Descriptive Statistics

Descriptive Statistics summarize the characteristics of a data set which helps to analyze the data in detail.

Measure of Central Tendency:

(a) Mean:

Mean is the most popular measure of central tendency. It is the average value determined by sum of all observations divided by total number of observations. It is represented by $\overline{x}$.

$$\overline{x} = \frac{\text{sum of all observations}}{n}$$

Important properties of mean:

1. Value of mean will change if extreme value is removed or added to the observations.
2. Sum of the deviations about the mean is equal to 0.

$$\Sigma (x - \overline{x}) = 0$$

3. Mean is least subject to sampling variation or most stable.

(b) Median (50th Percentile):

The median is the middle value for all the observations arranged in order of magnitude.
If number of observations is odd – median will be the middlemost value.
If number of observations is even – median will be the average of two centermost values.

Important properties of median:

1. It is less sensitive to extreme vales than the mean.
2. Median is more subject to sampling variation than the mean but less than the mode.

(c) Mode:

Mode is defined as the most frequent value in our dataset.

For ungrouped observations – mode is the observation will highest frequency.
For grouped observations – mode is midpoint of interval with highest frequency.

Important properties of mode:

1. It is the easiest of all central tendency measures.
2. There can be more than one mode.
   If one mode – Unimodal model
   If two modes – Bimodal model
   If more than two modes – Multimodal model
3. Mode is sensitive to sampling variations.

Measure of Variability:

This represents the extend of dispersion. Two distributions can have same central tendency but different degrees of dispersion. These measures are sensitive to each observation and are stable with regard to sampling variation.

(a) Range:

It is the difference between highest and lowest observation in the distribution.
It measures the spread of extreme scores.
Range = Highest observation – Lowest observation

(b) Variation:

Variation is the sum of squared deviation scores from the mean.
Variation(SS) = $\Sigma (x - \overline{x})^2$

(c) Variance:

It is the average squared deviation of scores from the mean. A high variance indicates that data points are spread widely whereas small variance indicates that the data points are closer to the mean of the data set. It is represented by s^2.

s^2 = SS / n (This is biased estimate of population variance)

To adjust for the bias, (n-1) is used instead of n.

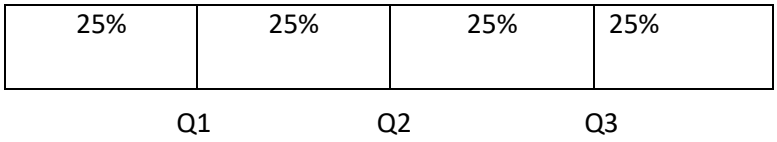
Therefore, $s^2 = SS / (n-1)$

(d) Standard Deviation:

Standard Deviation is the square root of variance. It is represented by SD and it can be interpret as the average distance of values from the mean.
SD or $s = \sqrt{s^2}$
The lower the SD, closer the scores to the mean of the dataset.

(e) Quartiles :

Quartiles divides the dataset into four equal parts. Q1, Q2 and Q3 are three quartiles.
Q1 : 25% of observations lie below this point and 75% lie above Q1.
Q2 : This point is median. 50% of observations are above it and 50% are below this point.
Q3 : 75% of observations are below this point and 25% above it.



Inter-quartile Range = (Q - Q1)

Shape:

(a) Skewness:

Skewness is the measure of symmetry of distribution or we should say the lack of symmetry. It measures the deviation of distribution from the normal distribution which is symmetric on both sides.

1. No Skewness : Symmetrical distribution. Here, Mean = Median = Mode
2. Positive Skew : Tail is on right side and curve is bigger on left side and value of skewness will be positive, also known as right-skewed. Here, Mean > Median > Mode

3. Negative Skew : Tail is on left side and curve is bigger on right side and value of skewness will be negative, also known as left-skewed. Here, Mean < Median < Mode

* Measure of Skewness :
Skewness index = $m_3 / (m_2)^{3/2}$

Where, m_3 is third moment and m_2 is second moment.

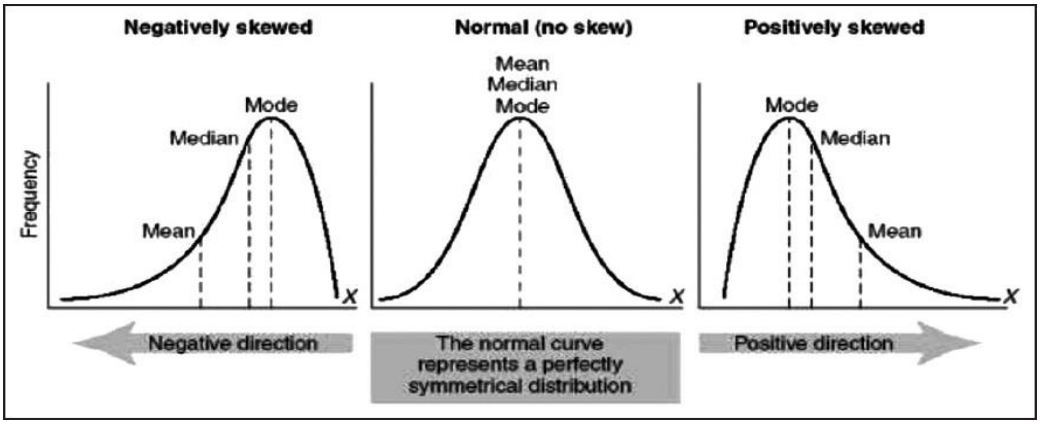
$$m_2 = \frac{\sum(x - \overline{x})^2}{n}$$

