Simple Random Sampling

Simple random sampling (SRS) is a method of selection of a sample comprising of n number of sampling units out of the population having N number of sampling units such that every sampling unit has an equal chance of being chosen.

The samples can be drawn in two possible ways.

- The sampling units are chosen without replacement in the sense that the units once chosen are not placed back in the population .
- The sampling units are chosen with replacement in the sense that the chosen units are placed back in the population.

1. Simple random sampling without replacement (SRSWOR):

SRSWOR is a method of selection of n units out of the N units one by one such that at any stage of selection, anyone of the remaining units have same chance of being selected, i.e. $1/N$.

2. Simple random sampling with replacement (SRSWR):

SRSWR is a method of selection of n units out of the N units one by one such that at each stage of selection each unit has equal chance of being selected, i.e., $1/N$..

Procedure of selection of a random sample:

The procedure of selection of a random sample follows the following steps:

1. Identify the N units in the population with the numbers 1 to N.
2. Choose any random number arbitrarily in the random number table and start reading numbers.
3. Choose the sampling unit whose serial number corresponds to the random number drawn from the table of random numbers.
4. In case of SRSWR, all the random numbers are accepted ever if repeated more than once.

In case of SRSWOR, if any random number is repeated, then it is ignored and more numbers are drawn.

Such process can be implemented through programming and using the discrete uniform distribution. Any number between 1 and N can be generated from this distribution and corresponding unit can be selected into the sample by associating an index with each sampling unit. Many statistical softwares like R, SAS, etc. have inbuilt functions for drawing a sample using SRSWOR or SRSWR.

Notations:

The following notations will be used in further notes:

N : Number of sampling units in the population (Population size).

n : Number of sampling units in the sample (sample size)

Y : The characteristic under consideration

Y_i : Value of the characteristic for the i^{th} unit of the population

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$: sample mean

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$: population mean

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

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

Probability of drawing a sample :

1.SRSWOR:

If n units are selected by SRSWOR, the total number of possible samples are $\binom{N}{n}$.

So the probability of selecting any one of these samples is $\frac{1}{\binom{N}{n}}$.

Note that a unit can be selected at any one of the n draws. Let u_i be the i^{th} unit selected in the sample. This unit can be selected in the sample either at first draw, second draw, …, or n^{th} draw.

Let $P_j(i)$ denotes the probability of selection of u_i at the j^{th} draw, $j = 1,2,...,n$. Then

$$P_j(i) = P_1(i) + P_2(i) + ... + P_n(i)$$
$$= \frac{1}{N} + \frac{1}{N} + ... + \frac{1}{N} \ (n \ times)$$
$$= \frac{n}{N}.$$

Now if $u_1, u_2, ..., u_n$ are the n units selected in the sample, then the probability of their selection is

$$P(u_1, u_2, ..., u_n) = P(u_1).P(u_2), ..., P(u_n).$$

Note that when the second unit is to be selected, then there are $(n - 1)$ units left to be selected in the sample from the population of $(N - 1)$ units. Similarly, when the third unit is to be selected, then there are $(n - 2)$ units left to be selected in the sample from the population of $(N - 2)$ units and so on.

If $P(u_1) = \frac{n}{N}$, then

$$P(u_2) = \frac{n-1}{N-1}, ..., P(u_n) = \frac{1}{N-n+1}.$$

Thus

$$P(u_1, u_2, .., u_n) = \frac{n}{N}.\frac{n-1}{N-1}.\frac{n-2}{N-2}...\frac{1}{N-n+1} = \frac{1}{\binom{N}{n}}.$$

2. SRSWR

When n units are selected with SRSWR, the total number of possible samples are N^n. The Probability of drawing a sample is $\frac{1}{N^n}$.

Alternatively, let u_i be the i^{th} unit selected in the sample. This unit can be selected in the sample either at first draw, second draw, …, or n^{th} draw. At any stage, there are always N units in the population in case of SRSWR, so the probability of selection of u_i at any stage is $1/N$ for all $i = 1,2,…,n$. Then the probability of selection of n units $u_1, u_2, ..., u_n$ in the sample is

$$P(u_1, u_2, .., u_n) = P(u_1).P(u_2)...P(u_n)$$
$$= \frac{1}{N}.\frac{1}{N}...\frac{1}{N}$$
$$= \frac{1}{N^n}$$

Probability of drawing an unit

1. SRSWOR

Let A_e denotes an event that a particular unit u_j is not selected at the ℓ^{th} draw. The probability of selecting, say, j^{th} unit at k^{th} draw is

$$P \text{ (selection of } u_j \text{ at } k^{th} \text{ draw)} = P(A_1 \cap A_2 \cap .... \cap A_{k-1} \cap \bar{A}_k)$$

$$= P(A_1)P(A_2|A_1)P(A_3|A_1A_2).....P(A_{k-1}|A_1, A_2.....A_{k-2})P(\bar{A}_k|A_1, A_2.....A_{k-1})$$

$$= \left(1-\frac{1}{N}\right)\left(1-\frac{1}{N-1}\right)\left(1-\frac{1}{N-2}\right)...\left(1-\frac{1}{N-k+2}\right)\frac{1}{N-k+1}$$

$$= \frac{N-1}{N}.\frac{N-2}{N-1}...\frac{N-k+1}{N-k+2}.\frac{1}{N-k+1}$$

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

2. SRSWR

$P[\text{selection of } u_j \text{ at } k^{\text{th}} \text{ draw}] = \frac{1}{N}$.

