Z-score

What is Z-score ?

Z-score is also known as standard score gives us idea of how far a data point is an element is from the mean. Hence, Z-score measured in terms of standard deviation from the mean. For example, a standard deviation of 2 indicates the value is 2 standard deviations away from the mean. In order to use a z-score, we need to know the population mean(μ) and also the population standard deviation(σ).

The Formula for Z-score

$$z = (X - \mu) / \sigma$$

where,
z = Z-Score,
X = The value of the element,
μ = The population mean, and
σ = The population standard deviation

Calculate Z-score?

The population mean(μ), the population standard deviation(σ),and the observed value (X) are provided in the problem statement, and substituting the same in the above Z-score equation yields us the Z-score is positive or negative , one makes use of the respective positive Z-score Table or negative Z-Table available online or on the back of your statistics textbook in the appendix.

Example: You take the GATE examination and score 500. The mean score for the GATE is 390 and the standard deviation is 45. How well did you score on the test compared to the average test taker?

Answer:

The following data is readily available in the above question statement

Raw score/observed value = X = 500

Mean score = μ = 390

Standard deviation = σ = 45

By applying the formula of z-score,

$$z = (X - \mu) / \sigma$$

$$z = (500 - 390) / 45$$

$$z = 110 / 45 = 2.44$$

This means that your z-score is **2.44.**

Since the z-score is positive 2.44, we will make use of the positive Z-Table.

Now let's take a look at Z-table to know how well you scored compared to the other test-takers.

Here, **z-score = 2.44**

1. Firstly, map the first two digits 2.4 on the Y-axis.
2. Then along the X-axis, map 0.04
3. Join both axes. The intersection of the two will provide you the Z-Score value you're looking for

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558
2	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972

As a result, you will get the final value which is **0.99266.**

Now, we need to compare how our original score of 500 on the GATE examination compares to the average score of the batch, To do that we need to convert th e Z-score into a percentage value.

0.99266 * 100 = 99.266%
Finally, you can say that you have performed well than almost **99%** of other test-takers.

T-Score

T score is the subtraction of individual standard deviation from individual mean and then divide the result with sample standard deviation whole result multiplied by sample size.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- $\bar{x}$ = sample mean
- μ_0 = population mean
- s = sample standard deviation
- n = sample size

	P						
one-tail	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	0.2	0.1	0.05	0.02	0.01	0.002	0.001
DF							
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.92	4.303	6.965	9.925	22.328	31.6
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.86	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.93	4.318
13	1.35	1.771	2.16	2.65	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.14
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.12	2.583	2.921	3.686	4.015
17	1.333	1.74	2.11	2.567	2.898	3.646	3.965
18	1.33	1.734	2.101	2.552	2.878	3.61	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.85
21	1.323	1.721	2.08	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.5	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.06	2.485	2.787	3.45	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.689
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.66
30	1.31	1.697	2.042	2.457	2.75	3.385	3.646
60	1.296	1.671	2	2.39	2.66	3.232	3.46
120	1.289	1.658	1.98	2.358	2.617	3.16	3.373
1000	1.282	1.646	1.962	2.33	2.581	3.098	3.3
Inf	1.282	1.645	1.96	2.326	2.576	3.091	3.291

Central limit theorem

The central limit theorem is a statistical theory that states that-if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.

