FUTURE VISION™
COMPUTER INSTITUTE

Lecture Notes

Inferential Statistics

Exploratory data analysis helped you understand how to **discover patterns in data** using various techniques and approaches. As you've learnt, EDA is one of the most important parts of the data analysis process. It is also the part on which data analysts spend most of their time.

However, sometimes, you may require a very large amount of data for your analysis which may need too much time and resources to acquire. In such situations, you are forced to work with a **smaller sample of the data**, instead of having the entire data to work with.

Situations like these arise all the time at big companies like Amazon. For example, say the Amazon QC department wants to know what proportion of the products in its warehouses are defective. Instead of going through all of its products (which would be a lot!), the Amazon QC team can just check a small sample of 1,000 products and then find, for this sample, the defect rate (i.e. the proportion of defective products). Then, based on this sample's defect rate, the team can **"infer"** what the defect rate is for all the products in the warehouses.

This process of "inferring" insights from sample data is called "Inferential Statistics".

Random Variables

Before performing any kind of statistical analysis on a problem, it is advisable to **quantify** its outcomes by using random variables.
So, the **random variable X** basically converts outcomes of experiments to something measurable.

For example, recall that we quantified the colours of the balls we would get after playing our game by assigning a value of X to each outcome. We did so by defining X as the number of red balls we would get after playing the game once.

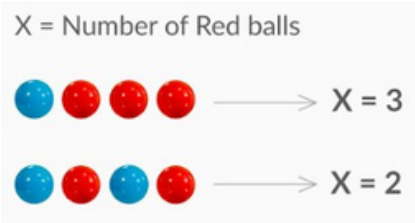


Figure 1 - Quantifying Using Random Variables

Probability Distribution

A **probability distribution** for X, basically, is ANY form of representation that tells us the probability for all possible values of X. It could be a table, a chart or an equation.