Estimation of population mean

One of the main objectives after the selection of a sample is to know about the tendency of the data to cluster around the central value and the scatterdness of the data around the central value. Among various indicators of central tendency, the popular choice is arithmetic mean. So the population mean is generally measured by the arithmetic mean (or weighted arithmetic mean). There are various popular estimators for estimating the population mean. Among them, sample arithmetic mean is more popular than other estimators. One of the reason to use this estimator is that they possess nice statistical properties.

*One may also consider other indicators like median, mode, geometric mean, harmonic mean for measuring the central tendency.

Stratified Sampling

An important objective in any estimation problem is to obtain an estimator of a population parameter which can take care of the salient features of the population. If the population is homogeneous with respect to the characteristic under study, then the method of simple random sampling will yield a homogeneous sample and in turn, the sample mean will serve as a good estimator of population mean. Thus, if the population is homogeneous with respect to the characteristic under study, then the sample drawn through simple random sampling is expected to provide a representative sample. Moreover, the variance of sample mean not only depends on the sample size and sampling fraction but also on the population variance. In order to increase the precision of an estimator, we need to use a sampling scheme which can reduces the heterogeneity in the population. If the population is heterogeneous with respect to the characteristic under study, then one such sampling procedure is stratified sampling.

The basic idea behind the stratified sampling is to

- divide the whole heterogeneous population into smaller groups or subpopulations, such that the sampling units are homogeneous with respect to the characteristic under study within the subpopulation and
- heterogeneous with respect to the characteristic under study between/among the subpopulations. Such subpopulations are termed as **strata.**
- Treat each subpopulation as separate population and draw a sample by SRS from each stratum.

[Note: 'Stratum' is singular and 'strata' is plural].

Example: In order to find the average height of the students in a school of class 1 to class 12, the height varies a lot as the students in class 1 are of age around 6 years and students in class 10 are of age around 16 years. So one can divide all the students into different subpopulations or strata such as

Students of class 1, 2 and 3: Stratum 1
Students of class 4, 5 and 6: Stratum 2
Students of class 7, 8 and 9: Stratum 3
Students of class 10, 11 and 12: Stratum 4

Now draw the samples by SRS from each of the strata 1, 2, 3 and 4. All the drawn samples combined together will constitute the final stratified sample for further analysis.

Notations:

We use the following symbols and notations:

