Statistical Features

In this lesson, various statistical features are discussed.

We'll cover the following

Introduction to statistical features

Features that provide numerical information about the given data are known as statistical features. They help to extensively explore the nature and properties of data. The following are some features that will be discussed here:

- **Mean/Median**
- **Standard deviation (STD)**
- **Quantiles**
- **Skewness**

The above properties of data provide information that helps in the examination, inference, and prediction. These properties can only be applied to quantitative parts of the data.

Mean/Median

- **Mean**: This is the average of the dataset computed by dividing the sum of numbers with their quantity.
- **Median**: This is the exact middle value of a dataset. The data needs to be sorted first to get this measure.

In statistics, the median value is preferred to be used over the mean value because sometimes the mean value can get affected by exceptionally small or large outliers which might bend the mean in the wrong direction. Therefore, the median value is considered as it provides a correct approximation of the middle value of the dataset.

Standard deviation (STD)

STD stands for standard deviation. This measure informs us how far the values of a dataset are dispersed from their mean value.

A low ***std*** value means that data points of the dataset are close to their mean value, and a high ***std*** value means that data points are widely spread and are far from the mean value. The square of **std** returns the variance of data.

Quantiles

Quantile is a statistical measure that divides the data into equal parts. **The main type of quantile is called quartile**, which divides data into **four** or less equal parts.

Three lines are dropped on data for this division. Each of these lines falls on specific values in the dataset which are explained below.

- The value that the first line hit is called the ***1st quartile*** and is denoted with **Q1**. This point of data indicates that **25%** of the data is below this point, and **75%** of the data is above this point. The data point that this line hits is the middle value between the smallest value of the dataset, and the median value of the dataset.

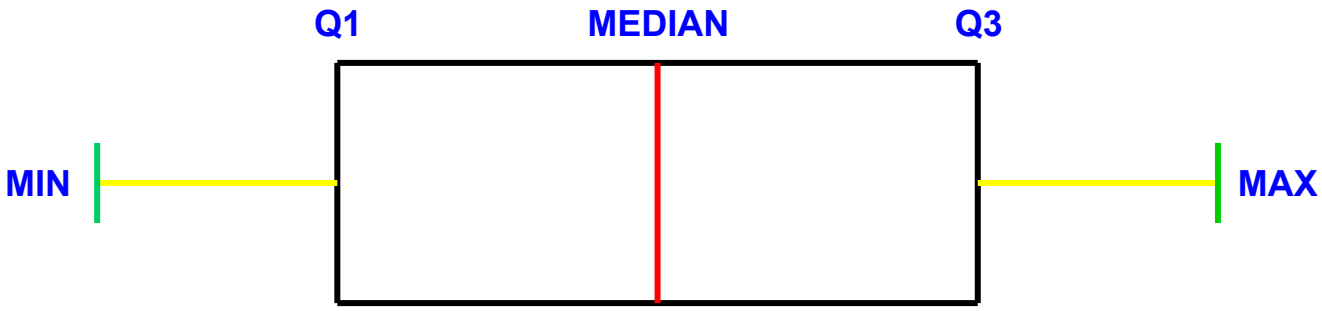
- The value that the second line hit is called the ***2nd quartile*** and is denoted with **Q2**. This point of data indicates that **50%** of the data is below this point, and **50%** of the data is above this point. The data point that this line hits is the median value of the dataset.

- The value that the third line hit is called the ***3rd quartile*** and is denoted with **Q3**. This point of data indicates that **75%** of the data is below this point, and **25%** of the data is above this point. The data point that this line hits is the middle value between the median value of the dataset and the largest value of

the dataset.

The following table summarizes this information:

Symbol	Names	Definition
Q1	First Quartile	Splits off the lowest 25% of data from the highest 75%
Q2	Second Quartile	Splits dataset in half
Q3	Third Quartile	Splits off the highest 25% of data from the lowest 75%



Five number summary

Five important values are obtained after applying this technique, including the **minimum**, **1st quartile**, **2nd quartile**, **3rd quartile**, and **maximum** values. These values help in understanding the growth of the data points among the dataset, like how far the values are from the mean or if there are any outliers or not. Another statistic, **IQR** can also be calculated to identify outliers as explained here.

Skewness

Skewness is another statistical measure, which informs us of the amount of asymmetry present in the data.

A perfectly symmetrical data graph has zero skewness, and the data is considered normally distributed. Skewness allows us to measure how much asymmetry is

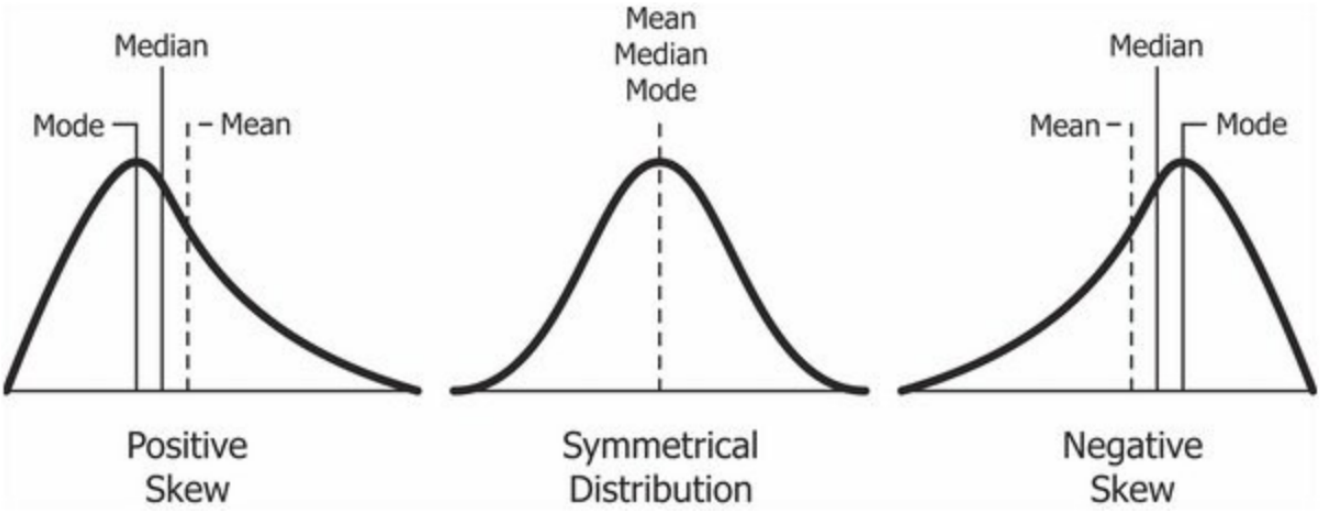
present in a dataset compared to a normally distributed dataset. A normally

distributed dataset is such whose values are distributed equally around its mean value, more on this is discussed in the next lesson.

The following two types of skewness can be found in a dataset:

1. **Negatively skewed**: In this, the mean value is less than the median because most of the data points are smaller and data distribution tends to favor the left side of the mean or the side below the mean value. The median value tends to be in the **3rd** quartile for this.
2. **Positively skewed**: In this, the mean value is greater than the median because most of the data points are greater and data distribution tends to favor the right side of the mean or the side below the mean value. The median value tends to be in the **1st** quartile for this.

The following figure might clear this difference.



Skewness comparison

As seen in the above figure, the negatively skewed graph has a tail that tends to be longer on the left side, and the positively skewed graph has a tail that tends to be longer on the right side correctly displaying the skewed distribution of data. The top point of each graph mostly represents *median* and *mode* values.

In the next lesson, the concepts of probability distributions are discussed.