensive to collect data on variables that are correlated with the variables of interest. For example, in forest surveys, it is very difficult and expensive to travel to remote areas to make on-ground determinations. However, aerial photographs are relatively inexpensive and determinations on forest type are strongly correlated with ground determinations. Two-phase sampling was called "double sampling" by Neyman. The problem was posed to him at the U.S. Department of Agriculture. A survey was to be conducted to estimate the total of a characteristic *y*. The determinations were very costly, but another variable *x* was known to be correlated with *y* and was cheap to observe. Two-phase samplings reduce the variance of the estimated total by using the correlation between *x* and *y* in constructing a total estimator. However, two-phase samplings are not always superior to one-phase designs. Given a fixed cost, selecting a first-phase sample reduces the number of observations on the response variable *y*. The two-phase framework can be applied in missing data problems, sampling at multiple occasions, and situations without a good frame.

(B) Sampling error and non sampling error:

If complete accuracy can be ensured in the procedures such as determination, identification and observation of sample units and the tabulation of collected data, then the total error would consist only of the error due to sampling, termed as sampling error. Measure of sampling error is mean squared error (MSE). The MSE is the difference between the estimator and the true value and has two components: - square of sampling bias. - sampling variance.

It is a general assumption in the sampling theory that the true value of each unit in the population can be obtained and tabulated without any errors. In

practice, this assumption may be violated due to several reasons and practical constraints. This results in errors in the observations as well as in the tabulation. Such errors which are due to the factors other than sampling are called non-sampling errors.

The non-sampling errors are unavoidable in census and surveys. The data collected by complete enumeration in census is free from sampling error but would not remain free from non-sampling errors. The data collected through sample surveys can have both – sampling errors as well as non-sampling errors. The non-sampling errors arise because of the factors other than the inductive process of inferring about the population from a sample. In general, the sampling errors decrease as the sample size increases whereas non-sampling error increases as the sample size increases. In some situations, the non-sampling errors may be large and deserve greater attention than the sampling errors.

(C)Describe sample survey and census survey:

Survey sampling describes the process of selecting a sample of elements from a target population to conduct a survey. The term "survey" may refer to many different types or techniques of observation. In survey sampling it most often involves a questionnaire used to measure the characteristics and/or attitudes of people. Different ways of contacting members of a sample once they have been selected is the subject of survey data collection. The purpose of sampling is to reduce the cost and/or the amount of work that it would take to survey the entire target population. A survey that measures the entire target population is called a census. A sample refers to a group or section of a population from which information is to be obtained

Survey samples can be broadly divided into two types: probability samples and super samples. Probability-based samples implement a sampling plan with specified probabilities (perhaps adapted probabilities specified by an adaptive procedure). Probability-based sampling allows design-based inference about the target population. The inferences are based on a known objective probability distribution that was specified in the study protocol. Inferences from probability-based surveys may still suffer from many types of bias.

Surveys that are not based on probability sampling have greater difficulty measuring their bias or sampling error Surveys based on non-probability samples often fail to represent the people in the target population.

Definition: The **Census Method** is also called as a **Complete Enumeration Survey Method** wherein each and every item in the universe is selected for the data collection. The universe might constitute a particular place, a group of people or any specific locality which is the complete set of items and which are of interest in any particular situation.

The census method is most commonly used by the government in connection with the national population, housing census, agriculture census, etc. where the vast knowledge about these fields is required. Whenever the entire population is studied to collect the detailed data about every unit, then the census method is applied.

One of the major advantages of census method is the **accuracy** as each and every unit of the population is studied before drawing any conclusions of the research. When more and more data are collected the degree of correctness of the information also increases. Also, the results based on this method are less biased.

The census method can be applied in a situation where the separate data for every unit in the population is to be collected, such that the separate actions for each is taken. For example, the preparation of the voter's list for election purposes, income tax assessment, recruitment of personnel, etc. are some of the areas where the census method is adopted. This method can be used where the population is comprised Though the census method provides a complete data of the population under study, it is very costly and time-consuming. Often, this method is dropped down because of these constraints and the **sampling method**, where certain items representative of the larger group, is selected to draw the conclusions of heterogeneous items, i.e. different characteristics.

(D) Write about probability proportional to size sampling.

Probability proportional to size (PPS) sampling is a method of sampling from a finite population in which a size measure is available for each population unit before sampling and where the probability of selecting a unit is proportional to its size.

Probability sampling requires that each member of the survey population have a chance of being included in the sample, but it does not require that this chance be the same for everyone. If information is available about the size of each unit (e.g., number of students for each school or classroom) and if those units vary in size, this information can be used in the sampling selection in order to increase the efficiency. This is known as sampling with probability proportional to size (PPS). With this method, the bigger the size of the unit, the higher the chance it has of being included in the sample. For this method to bring increased efficiency, the measure of size needs to be accurate.

Two methods of PPS sampling can be used (i) Cumulative total method and (ii) Lahiri's method.

(E)What do you understand by inverse sampling?.

The inverse sampling, first proposed by Haldane, suggests one continues to sample subjects until a pre-specified number of events of interest is observed. In contrast to the commonly used binomial sampling wherein the sample size is prefixed and the

number of events of interest observed is random, the number of events of interest observed is prefixed for inverse sampling and the sample size is a random variable follows a negative binomial distribution. Therefore, inverse sampling is also known as *negative binomial sampling*. It is generally considered to be more appropriate than the usual binomial sampling when the subjects come sequentially, when the response probability is rare, and when maximum likelihood estimators of some epidemiological measures are undefined under binomial sampling. For instance, in epidemiological investigations, the estimation of the prevalence of a given disease on public health in a community or the variation of a disease.

Section C

Q3. Describe different stages taken in planning and execution of a ample survey.

There are following points

(i) Objectives
(ii) Data to be gathered
(iii) Population under investigation
(iv) Sampling frame
(v) Methods of collecting data
(vi) Organization and supervision of field work
(vii) Tabulation and analysis of data
(viii) Precision of the survey
(ix) Writing of the reports

Q4.(A) Describe the principles and advantages of stratification for estimating the population mean $\bar{Y}_N$ in the case of heterogeneous population.

In stratified random sampling the population of N units is divided into sub-population of N1,N2,N3,…Nk units respectively in such away that the units within sub-population must be homogeneous. These sub-population are non-overlapping and together they can comprise the whole of the population, o that N1+N2+N3+…Nk=N These sub-population are called strata. The sample size within the strata are denoted by n1,n2,n3,…nk respectively.

Advantages of stratification:

(i) If the admissible error is given,a small samples should be taken so that our expenditure may be reduced.
(ii) There is a reduction of error due to stratification, if the cost of the survey is fixed.

(iii) Stratification provides the individual means of stratum and then for the whole population.

(iv) Stratification may provide the administrative convenience.

(B) Describe the proportional and Neyman allocation method to decide the sample sizes taken from different strata:

There are four types of choice of sample sizes in different strata.

(i) Equal allocation: ni=n/k where nis the total sample size and k is the number of strata.

