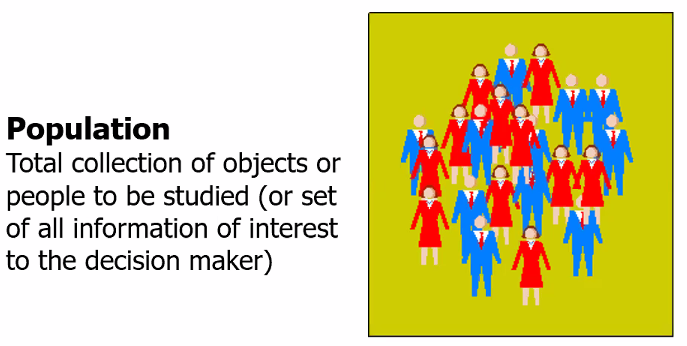
# Condition of a distribution being normal distribution

1. If mean ~ median ~ mode distribution is symmetric but cannot be called normal
2. For data to be normally distributed the Emperical rules apply:
   1. P(µ - σ ≤ X ≤ µ + σ ) = P(-1 ≤ Z ≤ +1) = 0.6827 i.e. ~ 68.2% of data points should lie between ±1σ
   2. P(µ - 2σ ≤ X ≤ µ + 2σ ) = P(-2 ≤ Z ≤ +2) = 0.9545 i.e. ~ 95.45% of data points should lie between ±2σ
   3. P(µ - 3σ ≤ X ≤ µ + 3σ ) = P(-3 ≤ Z ≤ +3) = 0.9974 i.e. ~ 99.74% of data points should lie between ±3σ
3. Other methods include:
   1. Visual inspection of the histogram (not very accurate)
   2. Numerical summaries like Skewness and Kurtosis
   3. Graphical summaries (Normal Quantile plot)

# Statistical Inference

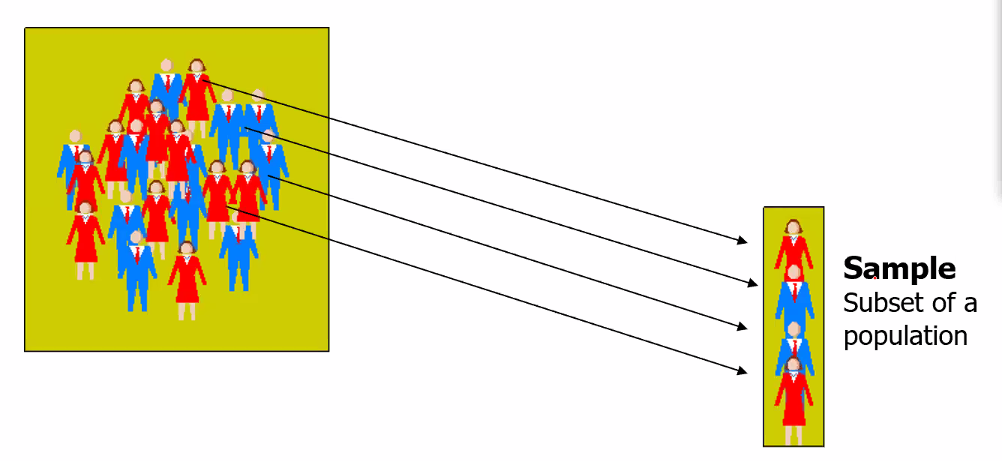
Definition: Drawing conclusion about the population from the sample data.



If ISB: all students are the population

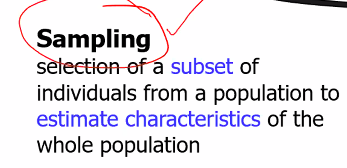
If India: The whole population of India is a population.

