Road to Data Science in 50 Days - Day 3 ¶

Statistics for Data Science

Previously we went through the descriptive statistics to describe the data we have in hand, however is alot more than just descriptive statistics when we talk about describing the data and we are going to learn more about them practically in future. Once we get a good hold of the data we have, the next step is to infer insights and observations about what lies outside the data (about Population).

Why do we need inferential statistics?

Inferential Statistics is one of the most important concepts in statistics for data science.

Suppose, you want to know the average salary of Data Science professionals in India.

- Which of the following methods can be used to calculate it?
- Meet every Data Science professional in India. Note down their salaries and then calculate the total average?

Or hand pick a number of professionals in a city like Gurgaon. Note down their salaries and use it to calculate the Indian average. Well, the first method is not impossible but it would require an enormous amount of resources and time. But today, companies want to make decisions swiftly and in a cost-effective way, so the first method doesn't stand a chance.

On the other hand, second method seems feasible. But, there is a caveat. What if the population of Gurgaon is not reflective of the entire population of India? There are then good chances of you making a very wrong estimate of the salary of Indian Data Science professionals.

Now, what method can be used to estimate the average salary of all data scientists across India?

This is where inferential statistics comes to the rescue.

Inferential Statistics

In simple language, Inferential Statistics is used to draw inferences beyond the immediate data available.

With the help of inferential statistics, we can answer the following questions:

- Making inferences about the population from the sample.
- Concluding whether a sample is significantly different from the population. For example, let's say you collected the salary details of Data Science professionals in Bangalore. And you observed that the average salary of Bangalore's data scientists is more than the average salary across India. Now, we can conclude if the difference is statistically significant.
- If adding or removing a feature from a model will really help to improve the model.
- If one model is significantly better than the other?
- Hypothesis testing in general.

But before we move forward with Inferential statistics, we need to understand the concept of probability.

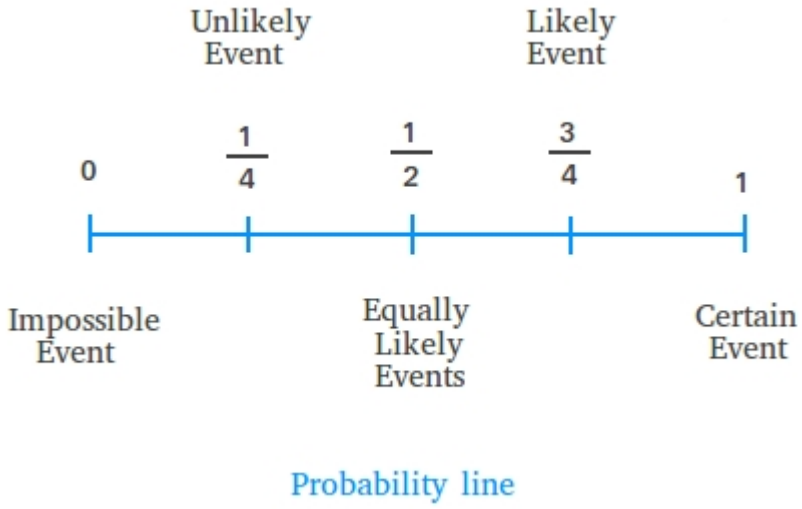
What is Probability?

Probability is a measure of the likelihood that an event will occur in a Random Experiment. Probability is quantified as a number between 0 and 1, where, we can say, 0 indicates uncertainty and 1 indicates certainty. The higher the probability of an event, the more likely it is that the event will occur.

Example

2 of the most common examples are "A coin toss" and "A dice roll".

While tossing a fair (unbiased) coin, there is a possibility of occurrence of two outcomes (“heads” and “tails”), which are equally probable; i.e, the probability of “heads” equals the probability of “tails”. The probability of either “heads” or “tails” is 1/2 (which could also be written as 0.5 or 50%).



Probability line

Before proceeding further we should be aware of the basic terms like:

Random Experiment:

A random experiment is a physical situation whose outcome cannot be predicted until it is observed.

Sample Space

A sample space is a set of all possible outcomes of a random experiment.

In the above example, we have:

Random Experiment: Tossing of a fair coin

Sample space: {Head, Tail}

As we got a little understanding of Probability, we will now read about Probability Distribution and its types with the help of examples and formulas wherever required.

Distribution

Came across this amazing blog by Mitali Singh on Acadgild - https://github.com/sidsaif/Road-to-Data-Science-in-50-Days (https://github.com/sidsaif/Road-to-Data-Science-in-50-Days) . And the following content will majorly from this blog, so do check it out if you want detailed explanations.

