Sampling Theory:

Often we are interested in drawing some valid inferences about a large group of individuals or objects called population in statistics. Instead of studying the entire population, which may be difficult or even impossible to study, we may study only a small portion of the population. Our objective is to draw valid inferences about certain facts for the population from results found in the sample; a process known as statistical inferences. The process of obtaining samples is called sampling and theory concerning the sampling is called sampling theory.

The sampling theory definition of the statistic is the creation of a sample set. This is recognized as one of the major processes. It retains the accuracy in bringing out the correct statistical information. The population tree is huge set and it turns out to be exhausting for the actual study and estimation process. Both money and time get exhausting in the process. The creation of the sample set saves time and effort and is a vital theory in the process of statistical data analysis.

PROCESS OF SAMPLING

In this part of the chapter, we will discuss a few details regarding the process of sampling. So the steps are mentioned in the steps below:

- The first step is a wise choice of the population set.
- The second step is focusing on the sample set and the size of it.
- Then, one needs to choose an identifiable property based on which the samples will be created out of the population set.
- Then, the samples can be chosen using any of the types of sampling theory – Simple random, systematic, or stratified. Each of them is thoroughly discussed in the article ahead.
- Checking the inaccuracy, if there is any.
- Hence, the set is achieved in the result.

Sampling can be done in their different method and they are given below:

1. Simple random type.
2. Systematic Sampling.
3. Stratified sampling.

6.3.1 Random Samples and Random Numbers:

Definition: Simple random sampling is defined as a sampling technique where every item in the population has an even chance and likelihood of being selected in the sample. Here the selection of items entirely depends on luck or probability, and therefore this sampling technique is also sometimes known as a method of chances. For e.g. Using the lottery method is one of the oldest ways and is a mechanical example of random sampling. In this method, the researcher gives each member of the population a number. Researchers draw numbers from the box randomly to

choose samples. 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 (SRS) is a method of selection of a sample comprising of *n* a 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.

Simple random sampling methods:

Researchers follow these methods to select a simple random sample:

1. They prepare a list of all the population members 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, researchers choose random samples using two ways: random number tables and random number generator software. Researchers prefer a random number generator software, as no human interference is necessary to generate samples.

Advantages of simple random sampling

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

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

3. The person conducting the research doesn't need to have prior knowledge of the data he/ she is collecting. Once can ask a question to gather the researcher need not be subject expert.

4. This sampling method is a fundamental method of collecting the data. You don't need any technical knowledge. You only require essential listening and recording skills.

5. Since the population size is vast in this type of sampling method, there is no restriction on the sample size that the researcher needs to create. From a larger population, you can get a small sample quite quickly.

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

Disadvantages:

1. Sampling is not feasible where knowledge about each element or unit or a statistical population is needed.

2. The sampling procedures must be correctly designed and followed otherwise, what we call as wild sample would crop up with mis-leading results.

3. Each type of sampling has got its own limitations.

4. There are numerous situations in which units, to be measured, are highly variable. Here a very large sample is required in order to yield enough cases for achieving statistically reliable information.

5. To know certain population characteristics like population growth rate, population density etc. census of population at regular intervals is more appropriate than studying by sampling.

6.3.2 Sampling With and Without Replacement:

Selection with Replacement (SWR): In this case, a unit is selected from a population with a known probability and the unit is returned to the population before the next selection is made (after recording its characteristic). Thus, in this method at each selection, the population size remains constant and the probability at each selection or draw remains the same. Under this sampling plan, a unit has chances of being selected more than once. For example a card is randomly drawn from a pack of cards and placed back in the pack, after noting its face value before the next card is drawn. Such a sampling method is known as sampling with replacement. There are N^n possible samples of size n from a population of N units in case of sampling with replacement.

Sampling with replacement (SWOR): In this selection procedure, if a unit from a population of size N selected, it is not returned to the population. Thus, for any subsequent selection, the population size is reduced by one. Obviously, at the time of the first selection, the population size is N and the probability of a unit being selected randomly is $\frac{1}{N}$; for the second unit to be randomly selected, the population size is $(N - 1)$ and the probability of selection of any one of the remaining sampling unit is $\frac{1}{(N-1)}$, similarly at the third draw, the probability of selection is $\frac{1}{(N-2)}$ and so on.

6.4 Sampling Distributions:

Sampling distribution is a statistic that determines the probability of an event based on data from a small group within a large population. Its primary purpose is to establish representative results of small samples of a comparatively larger population. Since the population is too large to analyze, the smaller group is selected and repeatedly sampled, or analyzed. The gathered data, or

statistic, is used to calculate the likely occurrence, or probability, of an event. Using a sampling distribution simplifies the process of making inferences, or conclusions, about large amounts of data.

The idea behind a sampling distribution is that when you have a large amount of data (gathered from a large group, the value of a statistic from random samples of a small group will inform you of that statistic's value for the entire group. Once the data is plotted on a graph, the values of any given statistic in random samples will make a normal distribution from which you can draw inferences.

Each random sample selected will have a different value assigned to the statistic being studied. For example, if you randomly sample data three times and determine the mean, or the average, of each sample, all three means are likely to be different and fall somewhere along the graph. That's variability. You do that many times, and eventually the data you plot should look like a bell curve. That process is a sampling distribution.

Factors that influence sampling distribution:

The sampling distribution's variability can be measured either by standard deviation, also called "standard error of the mean," or population variance, depending on the context and inferences you are trying to draw. They both are mathematical formulas that measure the spread of data points in relation to the mean.

