

# **STATISTICS FOR DATA SCIENCE BEGINNERS**

## **INTRODUCTION:**

In this article, we will learn all the important statistical concepts which are required for Data Science roles.

## **TABLE OF CONTENTS:**

- Difference between Parameter and Statistic
- Statistics and its types
- Data Types and Measurement levels
- Moments of Business Decision
- Central Limit Theorem (CLT)
- Probability Distributions
- Normal Distribution
- Regression
- Bias
- Graphical representations
- Hypothesis Testing

## **1. DIFFERENCE BETWEEN PARAMETER AND STATISTIC**

In our day in day out, we keep speaking about the Population and sample. So, it is very important to know the terminology to represent the population and the sample.

A parameter is a number that describes the data from the population. And, a statistic is a number that describes the data from a sample.

## **2. STATISTICS AND ITS TYPES**

The Wikipedia definition of Statistics states that “it is a discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.”

It means, as part of statistical analysis, we collect, organize, and draw meaningful insights from the data either through visualizations or mathematical explanations.

Statistics is broadly categorized into two types:

1. Descriptive Statistics
2. Inferential Statistics

## DESCRIPTIVE STATISTICS:

As the name suggests in Descriptive statistics, we describe the data using the Mean, Standard deviation, Charts, or Probability distributions.

Basically, as part of descriptive Statistics, we measure the following:

1. Frequency: no. of times a data point occurs
2. Central tendency: the centrality of the data – mean, median, and mode
3. Dispersion: the spread of the data – range, variance, and standard deviation
4. The measure of position: percentiles and quantile ranks

## **INFERENCEAL STATISTICS:**

In Inferential statistics, we estimate the population parameters. Or we run Hypothesis testing to assess the assumptions made about the population parameters.

In simple terms, we interpret the meaning of the descriptive statistics by inferring them to the population.

For example, we are conducting a survey on the number of two-wheelers in a city. Assume the city has a total population of 5L people. So, we take a sample of 1000 people as it is impossible to run an analysis on entire population data.

From the survey conducted, it is found that 800 people out of 1000 (800 out of 1000 is 80%) are two-wheelers. So, we can infer these results to the population and conclude that 4L people out of the 5L population are two-wheelers.

### 3. DATA TYPES AND LEVEL OF MEASUREMENT:

At a higher level, data is categorized into two types: **Qualitative** and **Quantitative**.

Qualitative data is non-numerical. Some of the examples are eye colour, car brand, city, etc.

On the other hand, Quantitative data is numerical, and it is again divided into Continuous and Discrete data.

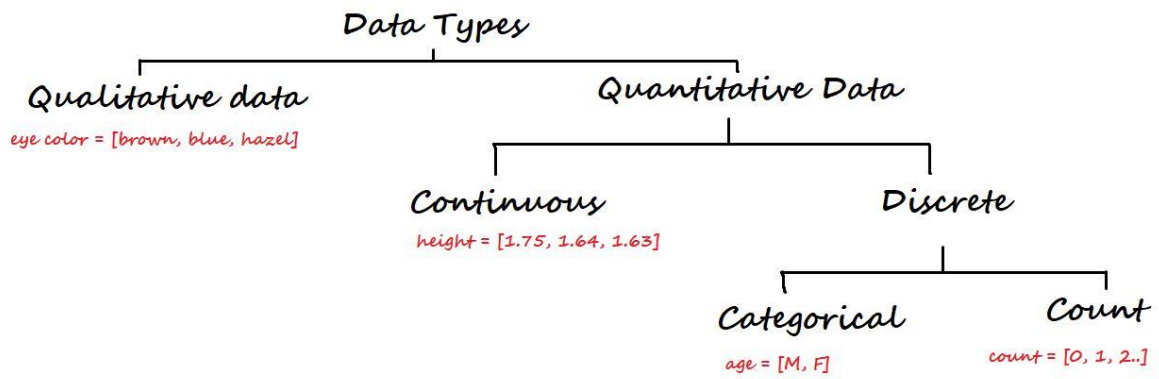
**CONTINUOUS DATA:** It can be represented in decimal format. Examples are height, weight, time, distance, etc.