(ii) Proportional allocation: ni=npi where pi= Ni/N stratum weight.

(iii) Optimum allocation: It involves cost per unit in the stratum

(iv) Neyman allocation: A particular case of optimum allocation when ci=c ,a constant cost for all the units ni= npisi/$\sum pisi$.

From above allocations, it can be concluded that (i) the larger the size of the stratum, the larger should be the size of the sample to be selected there from; (ii) the larger the variability within a stratum, the larger should be the size of the sample from that stratum; and the cheaper the labour in a stratum, the larger the sample from that stratum.

Q5(A) What do you mean by cluster sampling ?. Write its merits.

A sampling procedure, pre-supposes the division of the population into a finite number of distinct and identifiable units called the sampling units. Thus a population of fields under wheat in a given region might be regarded as composed of fields or groups of fields on the same holdings, villages, or other suitable segments. A human population might similarly be regarded as composed of individual persons, families, or groups of persons residing in houses and villages. The smallest units into which the population can be divided are called the elements of the population, and groups of elements the clusters. When the sampling unit is a cluster, the procedure of sampling is called cluster sampling. When the entire area containing the population under study is subdivided into smaller areas and each element in the population is associated with one and only one such small area, the procedure is alternatively called area sampling. For many types of population a list of elements is not available and the use of an element as the sampling unit is therefore not feasible. The method of cluster or area sampling is available in such cases. Thus, in a city a list of all the houses is readily available, but that of persons is rarely so. Again, lists of fields are not available, but those of villages are. Cluster sampling is, therefore, widely practiced in sample

surveys. The size of the cluster to be employed in sample surveys therefore requires consideration. In general, the smaller the cluster, the more accurate will usually be the estimate of the population character for a given number of elements in the sample. Thus, a sample of holdings independently and randomly selected is likely to be scattered over the entire area under the crop, and thereby provides a better cross-section of the population than an equivalent sample, i.e., a sample of the same number of holdings, clustered together in a few villages. On the other hand, it will cost more to survey a widely scattered sample of holdings than to survey an equivalent sample of clusters of holdings, since the additional cost of surveying a neighbouring holding is small as compared to the cost of locating a second independent holding and surveying it. The optimum cluster is one which gives an estimate of the character under study with the smallest standard error for a given proportion of the population sampled, or more generally, for a given cost.

(B) (i) Average yield per tree in the district= $\frac{1}{N}\sum_{i=1}^{N} \bar{yi} =$

$\frac{1}{15}\sum_{i=1}^{N}[6.71+15.81+22.56+14.57+44.66+14.79+5.87+23.01+15.88+24.78+33.19+20.43+9.19+26.40+12$

=290.52/15=19.368 Kg.

(ii) Sampling variance of cluster means= $\frac{1}{N-1}\sum_{i=1}^{N}(\bar{yi}-\bar{Y}_N)^2$ through this formula sampling variance can be determined.

Q6(A) Describe the procedure of selecting the systematic sampleof size n from a population of size N=kn

we have considered methods of sampling in which the successive units (whether elements or clusters) were selected with the help of random numbers. We shall now consider a method of sampling in which only the first unit is selected with the help of random numbers, the rest being selected automatically according to a predetermined pattern. The method is known as systematic sampling. The pattern usually followed in selecting a systematic sample is a simple pattern involving regular spacing of units. Thus, suppose a population consists of N units, serially numbered from 1 to N. Suppose further that N is expressible as a product of two integers k and II, so that N = kn. Draw a random number less than k, say i, and select the unit with the corresponding serial number and every k-th unit in the population thereafter. Clearly, the sample will contain the n units i, i + k, i + 2k, ... ,(i + n -1)k, and is

known as a systematic sample. The selection of every k-th strip in forest sampling for the estimation of timber, the selection of a corn field, every k-th mile apart, for observation on incidence of borers, the selection of every k-th time-interval for observing the number of fishing craft landing on the coast, the selection of every k-th punched card for advance tabulation or of every k-th village from a list of villages, after the first unit is chosen with the help of random numbers less than k, are all examples of systematic sampling. In the first three examples, the sequence of numbering is determined by Nature, the first two providing examples of distribution in space while the third that of distribution in time. In the fourth and the fifth, the ordering may be either alphabetical or arbitrary approximating to a random distribution. In the latter case, a systematic sample will obviously be equivalent to a random sample. The method is extensively used in practice on account of its low cost and simplicity in the selection of the sample. The latter consideration is particularly important in situations where the selection of a sample is carried out by the field staff themselves. A systematic sample also offers great advantages in organizing control over field work.

(B)(a)Relationship between systematic sampling and cluster sampling:

Systematic sampling and cluster sampling differ in how they pull sample points from the population included in the sample. Cluster sampling breaks the population down into clusters, while systematic sampling uses fixed intervals from the larger population to create the sample. Systematic sampling selects a random starting point from the population, and then a sample is taken from regular fixed intervals of the population depending on its size. Cluster sampling divides the population into clusters and then takes a simple random sample from each cluster.

Cluster sampling is considered less precise than other methods of sampling. However, it may save costs on obtaining a sample. Cluster sampling is a two-step sampling procedure. It may be used when completing a list of the entire population is difficult. For example, it could be difficult to construct the entire population of the customers of a grocery store to interview. However, a person could create a random subset of stores, which is the first step in the process. The second step is to interview a random sample of the customers of those stores. This is a simple manual process that can save time and money.

Systematic sampling is a type of probability sampling method in which sample members from a larger population are selected according to a random starting point and a fixed, periodic interval. Systematic sampling is simple and allows for a degree of process to be used in selecting the sample. This process also guarantees the entire population is evenly sampled. Systematic sampling is useful for certain purposes in finance.

(b)Relationship between systematic sampling and stratified random sampling:

A type of population frequently encountered in extensive samplings is one in which the variance within a group of elements increases steadily as the size of the group increases. This class of populations may be represented by a model in which the elements are serially correlated, the correlation between two elements being a positive and monotone decreasing function of the distance apart of the elements. For populations of this type, the relative efficiencies are compared for a systematic sample of every kth element, a stratified random sample with one element per stratum and a random sample. The stratified random sample is always at least as accurate on the average as the random sample and its relative efficiency is a monotone increasing function of the size of the sample. No general result is valid for the relative efficiency of the systematic sample. In fact, there are populations in the class in which the systematic sample is more accurate than the stratified sample for one sampling rate, but is less accurate than the random sample for another sampling rate.

Q7(A) RATIO AND REGRESSION METHOD OF ESTIMATION

In developing the theory of simple random sampling in the preceding chapters, we have considered only estimates based on simple arithmetic means of the observed values in the sample. We shall consider other methods of estimation which make use of the ancillary information and which, under certain conditions, give more reliable estimates of the population values than those based on the simple averages. Two of these methods are of particular importance. They are: (i) the ratio method of estimation, and (ii) the regression method of estimation.

Assumptions in Ratio Method of Estimation:

(i) The regression line of Y on X should be linear.
(ii) The regression lines passes through the point (0,0) i.e. origin.

