Day-3 Statistics for Data Science

Measures of Variability / Dispersion

Variability refers to how "spread out" a group of scores is. Variability and Dispersion of data inside a dataset are synonymous terms that refer to how spread out a distribution is.

Suppose, we have test score of a student ~~sty~~ studying in a school:

Scenario-1	Scenario-2
41, 15, 28, 37, 85, 93, 22, 39	75, 87, 24, 56, 73, 23, 12, 10
Mean $= \frac{360}{8} = 45$	Mean $= \frac{360}{8} = 45$

We can notice although the mean of both the scenarios are same, their distribution varies widely.

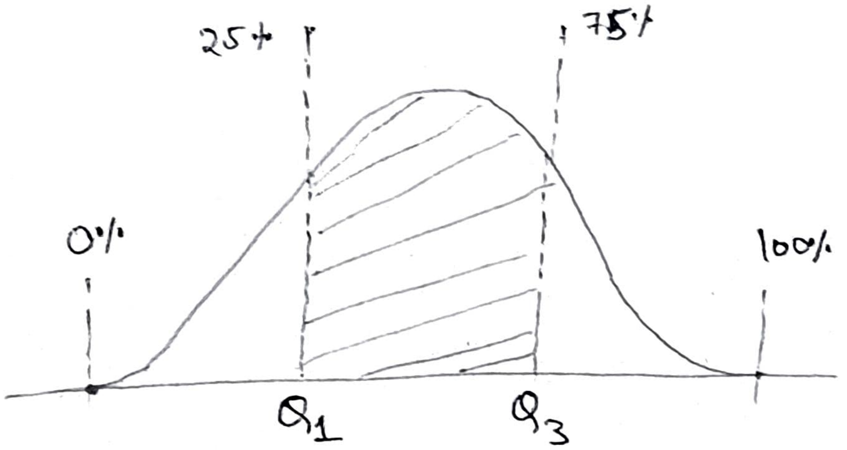
There are four measures of variability / dispersion.

(i) Range: It is the difference between the largest value and the smallest value in a distribution.

The range for scenario-1 scores is $= 93 - 15 = 78$

(ii) Interquartile Range (IQR): The interquartile range (IQR) is the range of middle 50% of scores in a distribution. It is calculated by taking the difference between 75% deviation and 25% deviation

IQR = 75th percentile - 25th percentile



$IQR = Q_3 - Q_1$

IQR is a robust measure of variability just like the median because neither is influenced by outliers because they don't depend on every value in the distribution.

IQR is also effective in case of skewed distributions.

(iii) Variance: Variance is defined as the average squared difference of the scores from the mean.

The formula for variance is given as-

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} ,$$

X_i → i^{th} observation
μ → Mean of distribution
N → distribution size

Note, when we select a sample from a population, where the population has mean (μ) and variance (σ^2), then variance of the sample is calculated by the formula,

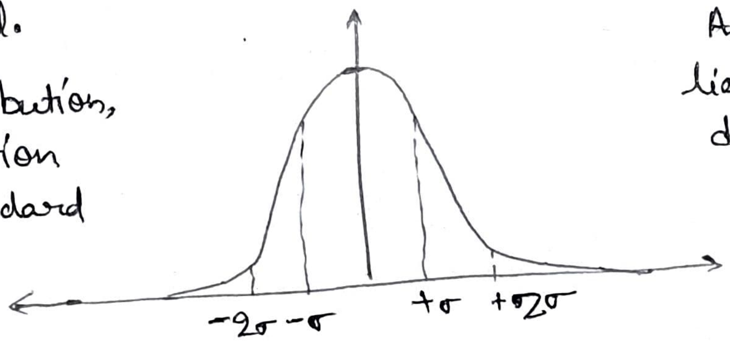
$$s^2 = \frac{\sum (x_i - M)^2}{N-1} ,$$

x_i → i^{th} observation in sample
M → Mean of sample distribution
N → distribution size of sample

(iv) Standard Deviation:

The standard deviation is simply the square root of the variance. The standard deviation is a very special measure of variability when our distribution is normal or approximately normal.

In a normal distribution, approx 68% of population lies inside one standard deviation of mean.



Also, 95% of population lies inside 2 standard deviation of mean.

* Why understanding of variability is important?

When we talk about a distribution having low variability it means, the data points inside the distribution will be consistent; On the other hand distribution having high variability will have more chances of extreme values. Thus, variability helps us to grasp the likelihood of unusual events.

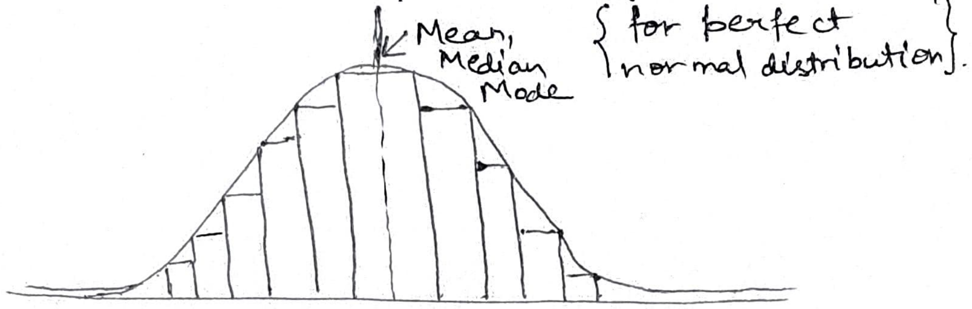
In some situations these extreme values can be very troublesome thus understanding of variability helps us to avoid those observations.

→ Range should be used mostly when the distribution size is small, as it only depends on highest and smallest value and can yield a lot of variance in case of large distribution.

Normal Distribution: The normal distribution is a continuous probability distribution that is symmetrical on both sides of a mean, so the right side of the center is a mirror image of the left side.

The area under the normal distribution curve represents probability and total area sums to one. Most of the continuous data values in a normal distribution tend to cluster around mean, meaning the further a value is from the mean, the less likely it is to occur in the distribution.

For a perfect normal distribution, the mean, median and mode will be the same value represented by peak of the curve.



The graph of normal distribution, is often called the bell curve because probability density graph looks like a bell. It is sometimes also referred to as the ~~bell~~ Gaussian distribution.

Standard Normal Distribution: A distribution having its mean centered at 0 and the standard deviation of 1 is called a standard normal distribution.

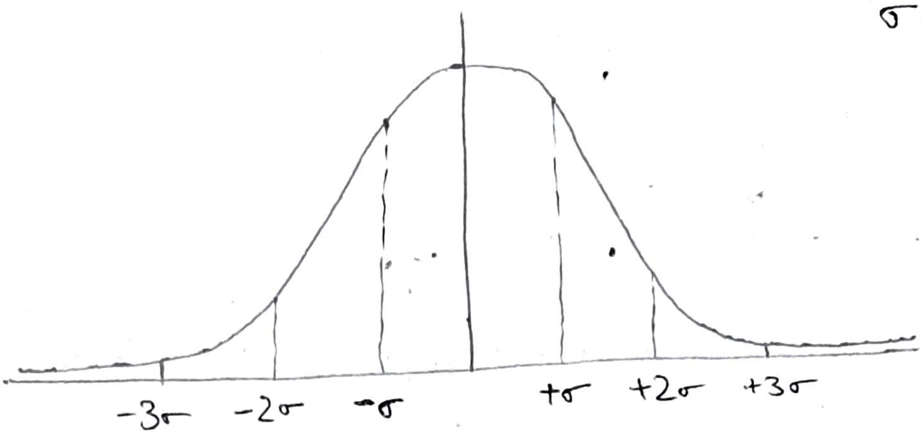
Why Normal Distribution is important?

The normal distribution is most common type of distribution that exists in the world. Either, we see the height of a population/ Weight of people in a population, etc. In all such cases, normal distribution is followed.

The normal distribution model is motivated by the **Central Limit Theorem**. (This we will read later).

→ Emperical Rule for Normal Distribution

σ → standard deviation



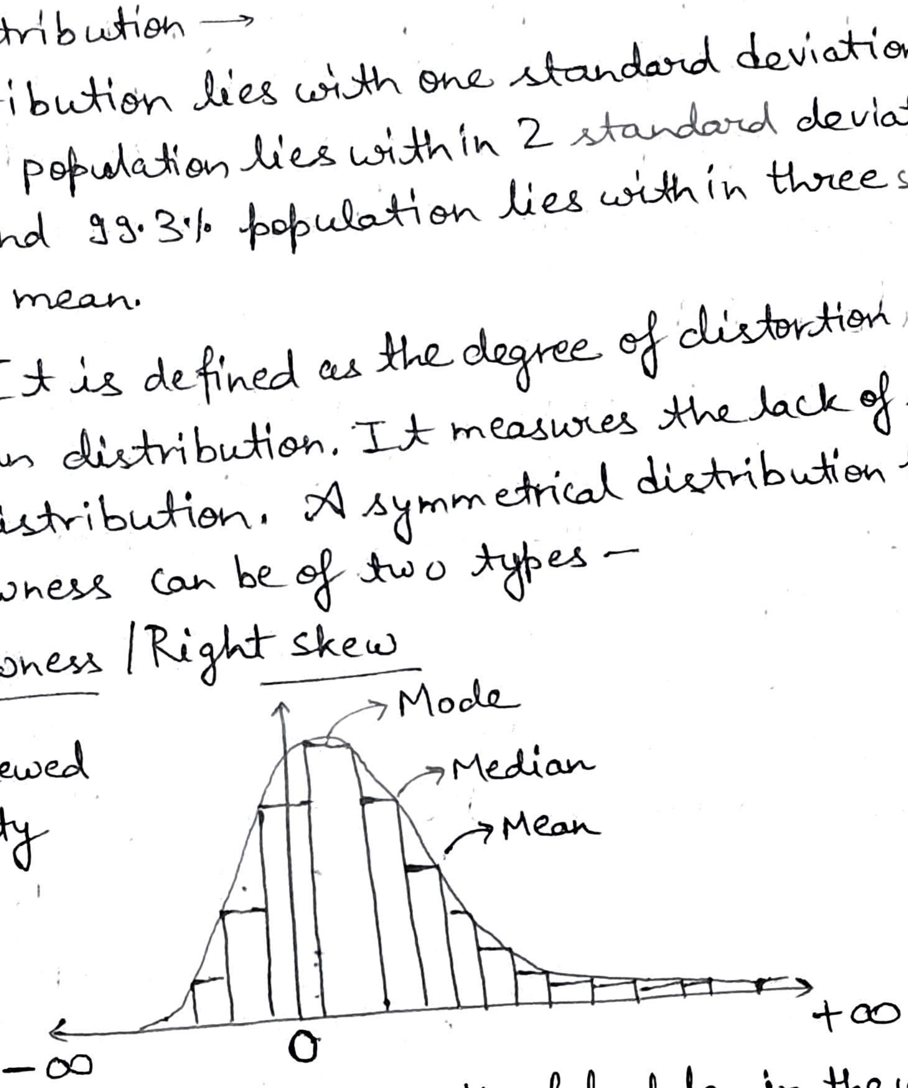
In a normal distribution →

* 68% of distribution lies with one standard deviation from mean. 95% population lies within 2 standard deviation from mean and 99.3% population lies within three standard deviation from mean.

*** Skewness** : It is defined as the degree of distortion from the normal/gaussian distribution. It measures the lack of symmetry in the data distribution. A symmetrical distribution has 0 skewness. Skewness can be of two types –

(i) Positive skewness / Right skew

In a positively skewed distribution, majority of data points lie towards the right hand side in positive direction.

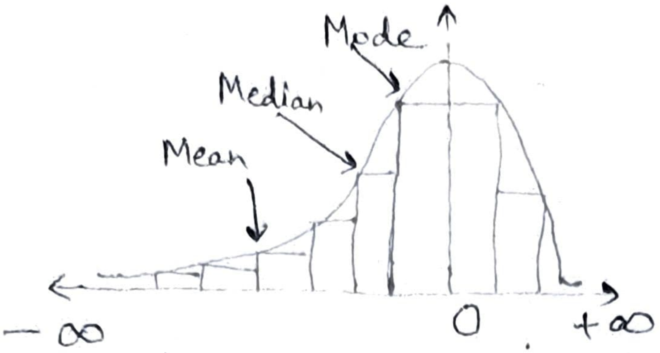


We can take an example of networth of people in the world. There will be very less no. of people in the world whose net worth will be in billions and there will be more people whose networth will be in lakhs, crores.

In such distribution,

$$\boxed{\text{Mean} > \text{Median} > \text{Mode}}$$

(iii) Negative skewness / Left skew



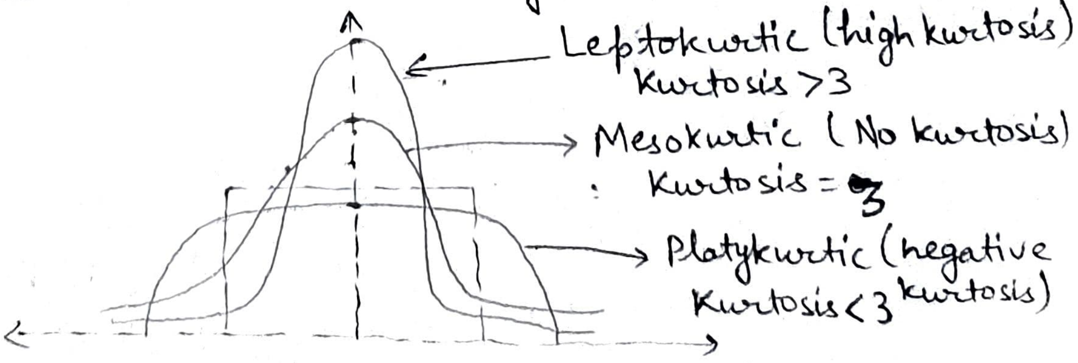
In negative / left skewed distribution, the left tail of distribution is longer to the right tail.

An example can be sleeping pattern of people, very less no. of people will sleep for less than 5 hours perday, but there will be large no. of people sleeping atleast 6-8 hours per day.

In such distribution, $\boxed{\text{Mean} < \text{Median} < \text{Mode}}$

*** Kurtosis:** Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution. In other words, kurtosis identifies whether the tails of a given distribution contain extreme values.

Skewness and kurtosis must not be confused with each other, as skewness measures the symmetry of the distribution, while kurtosis determines the heaviness of the distribution tails.



→ Leptokurtic ⇒ High kurtosis is an indicator that data has heavy tails and can have outliers.

→ Platykurtic ⇒ Low kurtosis is an indicator that data looks flatter and has presence of less outliers.

→ Mesokurtic ⇒ It generally represents normal distribution and has kurtosis value equals to 3.