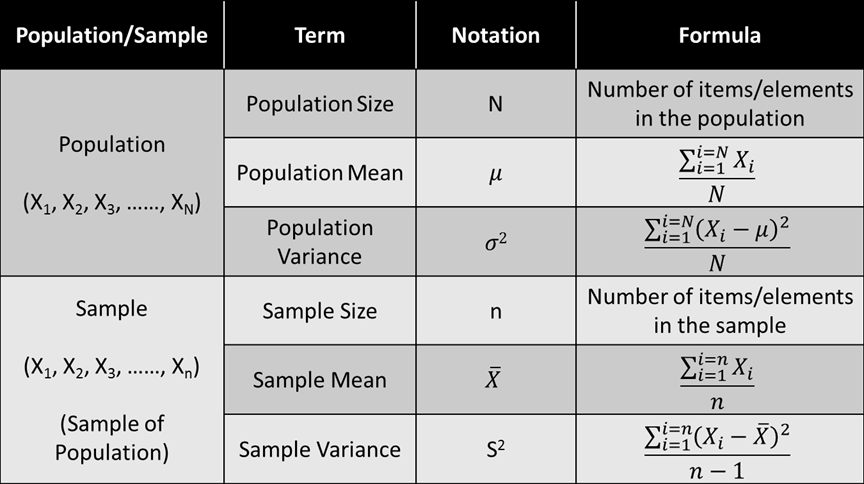
# CENTRAL LIMIT THEOREM

We are going to learn what a sample is and why it is so error prone. You will then understand how to quantify this error made in sampling using a popular theorem in statistics, called the central limit theorem.

## SAMPLES:

So far, we have conducted analyses for data on 75 people, 3,000 people, and so on. But what if you need to analyze a very large amount of data, e.g., data on 3,00,000 people? Or, what if you need to do this for, say, the entire Indian population?

Suppose for a business application, you want to find out the average number of times people in urban India visited malls last year. That’s 400 million (40 crore) people! You cannot possibly go and ask every single person how many times they visited the mall. That’s a costly and time-consuming process. How can you reduce the time and money spent on finding this number?



Degrees of freedom in the context of sample variance have to do with how much freedom you have to choose the values in your dataset. When you're working with a sample (a smaller group of data from a larger population), you usually have one less degree of freedom compared to working with the entire population. Here's why:

Population Variance: When you calculate the population variance, you use all the data points from the entire population to compute the variance. You calculate how each point deviates from the population mean. Since you have data for the entire population, you can calculate the mean precisely because there's no uncertainty about which values are included.

Sample Variance: In contrast, when working with a sample, you don't have the luxury of knowing all the data from the entire population. You can only calculate the sample mean based on the data you have in your sample. The reason you divide by n−1 in the sample variance formula is to account for the fact that you're using your sample mean instead of the population mean, and this introduces some uncertainty.

Let's break it down:

* If you had used the population mean to calculate the variance, you would have n degrees of freedom (no loss of freedom because you're certain about the population mean).
* However, you use the sample mean instead, which is calculated from the sample itself. This choice consumes one degree of freedom. So, to correct for this "lost" degree of freedom, you divide by n−1 instead of n to make your estimate of the variance more accurate.

In simpler terms, you're being a bit more conservative when calculating the sample variance by dividing by n−1 to ensure that your estimate of the population variance is less biased and closer to the true value, given that you're working with a sample rather than the whole population. This correction helps account for the variability that might exist between different samples from the same population.

## SAMPLING DISTRIBUTION:

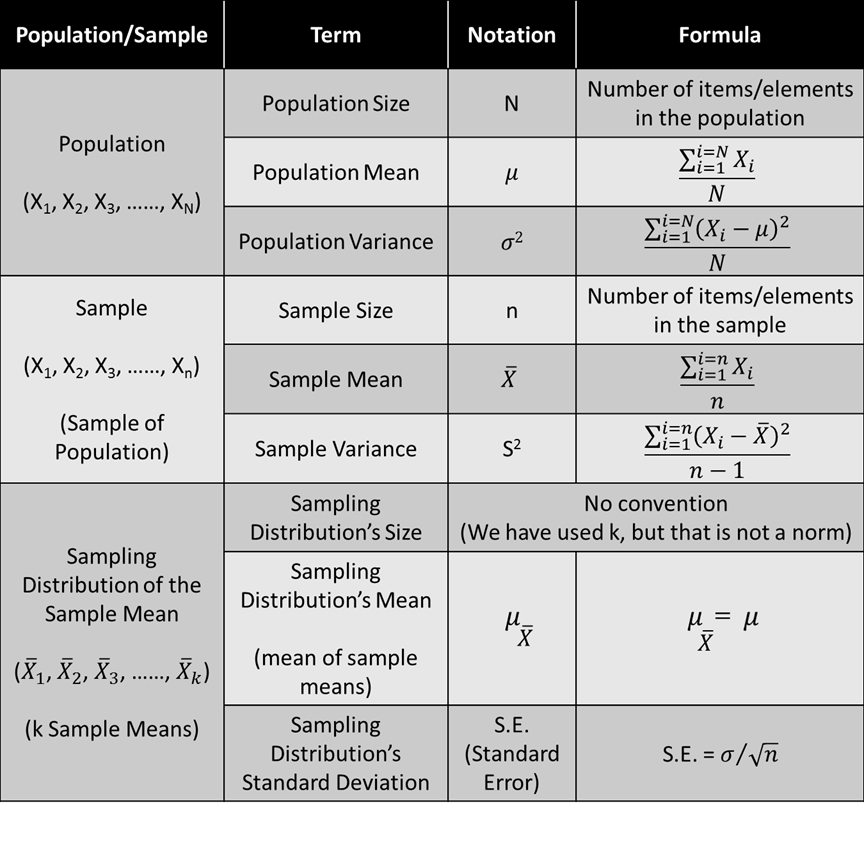
A sampling distribution is a probability distribution that describes the statistical properties of a particular statistic (such as the mean, variance, or proportion) calculated from a sample of data drawn from a larger population.

**Sampling Distribution of Sample Means:** The sampling distribution of the sample means, is a probability density function for the sample means of a population. The **sampling distribution mean is denoted by**μX**,** as it is the mean of the sampling means.

## PROPERTIES OF SAMPLING DISTRIBUTION:

there are two important properties of a sampling distribution of the mean:

1. **Sampling distribution’s mean** (μX) = **Population mean**(μ)
2. Sampling distribution’s standard deviation (**Standard Error**) = σ/√n, where σ is the population’s standard deviation and n is the sample size



## CENTRAL LIMIT THEOREM:

For a sufficiently large sample size, the sampling distribution of the sample mean (or the sum) of a random sample taken from any population, will be approximately normally distributed.

(or)

The central limit theorem says that for any kind of data, provided a high number of samples has been taken, the following properties hold true:

1. **Sampling distribution’s mean** (μX) = **Population mean** (μ),
2. Sampling distribution’s standard deviation (**standard error**) = σ/√n, and
3. **For n > 30**, the sampling distribution becomes a **normal distribution**.

Please refer to the jupyter notebook on CLT Demonstration to get a more detailed idea of how it works.