

DESCRIPTIVE STATISTICS

What is Statistics?

Statistics is the science of collecting data and analyzing them to infer proportions (sample) that are representative of the population. In other words, statistics is interpreting data in order to make predictions for the population.

Branches of Statistics:

There are two branches of Statistics.

- **DESCRIPTIVE STATISTICS** : Descriptive Statistics is a statistics or a measure that describes the data.
- **INFERENTIAL STATISTICS** : Using a random sample of data taken from a population to describe and make inferences about the population is called Inferential Statistics.

Descriptive Statistics

Descriptive Statistics is summarizing the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier. It does not involve any generalization or inference beyond what is available. This means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

Commonly Used Measures

1. Measures of Central Tendency
2. Measures of Dispersion (or Variability) or Spread

Measures of Central Tendency

A Measure of Central Tendency is a one number summary of the data that typically describes the center of the data. These one number summary is of three types.

1. **Mean** : Mean is defined as the ratio of the sum of all the observations in the data to the total number of observations. This is also known as Average. Thus mean is a number around which the entire data set is spread.
2. **Median** : Median is the point which divides the entire data into two equal halves. One-half of the data is less than the median, and the other half is greater than the same. Median is calculated by first arranging the data in either ascending or descending order.
 - If the number of observations are odd, median is given by the middle observation in the sorted form.
 - If the number of observations are even, median is given by the mean of the two middle observation in the sorted form.

An important point to note that the order of the data (ascending or descending) does not effect the median.

3. Mode : Mode is the number which has the maximum frequency in the entire data set, or in other words, mode is the number that appears the maximum number of times. A data can have one or more than one mode.

- If there is only one number that appears maximum number of times, the data has one mode, and is called **Uni-modal**.
- If there are two numbers that appear maximum number of times, the data has two modes, and is called **Bi-modal**.

- If there are more than two numbers that appear maximum number of times, the data has more than two modes, and is called **Multi-modal**.

Example to compute the Measures of Central Tendency

Consider the following data points.

17, 16, 21, 18, 15, 17, 21, 19, 11, 23

- Mean — Mean is calculated as

$$\text{Mean} = \frac{17 + 16 + 21 + 18 + 15 + 17 + 21 + 19 + 11 + 23}{10} = \frac{178}{10} = 17.8$$

- Median — To calculate Median, let's arrange the data in ascending order.

11, 15, 16, 17, 17, 18, 19, 21, 21, 23

Since the number of observations is even (10), median is given by the average of the two middle observations (5th and 6th here).

$$\text{Median} = \frac{5^{\text{th}} \text{ Obs} + 6^{\text{th}} \text{ Obs}}{2} = \frac{17 + 18}{2} = 17.5$$

- Mode — Mode is given by the number that occurs maximum number of times. Here, 17 and 21 both occur twice. Hence, this is a Bimodal data and the modes are 17 and 21.

Note-

1. Since Median and Mode does not take all the data points for calculations, these are robust to outliers, i.e. these are not effected by outliers.
2. At the same time, Mean shifts towards the outlier as it considers all the data points. This means if the outlier is big, mean overestimates the data and if it is small, the data is underestimated.
3. If the distribution is symmetrical, Mean = Median = Mode. Normal distribution is an example.

Measures of Dispersion (or Variability)

Measures of Dispersion describes the spread of the data around the central value (or the Measures of Central Tendency)

1. **Absolute Deviation from Mean** — The Absolute Deviation from Mean, also called Mean Absolute Deviation (MAD), describe the variation in the data set, in sense that it tells the average absolute distance of each data point in the set. It is calculated as

$$\text{Mean Absolute Deviation} = \frac{1}{N} \sum_{i=1}^N |X_i - \bar{X}|$$

2. **Variance** — Variance measures how far are data points spread out from the mean. A high variance indicates that data points are spread widely and a small variance indicates that the data points are closer to the mean of the data set. It is calculated as

$$\text{Variance} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$$

3. Standard Deviation — The square root of Variance is called the Standard Deviation. It is calculated as

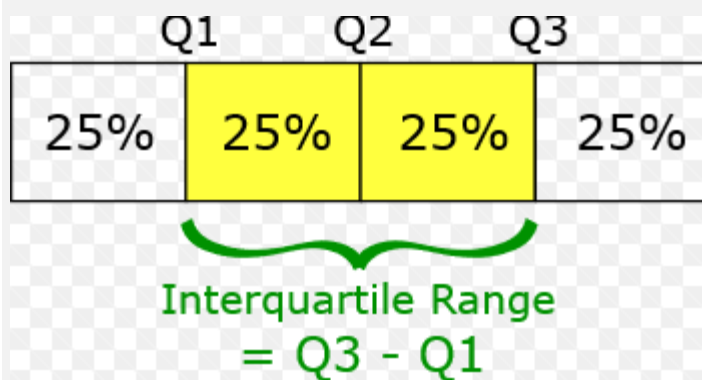
$$\text{Std Deviation} = \sqrt{\text{Variance}} = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2}$$