There are three primary factors that influence the variability of a sampling distribution. They are:

- **The number observed in a population:** This variable is represented by "*N*." It is the measure of observed activity in a given group of data.

- **The number observed in the sample:** This variable is represented by "*n*." It is the measure of observed activity in a random sample of data that is part of the larger grouping.

- **The method of choosing the sample:** How the samples were chosen can account for variability in some cases.

Types of distributions
There are three standard types of sampling distributions in statistics.

1. Sampling Distribution of Means.
2. Sampling Distribution of Proportions.
3. Sampling Distributions of Differences and Sums.

6.4.1 Sampling Distribution of Means:

The most common type of sampling distribution is of the mean. It focuses on calculating the mean of every sample group chosen from the population and plotting the data points. The graph shows a normal distribution where the center is the mean of the sampling distribution, which represents the mean of the entire population.

The mean of the sampling distribution of the mean is the mean of the population from which the scores were sampled. Therefore, if a population has a mean μ, then the mean of the sampling distribution of the mean is also μ. The symbol $\mu_{\bar{X}}$ is used to refer to the mean of the sampling distribution of the mean. Therefore, the formula for the mean of the sampling distribution of the mean with replacement can be written as:

$$\mu_{\bar{X}} = \mu$$

The standard deviation of the sampling distribution of the mean is computed as follows:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

That is, the standard deviation of the sampling distribution of the mean is the population Standard deviation divided by $\sqrt{N}$, the sample size (the number of scores used to compute a mean). Thus, the larger the sample size, the smaller the Standard deviation of the sampling distribution of the mean.

For sampling is drawn without replacement,

The mean of the sampling distribution of means given by

$$\mu_{\bar{X}} = \mu$$

The standard deviation of the sampling distribution of means is given by

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

Sampling Distribution of Proportions:

This sampling distribution focuses on proportions in a population. Samples are selected and their proportions are calculated. The mean of the sample proportions from each group represent the proportion of the entire population.

Suppose random samples of size n are drawn from a population in which the proportion with a characteristic of interest is p.

The Sampling Distribution of Proportion measures the proportion of success, i.e. a chance of occurrence of certain events, by dividing the number of successes i.e. chances by the sample size 'n'. Thus, the sample proportion is defined as

$$p = \frac{x}{n}$$

Therefore the mean μ_p and standard deviation σ_p are given by

$$\mu_p = p,$$

$$\sigma_p = \sqrt{\frac{pq}{n}}$$

Where q is probability of non-occurrence of event, which is given by $q = 1 - p$.

The following formula is used when population is finite, and the sampling is made without the replacement:

$$\sigma_p = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{pq}{n}}$$

If n is large, and p is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution. Practically, the Normal approximation can be used when both $np \geq 10$.

Once we have the mean and standard deviation of the survey data, we can find out the probability of a sample proportion. Here, the Z score conversion formula will be used to find out the required probability, i.e.

$$Z = \frac{X-\mu}{\sigma}.$$

Sampling Distributions of Differences and Sums:

Statistical analyses are very often concerned with the difference between means. A typical example is an experiment designed to compare the mean of a control group with the mean of an experimental group. Inferential statistics used in the analysis of this type of experiment depend on the sampling distribution of the difference between means.

The sampling distribution of the difference between means can be thought of as the distribution that would result if we repeated the following three steps over and over again:

1. sample n_1 scores from Population 1 and n_2 scores from Population 2.
2. compute the means of the two samples M_1 and M_2.
3. compute the difference between means, M_1 - M_2. The distribution of the differences between means is the sampling distribution of the difference between means.

As you might expect, the mean of the sampling distribution of the difference between means is:

$$\mu_{M_1 - M_2} = \mu_{M_1} - \mu_{M_2}$$

which says that the mean of the distribution of differences between sample means is equal to the difference between population means.

From the **variance sum law**, we know that:

$$\sigma^2_{M_1-M_2} = \sigma^2_{M_1} + \sigma^2_{M_2}$$

We can write the formula for the standard deviation of the sampling distribution of the difference between means as

$$\therefore \sigma_{M_1-M_2} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2}}$$

Similarly we say about the sampling distribution of the sum between means is given by:

$$\mu_{M_1+M_2} = \mu_{M_1} + \mu_{M_2}$$

$$\sigma_{M_1+M_2} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2}}$$

Standard Errors:

Another measure is standard error, which is the standard deviation of the sampling distribution of an estimator. The idea is that if we draw a number of repeated samples of fixed size n from a population having a mean μ and variance σ^2 , each simple mean, say $\bar{x}$, will have a different value. Here $\bar{x}$ it is a random variable and hence it has a distribution. The standard deviation of $\bar{x}$ is called **standard error**. It has been proved that the standard error $'\sigma_{\bar{x}}$ of the mean $\bar{x}$ based on a sample of size n is,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

From above formula, it is obvious that the larger the sample size, the smaller the standard error and vice-versa. The advantage of considering standard error instead of a standard deviation is

that this measure is not influenced by the extreme values present in a population under consideration.

In reality neither we use σ to calculate the standard error of $\bar{x}$ nor we take more than one sample. As a matter of fact, what we do is, that we select only one sample, find its standard deviation s and use the following formula to find out the standard error of $\bar{x}$ .i.e.

$$S.E.(\bar{x}) = \frac{s}{\sqrt{n}}$$

Standard error is commonly used in testing of hypothesis and interval estimation. Many distributions, which are originally not normally distributed, have been taken as normal by considering the distribution of mean $\bar{x}$ for a large n.