Assumptions in Regression Method of Estimation:

(i) The regression line of Y on X should be linear
(ii) The regression line does not pass through the origin.

In both ratio and regression method of estimation, the ratio and regression estimators are not an unbiased estimators of the population mean.

(B) Numerical already sent on ratio and regression method of estimation.

Q8(A) In SRSWOR, the selected units from the population is not replaced back into the population before the next draw and in SRSWR the selected units from the population is replaced back into the population before the next draw . Out of these two schemes,

SRSWOR is found to be more better rather than that of SRSWR because (i) the information on elements is always new and it will cover the whole population (ii) the variance of sample mean in SRSWOR is always less than SRSWR. In SRSWR, it may be possible that we have the information on only one unit of the population again and again resulting not appropriate sample from the population.

8(B). Unbiased estimate of population mean

$\frac{1}{n}\sum yi = \frac{1}{15}$ [sum all fifteen values]= the estimate of population mean

Objectives on Sampling (Stat 513)

(i) A sample consists of
(a) All units of the population
(b) 5% units of the population
(c) 10% units of the population
(d) Any fraction of the population

(ii) Sampling is used in the situations
(a) Blood test of the patients
(b) Populat
(c)

(iii) The number of possible samples of size n out of N population size in SRSWOR is equal to
(a) Nc_n
(b) N^n
(c) (N-n)/N
(d) n/N

(iv) The number of possible samples of size n out of N population size in SRSWOR is equal to

(a) Nc_n

(b)N^n

(c) (N-n)/N

(d) n/N

(v) The number of possible samples of size 2 out of 5 population size in SRSWOR is equal to
(a) 10
(b) 4
(c) 2
(d) 12

(vi) The number of possible samples of size 2 out of 5 population size in SRSWR is equal to
(a) 25

(b) 20

(c) 2

(d) 12

(vii) Probability of a drawing unit at each subsequent draw remains same in

(a) SRSWOR
(b) SRSWR
(c) Both (a) &(b)
(d) None

(viii) The sampling fraction in usual notation is expressed as
(a) n/N
(b) N/n
(c) 1-n/N
(d) None.

(ix) The finite population correction in usual notation is expressed as
(a) (N-n)/N
(b) 1-(n/N)
(c) Both(a)&(b)
(d) None

(x) A selection procedure of sampling having no involvement of probability is known as
(a) SRSWOR
(b) Purposive sampling
(c) SRSWR
(d) None

(xi) For gathering information on rare events,sampling is used
(a) SRSWOR
(b) Stratified random sampling
(c) Inverse sampling
(d) None

(xii) If a larger units have more probability of their inclusion in the sample, the sampling is known as
(a) SRSWOR
(b) PPS sampling
(c) Stratified random sampling
(d) None

(xiii) Simple random samples can be drawn with of help of
(a) Random numbers table
(b) Chit Method
(c) Roulette wheel
(d) All the above

(xiv) Sampling frame is a list of
(a) A list of units of a population
(b) A list of random numbers

(c) A list of natural numbers
(d) None

(xv) In SRSWR, the same sampling unit may be included in the sample
(a) Only once
(b) Two times
(c) More than once
(d) None

(xvi) The discrepancies between the estimate and the population parameter is known as
(a) Sampling error
(b) Non-sampling error
(c) Formula error
(d) None

(xvii) The error in a survey other than sampling error is known as
(a) Sampling error
(b) Non-sampling error
(c) Formula error
(d) None

(xviii) A function of sample observations is known as
(a) Statistic
(b) Estimator
(c) Both (a)&(b)
(d) None

(xix) If the sample sizes are large from the population, then which error will contribute more errors
(a) Sampling error
(b) Non-sampling error
(c) Both(a)&(b)
(d) None

(xx) If the sample sizes are large from the population, then which error will contribute less errors
(a) Sampling error
(b) Non-sampling error
(c) Both (a)&(b)
(d) None

Sampling for proportion and percentage:

We estimate the population mean or population total through considering samples of appropriate size from a population. Sometimes, we are usually interested to estimate the proportion or the number of units in the population possessing a particular attribute. For example in a household survey to assess the extent of damage caused by a hailstorm, earthquake, hurricane etc., it would be of more interest to know the number of households that have damaged extensively on the proportion of households that have spent more money on repairs. In others, one may estimate the number of persons engaged in different occupations.

We know that sample proportion provides an unbiased estimator of population proportion.

i.e. $E(pi) = Pi$

It has been shown with an example given below:

Let us consider a hypothetical population N=(1,2,3,4,5). Draw a sample proportion of size n=(3,5) using SRSWOR.

S.No.	No.of possible samples	Sample proportion pi
1	1,2	0
2	1,3	1/2
3	1,4	0
4	1,5	1/2
5	2,3	1/2
6	2,4	0
7	2,5	1/2
8	3,4	1/2
9	3,5	1
10	4,5	1/2
Total		4.0

Then,$E(pi)=4.0/10=0.4$

Population proportion $Pi=[0+0+1+0+1]/5=2/5=0.4$

It shows that sample proportion provides an unbiased estimator of population proportion.

If the sampling is done using SRSWR, then, $E(pi)=10.0/25=0.4$ and $Pi=0.4$.Here also sample proportion provides an unbiased estimator of population proportion.

6. Simple and Stratified Random Sampling

Simple Random Sampling(SRS): It is the process of selecting a sample from given population according to some law of chance in which each unit of population has an equal and independent chance of being included in the sample.

SRSWR(With Replacement): A selection process in which the unit selected at any draw is replaced to the population before the next subsequent draw is known as Simple random sampling with replacement. In this case the number of possible samples of size n selected from the population of size N is N^n. The samples selected through this method are not distinct.

SRSWOR(Without Replacement): A selection process in which the unit selected at any draw is not replaced to the population before the next subsequent draw and the next sample is selected from the remaining population is known as Simple random sampling without replacement. In this case the number of possible samples of size n selected from the population of size N is N_{c_n} . The samples selected through this method are distinct.

Note: Sample mean is an unbiased estimate of population mean in SRSWR and SRSWOR, whereas sample variance is an unbiased estimate of population variance in case of SRSWOR only.

SRSWOR is more efficient than SRSWR because $V(\bar{y}_n)_{SRSWOR} < V(\bar{y}_n)_{SRSWR.}$

Stratified Random Sampling: When the population is heterogeneous and we wish that every section of population is represented in the sample. We divide the whole population into different number of strata so that the one stratum is much different from one another whereas the samples within each stratum are more homogeneous. This technique of selecting a representative sample of whole population is known as stratified random sampling.

In stratified random sampling allocation of sample size to different strata is based on the staratum sizes (N_i), the variability within the stratum S_i^2 and the cost of surveying per sampling unit in the stratum.

Methods for allocation of sample size to different strata are

Equal Allocation : $n_i = \frac{n}{k}$

Proportional Allocation: $n_i = \frac{nNi}{N}$

Neyman Allocation: $n_i = n * \frac{N_i S_i}{\sum N_i S_i}$

Optimum Allocation (based on cost) : $n_i = n * \frac{N_i S_i \sqrt{C_i}}{\sum N_i S_i \sqrt{C_i}}$

Objective: In simple random sampling, show the sample mean and sample mean square is an unbiased estimate of population mean and population mean square with the help of an hypothetical population in SRSWOR and to determine its variances and S.E.