4. Range — Range is the difference between the Maximum value and the Minimum value in the data set. It is given as

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

5. Quartiles — Quartiles are the points in the data set that divides the data set into four equal parts. Q1, Q2 and Q3 are the first, second and third quartile of the data set.

- 25% of the data points lie below Q1 and 75% lie above it.
- 50% of the data points lie below Q2 and 50% lie above it. Q2 is nothing but Median.
- 75% of the data points lie below Q3 and 25% lie above it.



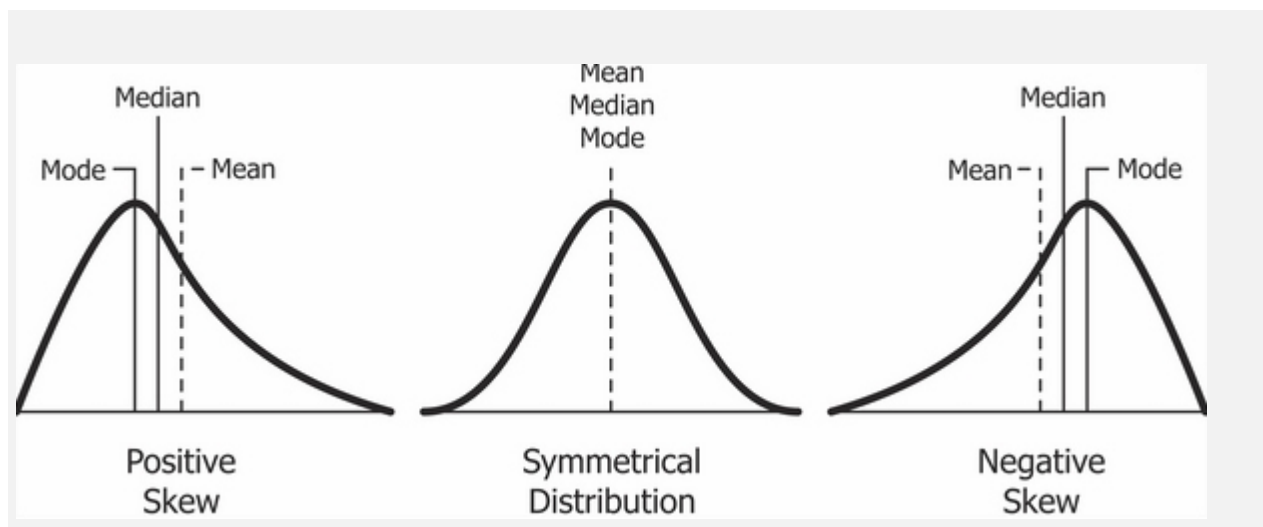
6. Skewness — The measure of asymmetry in a probability distribution is defined by Skewness. It can either be positive, negative or undefined.

- **Positive Skew** — This is the case when the tail on the right side of the curve is bigger than that on the left side. For these distributions, mean is greater than the mode.
- **Negative Skew** — This is the case when the tail on the left side of the curve is bigger than that on the right side. For these distributions, mean is smaller than the mode.

The most commonly used method of calculating Skewness is

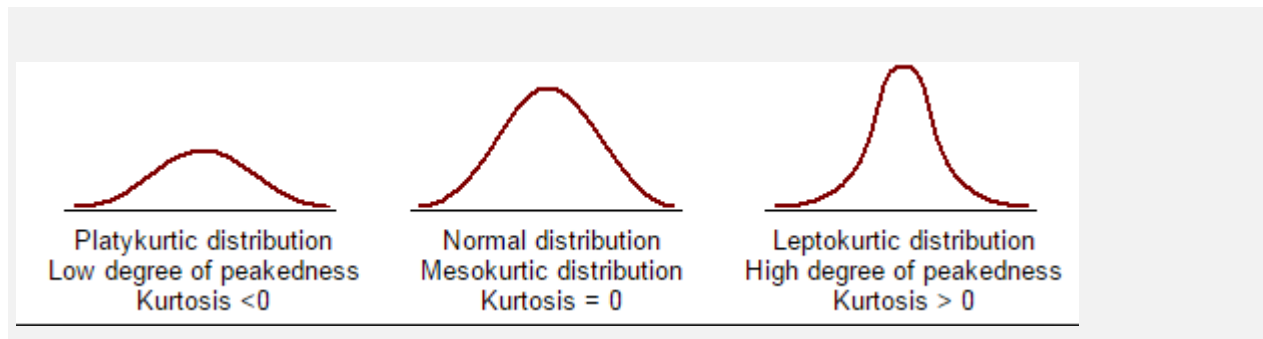
$$\text{Skewness} = \frac{3 (\text{Mean} - \text{Median})}{\text{Std Deviation}}$$