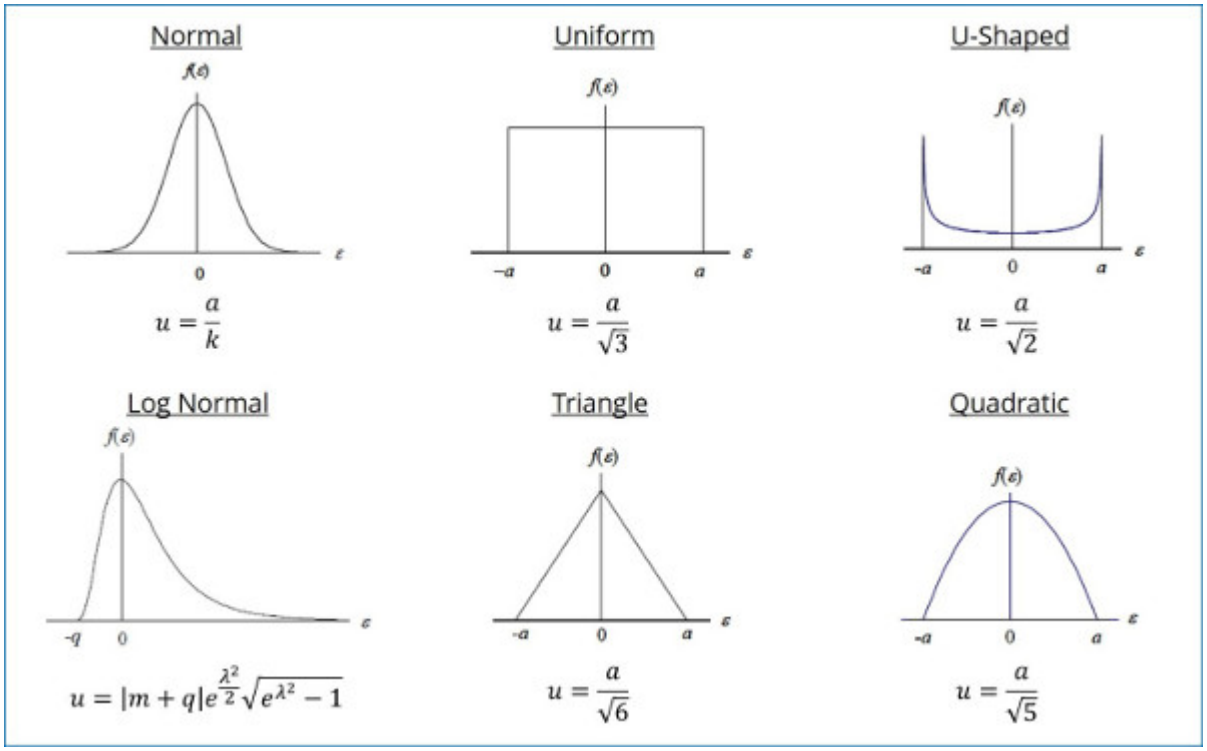
In statistics when we use the term Distribution it usually means Probability distribution.

A Distribution is a function that shows the possible values for a variable and how often they occur.

Or A Probability Distribution is a mathematical function that can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment.

Good examples are the

- Normal Distribution
- Binomial Distribution
- Uniform Distribution



To learn about various distribution in brief, refer: https://www.isobudgets.com/probability-distributions-for-measurement-uncertainty/ (https://www.isobudgets.com/probability-distributions-for-measurement-uncertainty/)

Types of Probability Distribution

There are many different types of probability distribution. Some of them that we will be covering in this blog is listed below:

- Normal Distribution
- Bernoulli's Distribution
- Binomial Distribution

- Uniform Distribution
- Student's T Distribution
- Poisson Distribution

Each probability distribution has a visual representation. It is a graph that describes the likelihood of occurrence of every event. The graph is just a visual representation of a distribution.

Do Not misunderstand that the Distribution is a graph. Distribution is defined by the underlying probability and not the graph.

1. Normal Distribution

The visual representation of Normal Distribution has already been seen above in the blog.

The Normal Distribution is a very common continuous probability distribution. This type of distributions are important in statistics and are often used to represent random variables whose distribution is not known.

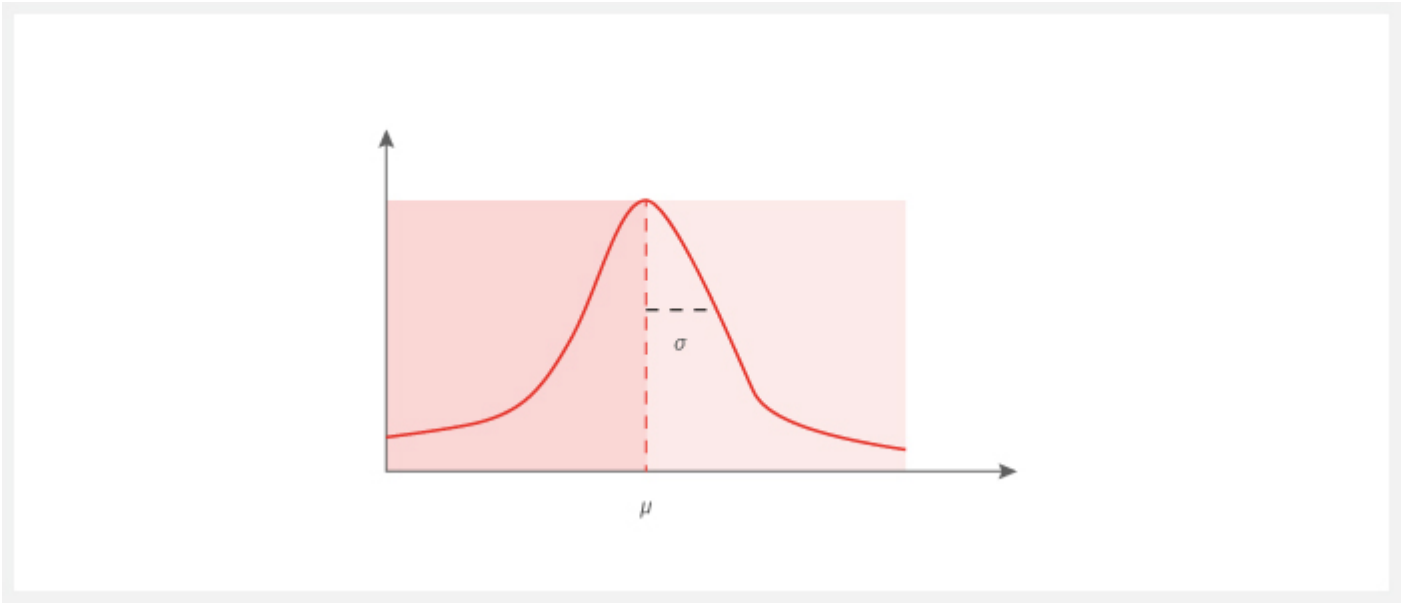
The statistical term for this type of distribution is Gaussian Distribution though many people call it Bell curve as it is shaped like one.