Kinds of data: The data relate to the hypothetical population whose units are 1, 2, 3, 4 and 5. Draw a sample of size n=3 using SRSWOR.

Solution: Number of all possible samples of size n=3 under SRSWOR is given by $N_{c_n} = 5_{c_3}$=10.

Compute the mean of each sample $\bar{y}_n = \frac{\sum y_i}{n}$ and sample mean square $s^2 = \frac{1}{n-1}\sum(y_i - \bar{y}_n)^2$.

Similarly the mean of population $\bar{Y} = \frac{\sum y_i}{N} = \frac{15}{5} = 3$ and population mean square $S^2 = \frac{1}{N-1}\sum(y_i - \bar{Y})^2$

$S^2 = \frac{1}{4}[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2] = \frac{10}{4} = 2.5$

The 10 possible samples are given below in the table.

S.No.	Possible samples	S in the t ean mple n $\bar{y}_n$	Sample mean square (s^2)	S (mpl error $\bar{y}_n - \bar{Y}$
1.	1,2,3	2.0	1.0	-1.0
2.	2,3,4	3.0	1.0	0.0
3.	3,4,5	4.0	1.0	1.0
4.	4,5,1	3.33	4.33	0.33
5.	5,1,2	2.67	4.33	-0.33
6.	1,3,4	2.67	2.33	-0.33
7.	2,4,5	3.67	2.33	0.67
8.	3,5,1	3.0	4.00	0.0
9.	4,1,2	2.33	2.33	-0.67
10.	5,2,3	3.33	2.33	0.33
Total		30.0	24.98=25	0.00

Now we have to check whether E ($\bar{y}_n$) = $\overline{y_N}$ and E (s^2) = S^2 ,

E ($\bar{y}_n$)= $\frac{\sum \overline{y_n}}{N_{c_n}} = \frac{30}{10} = 3 = \overline{y_N}$ and E (s^2)= $\frac{\sum s_i^2}{N_{c_n}} = \frac{25}{10} = 2.5 = S^2$,

then we can say that sample mean $\bar{y}_n$ and sample variance s^2 are an unbiased estimator of population mean $\bar{Y}$ and population variance S^2 respectively.

In order to find out the variance of sample mean in SRSWOR, we know that

$V(\bar{y}_n)_{SRSWOR} = \frac{N-n}{Nn} S^2 = \frac{5-3}{5*3} * 2.5 = 0.33$

We can verify that this variance is correct.

$V(\bar{y}_n) = \frac{\sum(\bar{y}_n - \bar{Y}(\bar{y}_n))^2}{N_{c_n}} = \frac{1}{N_{c_n}} [\sum \bar{y}_n^2 - \frac{(\sum \bar{y}_n)^2}{N_{c_n}}] = \frac{1}{10}$ [93.33-90]=0.33

This shows that $V(\bar{y}_n)_{SRSWOR}$ is correct.

Standard Error of $(\bar{y}_n) = \sqrt{V(\bar{y}_n)} = \sqrt{0.33}$ =0.57

We can also compare the two variances, one in SRSWOR and the other in SRSWR.

$V(\bar{y}_n)_{SRSWR} = \frac{N-1}{Nn} S^2 = \frac{5-1}{5*3} * 2.5 = 0.66$

Since $V(\bar{y}_n)_{SRSWOR} < V(\bar{y}_n)_{SRSWR}$

Hence we can say that SRSWOR is more efficient than SRSWR.

Objective: Showing the unbiased estimator for population mean and biased estimator for population mean square in simple random sampling with replacement (SRSWR) with the help of an hypothetical example and determination of its variance and standard error (S.E.)

Kind of data: Consider a finite population of size N=5 including the values of sampling units as (1,2,3,4,5). Enumerate all possible samples of size n=2 using SRSWR. find the estimate of $V(\bar{y}_n)$ in 9th sample.

Solution: Number of all possible samples of size n=3 under SRSWOR is given by $N^n = 5^2 = 25$.

Compute the mean of each sample $\bar{y}_n = \frac{\sum y_i}{n}$ and sample mean square $s^2 = \frac{1}{n-1}\sum(y_i - \bar{y}_n)^2$.

Similarly the mean of population $\bar{Y} = \frac{\sum y_i}{N} = \frac{15}{5} = 3$ and population mean square $S^2 = \frac{1}{N-1}\sum(y_i - \bar{Y})^2$

$S^2 = \frac{1}{4}[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2] = \frac{10}{4} = 2.5$

S.No.	Possible samples	Sample mean $\bar{y}_n$	Sample mean square (s^2)	Sample error $(\bar{y}_n - \bar{Y})$	S.No.	Possible samples	Sample mean $\bar{y}_n$	Sample mean square (s^2)	Sample error $(\bar{y}_n - \bar{Y})$
1	1,2	1.5	0.50	-1.5	13	4,1	2.5	4.50	-0.5
2	1,3	2.0	2.00	-1.0	14	5,1	3.0	8.00	0.0
3	1,4	2.5	4.50	-0.5	15	3,2	2.5	0.50	-0.5
4	1,5	3.0	8.00	0.0	16	4,2	3.0	2.00	0.0
5	2,3	2.5	0.50	-0.5	17	5,2	3.5	4.50	0.5
6	2,4	3.0	2.00	0.0	18	4,3	3.5	0.50	0.5
7	2,5	3.5	4.50	0.5	19	5,3	4.0	2.00	1.0
8	3,4	3.5	0.50	0.5	20	5,4	4.5	0.50	1.5
9	3,5	4.0	2.00	1.0	21	1,1	1.0	0.00	-2.0
10	4,5	4.5	0.50	1.5	22	2,2	2,0	0.00	- 1.0
11	2,1	1.5	0.50	-1.5	23	3,3	3.0	0.00	0.0
12	3,1	2.0	2.00	-1.0	24	4,4	4.0	0.00	1.0
					25	5,5	5.0	0.00	2.0
Total							75.0	50.00	

Now we have to check whether $E(\bar{y}_n) = \bar{Y}$ and $E(s^2) = S^2$,

E ($\bar{y}_n$)= $\frac{\sum \bar{y}_n}{N^n} = \frac{75}{25}$ =3 = $\bar{Y}$ and E (s^2)= $\frac{\sum s_i^2}{N^n} = \frac{50}{25}$ =2 ≠S^2,

then we can say that sample mean $\bar{y}_n$ is an unbiased estimate of population mean whereas and sample variance s^2 is not an unbiased estimate of population variance S^2 in case of SRSWR.

