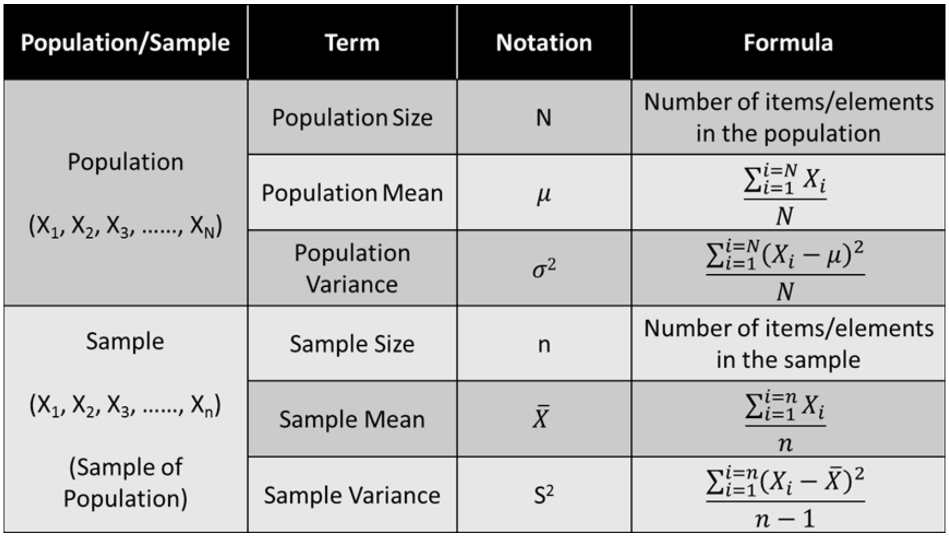
CENTRAL LIMIT THEOREM

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?



Population/Sample	Term	Notation	Formula
Population $(X_1, X_2, X_3, \ldots\ldots, X_N)$	Population Size	N	Number of items/elements in the population
	Population Mean	μ	$\frac{\sum_{i=1}^{i=N} X_i}{N}$
	Population Variance	σ^2	$\frac{\sum_{i=1}^{i=N} (X_i - \mu)^2}{N}$
Sample $(X_1, X_2, X_3, \ldots\ldots, X_n)$ (Sample of Population)	Sample Size	n	Number of items/elements in the sample
	Sample Mean	$\bar{X}$	$\frac{\sum_{i=1}^{i=n} X_i}{n}$
	Sample Variance	S^2	$\frac{\sum_{i=1}^{i=n} (X_i - \bar{X})^2}{n-1}$

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:

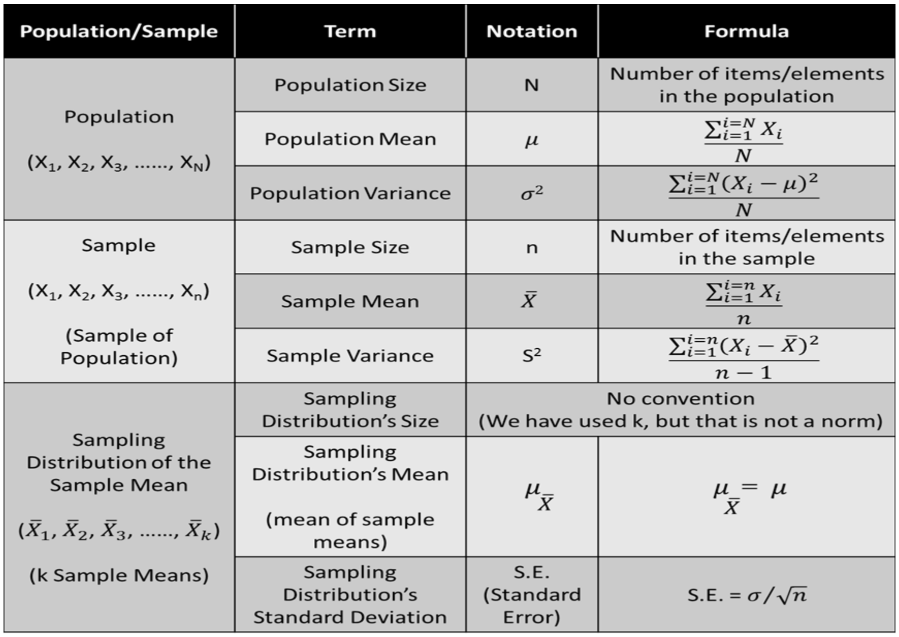
A sampling distribution is a probability distribution that describes the statistical properties of a particular statistic (such as the mean, variance, or proportion) calculated from a sample of data drawn from a larger population.