**DISCRETE DATA:** It cannot be represented in decimal format. Examples are the number of laptops, number of students in a class.

Discrete data is again divided into Categorical and Count Data.

**CATEGORICAL DATA:** represent the type of data that can be divided into groups. Examples are age, sex, etc.

**COUNT DATA:** This data contains non-negative integers. Example: number of children a couple has.



Data Types

## LEVEL OF MEASUREMENT:

In statistics, the level of measurement is a classification that describes the relationship between the values of a variable.

We have four fundamental levels of measurement. They are:

1. Nominal Scale
2. Ordinal Scale
3. Interval Scale
4. Ratio Scale

**1. NOMINAL SCALE** : This scale contains the least information since the data have names/labels only. It can be used for classification. We cannot perform mathematical operations on nominal data because there is no numerical value to the options (numbers associated with the names can only be used as tags).

Example: Which country do you belong to? India, Japan, Korea.

**2. ORDINAL SCALE** : In comparison to the nominal scale, the ordinal scale has more information because along with the labels, it has order/direction.

Example: Income level – High income, medium income, low income.

**3. INTERVAL SCALE** : It is a numerical scale. The Interval scale has more information than the nominal, ordinal scales. Along with the order, we know the difference between the two variables (interval indicates the distance between two entities).

Mean, median, and mode can be used to describe the data.

Example: Temperature, income, etc.

**4. RATIO SCALE** : The ratio scale has the most information about the data. Unlike the other three scales, the ratio scale can accommodate a true zero point. The ratio scale is simply said to be the combination of Nominal, Ordinal, and Interval scales.

Example: Current weight, height, etc.

## 4. MOMENTS OF BUSINESS DECISION

We have four moments of business decision that help us understand the data.

### 4.1. MEASURES OF CENTRAL TENDENCY

(It is also known as First Moment Business Decision)

Talks about the centrality of the data. To keep it simple, it is a part of descriptive statistical analysis where a single value at the centre represents the entire dataset.

The central tendency of a dataset can be measured using:

**MEAN:** It is the sum of all the data points divided by the total number of values in the data set. Mean cannot always be relied upon because it is influenced by outliers.

**MEDIAN:** It is the middlemost value of a sorted/ordered dataset. If the size of the dataset is even, then the median is calculated by taking the average of the two middle values.

**MODE:** It is the most repeated value in the dataset. Data with a single mode is called unimodal, data with two modes is called bimodal, and data with more than two modes is called multimodal.



## 4.2. MEASURES OF DISPERSION

(It is also known as Second Moment Business Decision)

Talks about the spread of data from its centre.

### DISPERSION CAN BE MEASURED USING:

**VARIANCE**: It is the average squared distance of all the data points from their mean. The problem with Variance is, the units will also get squared.

**STANDARD DEVIATION**: It is the square root of Variance. Helps in retrieving the original units.

**RANGE**: It is the difference between the maximum and the minimum values of a dataset.

MEASURE	POPULATION	SAMPLE
Mean	$\mu = (\sum X_i)/N$	$\bar{x} = (\sum x_i)/n$
Median	The middle value of the data	The middle value of the data
Mode	Most occurred value	Most occurred value
Variance	$\sigma^2 = (\sum X_i - \mu)^2/N$	$s^2 = (\sum x_i - \bar{x})^2 / (n-1)$
Standard Deviation	$\sigma = \text{sqrt} ((\sum X_i - \mu)^2/N)$	$s = \text{sqrt} ((\sum x_i - \bar{x})^2 / (n-1))$
Range	Max-Min	Max-Min