In order to find out the variance of sample mean in SRSWR, we known that

$$V(\bar{y}_n)=\frac{\sigma^2}{n} = \frac{N-1}{n} S^2 = \frac{5-1}{5*2}*2.5 = 1.0$$

Standard Error of ($\bar{y}_n$) = $\sqrt{V(\bar{y}_n)} = \sqrt{1}$ =1

In order to find the estimate of $V(\bar{y}_n)$ based on 9th sample, we have

$$V(\bar{y}_n)=\frac{N-1}{Nn} S^2 = \frac{5-1}{5*2}*2.0 = 0.8$$

Standard Error of ($\bar{y}_n$) = $\sqrt{V(\bar{y}_n)} = \sqrt{0.80}$ =0.894

Objective : Drawing of samples in stratified random sampling under different allocation along with determination of their variances and standard errors.

Kinds of data: A hypothetical population of N= 3000 is divided into four strata, their sizes of population and standard deviations are given as follows :

Strata	I	II	III	IV
Size N_i	400	600	900	1100
SD S_i	4	6	9	12

A stratified random sample of size 800 is to be selected from the population

Soultion : In case of

(i) Equal allocation the sizes of sample allocated to different strata will be the same. Hence the different sample sizes will be $n_i = \frac{n}{k} = \frac{total\ sample\ size}{number\ of\ strata} = \frac{800}{4}$ =200 samples from each allocation.

(ii) In case of proportional allocation n_i (i=1,2,3,4) is given by $n_i = np_i$ where $p_i = N_i/N$

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

Hence $n_1 = \frac{800*400}{3000}$ =106.67≈107 samples from stratum I

$n_2 = \frac{800*600}{3000}$ =160 samples from stratum II

$n_3 = \frac{800*900}{3000}$ =240 samples from stratum III

$n_4 = \frac{800*1100}{3000}$ =293 samples from stratum IV

Thus, $n_1 + n_2 + n_3 + n_4 = 800$ constitute the samples required from all the strata.

(iii) The sample size in Neyman allocation is given by $n_i = n * \frac{P_i S_i}{\sum P_i S_i} = n * \frac{N_i S_i}{\sum N_i S_i}$

Here $\sum N_i S_i$ =400*4+600*6+900*9+1100*12= 26500

Hence, $n_1 = 800 * \frac{400*4}{26500} = 48$, $n_2 = 800 * \frac{600*6}{26500} = 109$,

$n_3 = 800 * \frac{900*9}{26500} = 245$, $n_4 = 800 * \frac{1100*12}{26500} = 398$,

In Neyman allocation, the sample sizes from four strata are 48, 109, 245 and 398 which constitute the required sample size.

Variance of $\bar{y}_{st}$ in equal allocation $V(\bar{y}_t) = \frac{k\sum p_i^2 s_i^2}{n} - \frac{\sum p_i s_i^2}{N}$,

from above data $\sum p_i S_i = 8.83$, $\sum p_i s_i^2 = 86.43$ and $\sum p_i^2 s_i^2 = 28.37$

$V(\bar{y}_t) = \frac{4*86.43}{80} - \frac{28.37}{3000}$, $= .14 - .028 = 0.1130$

Standard Error of $(\bar{y}_{st}) = \sqrt{V(\bar{y}_t)} = \sqrt{0.1130} = 0.336$

Variance of $\bar{y}_t$ in proportional allocation $V(\bar{y}_t) = (\frac{1}{n} - \frac{1}{N})\sum p_i s_i^2 = (\frac{1}{800} - \frac{1}{3000})*86.43 = 0.0792$

Standard Error of $(\bar{y}_t)_{prop} = \sqrt{V(\bar{y}_{st})} = \sqrt{0.0792} = 0.2815$

Variance of $\bar{y}_t$ in Neyman allocation $V(\bar{y}_t) = \frac{(\sum p_i S_i)^2}{n} - \frac{\sum p_i s_i^2}{N} = \frac{8.83^2}{800} - \frac{86.43}{3000} = .068$

Standard Error of $(\bar{y}_t)_{ney} = \sqrt{V(\bar{y}_t)} = \sqrt{.068} = 0.262$

7. Ratio and Regression Estimator

Ratio Estimator: Ratio method of estimation is based on the information available for auxiliary variable. When the correlation coefficient between the study variable and the auxiliary variable is positive and high, the ratio method of estimation can be used to study the population parameters of study variable Y.

The equation of ratio estimator is given by $\bar{y}_R = \frac{\bar{y}_n}{\bar{x}_n}\bar{x}_N$, where $\bar{y}_n$ and $\bar{x}_n$ are sample means of y and x respectively and $\bar{x}_N$ is population mean.

In case of ratio estimator sample mean is not an unbiased estimate of population mean. The bias will be zero only when there is a perfect positive correlation between y and x.

The bias of ratio estimator to the first order of approximation is given by

$B_1(\bar{y}_R) = \frac{(N-n)}{Nn}\bar{x}_N (C_x^2 - \rho C_x C_y)$, where $C_x = \frac{S_x}{\bar{X}_N}$ and $C_y = \frac{S_Y}{\bar{Y}_N}$

The variance of ratio estimator is given by $V(\bar{y}_R) = \frac{(N-n)}{Nn}(S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y)$ where

$R = \frac{\bar{y}_N}{\bar{x}_N}$

Regression Estimator: Ratio estimator is used if y and x are linearly related and the line of regression between y and x are passes through origin. But when this is not the case and the variate y is approximately a constant multiple of an auxiliary variate x, the regression estimator is used.

The regression estimator can be defined as $\bar{y}_r = \bar{y}_n + b_{yx}(\bar{x}_N - \bar{x}_n)$

Regression estimator is also a biased estimate of population mean.

The variance of regression estimator is given by $V(\bar{Y}_{lr}) = \frac{(N-n)}{Nn} s_y^2(1-r_{xy}^2)$, here $r_{xy} = \frac{S_{xy}}{S_x S_y}$

and $b_{yx} = \frac{S_{xy}}{S_x^2}$ where $s_x^2 = \frac{1}{n-1}[\sum x_i^2 - \frac{(\sum x_i)^2}{n}]$ and $S_{xy} = \frac{1}{n-1}[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}]$

- Regression estimator is more efficient than Ratio Estimator $V(\bar{y}_r) < V(\bar{y}_R)$
- If correlation coefficient is equal to zero , we should not apply regression estimator.

Objective : Estimation of the average number of bullocks per acre using ratio estimator and show that it is a biased estimator of population mean. Compute bias and variance along with its standard error.

Kinds of data : A bivariate population of size N=6 is given below :

No. of bullocks(Y)	3	4	8	9	6	9
Farm Size (acre)(X)	15	20	40	45	25	42

Enumerate all possible samples of size n=2 using SRSWOR.

Solution : Here it is given that N=6 and n=2.