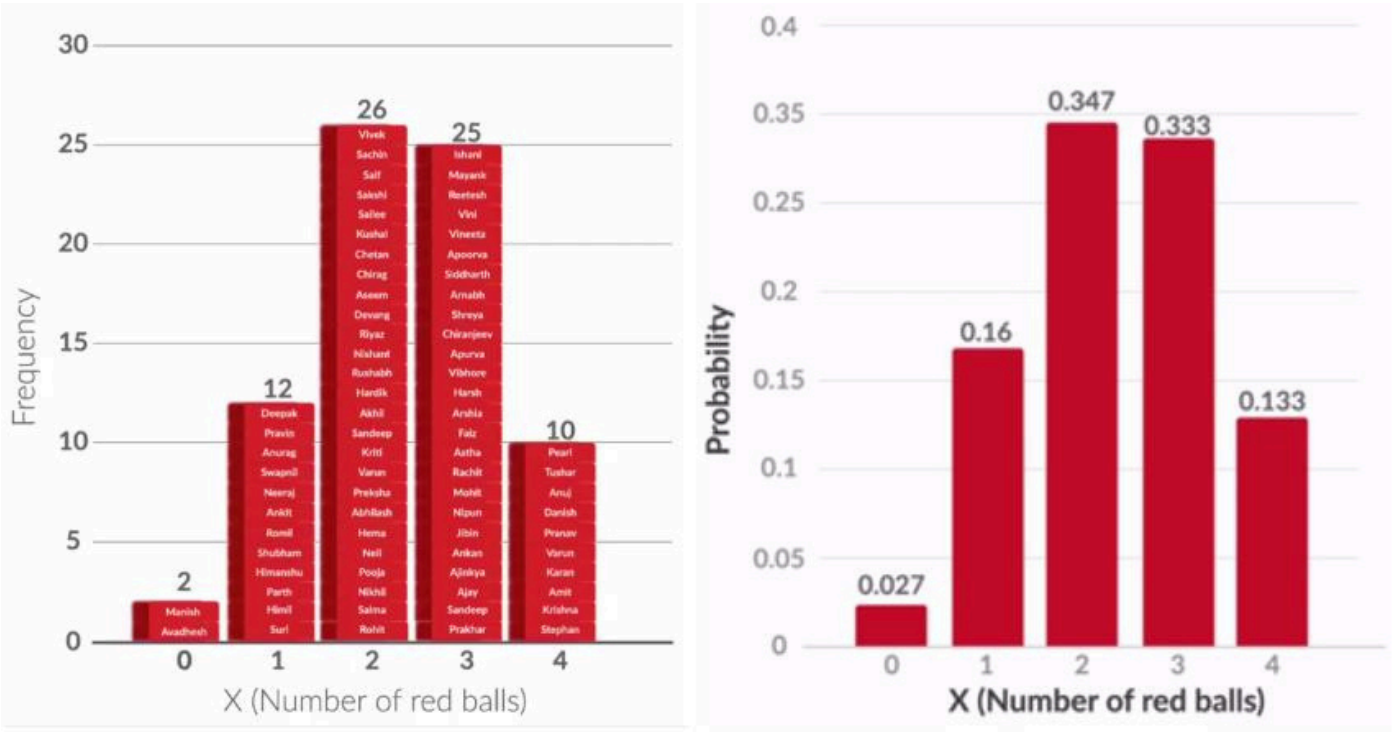


Figure 2 – Frequency Distribution (Left) vs Probability Distribution (Right)

Expected Value

The **expected value** for a variable X is the value of X we would "expect" to get after performing the experiment once. It is also called the expectation, average, and mean value. Mathematically speaking, for a random variable X that can take values x1, x2, x3, …… xn, the expected value (EV) is given by:

EV(X) = x1*P(X = x1) + x2*P(X = x2) + x3*P(X = x3) + ……. + xn*P(X = xn)

Where, P(X=xi) denotes the probability that the random variable will take the value xi.

For example, suppose you're trying to find the expected value of the number of red balls in our game. The random variable X, which is the number of red balls the player gets after playing the game once, can take values 0, 1, 2, 3 and 4. So, the expected value for the number of red balls would be –

EV(X) = 0*P(X = 0) + 1*P(X = 1) + 2*P(X = 2) + 3*P(X = 3) + 4*P(X = 4)
EV(X) = 0*(0.027) + 1*(0.160) + 2*(0.347) + 3*(0.333) + 4*(0.133) = 2.385

However, you can never get 2.385 red balls in one game, as the number of balls will be an integer, like 2 or 3. However, the expected value, the value you would "expect" to get after one experiment, does not have to be a value that will turn up in the experiment/game.

So, the expected value, actually, is the **average value** of X that you will get after playing the game an **infinite** number of times.

Probability Without Experiment

Using basic rules of probability, i.e., **addition rule** and **multiplication rule**, you saw how you could find the probabilities for our red ball game, without even playing the game once.

The probability distribution thus achieved (**theoretical probability distribution**), was very similar to the distribution achieved earlier via experiment (**observed probability distribution**).