This type of distribution is symmetric and its mean, median and mode are equal.

Mathematically, Gaussian Distribution is represented as:

N~(μ, σ2)

Where N stands for Normal, symbol ~ stands for distribution, symbol μ stands for mean and σ2 stands for variance.

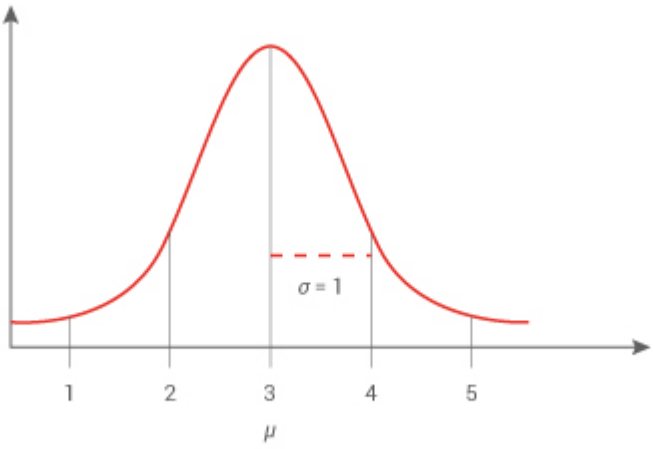


In the above image, we can see the highest point is located at the mean μ and the spread of the graph is determined by the standard deviation σ.

Let us understand this with the simplest example where we have a random variable X with distribution:

X = {1, 2, 3, 4, 5}

When we take the mean and standard deviation of the above data set we get mean(μ) = 3 and standard deviation(σ) = 1. When we plot it, we get some distribution like this:



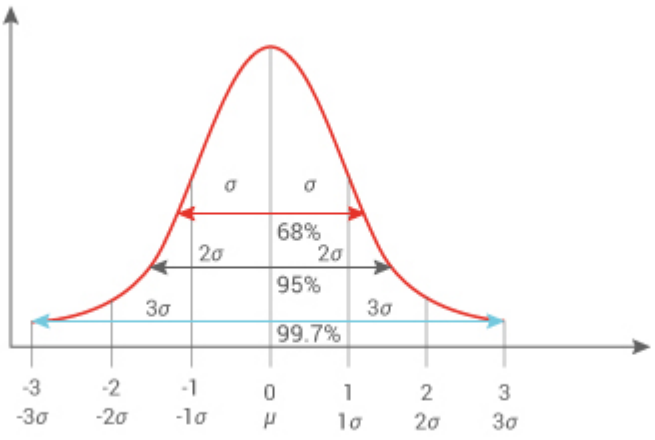
This Bell curve specifies the Gaussian/Normal Distribution.

When we talk about Gaussian Distribution or Normal Distribution we have often heard the term Empirical Formula. What exactly does this formula states, well is what we will be covering next.

1.1 Empirical Formula

The empirical rule states that for a Normal Distribution, nearly all of the data will fall within three range of standard deviations of the mean. The empirical rule can be understood when broken down into three parts:

- 68% of the data falls within the first standard deviation from the mean.
- 95% fall within two standard deviations.
- 99.7% fall within three standard deviations.



The rule is also called the 68-95-99.7 Rule or the Three Sigma Rule.

The Empirical Rule is often used in statistics for forecasting, especially when obtaining the right data is difficult or impossible to get. The rule can give you a rough estimate of what your data collection might look like.

When a Normal Distribution is standardized, the result is called a Standard Normal Distribution.

1.2 Standard Normal Distribution

Understanding Standardization in the context of statistics. Every distribution can be standardized. Let say if the mean and the variance of a variable are μ and σ2 respectively.

Standardization is the process of transforming a variable to one with a mean of 0 and a standard deviation of 1.

i.e., ~(μ, σ2) → ~ (0, 1)

When a Normal Distribution is standardized, the result is called a Standard Normal Distribution.

i.e., N~(μ, σ2) → ~ N(0, 1)

We use the following formula for standardization:

$$Z = \frac{x - \mu}{\sigma}$$ Z - score

Where x is data element, μ is mean and σ is the standard deviation

We use the letter Z to denote standardization. The standardized value i.e., Z is known as the z-score.

These Z scores are important because they tell you how far a value is from the mean. When you standardize a random variable, its ‘mean’ becomes 0 and its standard deviation becomes 1.

Let us understand the steps involved in Standardization with the help of a simple example.

Suppose we have a dataset with elements

X = { 1, 2, 2, 3, 3, 3, 4, 4, 5}

We get mean as 3, variance as 1.49 and std dev as 1.22 i.e., N ~ (3, 1.49).

Now we will subtract the mean from all data points, i.e., x – μ.

We will get a new data set as below:

X1 = {-2, -1, -1, 0, 0, 0, 1, 1, 2}

Now we get mean as 0, but variance and std dev still as 1.49 and 1.22 respectively i.e., N ~ (0, 1.49)

So far we have a new distribution but it is still normal and needs to standardized.