The total number of possible samples of size n=2 is $N_{c_n} = 6_{c_2} = 15$

S.No.	**Possible Samples (y_i)**	**Possible Samples ($\bar{x}_i$)**	**Sample mean**			
1.	3,4	15,20	3.5	17.5	6.233	0.20
2.	3,8	15,40	5.5	27.5	6.233	0.20
3.	3,9	15,45	6	30	6.233	0.20
4.	3,6	15,25	4.5	20	7.013	0.225
5.	3,9	15,42	6	28.5	6.561	0.211
6.	4,8	20,40	6	30	6.233	0.20
7.	4,9	20,45	6.5	32.5	6.233	0.20
8.	4,6	20,25	5	22.5	6.930	0.222
9.	4,9	20,42	6.5	31	6.535	0.210
10.	8,9	40,45	8.5	42.5	6.233	0.20
11.	8,6	40,25	7	32.5	6.713	0.215
12.	8,9	40,42	8.5	41	6.461	0.207

13.	9,6	45,25	7.5	35	6.679	0.214
14.	9,9	45,42	9	43.5	6.448	0.207
15.	6,9	45,42	7.5	33.5	6.978	0.224
Total				467.5	97.716	3.135

$\overline{Y_N} = \frac{\sum Y_i}{N} = \frac{187}{6}$=31.17, $\overline{Y_N} = \frac{\sum Y_i}{N} = \frac{39}{6}$ =6.50

$E(\overline{y_R}) = \frac{\sum \overline{Y_R}}{N_{Cn}} = \frac{97.716}{15} = 6.514,$

Since $E(\overline{y_R}) \neq \overline{y_N}$, *the* ratio estimator is not an unbiased estimator of population mean $\overline{Y_N}$.

The bias of ratio estimator to the first order of approximation is given by

$B_1(\overline{Y_R}) = \frac{(N-n)}{Nn} \overline{Y_N} (C_x{}^2 - \rho C_x C_y)$, where $C_x = \frac{S_x}{\overline{Y_N}}$ and $C_y = \frac{S_y}{\overline{Y_N}}$

Now, $S_X = \sqrt{\frac{1}{5} * (6639 - \frac{187^2}{6})} = 12.73$ and S_Y =2.588,

C_x= 0.408, C_y =0.397

To find out the value of ρ correlation coefficient between X and Y, we have to make the following values :

$\sum y$=39, $\sum x$=187, $\sum xy$=1378, $\sum x^2$=6639, $\sum y^2$=287

$\rho = \frac{1378 - \frac{187*39}{6*6}}{\sqrt{\frac{6639}{6} - (\frac{187}{6})^2} * \sqrt{\frac{287}{6} - (\frac{39}{6})^2}} = 0.9859$

Hence $B_1(\overline{Y_R}) = \frac{(6-2)}{6*2} * 6.50 * (0.408^2 - 0.9859 * 0.408 * 0.397)$=0.014

.

The variance of ratio estimator is given by

$V(\overline{Y_R}) = \frac{(N-n)}{Nn} (S_y{}^2 + R^2 S_x{}^2 - 2R\rho S_x S_y)$ where R = $\frac{\overline{Y_N}}{\overline{X_N}}$=0.208

$= \frac{(6-2)}{6*2}$ (2.58^2+ 0.208^2 * 12.73^2 -2*0.208*0.9859*12.73*2.58) =0.065 =0.0625.

The above formula of variance in terms of coefficient of variation can be written as :

$V(\overline{Y_R}) = \frac{(N-n)}{Nn} \overline{Y_N}^2 (C_y{}^2 + C_x{}^2 - 2\rho C_x C_y)$

$= (\frac{6-2}{6*2})$ * 6.50^2* (0.397^2 + 0.408^2 − 2 * 0.9859 * 0.408 * 0.397 = .0660

Both values of variances of ratio estimator are approximately equal.

Standard Error of Ratio Estimator ($\overline{Y_R}$) = $\sqrt{V(\overline{Y_R})} = \sqrt{0.0660}$ =0.256

Objective: Determination of the regression estimator, comparison with the ratio estimator, and its sampling variance and standard errors.

Kinds of data: A bi-variate population of size N=85 with population mean $\bar{Y}$ = 6.55 and $\bar{X}$ = 8.55, a random sample of size n=10 was drawn using SRSWOR scheme and was recorded as

Y	11	8	7	6	4	5	3	2	9	10
X	10	7	6	5	3	4	2	1	8	9

Solution: First we will calculate the $\bar{y}_n$ and $\bar{X}_N$

$\bar{y}_n = \frac{65}{10} =$ 6.5, $\bar{x}_n = \frac{55}{10} =$ 5.5, and $\bar{X}_N$ = 6.55

The equation of regression estimator $\bar{Y}_{lr} = \bar{y}_n + b_{yx}(\bar{X}_N - \bar{x}_n)$,

Where

$b_{yx} = \frac{S_{xy}}{S_x^2}$ where $s_x^2 = \frac{1}{n-1}[\sum x_i^2 - \frac{(\sum x_i)^2}{n}]$ and $S_{xy} = \frac{1}{n-1}[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}]$

											Total
Y	11	8	7	6	4	5	3	2	9	10	65
X	10	7	6	5	3	4	2	1	8	9	55
YX	110	56	42	30	12	20	6	2	72	90	440
Y^2	121	64	49	36	16	25	9	4	81	100	505
X^2	100	49	36	25	9	16	4	1	64	81	385

By putting the values we get $s_x^2 = \frac{1}{9}(385 - \frac{55^2}{10}) = \frac{82.5}{9} = 9.16$, $s_y^2 = 9.16$ and S_{xy} =9.16

So the value of $b_{yx} = \frac{S_{xy}}{S_x^2} = \frac{9.16}{9.16} = 1$

Now the equation of regression estimator $\bar{Y}_{lr} = \bar{y}_n + b_{yx}(\bar{x}_N - \bar{x}_n)$ = 6.5 + 1*(6.55-5.5)r=6.55

Estimate of the sampling variance of the estimator $\bar{Y}_r$

$V(\bar{Y}_r) = \frac{(N-n)}{Nn} s_y^2(1-r_{xy}^2)$, here $r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{9.16}{9.16} = 1$

By putting the values we get $V(\bar{Y}_r) = \frac{(85-10)}{85*10} * 9.16*(1-1^2) = 0$

Hence we can say that in case of perfect positive correlation the variance of linear regression estimator $\bar{Y}_{lr}$ for population mean $\bar{Y}_N$ will always be equal to zero.

The variance of ratio estimator is given by

$V(\bar{Y}_R) = \frac{(N-n)}{Nn}(S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y)$ where R = $\frac{\bar{y}_n}{\bar{x}_n} = \frac{6.5}{5.5} = 1.18$

$V(\bar{Y}_R) = \frac{(85-10)}{85*10}(9.16+1.18^2*9.16-2*1.18*1*9.16) = .0262$

The estimated sampling variance of the sample mean is given by

$V(\bar{y}_n)_{SRSWOR} = \frac{(N-n)}{Nn} = \frac{(85-10)}{85*10} * 9.16 = 0.808$

Here $V(\bar{Y}_r) < V(\bar{Y}_R) < V(\bar{y}_n)_{SRSWOR}$

This shows that sample mean $\bar{y}_n$ is less efficient than the ratio and regression estimator.

Simple Random Sampling:

Definition

Simple random sampling is a sampling technique where every item in the population has an equal chance and likelihood of being selected in the sample. Here the selection of items completely depends on chance or by probability and therefore this sampling technique is also sometimes known as a method of chances.