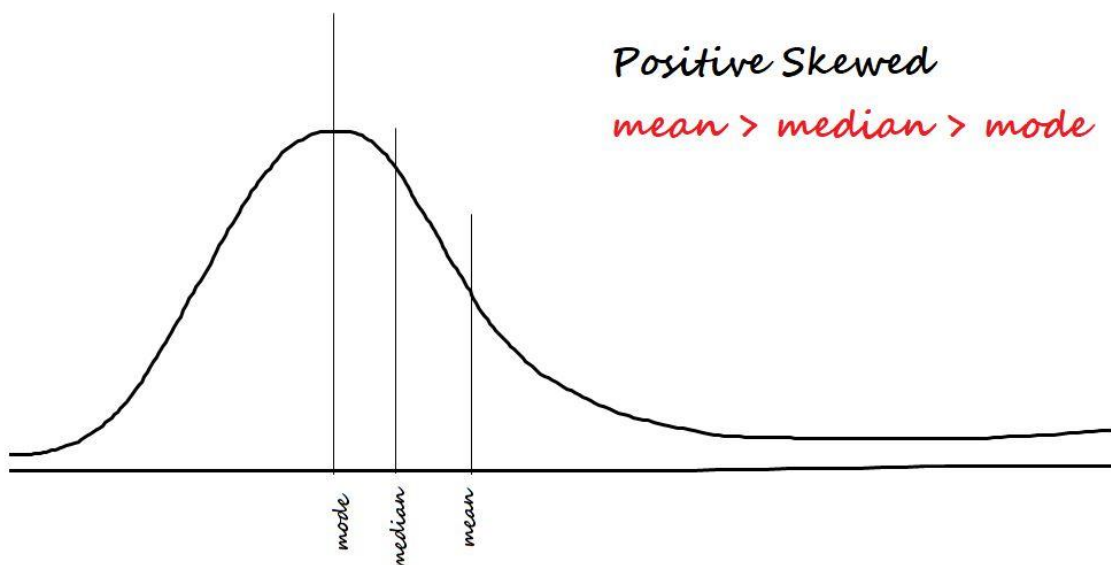
### 4.3. SKEWNESS

(It is also known as Third Moment Business Decision)

The formula of Skewness is  $E [(X - \mu) / \sigma ]^3 = Z^3$

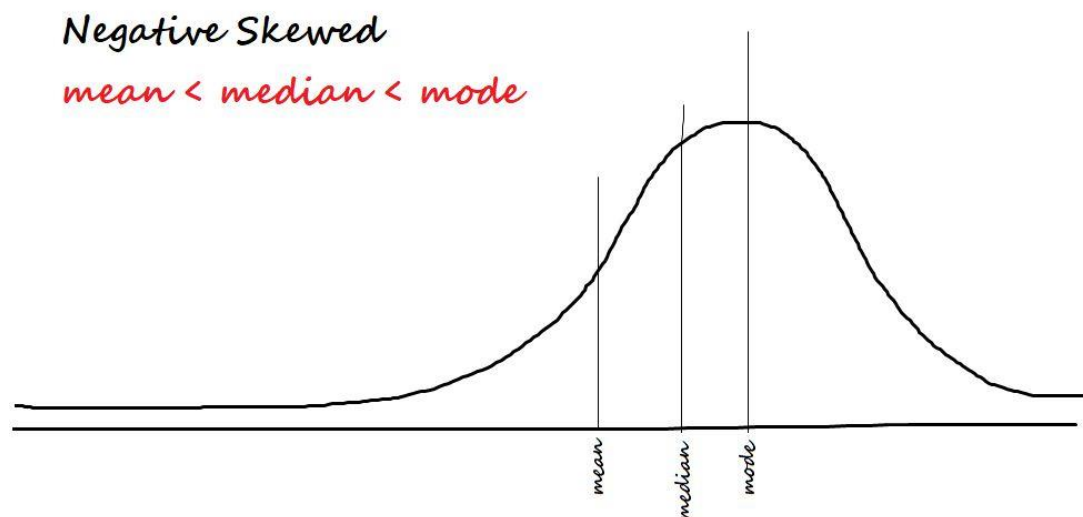
It measures the asymmetry in the data. The two types of Skewness are:

**POSITIVE/RIGHT-SKEWED**: Data is said to be positively skewed if most of the data is concentrated to the left side and has a tail towards the right.