N : Population size

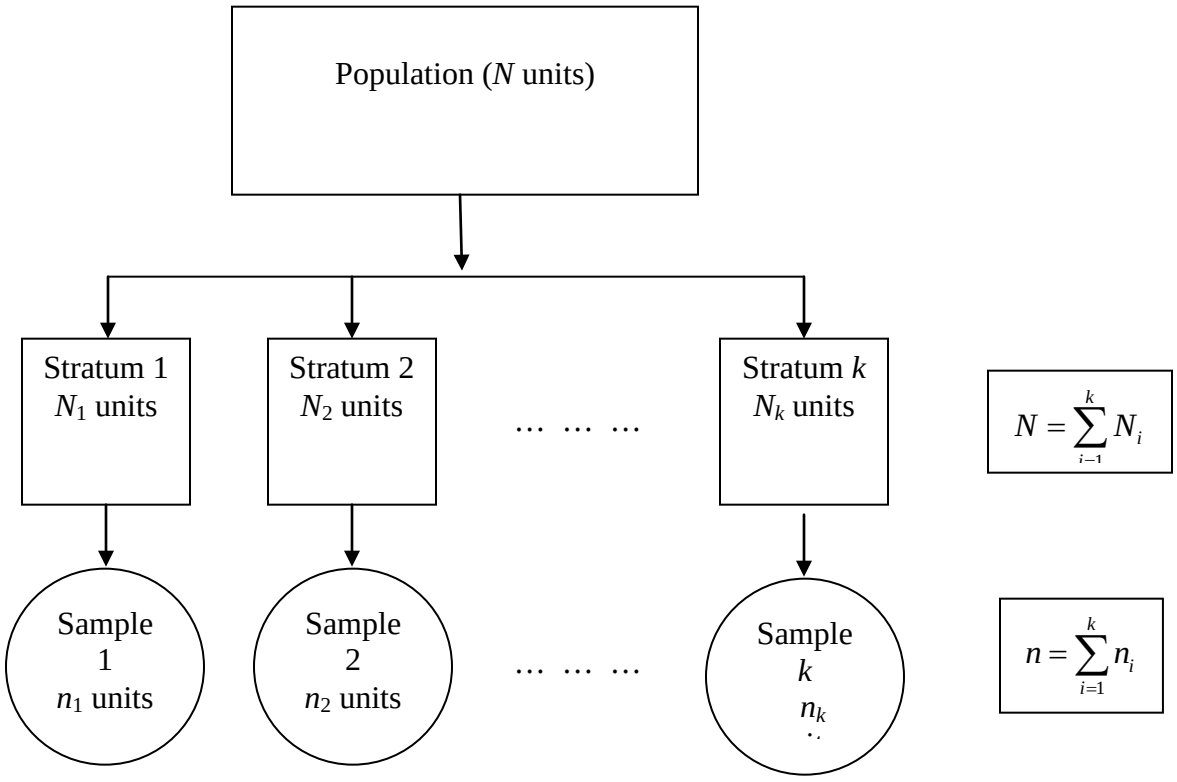
k : Number of strata

N_i : Number of sampling units in i^{th} strata

$N = \sum_{i=1}^{k} N_i$

n_i : Number of sampling units to be drawn from i^{th} stratum.

$n = \sum_{i=1}^{k} n_i$: Total sample size



Procedure of stratified sampling

Divide the population of *N* units into *k* strata. Let the i^{th} stratum has $N_1, i = 1, 2, ..., k$ number of units.

- Strata are constructed such that they are non-overlapping and homogeneous with respect to the characteristic under study such that $\sum_{i=1}^{k} N_i = N.$
- Draw a sample of size n_i from i^{th} $(i = 1, 2, ..., k)$ stratum using SRS (preferably WOR) independently from each stratum.
- All the sampling units drawn from each stratum will constitute a stratified sample of size $n = \sum_{i=1}^{k} n_i.$

Advantages of stratified sampling

1. Data of known precision may be required for certain parts of the population.

   This can be accomplished with a more careful investigation to few strata.

   Example: In order to know the direct impact of hike in petrol prices, the population can be divided into strata like lower income group, middle income group and higher income group. Obviously, the higher income group is more affected than the lower income group. So more careful investigation can be made in the higher income group strata.

2. Sampling problems may differ in different parts of the population.

   Example: To study the consumption pattern of households, the people living in houses, hotels, hospitals, prison etc. are to be treated differently.

3. Administrative convenience can be exercised in stratified sampling.

   Example: In taking a sample of villages from a big state, it is more administratively convenient to consider the districts as strata so that the administrative setup at district level may be used for this purpose. Such administrative convenience and the convenience in organization of field work are important aspects in national level surveys.

4. Full cross-section of population can be obtained through stratified sampling. It may be possible in SRS that some large part of the population may remain unrepresented. Stratified sampling enables one to draw a sample representing different segments of the population to any desired extent. The desired degree of representation of some specified parts of population is also possible.

5. Substantial gain in the efficiency is achieved if the strata are formed intelligently.

6. In case of skewed population, use of stratification is of importance since larger weight may have to be given for the few extremely large units which in turn reduces the sampling variability.

7. When estimates are required not only for the population but also for the subpopulations, then the stratified sampling is helpful.

8. When the sampling frame for subpopulations is more easily available than the sampling frame for whole population, then stratified sampling is helpful.

9. If population is large, then it is convenient to sample separately from the strata rather than the entire population.

10. The population mean or population total can be estimated with higher precision by suitably providing the weights to the estimates obtained from each stratum.

Allocation problem and choice of sample sizes is different strata

Question: How to choose the sample sizes $n_1, n_2, ..., n_k$ so that the available resources are used in an effective way?

There are two aspects of choosing the sample sizes:

(i) Minimize the cost of survey for a specified precision.

(ii) Maximize the precision for a given cost.

Note: The sample size cannot be determined by minimizing both the cost and variability simultaneously. The cost function is directly proportional to the sample size whereas variability is inversely proportional to the sample size.

Based on different ideas, some allocation procedures are as follows:

1. Equal allocation

Choose the sample size n_i to be the same for all the strata.

Draw samples of equal size from each strata.

Let *n* be the sample size and *k* be the number of strata, then

$n_i = \frac{n}{k}$ for all $i = 1, 2, ..., k.$

2. Proportional allocation

For fixed *k*, select n_i such that it is proportional to stratum size N_i, i.e.,

$n_i \propto N_i$

or $n_i = CN_i$

where *C* is the constant of proportionality.

$$\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} CN_i$$

or $n = CN$

$\Rightarrow C = \frac{n}{N}.$

Thus $n_i = \left(\frac{n}{N}\right) N_i.$

Such allocation arises from the considerations like operational convenience.

3. Neyman allocation

This allocation considers the size of strata as well as variability

$n_i \propto N_i S_i$

$n_i = C^* N_i S_i$

where C^* is the constant of proportionality.

$$\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} C^* N_i S_i$$

or $n = C^* \sum_{i=1}^{k} N_i S_i$

or $C^* = \dfrac{n}{\sum_{i=1}^{k} N_i S_i}$

Thus $n_i = \dfrac{n N_i S_i}{\sum_{i=1}^{k} N_i S_i}$.