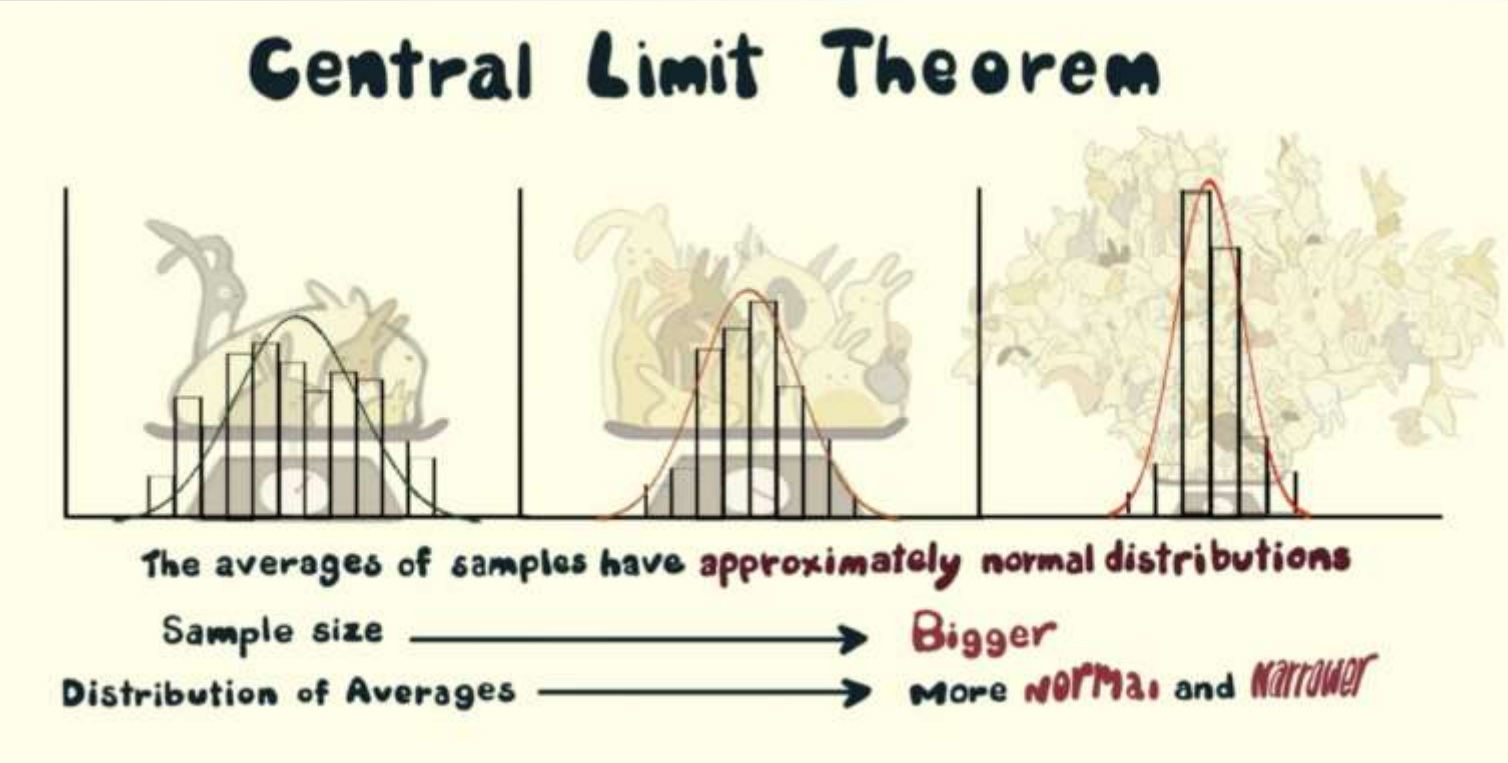
Consider there are 15 sections in class X, and each section has 50 students. Our task is to calculate the average marks of students in class X.

The standard approach will be to calculate the average simply:

- Calculate the total marks of all the students in Class X
- Add all the marks
- Divide the total marks by the total number of students

But what if the data is extremely large? Is this a good approach? No way, calculation marks of all the students will be a tedious and time-consuming process. So, what are the alternatives? Let's take a look at another approach.

•To begin, select groups of students from the class at random. This will be referred to as a sample. Create several samples, each with 30 students.
•Calculate each sample's individual mean.
•Calculate the average of these sample means.
•The value will give us the approximate average marks of the students in Class X.
•The histogram of the sample means marks of the students will resemble a bell curve or normal distribution.