Positively skewed data

**NEGATIVE/LEFT-SKEWED**: Data is said to be negatively skewed if most of the data is concentrated to the right side and has a tail towards the left.



Negatively Skewed data

## 4.4. KURTOSIS

(It is also known as Fourth Moment Business Decision)

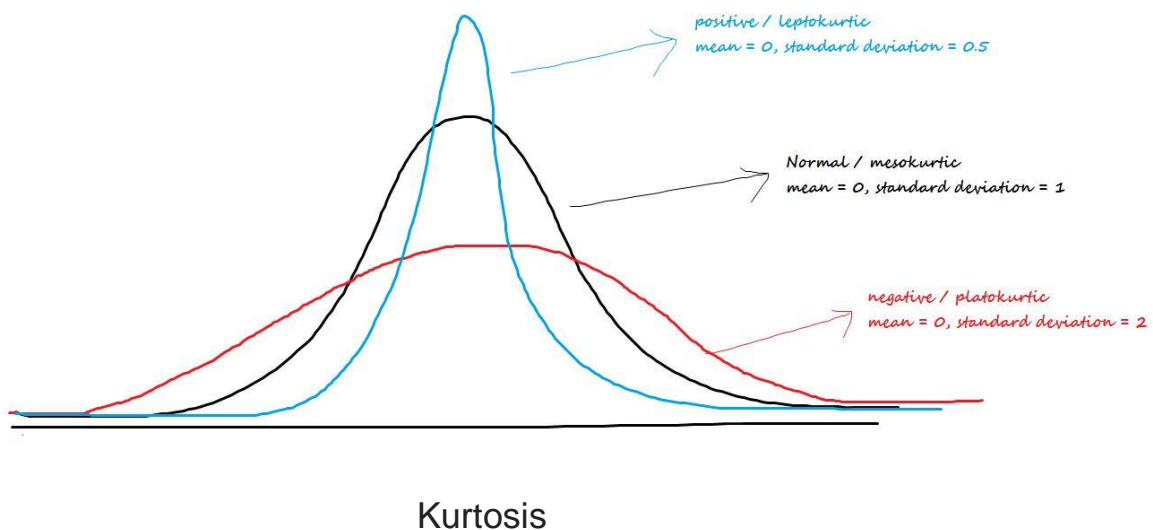
Talks about the central peakedness or fatness of tails. The three types of Kurtosis are:

**Positive/leptokurtic:** Has sharp peaks and lighter tails

**Negative/Platokurtic:** Has wide peaks and thicker tails

**MesoKurtic:** Normal distribution

The formula of Kurtosis is  $E [(X - \mu) / \sigma]^4 - 3 = Z^4 - 3$



**NOTE:** Together, Skewness and Kurtosis are called Shape Statistics.

## 5. CENTRAL LIMIT THEOREM (CLT)

Instead of analysing entire population data, we always take out a sample for analysis. The problem with sampling is that “sample means is a random variable – varies for different samples”. And random sample we draw can never be an exact representation of the population. This phenomenon is called sample variation.

To nullify the sample variation, we use the central limit theorem. And according to the Central Limit Theorem:

1. The distribution of sample means follows a normal distribution if the population is normal.
2. the distribution of sample means follows a normal distribution even though the population is not normal. But the sample size should be large enough.
3. The grand average of all the sample mean values give us the population mean.
4. Theoretically, the sample size should be 30. And practically, the condition on the sample size (n) is:

$n > 10(k_3)^2$ , where  $k_3$  is the sample skewness.