If the skewness is zero, the distribution is symmetrical. If it is negative, the distribution is Negatively Skewed and if it is positive, it is Positively Skewed.



7. Kurtosis — Kurtosis describes the whether the data is light tailed (lack of outliers) or heavy tailed (outliers present) when compared to a Normal distribution. There are three kinds of Kurtosis:

- Mesokurtic — This is the case when the kurtosis is zero, similar to the normal distributions.
- Leptokurtic — This is when the tail of the distribution is heavy (outlier present) and kurtosis is higher than that of the normal distribution.
- Platykurtic — This is when the tail of the distribution is light(no outlier) and kurtosis is lesser than that of the normal distribution.



Introduction:

Statistics is the building block for data science and it's important for a data scientist to have a hold on it. Learning and staying up to the mark is a tedious task and is something data scientists struggle with. This article is intended for people who are getting introduced to data science and need an overview. For others, this could be a refresher to the basics.

Outliers:

Outliers can be defined as values that fall outside normal range. For example in the series: 4,7,19,999,8,14, 999 will be the outlier. The analyst decides if value is an outlier or not, for a particular dataset.

Mode:

Mode can be defined as the most frequently occurring value in a distribution.

In the series 5,10,5,6,8,3,21 ,5 is the mode since it occurs the most number of times(twice).

All values are not important for a mode, since we only need to check for frequency of occurrence of numbers, which is also the reason why mode is robust to outliers. Let's say we add 900 to the series 5,10,5,6,8,3,21,900, the mode stays the same. Mode is generally used for categorical variables, similar to the example below.

Let's consider the distribution of 5 color balls: Red,Red,Green,Blue,Red

Here, Red is the mode, as it occurs most frequently,i.e 3 times.

A distribution can have 1 or more than one modes. A single mode distribution is unimodal, two mode distribution is bi-modal and distribution having many modes is a multi-modal distribution.

Mean:

Mean is the average of numbers in a distribution, or generally speaking:

$$\text{Mean} = (\text{Sum of terms})/(\text{number of terms})$$

Mean is sensitive to outliers, therefore it's not a very robust measure.

Example: Let us consider previous distribution

1. Without outlier: 5,10,5,6,8,3,21, mean = $58/7 = 8.29$
2. With Outlier: 5,10,6,8,3,21,900,5 mean = $958/8 = 119.75$.

As seen in the example above, adding outliers can drastically change the mean value. Mean is generally used for continuous variables.

Median:

Median can be defined as the absolute central value of a numeric distribution sorted in ascending order. The median for an odd length series is the middle most element and for even length series it's the mean of the middle two elements

Examples:

1. 3,5,5,6,8,10,21. Here the series length is odd and the middle element is 6, so 6 is the median.

2. 3,5,5,6,8,10,21,900. Here the series length is even and the middle elements are 6 and 8, so mean of 6 & 8, i.e 7 is the median.

We can also observe from the above examples that addition of outlier in the second example did not affect the mean. Thus, median can be used as a more robust alternative to mean

Median is also generally used for continuous variables.

Quantile and Quartile:

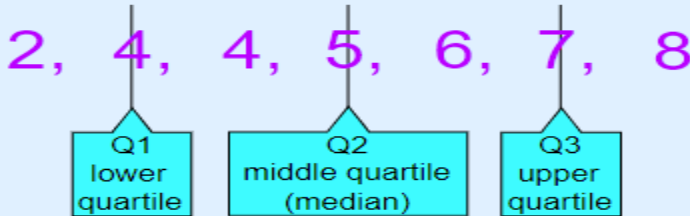
A Quantile is an arbitrary point of data, while quartiles are values dividing dataset into quarters While we will deal with quartiles mostly while working on a data problem, it is better to understand the difference between both and clear out the confusion.