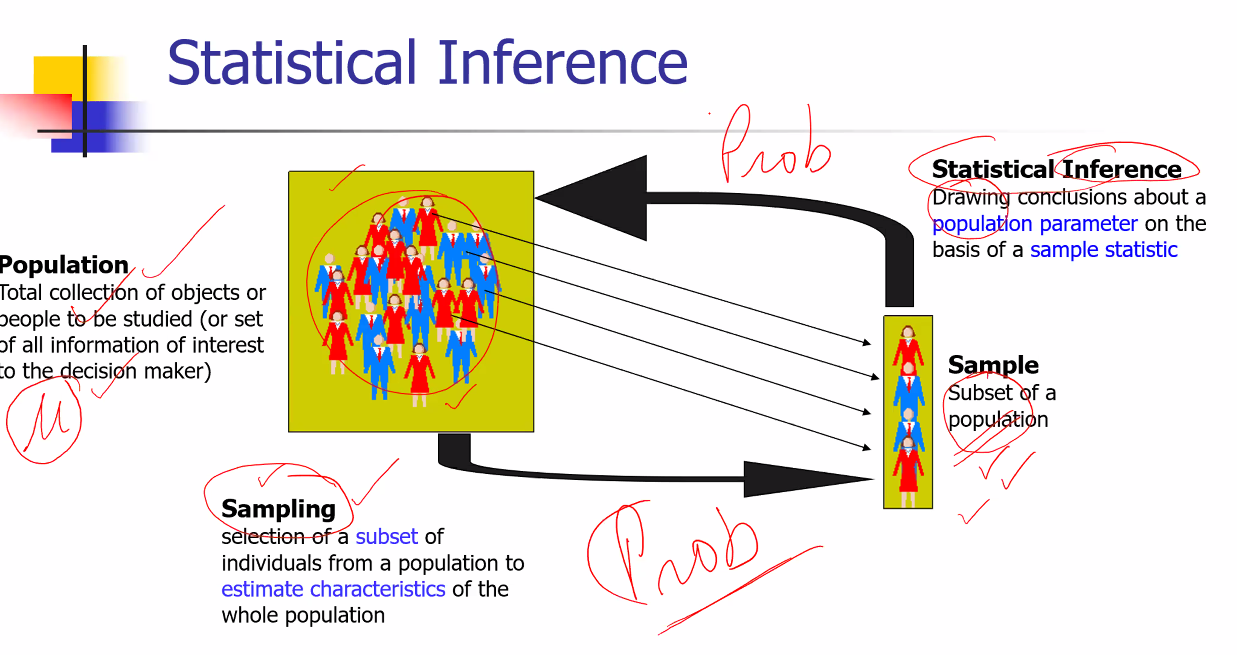
The end idea is study the characteristics of this population.



The sample should be representing the population.

The process of taking the sample is called **sampling**.

**Sampling: **

****

**There is a certain amount of uncertainty with sampling.** The main part of the statistical inference is how can we factoring the uncertainty.

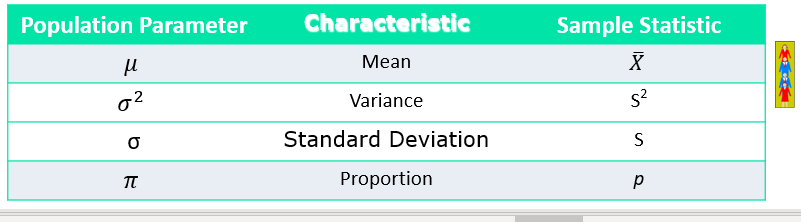
Both population 🡨🡪 sampling involves probability

# Terms for sample and population

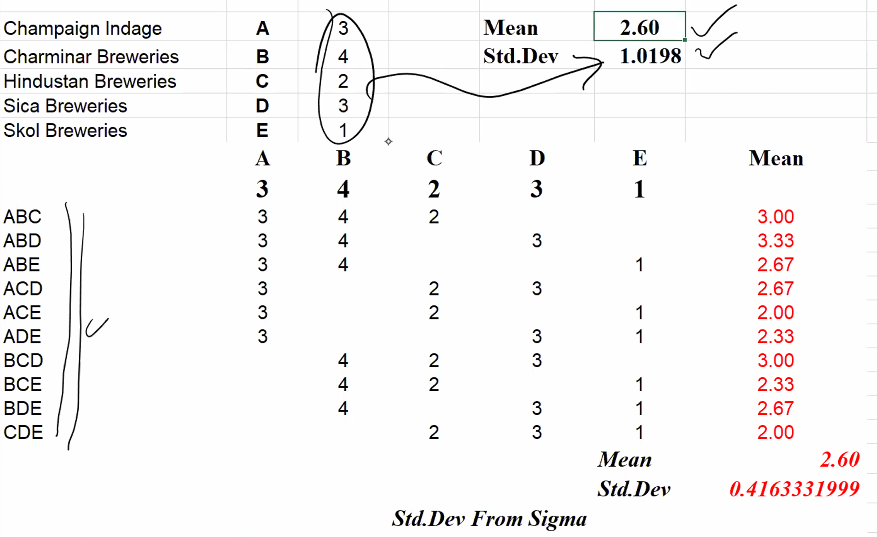
1. Sample terms with statistics
2. Population terms with parameters.