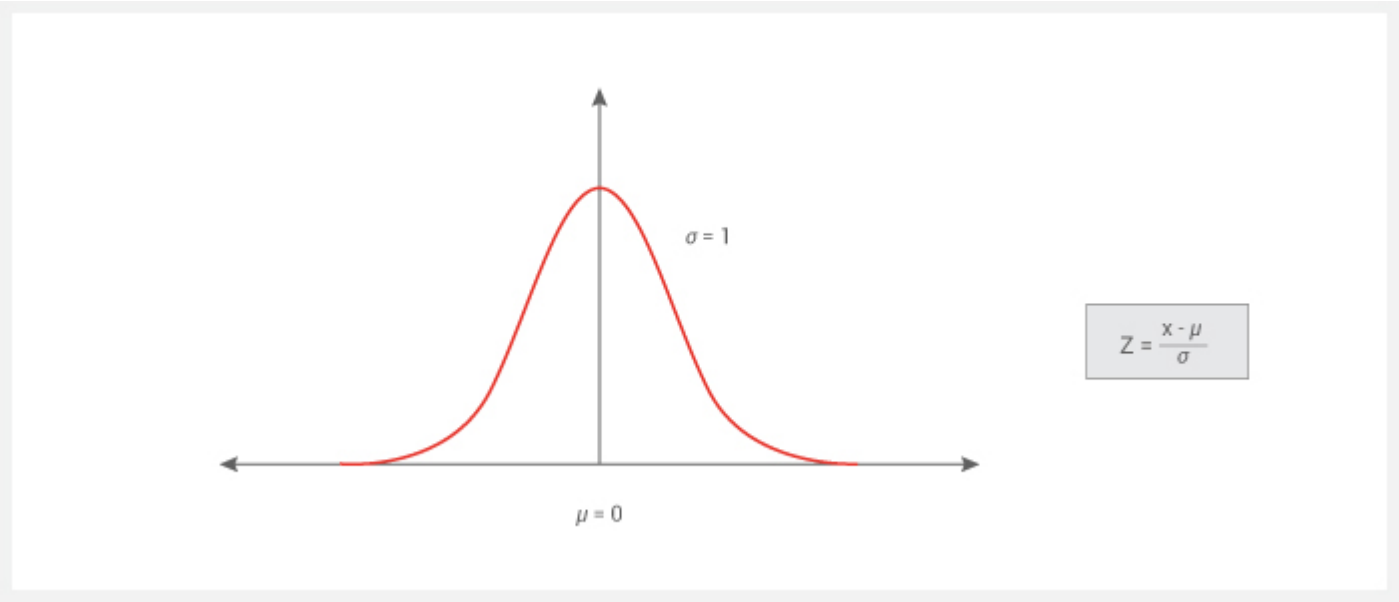
So the next step of standardization is to divide all the data points by the standard deviation, i.e., (x – μ)/σ.

Dividing each datapoint by 1.22(std dev) we get a new data set as :

X2 = {-1.6, -0.82, -0.82, 0, 0, 0, 0.82, 0.82, and 1.63.}

Now if we calculate the mean we get as 0 and standard deviation as 1 i.e., N ~ (0, 1)

Plotting it on a graph we get something like this



This is how we can obtain Standard Normal Distribution from any normally distributed dataset.

Using this standardized normal distribution makes inferences and predictions much easier.

1.3 Probability Density Function and Probability Mass Function

Probability density function and Probability mass function is a statistical expression that defines a Probability Distribution for a random variable.

Do not get confused between the two terms. Probability density function(PDF) is used to determine the probability distribution for a Continuous Random Variable. When the PDF is graphically plotted the area under the curve indicates the interval in which the variable will fall.

Whereas the Probability Mass Function(PMF) is used to determine the probability distribution for a Discrete Random Variable.

"As we know Continuous Random Variables are the one which takes an infinite number of possible values eg: the weight of a person can be 50.2, 44.5, 60.7, etc and Discrete Random Variables are the one which may take on only a countable number of distinct values such as 0,1,2,3,4…"

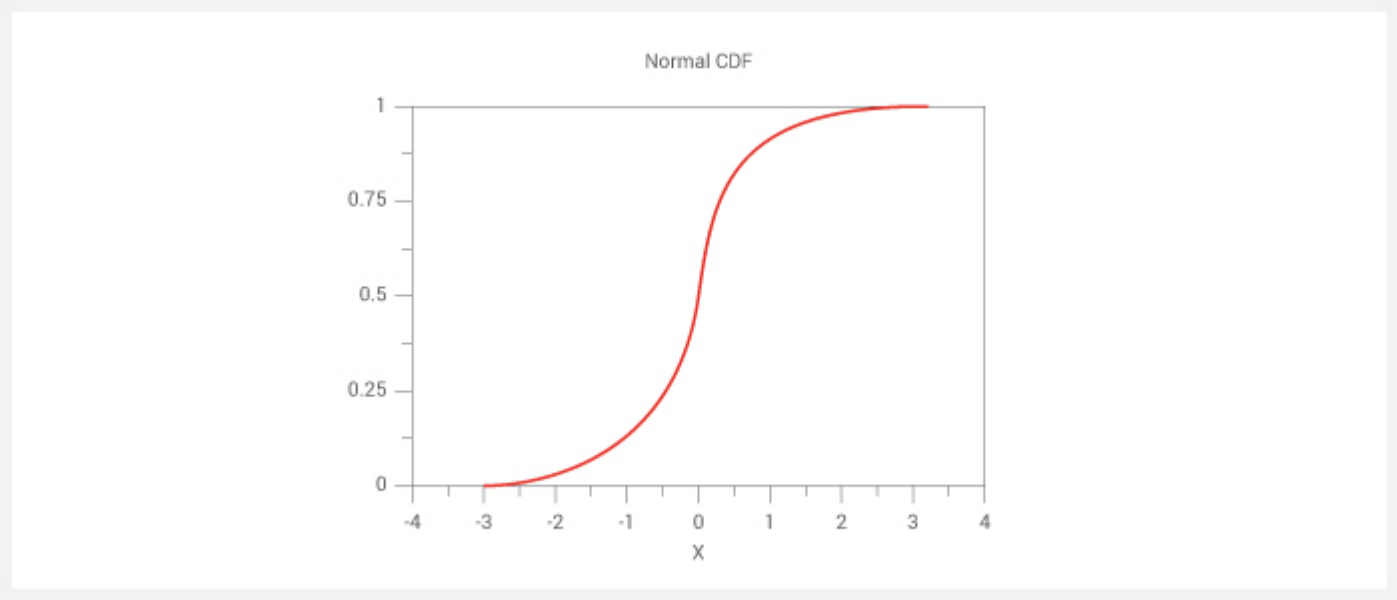
If we know the mean and variance of our dataset we can compute the PDF and PMF. PDF and PMF tell how well our data has been distributed with respect to the mean and standard deviation within a particular curve.