$$m_3 = \frac{\sum(x - \overline{x})^3}{n}$$

Skewness can affect test of mean difference, correlation and results of regression model, therefore sometimes transformation to the normal curve can help.

If -0.5 < Skewness < 0.5, distribution is fairly symmetrical.
If -1 <= Skewness <= -0.5 or 0.5 <= Skewness <= 1, distribution is moderately skewed.
If Skewness>1 or Skewness < -1, distribution is highly skewed.

|Skewness|>2 is a cause for concern.



(b) Kurtosis:

Kurtosis is a measure of peakedness or flatness of distribution when compared to the normal distribution.

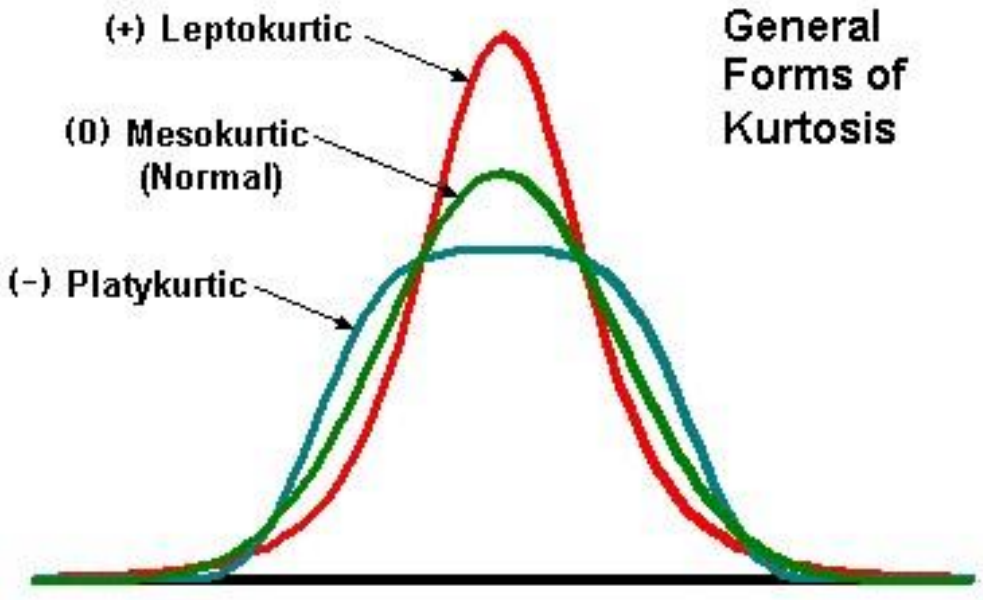
Kurtosis = $m_4 / (m_2)^2$

Where m_4 is fourth moment and m_2 is second moment.

$$m_2 = \frac{\sum(x - \overline{x})^2}{n}$$

$$m_4 = \frac{\sum(x - \overline{x})^4}{n}$$

Mesokurtic (Zero kurtosis): Value of kurtosis is 3, just like the normal distribution.
Leptokurtic (Positive kurtosis): Value of kurtosis >3, high peak, low shoulders and a long tail.
Playtkurtic (Negative kurtosis): Value of kurtosis <3, low peak, high shoulders, and a short tail.



Transformation:

When distribution is highly skewed, non-linear transformations are used to normalize the distribution. Many times, transformation used to decrease the skewness will also decrease the kurtosis value.

1. Transformations used for Positive Skewed Distribution:

Distribution	Transformation
Moderate skewness	NEWX=SQRT(X)
Substantial skewness	NEWX=LOG10(X)
Substantial skewness with Zero	NEWX=LOG10(X + C)
Severe skewness L shaped	NEWX=1/X
Severe skewness L shaped with zero	NEWX=1/(X+C)

C is a constant added to each score so that the smallest score is 1. usually equal to the 1- the smallest score.

2. Transformations used for Negative Skewed Distribution:

Distribution	Transformation
Moderate negative skewness	NEWX=SQRT(K-X)
Substantial negative skewness	NEWX=LOG10(K-X)
Severe negative skewness J shaped	NEWX=1/(K-X)

Use K – X to change the distribution from negative skewness to positive skewness. K is a constant form which each score is subtracted so that the smallest score is 1; usually equal to the 1+largest score.