1. What is Descriptive Statistics?
2. Types of Descriptive Statistics?
3. Measure of Central Tendency (Mean, Median, Mode)
4. Measure of Spread / Dispersion (Standard Deviation, Mean Deviation, Variance, Percentile, Quartiles, Interquartile Range)
5. What is Skewness?
6. What is Kurtosis?
7. What is Correlation?

**What is Descriptive Statistics?**

Descriptive statistics involves summarizing and organizing the data so they can be easily understood. Descriptive statistics, unlike inferential statistics, seeks to describe the data, but does not attempt to make inferences from the sample to the whole population. Here, we typically describe the data in a sample. This generally means that descriptive statistics, unlike inferential statistics, is not developed on the basis of probability theory.

**Types of Descriptive Statistics?**

Descriptive statistics are broken down into two categories. Measures of central tendency and measures of variability (spread).

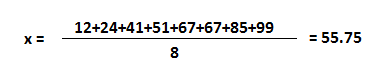
**The Measure of Central Tendency**

Central tendency refers to the idea that there is one number that best summarizes the entire set of measurements, a number that is in some way “central” to the set.

**Mean / Average**

Mean or Average is a central tendency of the data i.e. a number around which a whole data is spread out. In a way, it is a single number that can estimate the value of the whole data set.

Let’s calculate the mean of the data set having 8 integers.



**Median**

Median is the value that divides the data into 2 equal parts i.e. number of terms on the right side of it is the same as a number of terms on the left side of it when data is arranged in either ascending or descending order.

Note: If you sort data in descending order, it won’t affect the median but IQR will be negative. We will talk about IQR later in this blog.

Median will be a middle term if the number of terms is odd

Median will be the average of the middle 2 terms if a number of terms is even.



The median is 59 which will divide a set of numbers into equal two parts. Since there are even numbers in the set, the answer is the average of middle numbers 51 and 67.

Note: When values are in arithmetic progression (difference between the consecutive terms is constant. Here it is 2.), the median is always equal to the mean.



The mean of these 5 numbers is 6 and so median.

**Mode**

Mode is the term appearing maximum time in data set i.e. term that has the highest frequency.



In this data set, the mode is 67 because it has more than the rest of the values, i.e. twice.

But there could be a data set where there is no mode at all as all values appear same number of times. If two values appeared same time and more than the rest of the values then the data set is bimodal. If three values appeared same time and more than the rest of the values then the data set is trimodal and for n modes, that data set is multimodal.

**Measure of Spread / Dispersion**

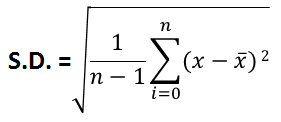
Measure of Spread refers to the idea of variability within your data.

**Standard deviation**

Standard deviation is the measurement of the average distance between each quantity and mean. That is, how data is spread out from the mean. A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

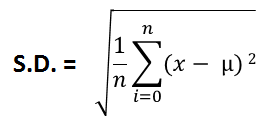
There are situations when we have to choose between sample or population Standard Deviation.

When we are asked to find SD of some part of a population, a segment of population; then we use sample Standard Deviation.



where x̅ is mean of a sample.

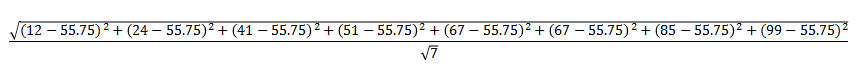
But when we have to deal with a whole population, then we use population Standard Deviation.



where µ is mean of a population.

Though sample is a part of a population, their SD formulas should have been same, but it is not. To find out more about it, refer this link

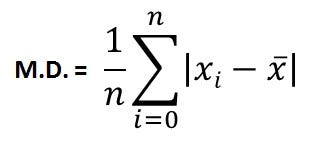
As you know, in descriptive statistics, we generally deal with data available in a sample, not in a population. So if we use the previous data set, and substitute the values in the sample formula,



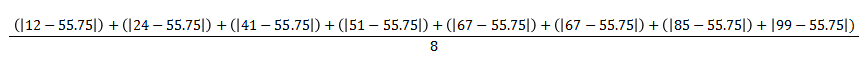
And the answer would be 29.62.

**Mean Deviation / Mean Absolute Deviation**

It is an average of absolute differences between each value in a set of values, and the average of all values of that set.



So, if we use the previous data set, and substitute the values,

And the answer would be 23.75.

**Variance**

Variance is a square of average distance between each quantity and mean. That is it is square of standard deviation.



And the answer would be 877.34.

**Range**

Range is one of the simplest techniques of descriptive statistics. It is the difference between the lowest and highest value.



Range is 99–12 = 87

**Percentile**

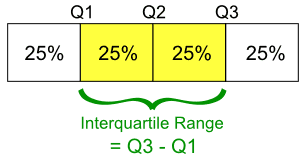
Percentile is a way to represent position of a values in data set. To calculate percentile, values in data set should always be in ascending order.



The median 59 has 4 values less than itself out of 8. It can also be said as: In data set, 59 is 50th percentile because 50% of the total terms are less than 59. In general, if k is nth percentile, it implies that n% of the total terms are less than k.

**Quartiles**

In statistics and probability, quartiles are values that divide your data into quarters provided data is sorted in an ascending order.