Greek letter to denote population. All sample are denoted by English letters.

**All population belong to Greece and all samples belong to England.**



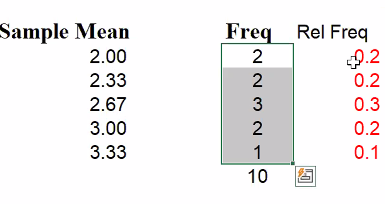
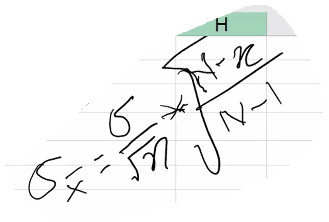
# The exercise on sampling 5 companies into groups of 3



So below are the results:

1. Population mean = 2.6
2. Population standard deviation = 1.0198
3. Mean(Sample means) = 2.6
4. Std(Sample Means) = 0.4163

Below are the observations:

1. Out of the samples there no sample mean = population mean. So P(sample mean=population mean) = 0. So it is least likely that none of the sample mean = population mean. But mean of means = population mean.
2. If we see the sample mean frequency distribution. The sample means becomes a random variable as it has value and probability associated with it. But as the sample mean takes only the 5 values. So it is discrete. Now the E(Sample Mean) = Req Freq o Sample Mean (o = sum product). Calculating this gives 2.6 = population mean.
3. Why this happens? What is the probability of A or B or C is included in our samples? It is equal to 0.6. In real life what happens is that no one will be doing study of all the combinations. They will take only one sample. So why the E(Sample mean) = E(Population). Now what is the probability of A getting selected? 1/5. B getting selected? ¼. C getting selected? 1/3 if selection is done without replacement. While P(A being included in the sampling) = 0.6, P(B being included in the sampling) = 0.6
4. σ = 1.0198 σ(x\_bar) = 0.416. The relation between is given as: σ(x\_bar) and σ

Here N = Population size = 5, n = sample size = 3, σ = 1.0198.

Now in our case N is small in size. But if we consider the population becomes large enough and sample size is small enough then the

# Impact of selecting number of samples

Suppose there are 4000 companies in BSE. We are selecting sample sizes of 100. This gives us a combination of 4000C100 = 4.94e201. Thus X\_bar becomes a continuous variable.

If you take a sample and the sample is taken from simple random sampling.

If you do that x\_bar will follow normal distribution.

E(x\_bar) = µ

σx\_bar = σ/root(n)

x\_bar is a continuous random variable.

This is the essence of Central Limit theorem.

Here we will term σx\_bar = σ/root(n) as **std. error.**

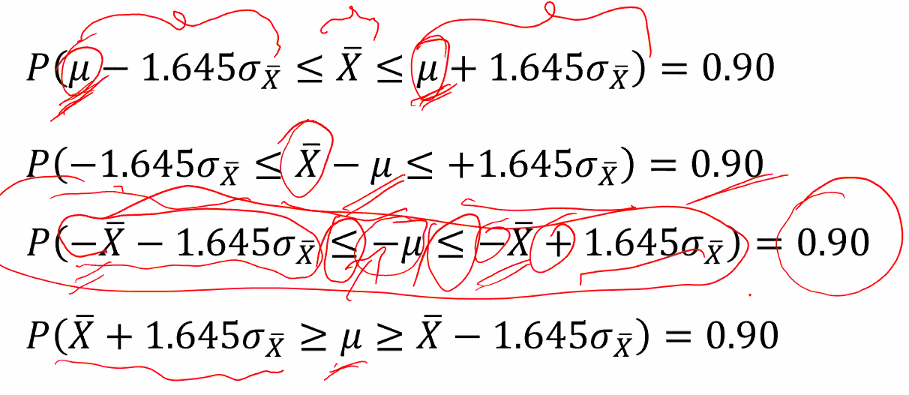
# Confidence intervals

## Business case

The business case is defined as the following:

* SariSagar, an online retailer for designer saris. They are planning to introduce a new line of “tussar” silk saris, ranging between Rs. 2,180 and Rs. 8,420.
* The success of the new line depends on the order size (in Rs.)
* A sample of 2,000 potential customers were randomly selected from the database and a campaign is launched.
* SariSagar received orders from 144 customers
* Sample average (average order size) = ₹ 6,540
* Based on the previous campaigns, it was assumed that the population standard deviation (σ) is ₹ 4,608 from previous such launches.
* What can we say about the average order size that will be held after a full-fledged market launch?