This process and technique is known as simple random sampling, and should not be confused with systematic random sampling. A simple random sample is a fair sampling technique.

Simple random sampling is a very basic type of sampling method and can easily be a component of a more complex sampling method. The main attribute of this sampling method is that every sample has the same probability of being chosen.

The sample size in this sampling method should ideally be more than a few hundred so that simple random sampling can be applied in an appropriate manner. It is sometimes argued that this method is theoretically simple to understand but difficult to practically implement. Working with large sample size isn't an easy task and it can sometimes be a challenge finding a realistic sampling frame.

Simple random sampling methods

The following steps are involved in selecting simple random sampling:

1. A list of all the members of the population is prepared initially and then each member is marked with a specific number (for example, there are nth members then they will be numbered from 1 to N).

2. From this population, random samples are chosen using two ways: random number tables and random number generator software. A random number generator software is preferred more as the sample numbers can be generated randomly without human interference.

There are two approaches that aim to minimize any biases in the process of simple random sampling:

- **Method of lottery**

Using the method of the lottery is one of the oldest methods and is a mechanical example of random sampling. In this method, each member of the population has to number systematically and in a consequent manner by writing each number on a

separate piece of paper. These pieces of paper are mixed and put into a box and then numbers are drawn out of the box in a random manner.

- **Use of random numbers**

The use of random numbers is an alternative method that also involves numbering the population. The use of a number table similar to the one below can help with this sampling technique.

Simple Random Sampling Example

An organization has 500 employees. We want to extract a sample of 100 from them.

- Step 1: **Make a list** of all the employees working in the organization. (as mentioned above there are 500 employees in the organization, the list must contain 500 names).
- Step 2: **Assign a sequential number** to each employee (1,2,3…n). This is your sampling frame (the list from which you draw your simple random sample).
- Step 3: **Figure out what your sample size is going to be**. (In this case, the sample size is 100).
- Step 4: **Use a random number generator** to select the sample, using your sampling frame (population size) from Step 2 and your sample size from Step 3. For example, if your sample size is 100 and your population is 500, generate 100 random numbers between 1 and 500.

Simple Random Sampling in research

Today, the market research projects are much larger and sometimes an indefinite number of item are involved. It is practically not possible to study the thought process of every member of the population and derive interference from the study.

If as a researcher, you want to save your time and money simple random sampling is one of the best probability sampling methods that you can use. Getting data from a sample is more advisable and practical

Whether to use a census or a sample depends on a number of factors, such as the type of census, the degree of homogeneity/heterogeneity, costs, time, feasibility to study, the degree of accuracy needed, and some others.

Advantages of Simple Random Sampling

1.It is a fair method of sampling and if applied appropriately it helps to reduce any bias involved as compared to any other sampling method involved.

2. Since it involves a large sample frame it is usually easy to pick smaller sample size from the existing larger population.

3. The person who is conducting the research doesn't need to have a prior knowledge of the data being collected. One can simply ask a question to gather the researcher need not be a subject expert.

4. This sampling method is a very basic method of collecting the data. There is no technical knowledge required and need basic listening and recording skills.

5. Since the population size is large in this type of sampling method there is no restriction on the sample size that needs to be created. From a larger population, you can get a small sample quite easily.

6. The data collected through this sampling method is well informed, more the samples better is the quality of the data.

Disadvantages of Simple Random Sampling

1. It is a costlier method of sampling as it requires a complete list of all potential respondents to be available beforehand.

2. This sampling method is not suitable for studies involving face-to-face interviews as covering large geographical areas have cost and time constraints.

3. A sample size that is too large is also problematic since every member of the population has an equal chance of selection. The larger population means a larger sample frame. It is difficult to manage the large population.

4. The quality of the data depends on the researcher and his/her perspective. If the researcher is experienced then there are fair chances the quality of data collected is of a superior quality. But if the researcher is inexperienced then the data collected may or may not be upto the mark.

Difference between SRSWOR and SRSWR (Simple random sampling without replacement and simple random sampling with replacement)

1. If the selected units are not being replaced in the population before the next draw, it is called SRSWOR ,
2. If the selected units are being replaced in the population before the next draw, it is called SRSWR ,
3. There may be some chances of having the same units of the population in the sample,
4. Therefore, SRSWOR is superior to SRSWR, because the variance of sample mean in SRSWOR is found to be always less rather than that of SRSWR and no chances of repetitions of the units in the selection.

Total number of Samples in SRSWOR: $\left(N_{C_n}\right)$

Total number of Samples in SRSWR : N^n

Where N= Population size and n= Sample size

Example 1:

Objective: Showing the unbiased estimator for population mean and biased estimator for population mean square in SRSWR with the help of an example

Kinds of data: Consider a finite population of size N=4 including the values of sampling units as (1,2,3,4). Enumerate all possible samples of size n=2 using SRSWR.

S.No.	Possible samples	Sample mean y_n	Sample mean square s^2	Sampling error
1.	1,2	1.5	0.50	-1.0
2.	1,3	2.0	2.00	-0.5
3.	1,4	2.5	4.50	0.0
4.	2,3	2.5	0.50	0.0
5.	2,4	3.0	2.00	0.5
6.	3,4	3.5	0.50	1.0
7.	2,1	1.5	0.50	-1.0
8.	3,1	2.0	2.00	-0.5
9.	4,1	2.5	4.50	0.0
10,	3,2	2.5	0.50	0.0
11.	4,2	3.0	2.00	0.5
12.	4,3	3.5	0.50	1.0
13.	1,1	1.0	0.00	1.5
14.	2,2	2.0	0.00	-0.5
15.	3,3	3.0	0.00	0.5
16.	4,4	4.0	0.00	1.5
Total		40.0	20.00	0.0

$E[y_n]$ =40.0/16=2.5

Population mean= Y_N =(1+2+3+4)/4 = 2.5

It indicates that sample mean in SRSWR provides an unbiased estimator of population mean i.e. sample mean is so strong and true that we can have true information regarding population mean. Similarly

$E[s^2]$ =20/16=1.25

And population mean square $S^2=\frac{1}{N-1}(Yi-\overline{Y_N})^2$ =5/3=1.67

It shows that in SRSWR, sample mean square is a biased estimator of population mean square, because 1.25 is not equal to 1.67.

Example 2:

Objective: Showing the unbiased estimator for population mean and population mean square in SRSWOR with the help of same example given above

Kinds of data: Consider a finite population of size N=4 including the values of sampling units as (1,2,3,4). Enumerate all possible samples of size n=2 using SRSWOR

S.No.	Possible samples	Sample mean y_n	Sample mean square s^2	Sampling error
1.	1,2	1.5	0.50	-1.0
2.	1,3	2.0	2.00	-0.5
3.	1,4	2.5	4.50	0.0
4.	2,3	2.5	0.50	0.0
5.	2,4	3.0	2.00	0.5
6.	3,4	3.5	0.50	1.0
Total		15.0	10.0	0.0

$E[y_n]$ =15.0/ 6=2.5

Population mean= Y_N =(1+2+3+4)/4 = 2.5

It indicates that sample mean in SRSWOR provides an unbiased estimator of population mean i.e. sample mean is so strong and true that we can have true information regarding population mean. Similarly

$E[s^2]$ =10/6=1.67

And population mean square $S^2=\frac{1}{N-1}(Yi-\overline{Y_N})^2$ =5/3=1.67

It shows that in SRSWOR, sample mean square is an unbiased estimator of population mean square, because 1.67 is equal to 1.67.