Common Data Types

Before we jump on to the explanation of distributions, let's see what kind of a data can we encounter. The data can be discrete or continuous.

Discrete Data: as the name suggests, can take only specified values. For example, what you roll a die the possible outcomes are 1,2,3,4,5, or 6 and not 1.5 or 2.45.

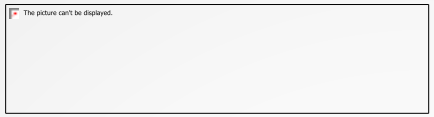
Continuous Data can take any values within a given range. The range may be finite or infinite. For example, A girl's weight or height, the length of the road. The weight of a girl can be any value from 54kgs, or 54.5 kgs, or 54.5436kgs.

Types of Distributions

Binomial Distribution has only two possible outcomes, namely 1(success) and 0(failure), and single trail. So the random variable X which has a Binomial distribution can take value !with the probability of **success, say p and the value 0 wit the probability of failure, say q or 1- p.**

Here, the occurrence of a head denote success, and the occurrence of a tail denotes failure. Probability of getting a head = 0.5 = Probability of getting a tail since there are only two possible outcomes.

The probability mass function is given by: $p^x(1-p)^{1-x}$ where x € (0, 1).
It can also be written as



The probabilities of success and failure need not be equally likely the result of a fight between me and undertaker. He is pretty much certain to win . So in this case probability of my success is 0.15 while my failure is 0.85.

Here, the probability of success = 0.15 and probability of failure = 0.85. The expected value is exactly what it sounds. If I punch you, I may expect you to punch me back. Basically expected value of any distribution is the mean of the distribution. The expected value of a random variable X from a Binomial distribution is found as follows:

$$E(X) = 1*p + 0*(1-p) = p$$

The variance of a random variable from a binomial distribution is:

$$V(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1-p)$$

There are many examples of Bernoulli distribution such as whether it's going to rain tomorrow or not where rain denotes success and no rain denotes failure and Winning (success) or losing (failure) the game.

Uniform Distribution

when you roll a fair die, The outcomes are 1 to 6. The probabilities pf getting these out comes are equally likely and that is the basis of a unform distribution. Unlike binomial Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.

A variable X is said to be uniform distribution if the density function is:

$$f(x) = \frac{1}{b-a} \quad for\ -\infty < a \leq x \leq b < \infty$$

The number of bouquets sold daily at a flower shop is uniformly distribution with maximum of 40 and a minimum of 10.
Let's try calculating the probability that the daily sales will fall between 15 and 30.

The probability that daily sales will fall between 15 and 30 is (30-15)*(1/(40-10)) = 0.5

Similarly, the probability that daily sales are greater than 20 is = 0.667
The mean and variance of X following a uniform distribution is:
Mean -> E(X) = (a+b)/2
Variance -> V(X) = $(b-a)^2/12$

Binomial distribution

Suppose that you won the loss today and this indicates a successful event.You toss again but you lost this time. If you win a toss today, this does not necessitate that you will the toss tomorrow. Let's assign a random variable, say X, to the number of times you won the toss. What can be possible value of X? It can be any number depending on the number of times you tossed a coin.

There are only two possible outcomes . Head denoting success and tail denoting failure. Therefore, probability of getting a head = 0.5 and the probability of failure can be sassily computed as q=1-p=0.5

A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.

The outcomes need not be equally likely. Remember the example of a fight between me and Undertaker? So, if the probability of success in an experiment is 0.2 then the probability of failure can be easily computed as q = 1 – 0.2 = 0.8.

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.

On the basis of the above explanation, the properties of a Binomial Distribution are

1. Each trial is independent.
2. There are only two possible outcomes in a trial- either a success or a failure.
3. A total number of n identical trials are conducted.
4. The probability of success and failure is same for all trials. (Trials are identical.)