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** μ_X**,** 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** (μ_X) = **Population mean** (μ)
2. Sampling distribution's standard deviation (**Standard Error**) = $\sigma/\sqrt{n}$, where σ is the population's standard deviation and n is the sample size



Population/Sample	Term	Notation	Formula
Population $(X_1, X_2, X_3, \ldots\ldots, X_N)$	Population Size	N	Number of items/elements in the population
	Population Mean	μ	$\frac{\sum_{i=1}^{i=N} X_i}{N}$
	Population Variance	σ^2	$\frac{\sum_{i=1}^{i=N}(X_i - \mu)^2}{N}$
Sample $(X_1, X_2, X_3, \ldots\ldots, X_n)$ (Sample of Population)	Sample Size	n	Number of items/elements in the sample
	Sample Mean	$\bar{X}$	$\frac{\sum_{i=1}^{i=n} X_i}{n}$
	Sample Variance	S^2	$\frac{\sum_{i=1}^{i=n}(X_i - \bar{X})^2}{n-1}$
Sampling Distribution of the Sample Mean $(\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots\ldots, \bar{X}_k)$ (k Sample Means)	Sampling Distribution's Size	No convention (We have used k, but that is not a norm)	
	Sampling Distribution's Mean (mean of sample means)	$\mu_{\bar{X}}$	$\mu_{\bar{X}} = \mu$
	Sampling Distribution's Standard Deviation	S.E. (Standard Error)	S.E. = $\sigma/\sqrt{n}$

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** (μ_X) = **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.

PRACTICE QUESTIONS:

Comprehension 1: Let's say that you work for a news agency, which is conducting an exit poll for the MCD (Municipal Corporation of Delhi) elections. You have been tasked with predicting the winner for ward 75N (Ashok Vihar). You ask 100 randomly selected voters from this ward to name the party they had voted for.

The data thus collected is as given in the following table.

Party	No. of Votes
BJP	58
INC	42

From this sample, you have to estimate the percentage of voters that may have voted for BJP.

Define X as the proportion of people that voted for BJP. Then, the frequency distribution for X would be as follows:

X	Frequency
1	58
0	42

Mean for this data: (1*58+0*48)/100 = 0.58 or 58%

Variance = $58*(1-0.58)^2 + 42*(0-0.58)^2/(100-1) = 0.246$

SD = 0.496

ESTIMATING MEAN USING CLT:

Population mean = Sample mean +/- margin

We can find this margin using CLT.

let's say that you have a sample with sample size n, mean $\bar{X}$ and standard deviation S. Now, the **y% confidence interval** (i.e., the confidence interval corresponding to a y% confidence level) for μ would be given by the range: Confidence interval = $\boxed{(\bar{X} - \frac{Z^*S}{\sqrt{n}}, \bar{X} + \frac{Z^*S}{\sqrt{n}})}$

Z* is the Z-score associated with a y% confidence level. In other words, the population mean and the sample mean differ by a **margin of error** given by $\frac{Z^*S}{\sqrt{n}}$. Some commonly used Z* values are given below:

Confidence Level	Z (+/-)	Risk of error	Interval Width	Common Usage	
90	1.65	10%	Narrow	Reasonable level of confidence with some risk of being wrong.	90% sure that true value lies in between the range
95	1.96	5%	Moderate	For Scientific research. Balance between precision and confidence.	95% sure that true value lies in between the range
99	2.58	1%	Wide	For critical decisions	99% sure that true value lies in between the range

In simple terms, the confidence interval indicates how confident you are on the population mean in terms of calculated sample mean. Example: 95% sure that the population mean is in the interval 160cm – 170cm, while the sample mean is 165cm.

The wider the confidence interval, the less certain you are about the true value. The narrower it is, the more certain you are. Confidence intervals are a way of acknowledging the uncertainty in statistics and expressing it in a user-friendly manner. (99% is most conservative with the wider confidence interval, while 90% confidence interval is less conservative with narrow interval and low certainty)