1.4 Cumulative Density Function

The cumulative distribution function (CDF) of a random variable is another method to describe the distribution of random variables.

The cumulative frequency is the sum of the relative frequencies. It starts at the frequency of the first brand, then we add the second, the third and so on until it finishes at 100%.

The advantage of the CDF is that it can be defined for any kind of random variable (discrete, continuous, and mixed).



1.5 Central Limit Theorem

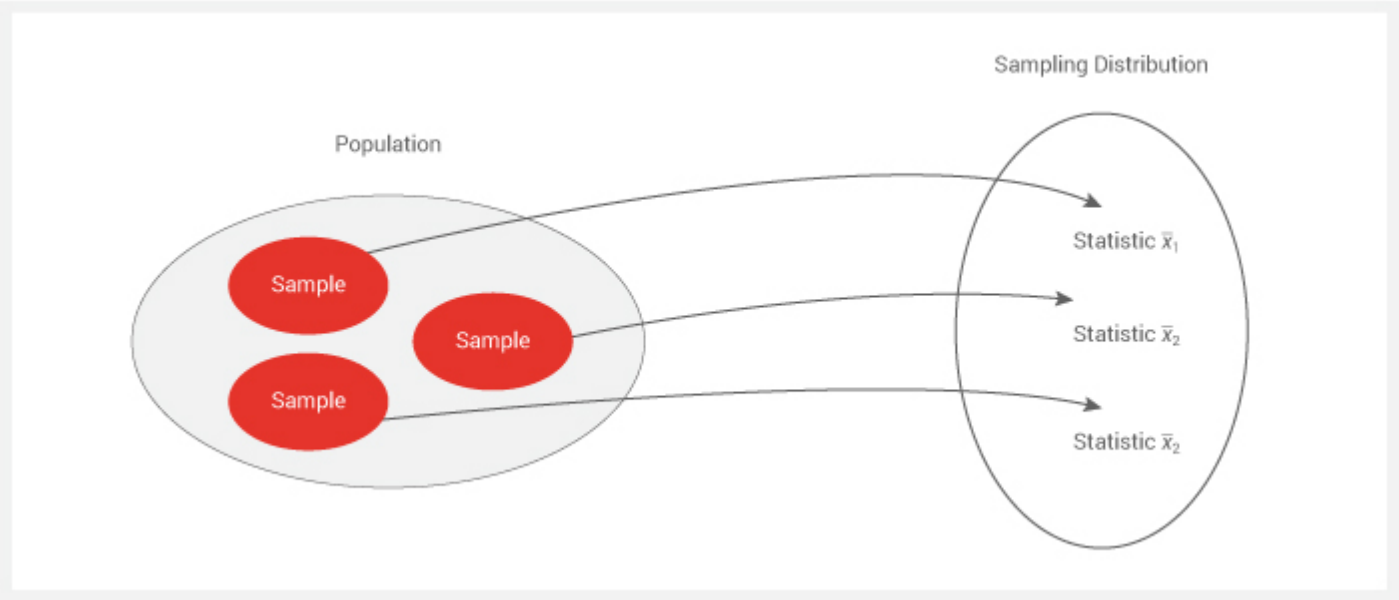
The Central Limit Theorem is one of the most important concepts in Statistics.

This theorem states that as the number of samples taken which are maximum in number, the distribution of the average of the sample means when plotted tends to be a normal distribution.

Don’t worry, let’s study it briefly with pictorial representation.

Suppose we have a very large dataset, **whose distribution doesn’t matter and could be normal, uniform, binomial or random**.

The first thing we do is take out the subsets from the large data sets, that means, we will fetch smaller datasets of size 30 or more and create different subsets. **This process is called sampling and the subsets are called samples. We do this to gain a better idea of how the entire dataset is made.**



After fetching different samples, which is sufficient in number, we will then calculate the mean of each sample and then plot this different distribution. Surprisingly, our graph of the sample means look more like a Normal/Gaussian Distribution.

Also, if we take the average of all sample means it will be nearly equal to the actual population mean and the standard deviation equals σ/√n.

Where: σ = the population standard deviation

n = the sample size(i.e., number of observations in our sample)

Let us revise the key summary that should be kept in mind while applying the Central Limit Theorem.

Here the probability of both the outcomes is the same for all the trials.

Each trial is independent since the outcome of the previous toss doesn't determine or affect the outcome of the current toss. An experiment with only two possible outcomes repeated n number of times is called binomial. The parameters of a binomial distribution are n and p where n is the total number of trials and p is the probability of success in each trial.