The mathematical representation of binomial distribution is given by:

$$P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

The mean and variance of a binomial distribution are given by:

Mean -> $\mu = n*p$

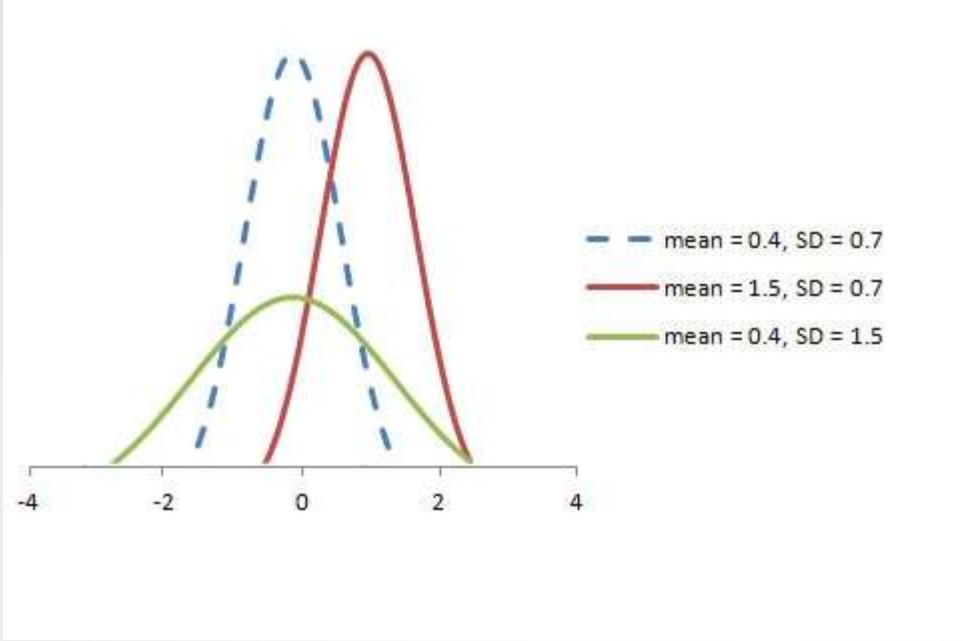
Variance -> $Var(X) = n*p*q$

Normal Distribution represents the behavior of most of the situations in the universe (That is why it's called a"normal" distribution .I guess!).the large sum of (small)random variable often turns out to be normally distributed , contributing to its widespread application . Any distribution is known as Normal distribution if it has the following.

characteristics:
1.The mean, median and mode of the distribution coincide.
2.The curve of the distribution is bell-shaped and symmetrical about the line x=μ.
3.The total area under the curve is 1.
4.Exactly half of the values are to the left of the center and the other half to the right.
A normal distribution is highly different from Binomial Distribution. However, if the number of trials approaches infinity then the shapes will be quite similar.
The PDF of a random variable X following a normal distribution is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\{-\frac{1}{2}(\frac{x-u}{\sigma})^2\}} \qquad for -\infty < x < \infty.$$

The mean and variance of a random variable X which is said to be normally distributed is given by:
Mean -> E(X) = μ
Variance -> Var(X) = σ^2
Here, μ (mean) and σ (standard deviation) are the parameters.
The graph of a random variable X ~ N (μ, σ) is shown below.



A standard normal distribution is defined as the distribution with mean 0 and standard deviation 1. For such a case, the PDF becomes:

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \qquad for\ -\infty < x < \infty$$

Poisson Distribution Suppose you work at a call center, approximately how many calls do you get in a day ? It can be any number . Now ,the entire number of calls at a call center in a day is modeled by Poisson distribution . Some more examples are

- The number of calls recorded at a hospital in a day.
- The number of thefts reported in an area on a day.
- The number of customers arriving at a salon in an hour.
- The number of suicides reported in a particular city.
- The number of printing errors at each page of the book.