Median divides the dataset into 2 parts. Median of the data on the left of the median is the 1st quartile, and that to the right of the mean is the 3rd quartile of the distribution. This can be clearly understood with the example below:

Example: 5, 7, 4, 4, 6, 2, 8

Put them in order: 2, 4, 4, 5, 6, 7, 8

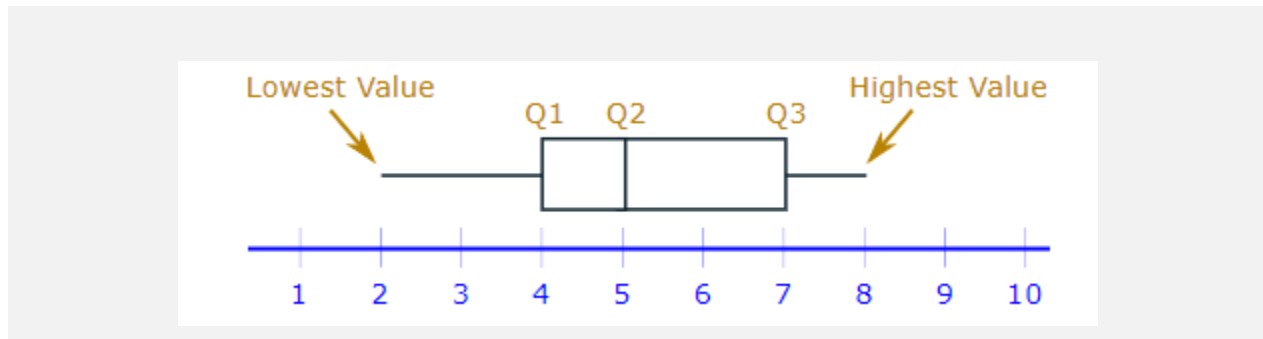
Cut the list into quarters:



And the result is:

- Quartile 1 (Q1) = 4
- Quartile 2 (Q2), which is also the Median, = 5
- Quartile 3 (Q3) = 7

The quartiles and other important values can be represented by a box plot as shown below:



<https://www.mathsisfun.com/data/quartiles.html>

Spread of Data:

We may need to check how similar or varied our set of observations are, while working on a data science project. There are 2 measures to calculate this:

1. Range: It is the difference between maximum and minimum values. It is directly proportional to the spread of data. Range is sensitive to outliers
2. Interquartile Range(IQR): It is the difference between 3rd quartile and 1st quartile. It is robust to outliers, since it takes into account the quartiles, which as we know are derived from medians, which are robust to outliers

Note that, we use a similar approach to calculating the quartiles as covered before.

Example:

1. Without outlier: 3,5,5,6,8,10,21

Quartile 1: 5

Quartile 2(Median): 6

Quartile 3: 10

Range: $21 - 3 = 18$

IQR: $10 - 5 = 5$

2. With Outlier: 3,5,5,6,8,10,21,900

Quartile 1: 5

Quartile 2(Median): $(6+8)/2 = 7$

Quartile 3: 21

Range: $900 - 3 = 897$

IQR: $21 - 5 = 16$

3. Variance and standard deviation can also be used to measure the spread of the data. We'll cover them later in this article.

Below is a table of contents showing measures and their sensitivity to outliers.

--

Measure	Sensitive to Outliers
Mean	Yes
Median	No
Mode	No
Range	Yes
Interquartile Range	No

Variance and Standard Deviation:

Let us have a look at the wikipedia definitions for both these terms

Variance: The expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of (random) numbers are spread out from their average value.

Standard Deviation: A measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range.

Below are the formulas:

$$\text{variance} = \sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$$

$$\text{standard deviation } \sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

μ = mean

<https://towardsdatascience.com/intro-to-descriptive-statistics-and-probability-for-data-science-8effec826488>

We use squares of deviation for variance ensure that deviation above and below the mean do not nullify each other, this can be understood by the little example below

Value	Mean	Deviation from mean(value - mean)	Square difference of deviation from mean
5	10	-5	25
10	10	0	0
15	10	5	25

Adding the absolute differences between deviations we get: $-5+0+5 = 0$.

Adding squared difference of deviation from mean we get: $25+0+25 = 50$.

The added benefit is that we penalize the outliers heavily. However, because of the squaring, variance is not in the same unit of measurement as the original data. This is the reason we generally use standard deviation, the square root of the variance for calculation purposes.