$n > 10(k_4)$ , where  $K_4$  is the sample Kurtosis.

## 6. PROBABILITY DISTRIBUTIONS

In statistical terms, a distribution function is a mathematical expression that describes the probability of different possible outcomes for an experiment. In simple terms, an event refers to the result of an experiment like tossing a coin. Events are of two types dependent and independent.

**INDEPENDENT EVENT:** The event is said to be an independent event when it is not affected by the earlier events. For example, tossing a coin, let us consider a coin is tossed the first outcome is head when the coin is tossed again the outcome may be head or tail. But this is entirely independent of the first trial.

**DEPENDENT EVENT:** The event is said to be dependent when the occurrence of the event is dependent on the earlier events. For example, when a ball is drawn from a bag that contains red and blue balls. If the first ball drawn is red, then the second ball may be red or blue; this depends on the first trial.

The probability of independent events is calculated by simply multiplying the probability of each event and for a dependent event is calculated by conditional probability.

## 7. NORMAL / GAUSSIAN DISTRIBUTION

Normal is used to define the probability density function for a continuous random variable in a system. The standard normal distribution has two parameters – mean and standard deviation that are discussed above. When the distribution of random variables is unknown, the normal distribution is used. The central limit theorem justifies why normal distribution is used in such cases.

## 8. REGRESSION

It is a method that is used to determine the relationship between one or more independent variables and a dependent variable. Regression is mainly of two types:

LINEAR REGRESSION: It is used to fit the regression model that explains the relationship between a numeric predictor variable and one or more predictor variables.

LOGISTIC REGRESSION: It is used to fit a regression model that explains the relationship between the binary response variable and one or more predictor variables.

## 9. BIAS

In statistical terms, it means when a model is representative of a complete population. This needs to be minimized to get the desired outcome.

The three most common types of bias are:

**SELECTION BIAS**: It is a phenomenon of selecting a group of data for statistical analysis, the selection in such a way that data is not randomized resulting in the data being unrepresentative of the whole population.

**CONFIRMATION BIAS**: It occurs when the person performing the statistical analysis has some predefined assumption.

**TIME INTERVAL BIAS**: It is caused intentionally by specifying a certain time range to favour a particular outcome.



## **10. GRAPHICAL REPRESENTATIONS**

Graphical representation refers to the use of charts or graphs to visualize, analyse and interpret numerical data.

For a single variable (Univariate analysis), we have a bar plot, line plot, frequency plot, dot plot, boxplot, and the Normal Q-Q plot.

We will be discussing the Boxplot and the Normal Q-Q plot.

## 10.1. BOXPLOT

A boxplot is a way of visualizing the distribution of data based on a five-number summary. It is used to identify the outliers in the data.

The five numbers are minimum, first Quartile (Q1), median (Q2), third Quartile (Q3), and maximum.

The box region will contain 50% of the data. The lower 25% of the data region is called the Lower whisker and the upper 25% of the data region is called the Upper Whisker.

