**Statistics & Distribution**

**Calculate the measures of central tendency, asymmetry, and variability. Type of Distribution**

To understand a distribution completely and properly we need the following measures:

1. Measures of central tendency

2. Measures of dispersion

3. Measures to describe shape of distribution

**Measures of central tendency:**This measure tries to describe the entire dataset with a single value or metric which represents the middle or center of distribution. It is also known as measure of center or central location. These measures include:

**Mean:**Computed by taking the sum of all the values in the dataset divided by the total number of values.

**Median**: It’s the value which lies in the middle of the dataset when arranged in ascending or descending order.

**Mode:**It is the most occurring value in the dataset or the value which occurs very frequently.

Measures of central tendency

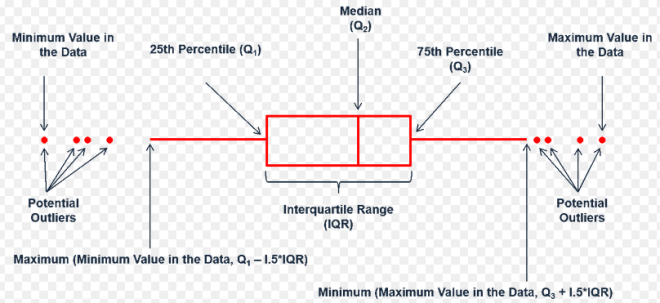
**Measures of dispersion:**As the measure of central value gives out the central value of the distribution, measures of dispersion describe the spread of data or the variation of data around the central value. Two different distributions can have the same mean or median but different levels of variability. There are many different measures to measure the dispersion as described below:

**Range:**Provides the difference between the largest and smallest value of the dataset. It’s the simplest method to calculate and see the spread of our data but as it’s the difference between the extreme values it doesn’t tell us how our data is distributed.

**Quartile/ Percentile**: makes it easy to work with data which is not symmetrically distributed and has outliers.

Mean, Median and mode is the numerical summary of the entire dataset which is symmetrically distributed whereas quartiles divide our dataset into four equally sized groups based on five number summary:

Minimum, first quartile, median, third quartile and maximum. The box in the box plot represents the 50 percent of the data values known as interquartile range (IQR). IQR indicates the variability in the set of values. Large IQR means a large spread in values. Small IQR indicates most of the values fall near the center of data. Box plot shows minimum and maximum values through the whiskers which extends both the sides and also outlier points which extends beyond the whiskers.



**Box plot**

**Interquartile Range (IQR):** It’s the difference between the third quartile and the first quartile. 50% of the population data lies here.

**Variance:**Calculated by taking the deviation of each value in the dataset from the mean and then squaring it. In other words, Variance is the average of all squared deviations.

**Standard Deviation**gives us an idea about the concentration of the data around the mean of the dataset. Standard deviation is low if the data is highly concentrated around the mean and vice versa. It is very sensitive to outliers as outliers impact the mean of the dataset which in turn affects the standard deviation. It is the square root of variance.

Measures of dispersion

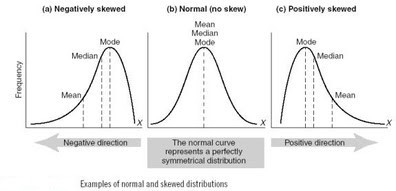
**Measures to describe shape of distribution:**

Measures of center and spread tells us just about the central values and spread. How do we describe the shape of the distribution?

Histogram will give us a general idea, but two numerical measures of shape will help us with the precise evaluation of the shape of the distribution.

**Skewness**is the asymmetry in the distribution because of which the curve appears distorted or skewed either to left or right of the normal distribution in a dataset.

In other words skewness is the extent to which a distribution differs from a normal distribution.



**Positively and negatively skewed distribution**

**Skewed distributions can be:**

**Positively skewed:**Most frequent values are low and tail is towards high values.

**Negatively skewed:**Most frequent values are high and tail is towards low values.

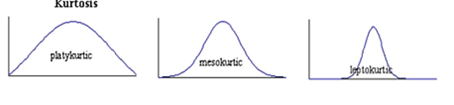
**Measures of central tendency can also be used to detect skewness in a distribution.**

Central tendency values will not be the same if the distribution is skewed.

If Mode< Median< Mean then the distribution is positively skewed.

If Mode> Median> Mean then the distribution is negatively skewed.

**Kurtosis**is the measure of the combined weight of the tails of the distribution relative to the center of the distribution. When a normal distribution is represented via histogram, it shows a bell curve with + and — standard deviations from mean. However, when kurtosis is present then the tails further extend farther than the + and — standard deviations of the normal bell-curved distribution. Kurtosis makes the data to look flatter (or less flat as compared to normal distribution). The standard normal distribution has kurtosis of 3.

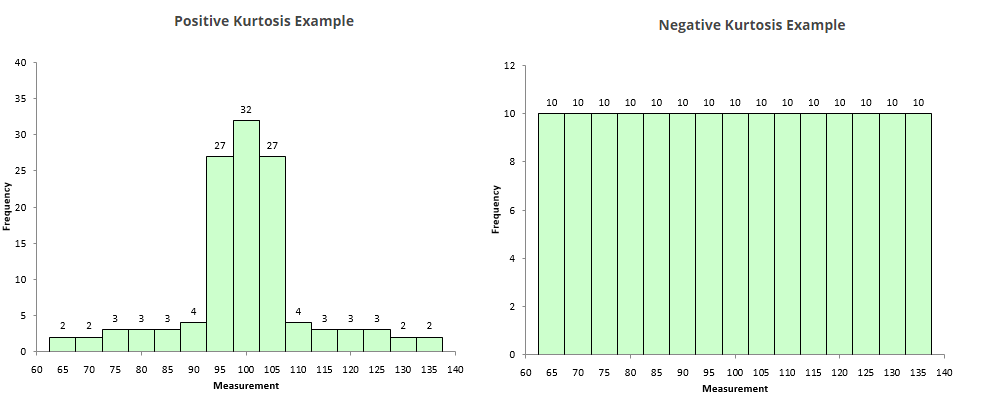


**Types of Kurtosis**

When the distribution of the data is similar to the normal distribution or the kurtosis of the distribution is 3, it is called as Mesokurtic distribution.

Any distribution which has kurtosis more than Normal distribution (K>3), it is called as leptokurtic (thin) distribution. This type distribution has positive kurtosis.

Distribution which has kurtosis less than Normal distribution (K < 3), it is called as platykurtic(flat) distribution. This type distribution has negative kurtosis.



So this is how can actually describe any population distribution using the different measures and see graphically how it looks and understand it completely.