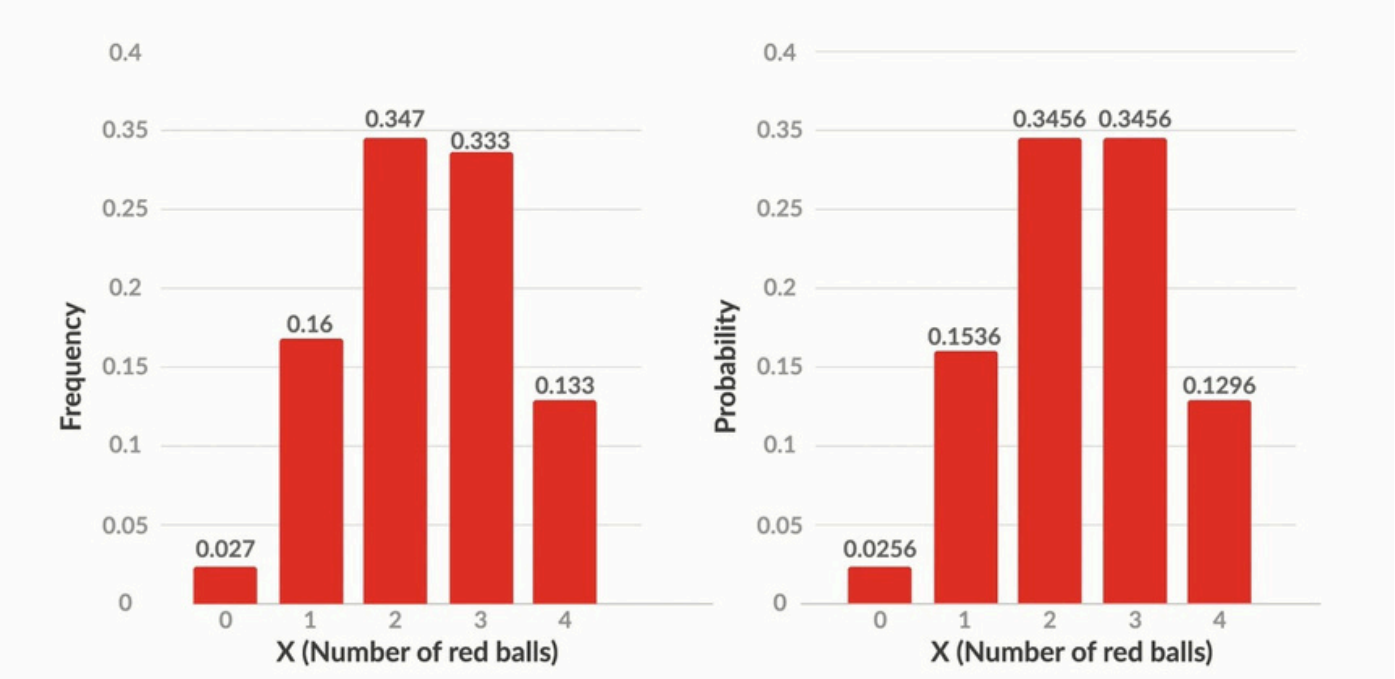


Figure 3 – Observed Probability Distribution (Left) vs Theoretical Probability Distribution (Right)

Notice that the values of P(X = 0) are very close in both graphs, as are the values of P(X = 1), P(X = 2), P(X = 3) and P(X = 4). If the number of experiments conducted would have been more than 75, the values would have been even closer. In fact, for an infinite number of experiments, the values will be exactly same for both graphs.

Binomial Distribution

The **binomial distribution** can be used to calculate the probability of an event, if it follows the following conditions –

1. The **total** number of trials is **fixed**
2. Each trial is **binary**, i.e. has only two possible outcomes, success and failure

3. The **probability of success** is the **same** for all the trials

Basically, it should be a series of yes or no questions, with the probability of yes remaining same for all questions. Examples of such situations are –

{balls are put back after drawing})

3. Finding the probability of 9 out of the next 20 coin tosses resulting in a heads.

For such a situation, the probability of r successes, is given by –

$$P(X = r) = {}_nC_r (p)^r (1-p)^{n-r}$$

Where,

n is the total number of trials/questions
p is the probability of success in 1 trial
r is the number of successes after n trials

For example, in our game –

Total number of trials, n = 4
Probability of getting a red ball in 1 trial, p = 0.6

So, the probability of getting r red balls is given by –

$$P(X = r) = {}_4C_r (0.6)^r (0.4)^{4-r}$$

Using this, we get $P(X = 0) = {}_4C_0(0.6)^0(0.4)^4 = 0.0256$. Also, $P(X = 1) = {}_4C_1(0.6)^1(0.4)^3 = 0.1536$. Similarly, we can find P(X = 2), P(X = 3) and P(X = 4).

Cumulative Probability

Cumulative probability of x, generally denoted by **F(x)**, is the probability of the random variable X, taking a value lesser than x. Mathematically speaking, we'd say –

$$F(x) = P(X \leq x)$$

For example, for our game,

$F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.0256 + 0.1536 + 0.3456 = 0.5238.$

Probability Density Functions

For a **continuous random variable**, the probability of getting an **exact value** is very low, almost **zero**. Hence, when talking about the probability of continuous random variables, you can only talk **in terms of intervals**.