You can now think of many examples following the same course. Poisson Distribution is applicable in situations where events occur at random points of time and space wherein our interest lies only in the number of occurrences of the event.
A distribution is called **Poisson distribution** when the following assumptions are valid:

1. Any successful event should not influence the outcome of another successful event.
2. The probability of success over a short interval must equal the probability of success over a longer interval.
3. The probability of success in an interval approaches zero as the interval becomes smaller.

Now, if any distribution validates the above assumptions then it is a Poisson distribution. Some notations used in Poisson distribution are:

•λ is the rate at which an event occurs,

•t is the length of a time interval,

•And X is the number of events in that time interval.

Here, X is called a Poisson Random Variable and the probability distribution of X is called Poisson distribution.

Let μ denote the mean number of events in an interval of length t. Then, μ = λ*t.

The PMF of X following a Poisson distribution is given by:

$$P(X = x) = e^{-\mu}\frac{\mu^x}{x!} \qquad for\ x = 0, 1, 2, \ldots\ldots$$

Exponential Distribution

Let's consider the call center example one more time. What about the interval of time between the calls ? Here, exponential distribution comes to our rescue. Exponential distribution models the interval of time between the calls.
Other examples are:
1. Length of time between metro arrivals,
2. Length of time between arrivals at a gas station
3. The life of an Air Conditioner

Exponential distribution is widely used for survival analysis. From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result.
A random variable X is said to have an **exponential distribution** with PDF:

$$f(x) = \{ \lambda e^{-\lambda x}, \quad x \geq 0$$

and parameter **$\lambda > 0$** which is also called the rate.
For survival analysis, λ is called the failure rate of a device at any time t, given that it has survived up to t.
Mean and Variance of a random variable X following an exponential distribution:

$$\text{Mean} \rightarrow E(X) = 1/\lambda$$

$$\text{Variance} \rightarrow Var(X) = (1/\lambda)^2$$

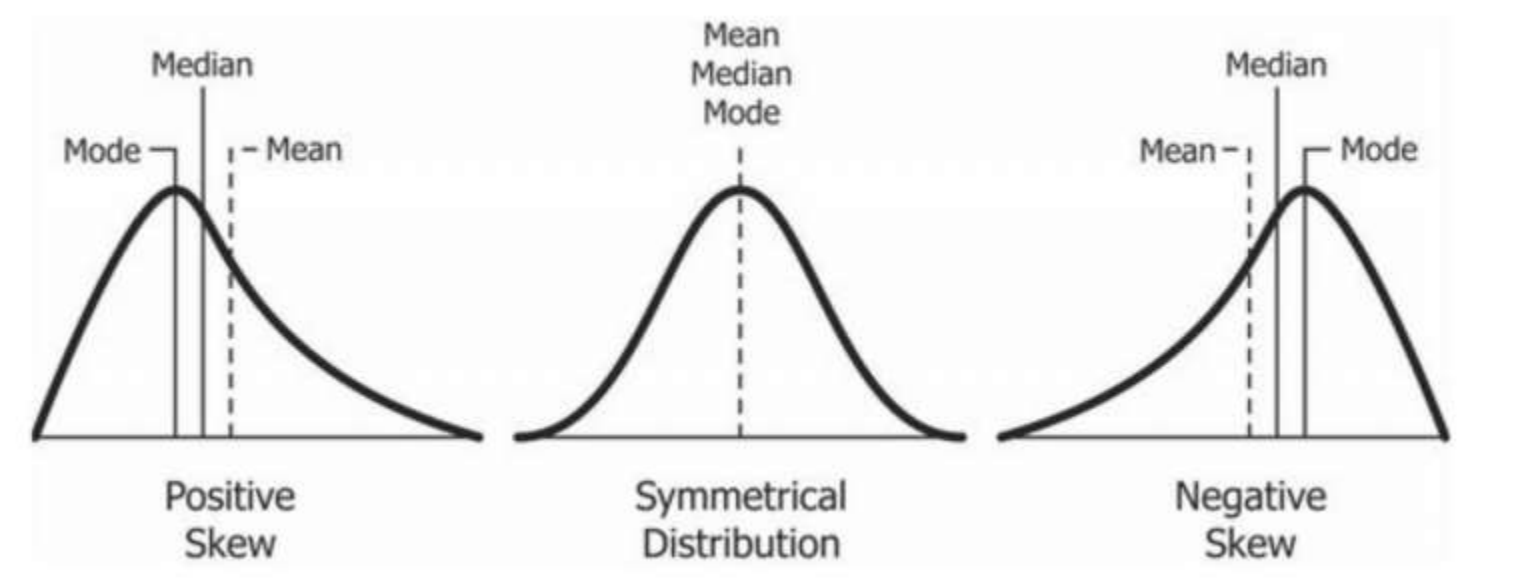
Skewed data

Skewness is the measure of the asymmetry of an ideally symmetric probability distribution and is given by the third standardized moment. If that sounds way too complex, don't worry! Let me break it down for you.

In simple words, skewness is the measure of how much the probability distribution of a random variable deviates from the normal distribution. Now,you might be thinking – why am I talking about normal distribution here?

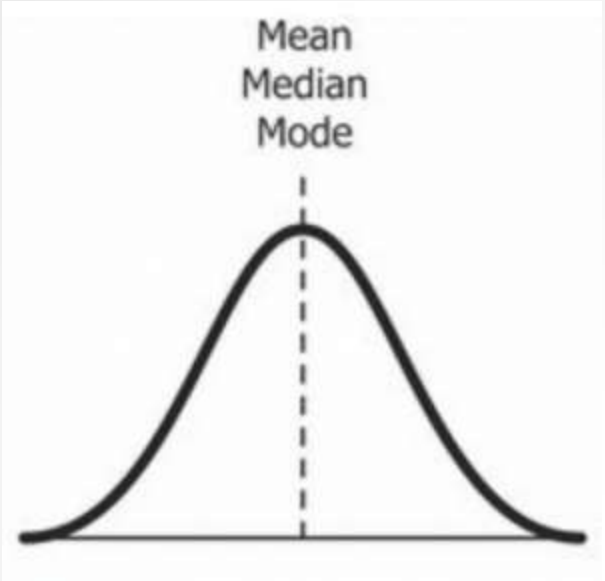
Well, the normal distribution is the probability distribution without any skewness. You can look at the image below which shows symmetrical distribution that's basically a normal distribution and you can see that it is symmetrical on both sides of the dashed line. A part this, there are two types of skewness.

- Positive Skewness
- Negative Skewness



The probability distribution with its tail on the right side is a positively skewed distribution and the one with tail on the left side a negatively skewed distribution. If you ‘ re finding the above figure confusing that’s alright. We’ll understand this in more detail later.

Symmetric/ Normal Distribution

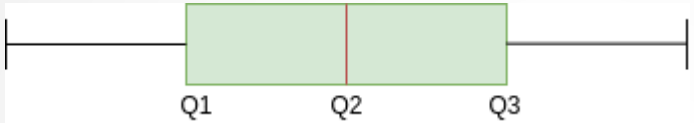


It is used as a reference for determining the skewness of a distribution . As I mentioned earlier, the ideal normal distribution is the probability distribution with almost no skewness.It is nearly perfectly symmetrical.Due to this , the valu eof skewness for a anormal distribution is zero.

But, why is it nearly perfectly symmetrical and not absolutely symmetrical?
That's because, in reality, no real word data has a perfectly normal distribution. Therefore, **even the value of skewness is not exactly zero; it is nearly zero.** Although the value of zero is used as a reference for determining the skewness of a distribution.

You can see in the above image that the same line represents the mean, median, and mode. It is because the mean, median, and mode of a perfectly normal distribution are equal.

