Inferential Statistics

One of the most important application of statistics is making estimations about an entire population based on the information from a small sample. This process is known as **statistical inference**. This can be achieved only if we are confident that our sample accurately reflects the desired population. For example, making exit poll results of public opinions using a small group of thousand voters and exactly predicting the outcome of an election in which millions of votes are cast.

This chapter on *inferential statistics* will take you to see how to draw conclusions from a sample and generalize them to a larger population.

Population and Sample:

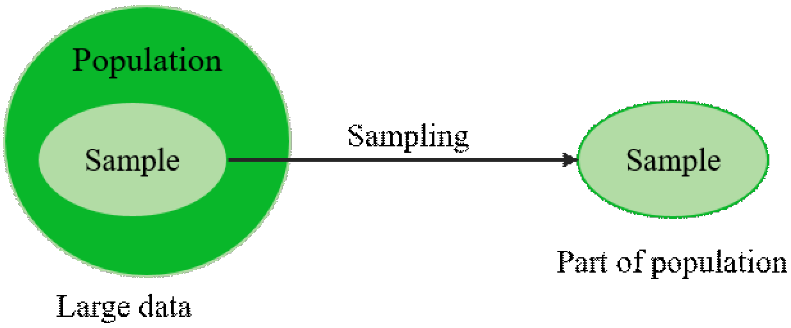
Several real-life problems are statistical in nature. Let's take some examples;

1. You are a part of a fitness campaign in your school. You are concerned about the overall wellbeing of fellow students and want to know that what proportion of students regularly do exercises.
2. As a quality control expert, you want to know what percentage of good computer chips are produced by the manufacturing unit of your company in a week.

In example 1, the population under study is total number of students enrolled in school as you want to conduct study on them. In example 2, the population is the total number of computer chips produced by manufacturing unit in a week then out of it you will see what proportion is good.

Thereby a *population* is a group of all distinct individuals or objects that you want to draw conclusions about. The number of individuals/objects in a population is called population size.

In statistics, we commonly use a *sample* that is a small subset of a larger set of data for making inferences about the large set. Here larger set is population out of which sample is drawn



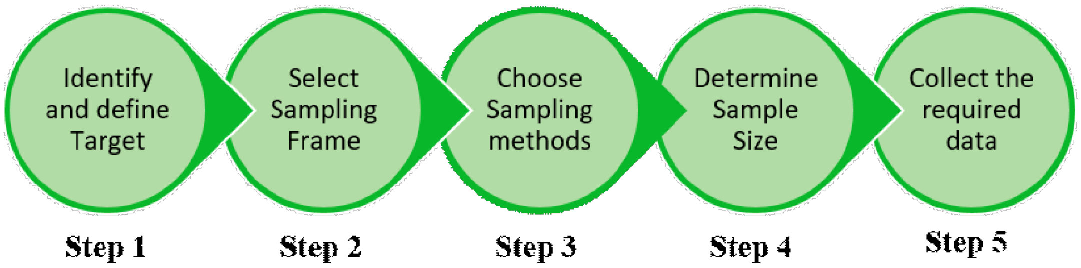
***NOTE** : ● Every time the sample size is smaller than the population's total size.*

The population refers to the entire group from which you want to draw conclusions.

Sampling is a technique of selecting small group (subset) of population for estimating the characteristics, without having to investigate every individual. It includes selecting a group of people, events, behaviors, or other elements with which we are concerned to make our conclusions. We can extend our results obtained from sample group to the entire population.

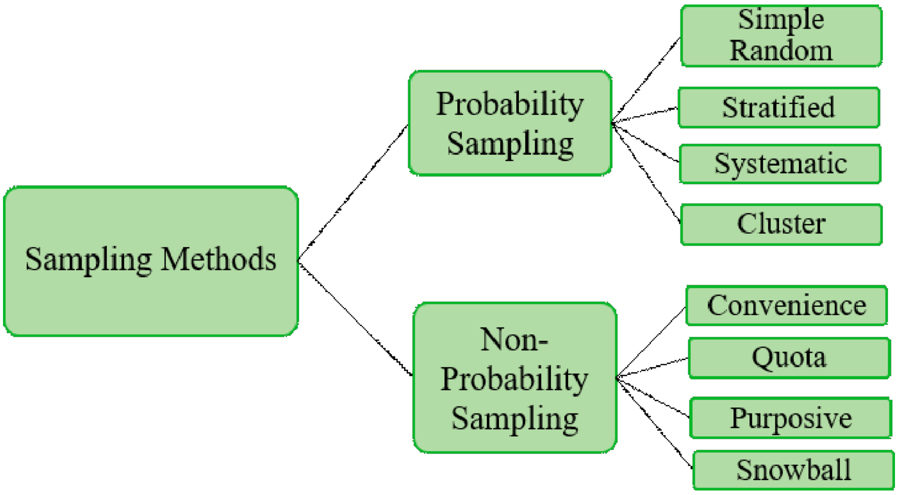
Let us suppose a vaccine company has manufactured a new vaccine for COVID-19 and would like to see its adverse effects on country's population, then it is almost impossible to perform clinical trials that includes all. So in this scenario, researchers select a group of people from each demographic for conducting the tests on them and estimates the impact on whole population.

Steps involved in Sampling



here are number of ways in which the sampling process can be carried out. But in this chapter, we shall limit ourselves to simple random sampling and systematic random sampling only.

BONUS if you want to learn more on the sampling techniques these are some methods:



1. **Probability Sampling:** Randomization (choosing something at random) is used in this sampling method to ensure that every member of the population has an equal chance of being included in the selected sample.
2. **Non-Probability Sampling:** Randomization is not used in non-probability sampling. The result of this method can be biased, making it difficult for all the elements of population to be included in the sample equally.