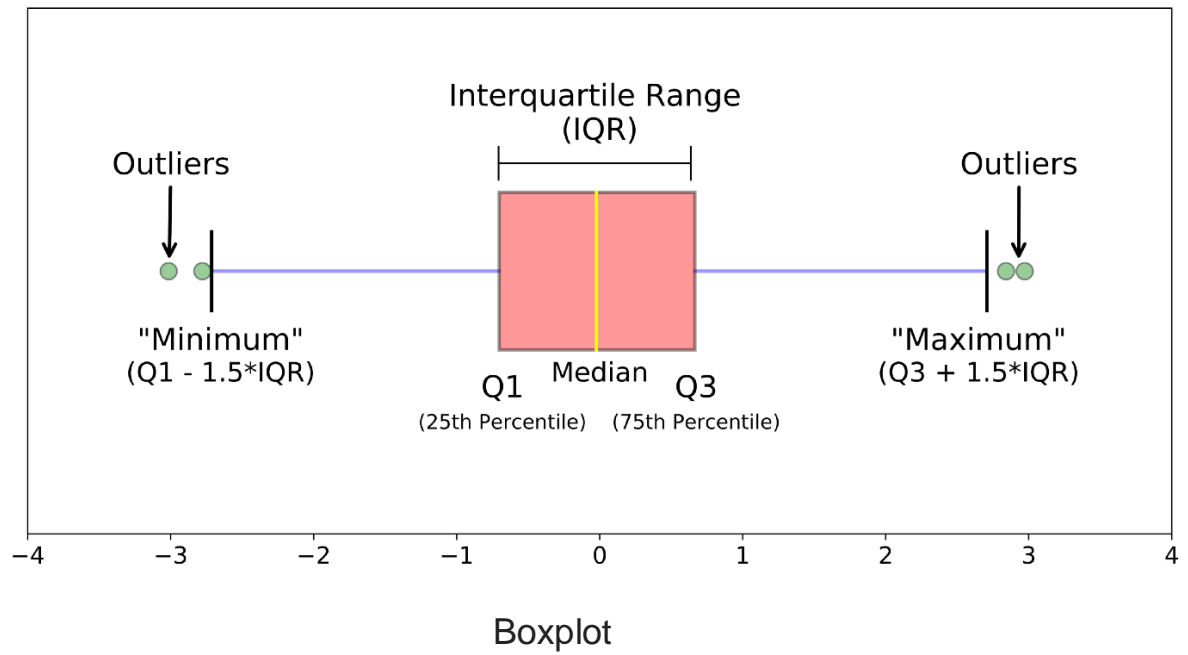
The Interquartile region (IQR) is the difference between the third and first quartiles.  $IQR = Q3 - Q1$ .

Outliers are the data points that lie below the lower whisker and beyond the upper whisker.

The formula to find the outliers is  **$Outlier = Q \pm 1.5 * (IQR)$**

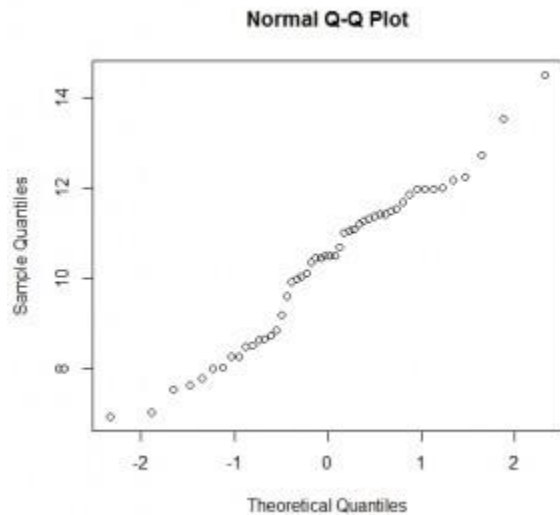
The outliers that lie below the lower whisker are given as  **$Q1 - 1.5 * (IQR)$**

The outliers that lie beyond the upper whisker are given as  **$Q3 + 1.5 * (IQR)$**



## 10.2. NORMAL Q-Q PLOT

A Normal Q-Q plot is a kind of scatter plot that is plotted by creating two sets of quantiles. It is used to check if the data is following normality or not.



Normal Q-Q plot

On the x-axis, we have the Z-scores and on the y-axis, we have the actual sample quantiles. If the scatter plot forms a straight line, data is said to be normal.

## 11. HYPOTHESIS TESTING

Hypothesis testing in statistics is a way to test the assumptions made on the population parameters.

End Notes:

Thank you for reading till the conclusion. By the end of this article, we are familiar with the important statistical concepts.

I hope this article is informative. Feel free to share it with your study buddies.