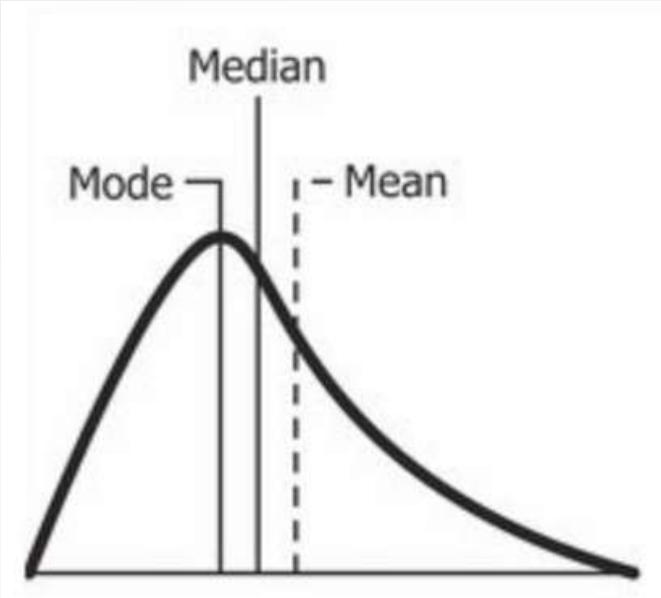
So far, we've understood the skewness of normal distribution using a probability or frequency distribution. Now, let's understand it in terms of a boxplot because that's the most common way of looking at a distribution in the data science space.



The above image is a boxplot of symmetric distribution. You'll notice here that the distance between Q1 and Q2 and Q2 and Q3 is equal i.e.:

But that's not enough for concluding if a distribution is skewed or not. We also take a look at the length of the whisker; if they are equal, then we can say that the distribution is symmetric, i.e. it is not skewed.

Positive skewed

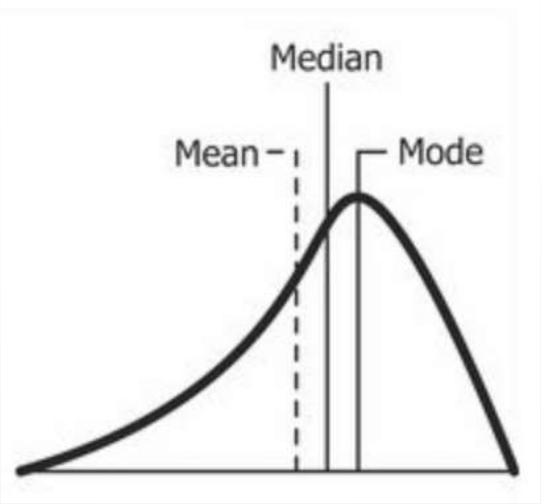


A positively skewed distribution is the distribution with the tail on its right side. The value of skewness for a positively skewed distribution is greater than zero. As you might have already understood by looking at the figure, the value of mean is the greatest one followed by median and then by mode.

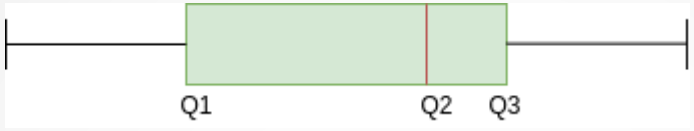
Well, the answer to that is that the skewness of the distribution is on the right; it causes the mean to be greater than the median and eventually move to the right. Also, the mode occurs at the highest frequency of the distribution which is on the left side of the median. Therefore, **mode < median < mean.**

In the above boxplot, you can see that Q2 is present nearer to Q1. This represents a positively skewed distribution. In terms of quartiles, it can be given by:

Negatively Skewed Distribution



As you might have already guessed, a negatively skewed distribution is the distribution with the tail on its left side. The value of skewness for a negatively skewed distribution is less than zero. You can also see in the above figure that the **mean < median < mode**.



In the boxplot, the relationship between quartiles for a negative skewness is given by: