Descriptive Statistics

Descriptive statistics are methods used to describe and summarize a dataset. This can be done in three ways:

1. By measuring the central tendency of a dataset.
2. By measuring the extent of dispersion in a dataset.
3. By visualizing the frequency distribution of a dataset.

Central Tendency Measures

This type of statistical methods includes mean, median and mode of a dataset.

a) Mean: It is the average value of the data. It gets greatly affected if there are outliers present. $\frac{\sum_{i=0}^{n} x_i}{n}$ where, n = number of observations, x_i = ith observation.
b) Median: It is the middlemost value of the data, also known as, 50th percentile. Outliers have no impact on the median.
c) Mode: It is the most frequent value of the data and hence there can be multiple modes. Outliers have no impact on mode. It can be calculated for both quantitative as well as qualitative data

Dispersion Measures

This type of statistical methods measure the spread of the data.

a) Range: Difference between largest and smallest value in the dataset.
b) Inter Quartile Range (IQR): Quartiles are three points that divide the dataset into 4 equal parts. The three points are media of the data, median of the leftmost half part of the data and the median of the rightmost half part of the data. IQR is the difference between 3rd quartile and 1st quartile value.
c) Standard Deviation: It shows how far the data spreads from the mean value.
    a. *Sample Standard Deviation*: Measures deviation of data from sample mean.

$$\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

$$where, x = value\ in\ a\ sample$$

$\bar{x} = sample\ mean$

$n = number\ of\ samples.$

b. *Population Standard Deviation*: Measures deviation of data in a population.

$$\sqrt{\frac{\sum(x - \bar{\mu})^2}{N}}$$

$where, x = value\ in\ a\ population$

$\bar{\mu} = population\ mean$

$N = total\ nmber\ of\ data\ points\ in\ a\ population$

d) Variance: It measures variability in the data. It is calculated by squaring the standard deviation.

a. Sample Variance: $\frac{\sum(x-\bar{x})^2}{n-1}$

b. Population Variance: $\frac{\sum(x-\bar{\mu})^2}{N}$

Frequency Distribution

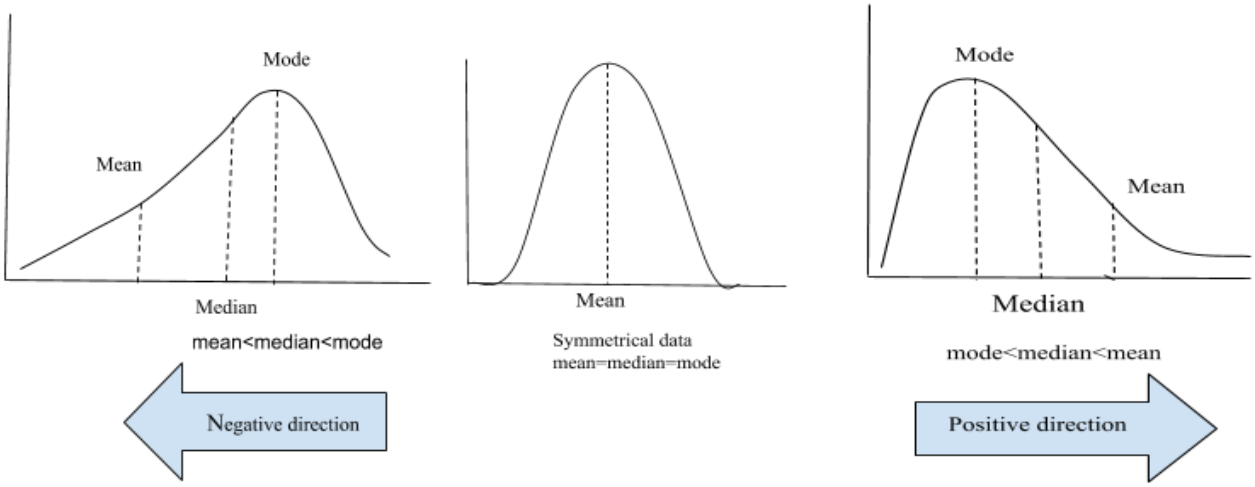
Shape of a data gives idea about probability distribution. It helps us understand the following:

a) Symmetry: Data is said to be symmetric when it follows normal distribution, i.e., data is distributed on either side of the mean. Mean and median are close together.

b) Skewness: Data is said to be skewed when data falls dominantly on one side. It is a measure of asymmetry.

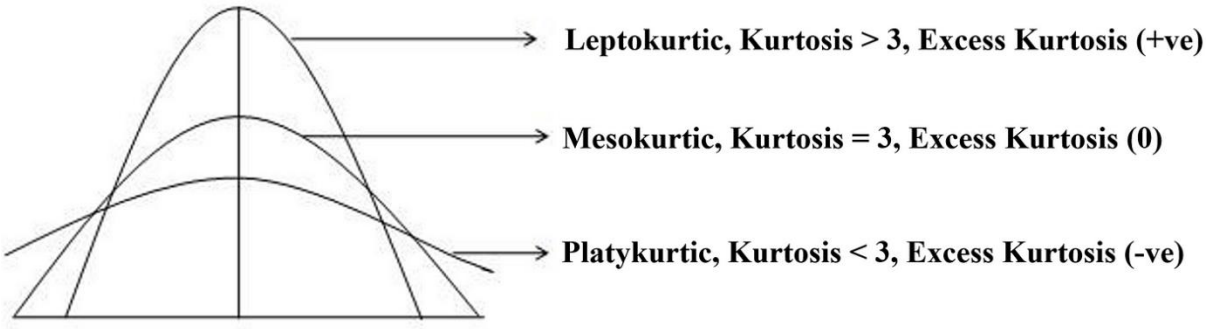
a. *Positive Skewness*: Data is positively skewed when data points are clustered on the left side of the distribution. Mean and median are greater than mode.

b. *Negative Skewness*: Data is negatively skewed when data points are clustered on the right side of the distribution. Mean and median are less than mode.



Skewness

c) Kurtosis: It measures degree of peakness in a frequency distribution. It tells us about the outliers present in a data.
    a. *Platykurtic*: The distribution is platykurtic when it has flat tails and the value of kurtosis is less than 3. It is aka negative kurtosis. The data is fairly distributed with few outliers at the end of the tail on either side.
    b. *Mesokurtic*: The distribution is mesokurtic when it is normally distributed and the value of kurtosis is equal to 3.
    c. *Leptokurtic*: The distribution is said to be leptokurtic when is has a sharp peak and the value of kurtosis is greater than 3. It is aka negative kurtosis. The tails are heavier and can be indicative of outliers in the data.



Kurtosis

Univariate and Bivariate Data Statistics

Univariate: When the focus of exploratory data analysis is on a single variable it is called univariate data statistics. The main point of analysis is to understand the characteristics of a single variable which is done by measuring the central tendency, distribution and the frequency distribution of that variable.

Bivariate: When the focus of exploratory data analysis is on understanding the relationship between two variables along with their individual characteristics, it is known as bivariate statistics. It involves the basic descriptive statistics as well as calculation of correlation coefficients.