For example, for a particular company, the probability of an employee's commute time being exactly equal to 35 minutes was zero, but the probability of an employee having a commute time between 35 and 40 minutes was 0.2.

Hence, for continuous random variables, probability density functions (**PDF**s) and cumulative distribution functions (**CDF**s) are used, instead of the bar chart type of distribution used for the probability of discrete random variables. These functions are preferred because they talk about probability in terms of intervals.

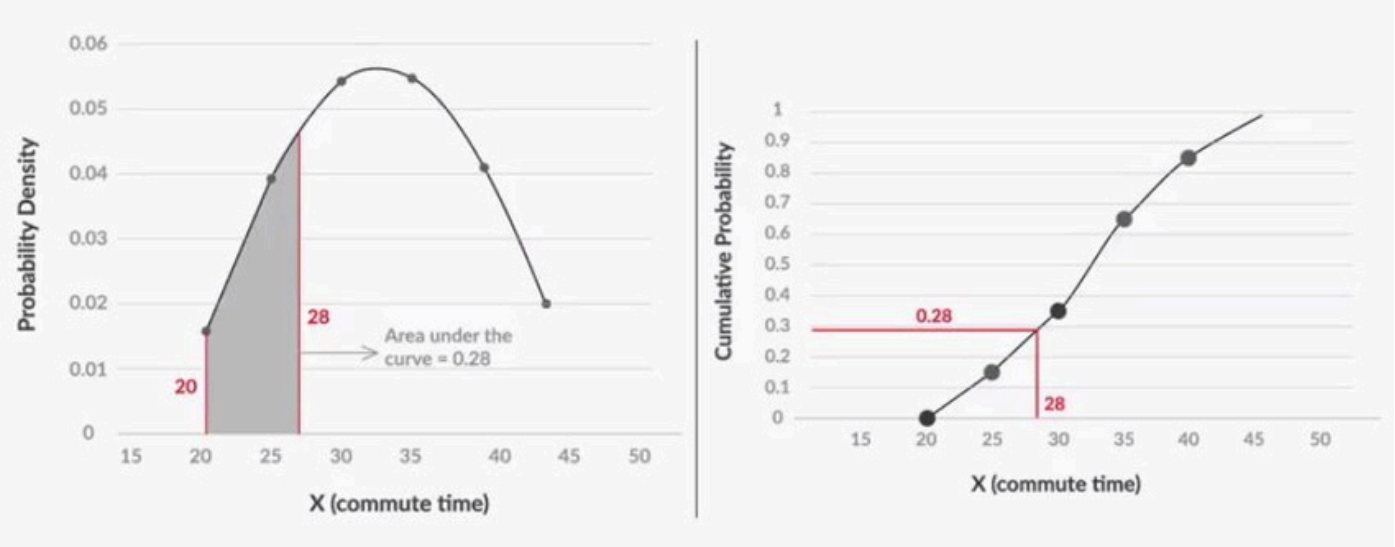


Figure 4 – PDFs vs. CDFs (X = commute time)

To find the cumulative probability using a CDF, you just have to check the value of the graph. For example, F(28), i.e., the probability of an employee having a commute time less than or equal to 28 minutes, is given by the value of the CDF at X = 28. In the PDF, it is given by the area under the graph, between X = 20, the lowest value and X = 28.

Normal Distribution

A very commonly used probability density function is the **normal distribution**. It is a **symmetric** distribution, and its **mean**, **median** and **mode** lie at the center.

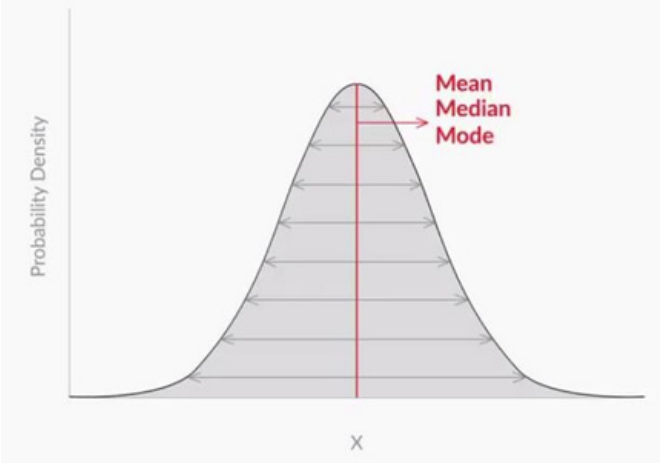


Figure 5 – Normal Distribution

Also, a variable that is normally distributed, follows the **1-2-3** rule, which states that there is a –