We define Binomial Distribution with the below formula:

$$P_p(n|N) = \binom{N}{n} p^n q^{N-n}$$

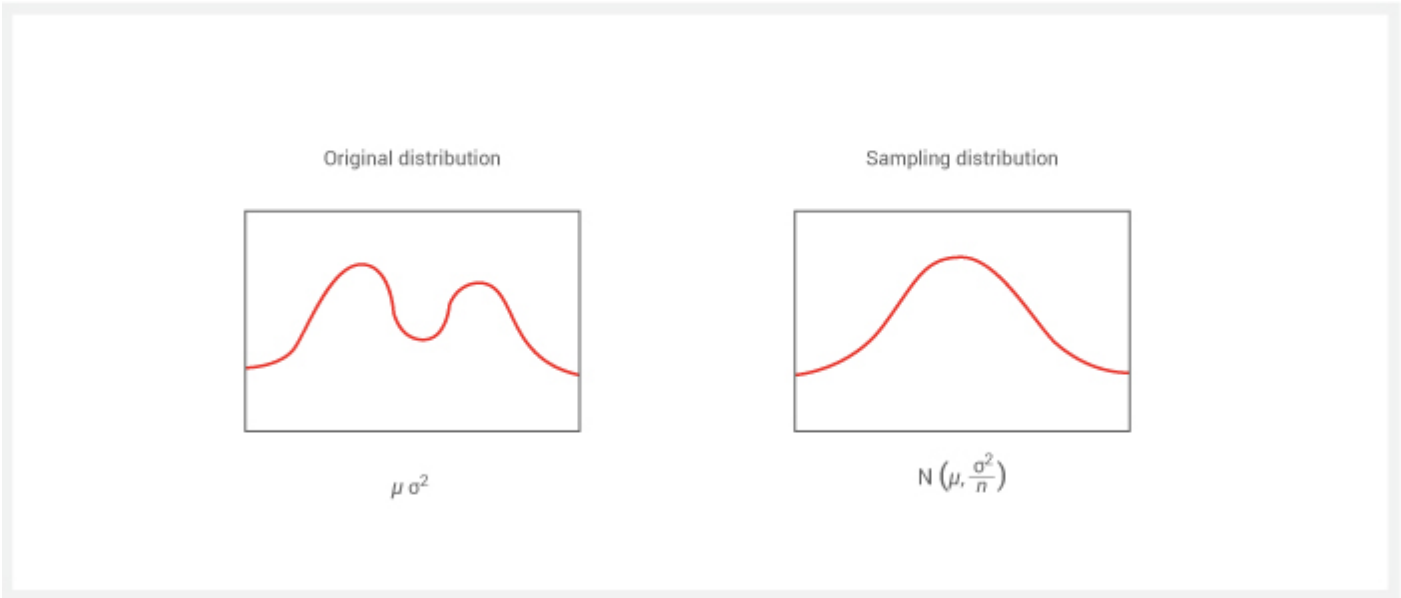
$$= \frac{N!}{n!\,(N-n)!}\; p^n (1-p)^{N-n}$$

where $\binom{N}{n}$ is a binomial coefficient, and p and q referes to succeess and failure, respectively.

$P_p(n|N)$ gives the discrete probability distribution, obtaining n successes out of N trials.

1. Bernoulli's Distribution* The distribution of the original(population) dataset doesn't matter. It could be normal, uniform, binomial, etc.

- The distribution of the sample means would always be Normal Distribution



2. Binomial Distribution

This type of distribution is used when there are exactly two outcomes of a trial. These outcomes are labeled as "Success" and "Failure".

Here the probability of both the outcomes is the same for all the trials.

Each trial is independent since the outcome of the previous toss doesn't determine or affect the outcome of the current toss. An experiment with only two possible outcomes repeated n number of times is called binomial. The parameters of a binomial distribution are n and p where n is the total number of trials and p is the probability of success in each trial.

We define Binomial Distribution with the below formula:

$$P_p(n|N) = \binom{N}{n} p^n q^{N-n}$$

$$= \frac{N!}{n!\,(N-n)!}\; p^n (1-p)^{N-n}$$

where $\binom{N}{n}$ is a binomial coefficient, and p and q referes to succeess and failure, respectively.

$P_p(n|N)$ gives the discrete probability distribution, obtaining n successes out of N trials.

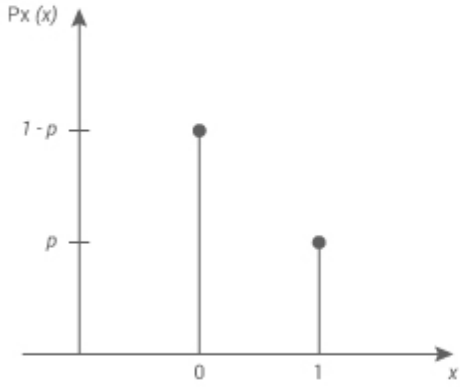
3. Bernoulli's Distribution

Binomial Distribution is closely related to Bernoulli's Distribution.

Bernoulli Distribution is a special case of Binomial Distribution with a single trial.

The Bernoulli distribution is a discrete distribution having two possible outcomes that is, 0 and 1, where n = 1 (usually called a "success") occurs with probability p and n = 0 (usually called a "failure") occurs with probability q = 1 – p, where 0 < p < 1.

Therefore the probability density function(pdf) and the graph for Bernoulli's Distribution is shown in the figure below:



X ~ Bennoulli (p)

$$P(n) = \begin{cases} 1-p & \text{for } n = 0 \\ p & \text{for } n = 1 \end{cases}$$

wich can also be written

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

In the above graph, 1 refers to success and 0 specifies the failure.

The head and tail distribution in coin tossing is an example of Bernoulli's Distribution with p = q = ½.

4. Uniform Distribution

A uniform distribution is a distribution that has a constant probability.

We have already seen the graphical representation of uniform distribution above. Let us understand this with the help of an example.

EXAMPLE: If we roll a die(numbered from 1 to 6), then the probability of getting 1 is one out of six i.e., 1/6

Similarly, the probability of getting 2, 3, 4, 5 and 6 also is ⅙. There is an equal chance of getting each of the 6 outcomes.

Now, if we check for the probability of getting 7, then it is 0 since it is impossible to get a 0 when rolling a die.

For the probability of outcomes for 1 to 6, we have an equal chance of occurrence and this is what we call a Discrete Uniform Distribution.

Remember that the sum of their probabilities is equal to 1 or 100%.

5. Student's T-Distribution

T Distribution or Student's T Distribution is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.