Parameter	Statistic
➢ It is a characteristic of a population.	➢ It is a characteristic of a sample.
➢ A parameter is a numerical value that is taken from the entire population, such as the population mean.	➢ A statistic is the numerical value taken from a sample and calculated from the sample observations alone, i.e. some subset of the entire population.
➢ The value of a parameter is computed from all the population observations.	➢ The value of a statistic is computed from portion of population (sample).
➢ Generally denoted by Greek alphabets (mean-μ, S.D.-σ, Variance- σ^2 etc.)	➢ Generally denoted by english alphabets (Mean –X, S.D. –S, Variance –S^2, etc.
Example: Under a study of calculating the average income of people of some specific region, the mean income and standard deviation of these incomes are parameters. Knowing the average height of adults in India is a parameter which is nearly impossible to calculate.	***Example:*** Mean and standard deviation of income of 1000 residents from South Delhi. Mean and standard deviation of height of 50 Indian adults.

Thus, Parameter and statistic both are related yet distinct measures. The first refers to the whole population, while the second refers to part of the population.

Statistical Significance and Sampling distribution

Statistical Significance

Statistical significance is a measure of reliability of findings which establishes that when a finding is significant, it simply means we are confident that it is real and sample was framed wisely.

Sampling distribution

The sampling distribution of a statistic is the distribution of all possible values taken by the statistic when all possible samples of a fixed size n are taken from the population. It is a theoretical idea—we do not actually build it.

To put it another way, suppose we are regularly taking samples of the same sample size from the population, compute the statistics (Mean, S.D. mean), and then draw a histogram of those statistics, the distribution of that histogram tends to have is called the sample distribution of that particular statistics (Mean, S.D.).

Central Limit Theorem:

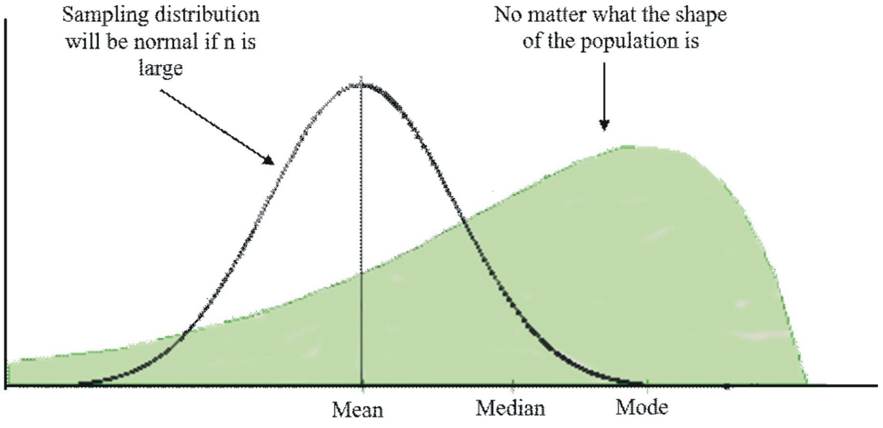
Refer: https://onlinestatbook.com/stat_sim/sampling_dist/

Central limit theorem (CLT) implies that the distribution of a sample leads to become a normal distribution (bell curve shaped) as the sample size becomes larger, considering that all the sizes of samples are identical, whatever be the shape of the population distribution.

A sample size of 30 or more is considered to be sufficient to hold CLT and as the sample size becomes large the prediction of characteristics of population becomes more accurate.

NOTE *: As per CLT, when sample size increases the mean of a sample of data becomes close to mean of overall population.*

The interesting thing about CLT is that as N increases, the sampling distribution of the mean approaches a normal distribution, regardless of the shape of the parent population.

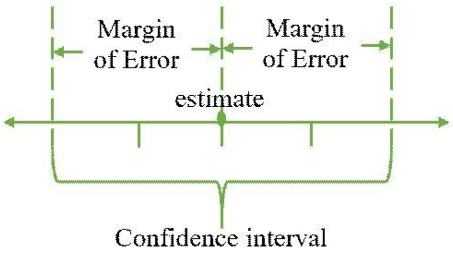


Central Limit Theorem

Confidence Interval:

We use Confidence Interval (CI) to express the precision and uncertainty of a sampling process. A confidence level, a statistic and a margin of error are the three components of it. The margin of error describes the accuracy of a sampling method, while the confidence level explains its uncertainty.

Consider the case where we are computing an interval estimate of a population parameter with a 95% confidence interval. It means that 95% of the time, by using the same sampling method to pick different samples and computing different interval estimates, the true population parameter would fall within the margin of error specified by the sample statistic.



For example: Assume a news channel conducts pre-election survey and predicts that the candidate A will get 30% of the vote. According to news channel the survey had margin of error of 5% and a confidence level of 95%. This means that we are 95% sure that the candidate A will receive between 25% and 35% of the vote.

Standard Error of Mean:

When we take a sample from a population, we pick up one of many samples. Some of them will have the same mean whereas some will have very different means. Standard error of the mean (SEM) measures how much dispersion there is likely to be in a sample's mean compared to the population mean i.e it measures the standard deviation of sampling distribution about the mean.

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

σ = Standard deviation of original

distribution

N = Sample size

- **Small SEM:** Having large number of observations and all of them being close to the sample mean (large N, small SD) gives us confidence that our estimation of the population means (i.e., that it equals the sample mean) is relatively accurate.
- **Large SEM:** Having small number of observations and they vary a lot (small N, large SD), then population estimation is likely to be quite inaccurate.