There are three quartile values. First quartile value is at 25 percentile. Second quartile is 50 percentile and the third quartile is 75 percentile. Second quartile (Q2) is median of the whole data. First quartile (Q1) is median of upper half of the data. And Third Quartile (Q3) is median of lower half of the data.



So here, by analogy,

Q2 = 67: is 50 percentile of the whole data and is median.

Q1 = 41: is 25 percentile of the data.

Q3 = 85: is 75 percentile of the date.

Interquartile range (IQR) = Q3 - Q1 = 85 - 41 = 44

Note: If you sort data in descending order, IQR will be -44. The magnitude will be same, just sign will differ. Negative IQR is fine, if your data is in descending order. It just we negate smaller values from larger values, we prefer ascending order (Q3 - Q1).

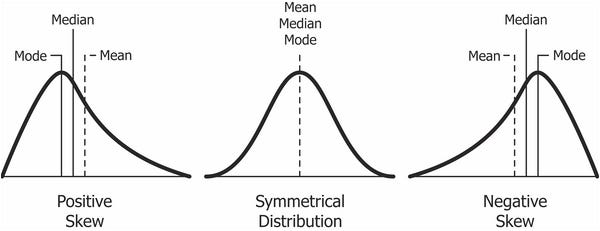
**Skewness**

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or undefined.

In a perfect normal distribution, the tails on either side of the curve are exact mirror images of each other.

When a distribution is skewed to the left, the tail on the curve’s left-hand side is longer than the tail on the right-hand side, and the mean is less than the mode. This situation is also called negative skewness.

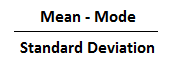
When a distribution is skewed to the right, the tail on the curve’s right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode. This situation is also called positive skewness.



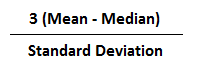
**How to the skewness coefficient?**

To calculate skewness coefficient of the sample, there are two methods:

1] Pearson First Coefficient of Skewness (Mode skewness)



2] Pearson Second Coefficient of Skewness (Median skewness)



***Interpretations***

The direction of skewness is given by the sign. A zero means no skewness at all.

A negative value means the distribution is negatively skewed. A positive value means the distribution is positively skewed.

The coefficient compares the sample distribution with a normal distribution. The larger the value, the larger the distribution differs from a normal distribution.

Sample problem: Use Pearson’s Coefficient #1 and #2 to find the skewness for data with the following characteristics:

Mean = 50.

Median = 56.

Mode = 60.

Standard deviation = 8.5.

Pearson’s First Coefficient of Skewness: -1.17.

Pearson’s Second Coefficient of Skewness: -2.117.

*Note*: Pearson’s first coefficient of skewness uses the mode. Therefore, if frequency of values is very low then it will not give a stable measure of central tendency. For example, the mode in both these sets of data is 9:

1, 2, 3, 4, 4, 5, 6, 7, 8, 9.

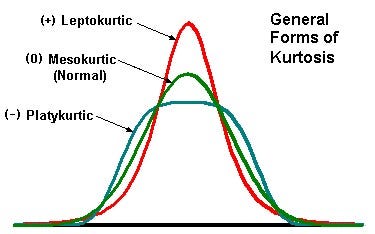
In the first set of data, the mode only appears twice. So it is not a good idea to use Pearson’s First Coefficient of Skewness. But in the second set,

1, 2, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 6, 7, 8, 9, 10, 12, 12, 13.

mode 4 appears 8 times. Therefore, Pearson’s Second Coefficient of Skewness will likely give you a reasonable result.

**Kurtosis**

The exact interpretation of the measure of Kurtosis used to be disputed but is now settled. It's about the existence of outliers. Kurtosis is a measure of whether the data are heavy-tailed (profusion of outliers) or light-tailed (lack of outliers) relative to a normal distribution.



There are three types of Kurtosis

1. **Mesokurtic**

Mesokurtic is the distribution that has similar kurtosis as normal distribution kurtosis, which is zero.

1. **Leptokurtic**

Distribution is the distribution that has kurtosis greater than a Mesokurtic distribution. Tails of such distributions are thick and heavy. If the curve of distribution is more peaked than the Mesokurtic curve, it is referred to as a Leptokurtic curve.

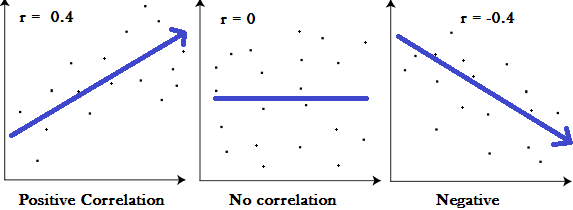
1. **Platykurtic**

Distribution is the distribution that has kurtosis lesser than a Mesokurtic distribution. Tails of such distributions thinner. If a curve of a distribution is less peaked than a Mesokurtic curve, it is referred to as a Platykurtic curve.

The main difference between skewness and kurtosis is that the skewness refers to the degree of symmetry, whereas the kurtosis refers to the degree of presence of outliers in the distribution.

**Correlation**

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.



The main result of a correlation is called the correlation coefficient (or “r”). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.

If r is close to 0, it means there is no relationship between the variables. If r is positive, it means that as one variable gets larger the other gets larger. If r is negative it means that as one gets larger, the other gets smaller (often called an “inverse” correlation).