1. **68%** probability of the variable lying **within 1 standard deviation** of the mean
2. **95%** probability of the variable lying **within 2 standard deviations** of the mean
3. **99.7%** probability of the variable lying **within 3 standard deviations** of the mean

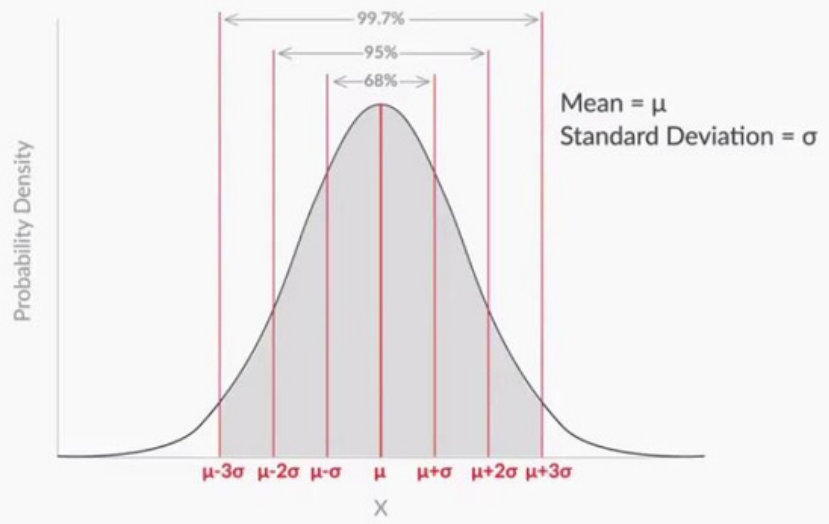


Figure 6 – 1-2-3 Rule for Normal Distribution

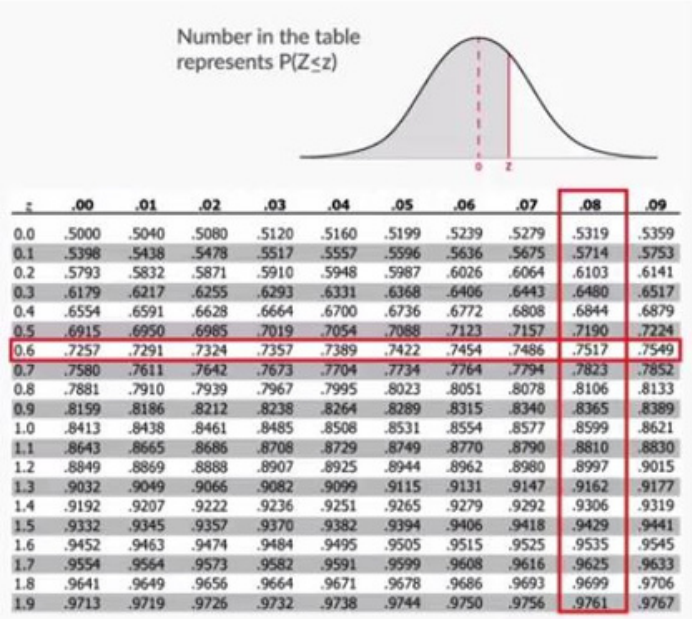
Standard Normal Distribution

In order to find the probability for a normal variable, you actually do not need to know the value of the mean or the standard deviation — it is enough to know **the number of standard deviations away from the mean** your random variable is. That is given by:

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

This is called the **Z score**, or the **standard normal variable**.