Visually, the Student's T distribution looks much like a Normal distribution but generally has fatter tails. Fatter tails, allow for a higher dispersion of variables, as there is more uncertainty.

As the z-statistic is related to the standard Normal distribution, the t-statistic is related to the Student's T distribution.

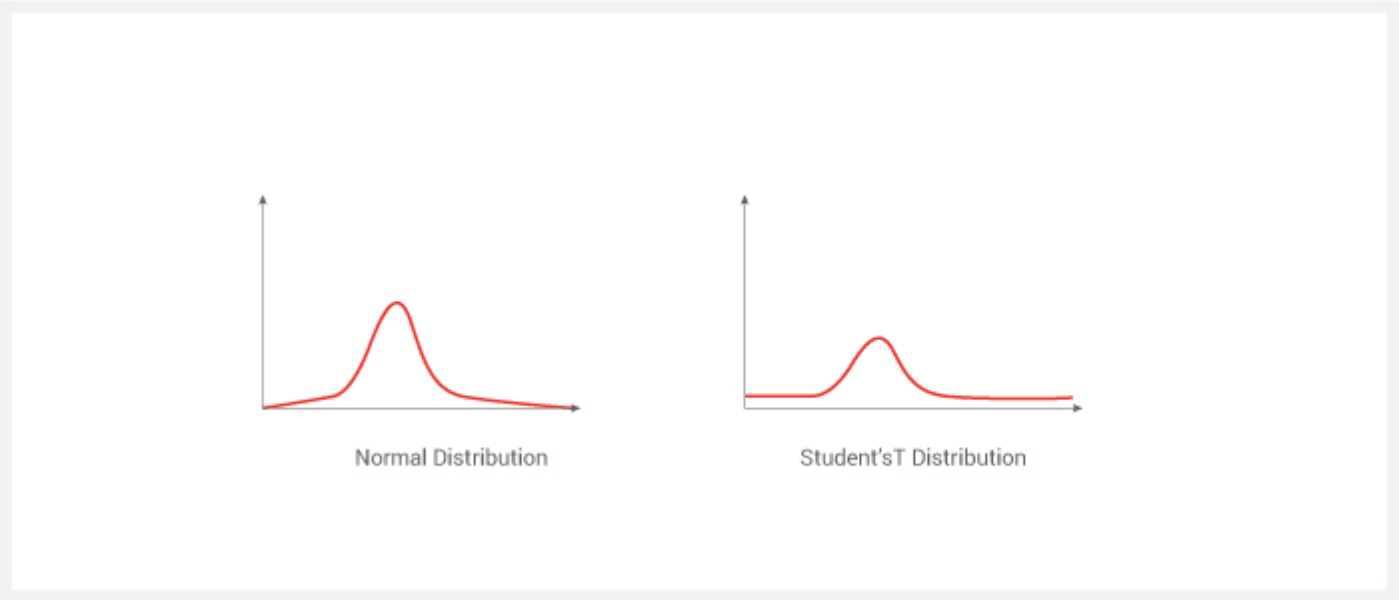
The formula that allows us to calculate it is:

$t = [\bar{x} – \mu] / (s / \sqrt{n})$

t with n-1 degrees of freedom equals the sample mean minus the population means, divided by the standard deviation of the sample by n which refers to the sample size.

The degrees of freedom refers to the number of independent observations in a set of data.

Now we will see the graph for Student's T Distribution and will also see how it is different from Normal Distribution.



Why use T-Distribution?

According to the Central Limit Theorem, the distribution follows Normal Distribution when the sample size is sufficiently large. Here we know the standard deviation and can calculate the z-score and can plot the Normal Distribution.

But sometimes the sample sizes are small and also we do not know the standard deviation of the population. This is where statistician prefer on the distribution of T-Distribution (also known as t-score).

6. Poisson Distribution

The Poisson Distribution is a discrete probability distribution which states that the number of events occurring in a fixed interval of time or space conditionally that the value of an average number of occurrence of the event is known.

For instance, If the average number of diners for seven days is 500, we can predict the probability of a certain day having more customers.

The Poisson Distribution results from a Poisson's Experiments which states that for a series of discrete event where the average time between events is known, but the exact timing of events is random.

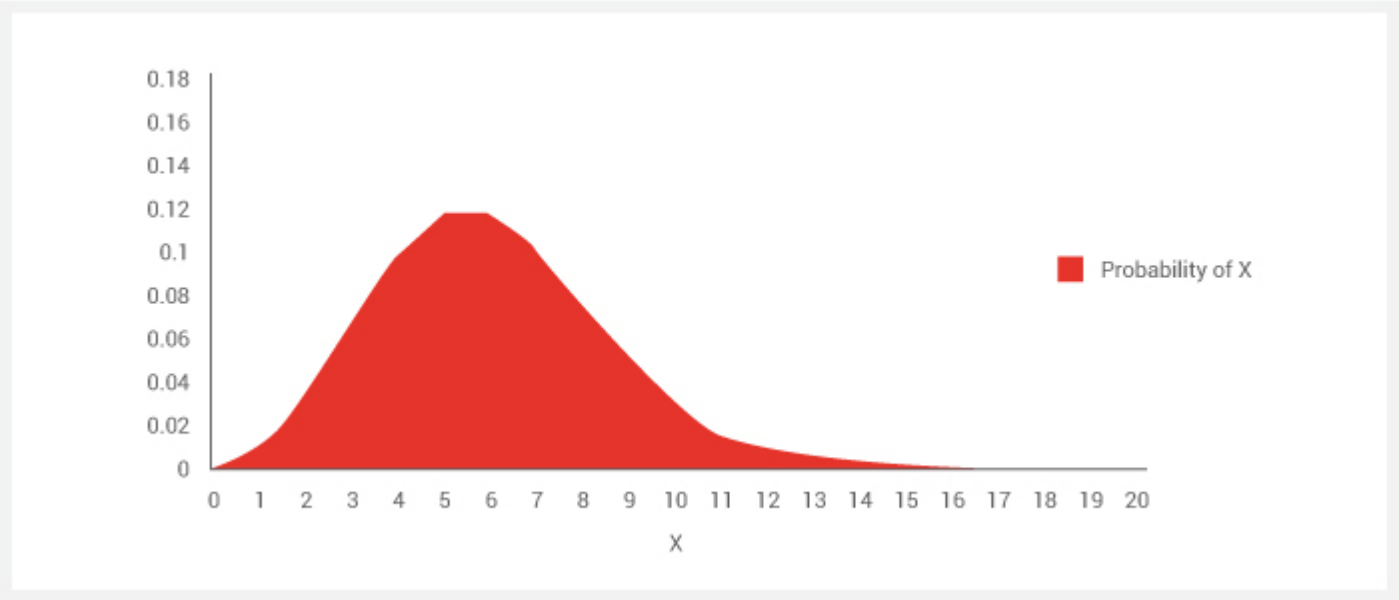
Suppose we conduct a Poisson experiment, in which the average number of successes within a given region is μ.