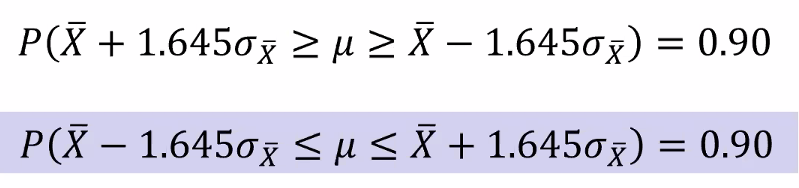
The saree problem statement calculations.



The normal distribution gets flipped by 180o.

I don’t like >= signs

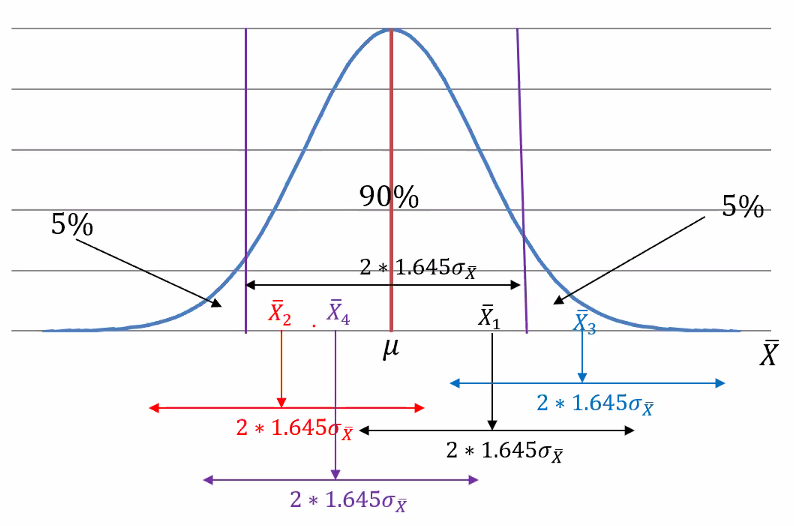
Flip the left portion with the right portion.



Confidence interval: Consider the three statements:

1. This AMPBA group has average experience between 6 to 7 years – This is the most precise but lower confidence level.
2. This AMPBA group has average experience between 3 to 10 years
3. This AMPBA group has average experience between 0 to 20 years
4. Question is can we achieve high confidence and precision? Yes by increasing the sample size.
5. Another way to achieve high confidence and precision? By reducing sigma. How? Using IS:1900. This is the variation in our process. This is done by Quality Control. Make people do exactly the same thing the same way again and again. Many situations sigma is changed. But will cost me more than increasing sample size. So it is actually a strategy not a method. ISO 9000 does that.

## What is the random variable in confidence intervals



Here we can see:

1. X\_bar1 ± 2\*1.645σx\_bar contains µ
2. X\_bar2 ± 2\*1.645σx\_bar contains µ
3. X\_bar3 ± 2\*1.645σx\_bar doesn’t contains µ
4. X\_bar4 ± 2\*1.645σx\_bar contains µ

By 90% confidence level we are saying that if we go on taking 1000 such samples and go on building 2\*1.645σx\_bar window around each sample means X\_bar, 900 times we will be able to get a window which will contain µ. Thus the random variable is actually the range here.

**Visualize:** [**https://github.com/sandipto-sanyal/ampba/blob/conceptual\_examples/statistics/Confidence%20Interval%20Concepts.ipynb**](https://github.com/sandipto-sanyal/ampba/blob/conceptual_examples/statistics/Confidence%20Interval%20Concepts.ipynb)

## Confidence interval for the population mean

When we take a sample there is always an uncertainty associated. Suppose if a light pole is being installed right next to our house? What should be a tolerance for error level? Say 1%.

But suppose if someone is building a nuclear power plant beside my house? What should be the tolerance limit for the error? It should be close to 0%. Application is different and thus error tolerance level.

If confidence level is 90% we are taking an error tolerance is 10%.