In fact, you can use the **Z table** to find the cumulative probability for various values of Z. For example, say, you want to find the cumulative probability for Z = 0.68 using the Z table.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Figure 7 – Z Table

The intersection of row “0.6” and column “0.08” is 0.7517, which is our answer.

Samples

Instead of finding the mean and standard deviation for the entire population, it is sometimes beneficial to find the mean and standard deviation for only a **small representative sample**. You may have to do this because of time and/or money constraints.

For example, for an office of 30,000 employees, we wanted to find the average commute time. So, instead of asking all employees, we asked only 100 of them and found that for them, the mean was equal to 36.6 minutes and the standard deviation was equal to 10 minutes.

However, we said that it would not be fair to infer that the population mean is exactly equal to the sample mean. This is because the flaws of the sampling process must have led to some error. Hence, the sample mean’s value has to be reported with some error margin.

For example, the mean commute time for the office of 30,000 employees would be equal to 36.6 + 3 minutes, 36.6 + 1 minutes or 36.6 + 10 minutes or, for that matter, 36.6 minutes + some error margin

However, in order to find this margin, it would be necessary to understand what sampling distributions are, as there properties help in finding this margin.

Sampling Distributions & Central Limit Theorem

The **sampling distribution**, which is basically the distribution of sample means of a population, has some interesting properties which are collectively called the **central limit theorem**, which states that no matter how the original population is distributed, the sampling distribution will follow these three properties –

1. **Sampling Distribution’s Mean** (μ??) = **Population Mean** (μ)
2. Sampling Distribution’s Standard Deviation (**Standard Error**) = $\frac{\sigma}{\sqrt{n}}$, where σ is the population’s standard deviation and n is the sample size
3. For **n > 30**, the sampling distribution becomes a **normal** distribution

To verify these properties, we performed sampling using data collected for our U game from Session 1. The values for the sampling distribution thus created ($\mu\bar{?}\bar{?}$ = 2.348, S.E. = 0.4248) were pretty close to the values predicted by theory ($\mu\bar{?}\bar{?}$ = 2.385, S.E. = 0.44).

To summarise, the notation and formulae related to samples, populations and sampling distributions are –

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. = $\sigma/\sqrt{n}$

Mean Estimation Using CLT

Using CLT, you can estimate the population mean from the sample mean and standard deviation.

For example, to estimate the mean commute time of 30,000 employees of an office, you took a sample of 100 employees and found their mean commute time. For this sample, the sample mean ?? = 36.6 minutes, sample standard deviation S = 10 minutes.

Using CLT, you can say that the sampling distribution for mean commute time will have -

1. Mean = μ {unknown}
2. Standard error = $\frac{\sigma}{\sqrt{n}} = \frac{S}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$
3. Since n(100) > 30, the sampling distribution is a normal distribution

Using these properties, you can **claim** that the probability that the population mean μ lies between 34.6 (36.6-2) mins and 38.6 (36.6+2) mins, is 95.4%.

Also, there is some terminology related to the claim -

1. Probability associated with the claim is called **confidence level** (Here it is 95.4%)
2. Maximum error made in sample mean is called **margin of error** (Here it is 2 minutes)
3. Final interval of values is called **confidence interval** {Here it is the range – (34.6, 38.6)}

In fact, you can generalise the entire process. Let's say you have a sample with sample size n, mean $\bar{??}$ and standard deviation S. Now, the y% confidence interval (i.e., confidence interval corresponding to y% confidence level) for μ will be given by the range –

$$\text{Confidence Interval} = (\bar{??} - \frac{Z*S}{\sqrt{n}}, \bar{??} + \frac{Z*S}{\sqrt{n}})$$

Where, **Z*** is the Z-score associated with a y% confidence level.

For example, the 90% confidence interval for the mean commute time will be –

$$\mu = (\bar{??} - \frac{Z*S}{\sqrt{n}}, \bar{??} + \frac{Z*S}{\sqrt{n}})$$

Here,

$\bar{??}$ = 36.6 minutes

S = 10 minutes

n = 100

Z* = 1.65 (Z* corresponding to 90% confidence level)

So, the confidence interval is –

$$\mu = (34.95 \text{ mins}, 38.25 \text{ mins})$$