Then, the Poisson probability is:

$P(x; \mu) = (e-\mu) (\mu x) / x!$

where x is the actual number of successes that result from the experiment, and e is approximately equal to 2.71828.

The graph of Poisson Distribution is as shown below:



The mean and variance of x following a Poisson distribution:

Mean → E(x) = μ

Variance → Var(x) = μ

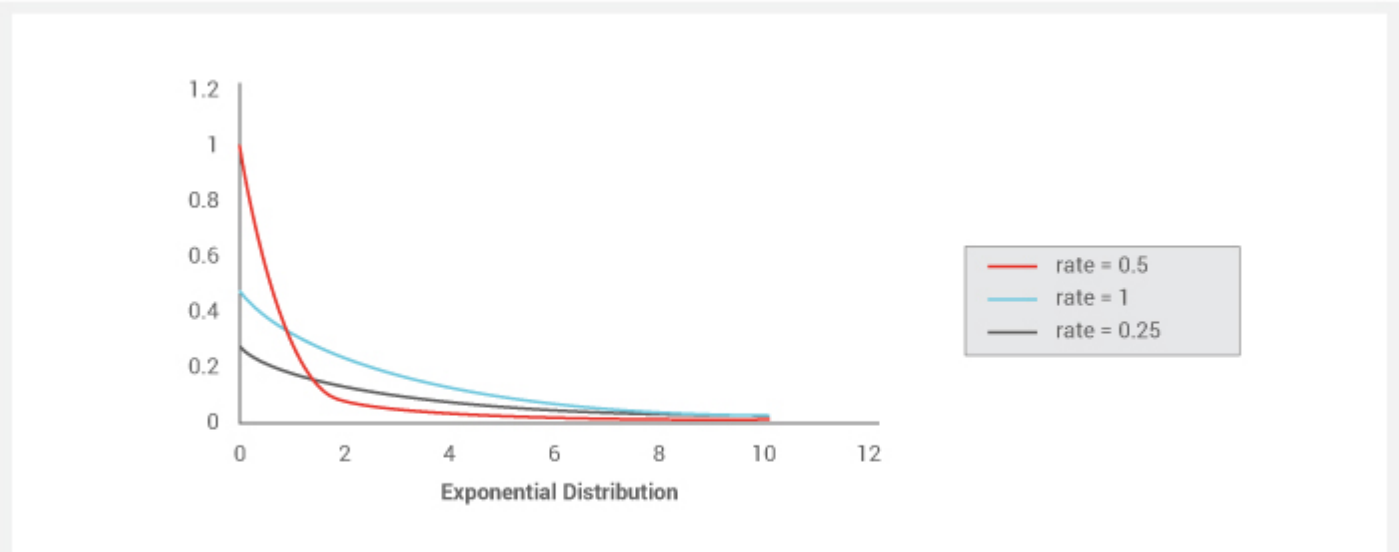
7. Exponential Distribution

Exponential Distribution is one of the most widely used continuous distributions. It measures the expected time of an event to occur.

The exponential distribution is highly used for survival analysis purposes. An example of an exponential distribution is the lifespan of a machine.

It basically answers our query as to how much time do we need to wait before a given event occurs.

The graph of Exponential Distribution is shown below:



Important Links to refer:

https://www.youtube.com/watch?v=qBigTkBLU6g&list=PLblh5JKOoLUK0FLuzwntyYI10UQFUhsY9 (https://www.youtube.com/watch?v=qBigTkBLU6g&list=PLblh5JKOoLUK0FLuzwntyYI10UQFUhsY9) - Amazing tutorials on Statistics fundamentals by none other than **Joshua Starmer** (Personal Recommendation)

https://www.analyticsvidhya.com/blog/2017/02/basic-probability-data-science-with-examples/ (https://www.analyticsvidhya.com/blog/2017/02/basic-probability-data-science-with-examples/) - By Dishashree gupta on Analytics Vidhya

https://www.youtube.com/watch?v=XcLO4f1i4Yo (https://www.youtube.com/watch?v=XcLO4f1i4Yo) - Statistics and Probility for Data Science By Edureka

https://www.analyticsvidhya.com/blog/2017/04/40-questions-on-probability-for-all-aspiring-data-scientists/ (https://www.analyticsvidhya.com/blog/2017/04/40-questions-on-probability-for-all-aspiring-data-scientists/) - 40 Interview questions to solve on probability.

That's all for the day, i hope we have got a fair idea about what is probability, and different types of distribution. Tomorrow we will cover estimations and inferential statistics such as hypothesis testing, chi-square, ANOVA, etc. *</i>*