Stratified Random Sampling: Definition

Stratified random sampling is a type of probability sampling where the entire population is divided into non-overlapping, homogeneous groups (strata) and randomly choose samples from the various strata for research which reduces cost and improves efficiency. Members in each of these groups should be distinct so that every member of all groups get equal opportunity to be selected using SRSWOR. This sampling method is also called “restricted sampling”.

Age, socioeconomic divisions, nationality, religion, educational achievements and other such classifications fall under stratified random sampling.

Let’s consider a situation where a research team is seeking opinions about religion amongst various age groups. Instead of collecting feedback from 326,044,985 U.S citizens, random samples of around 10000 can be selected for research. These 10000 citizens can be divided into strata according to age,i.e, groups of 18-29, 30-39, 40-49, 50-59, and 60 and above. Each stratum will have distinct members.

Steps to select a stratified random sample:

1. Define the target population.
2. Recognize the stratification variable or variables and figure out the number of strata to be used. These stratification variables should be in line with the objective of the research. Every additional information decides the stratification variables. For instance, if the objective of research to understand all the subgroups, the variables will be related to the subgroups and all the information regarding these subgroups will impact the variables. Ideally, no more than 4-6 stratification variables and no more than 6 strata should be used in a sample because an increase in stratification variables will increase the chances of some variables canceling out the impact of other variables.
3. Use an already existent sampling frame or create a frame that’s inclusive of all the information of the stratification variable for all the elements in the target audience.
4. Make changes after evaluating the sampling frame on the basis of lack of coverage, over-coverage, or grouping.
5. Considering the entire population, each stratum should be unique and should cover each and every member of the population. Within the stratum, the differences should be minimum whereas each stratum should be extremely different from one another. Each element of the population should belong to just one stratum.
6. Assign a random, unique number to each element.
7. The researcher can then select random elements from each stratum to form the sample. Minimum one element must be chosen from each stratum so that there’s representation from every stratum but if two elements from each stratum are selected, to easily calculate the error margins of the calculation of collected data.

Types of Choice of Sample Sizes in Stratified Random Sampling:

- Equal allocation: ni=n/k

- **Proportional allocation:** $n_i = npi=(N_i / N) * n$

n_i= Sample size for i^{th} stratum

N_i= Population size for i^{th} stratum

N = Size of entire population

n = Size of entire sample

Example: A hypothetical population of N=2000 is divided into four strata. Their sizes of population and standard deviations are as under:

Strata	I	II	III	IV
Size(Ni)	300	400	600	700
S.D.(Si)	6	10	12	15
Sample means(yni)	5	10	15	20

A stratified random sample of size 400 is to be chosen from the population using SRSWOR.

(i) Equal allocation; ni=n/k, means 400/4=100 samples will be tken from each stratum.

(ii) Proportional allocation: ni=npi
n1=400x300/2000=60 samples from I stratum
n2=400x400/2000=80 samples from II stratum
n3=400x600/2000=120 samples from III stratum
n4=400x700/2000=140 samples from IV stratum

- Neyman allocation:ni=npisi/Σpisi
-

n1=400x(300/2000)x6/(241/20)=40 samples from I stratum
n2=400x(400/2000)x10/(241/20)=66 samples from II stratum
n3=400x(600/2000)x12/(241/20)=120 samples from III stratum
n4=400x(700/2000)x15/(241/20)=174 samples from IV stratum

In this way ,we have 400 sample from the entire population using three allocations.

Advantages of Stratified Random Sampling:

- Better accuracy in results in comparison to other probability sampling methods such as cluster sampling, simple random sampling, and systematic sampling .
- Convenient to train a team to stratify a sample due to the exactness of the nature of this sampling technique.
- Due to statistical accuracy of this method, smaller sample sizes can also retrieve highly useful results for a researcher.
- This sampling technique covers maximum population as the researchers have complete charge over the strata division.

When to use Stratified Random Sampling?

- Stratified random sampling is an extremely productive method of sampling in situations where the researcher intends to focus only on specific strata from the available population data.
- Researchers rely on this sampling method in cases where they intend to establish a relationship between two or more different strata. If this comparison is conducted using SRSWOR there is a higher likelihood of the target groups .
- Samples with a population which are difficult to access or contact, can be easily be involved in the research process using the stratified random sampling technique.
- The accuracy of statistical results is higher than simple random sampling since the elements of the sample and chosen from relevant strata. The diversification within the strata will be much lesser than the diversification which exists in the target population. Due to the accuracy involved, it is highly probable that the required sample size will be much lesser and that will help researchers in saving time and efforts.

Important Theorems in SRSWOR:

Theorem1: In SRSWOR, the probability of drawing any specified unit at rth draw is equal to the probability of drawing it at the first draw.

Proof: In order to prove this property two probability statements will be multiplied together since they are mutually exclusive.

(i) The unit is selected at the previous (r-1)th draw, and
(ii) The unit is selected at the rth draw.

Let us suppose that there are N units in the population. The probability of selecting a unit at the first draw is 1/N and the probability of its not selection at this draw is 1-(1/N)= (N-1)/N. Similarly, the probability of not selecting the unit at the second draw is (N-2)/(N-1)

In the same way, the probability of not selecting the unit at the third draw is (N-3)/(N-2).

In general, above statements say that the probability of not selecting the unit at the (r-1)th draw is (N-r+1)/(N-r+2).

The second statement was that the unit should be selected at the rth draw.

If we proceed in the same way, we can deduce this probability. The probability of selecting a unit at the first draw is 1/N. The probability of selecting a unit at the second draw is 1/(N-1). Similarly, the probability of selecting a unit at the third draw is 1/(N-2).

Then, the probability of selecting a unit at the rth draw is 1/(N-r+1).

Multiplying the above two statements we have,

$\frac{(N-1)}{N}\frac{(N-2)}{(N-1)}\frac{(N-3)}{(N-2)}.....\frac{(N-r+1)}{(N-r+2)}\frac{1}{(N-r+1)}$ which is equal to 1/N

It shows that in SRSWOR, the probability of drawing any specified unit at rth draw is equal to the probability of drawing it at the first draw.

Hence proved.

Theorem 2: In SRSWOR , the probability of inclusion any specified unit in the sample is equal to n/N.

Proof: In order to prove this property,we have to sum all the probability of selecting the units. The probability of selecting the unit at the first draw ,then for the second draw , third draw and so on and so forth

Thus, $\frac{1}{N}+\frac{1}{N}+\frac{1}{N}+\frac{1}{N}$ ……n samples which is equal to n/N.

Hence proved.