# Introduction to Statistics

Statistics is a type of mathematical analysis that employs quantified models and representations to analyse a set of experimental data or real-world studies. The main benefit of statistics is that information is presented in an easy-to-understand format. Data processing is the most important aspect of any Data Science plan. When we speak about gaining insights from data, we're basically talking about exploring the chances. In Data Science, these possibilities are referred to as Statistical Analysis.

## Importance of Statistics

**1)** Using various statistical tests, determine the relevance of features.

**2)** To avoid the risk of duplicate features, find the relationship between features.

**3)** Putting the features into the proper format.

**4)** Data normalization and scaling This step also entails determining the distribution of data as well as the nature of data.

**5)** Taking the data for further processing and making the necessary modifications.

## 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

**Inferential 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.

**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.

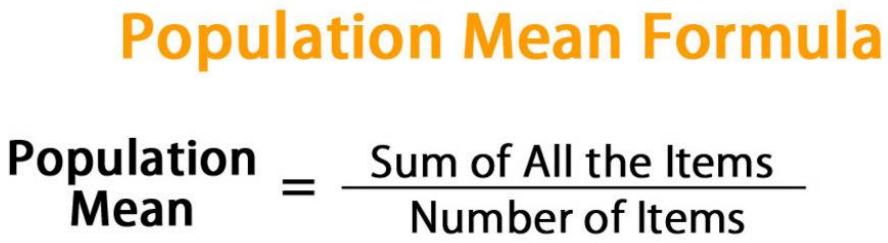
Example: Temperature, income, etc.

**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 Intercal scales.

Example: Current weight, height, etc.

## Central Tendency in Statistics

**1) Mean**: The mean (or average) is that the most generally used and well-known measure of central tendency. It will be used with both discrete and continuous data, though it's most typically used with continuous data (see our styles of Variable guide for data types). The mean is adequate the sum of all the values within the data set divided by the number of values within the data set. So, if we have n values in a data set and they have values x1,x2, …,xn, the sample mean, usually denoted by **"x bar"**, is:




**2) Median**: The median value of a dataset is the value in the middle of the dataset when it is arranged in ascending or descending order. When the dataset has an even number of values, the median value can be calculated by taking the mean of the middle two values.

The following image gives an example for finding the median for odd and even numbers of samples in the dataset.

1, 3, 3, **6**, 7, 8, 9

Median = **6**

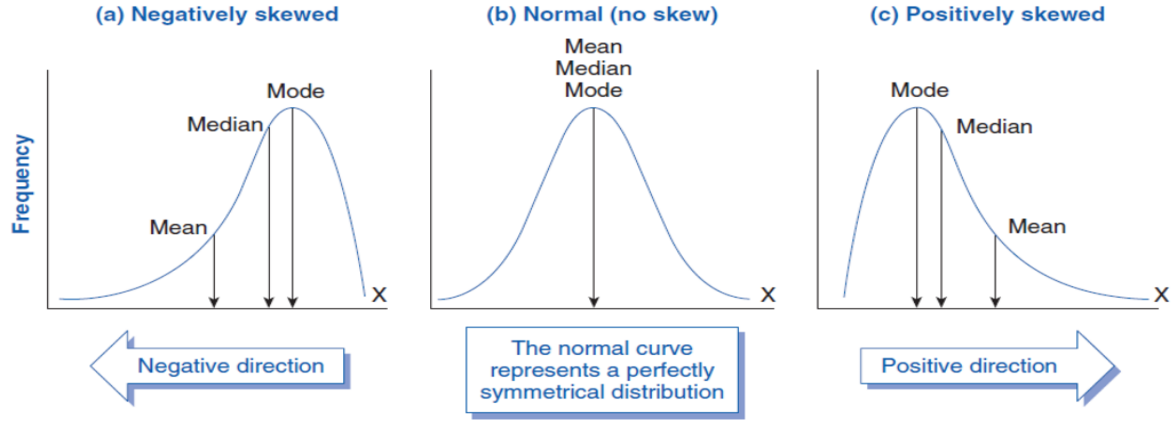
1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= **4.5**

**3) Mode**: The mode is the value that appears the most frequently in your data set. The mode is the highest bar in a bar chart. A multimodal distribution exists when the data contains multiple values that are tied for the most frequently occurring. If no value repeats, the data does not have a mode.

**4) Skewness**: Skewness is a metric for symmetry, or more specifically, the lack of it. If a distribution, or data collection, looks the same to the left and ight of the centre point, it is said to be symmetric.




**Measures of Dispersion**

# Variability in Statistics

**Range**: In statistics, the range is the smallest of all dispersion measures. It is the difference between the distribution's two extreme conclusions. In other words, the range is the difference between the distribution's maximum and minimum observations.

**Range = Xmax – Xmin**

Where Xmax represents the largest observation and Xmin represents the smallest observation of the variable values.

## *Percentiles, Quartiles and Interquartile Range (IQR)*

· **Percentiles** — It is a statistician's unit of measurement that indicates the value below which a given percentage of observations in a group of observations fall.

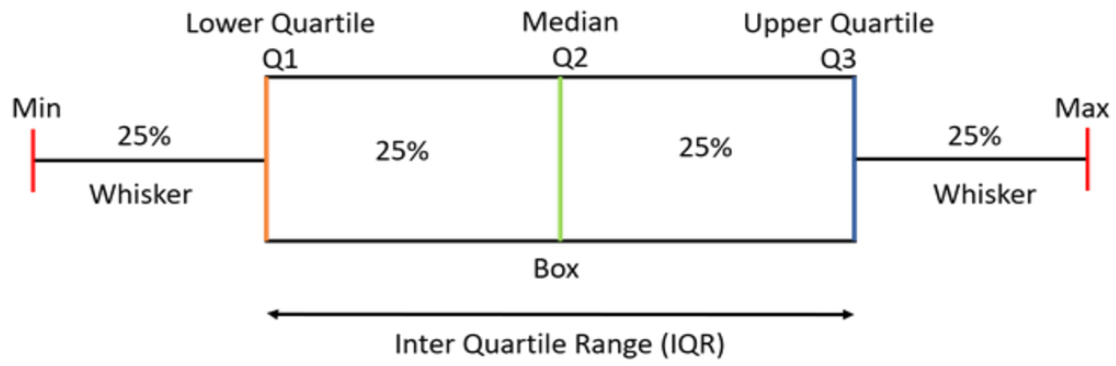
**For instance, the value QX represents the 40th percentile of XX (0.40)**

· **Quantiles**— Values that divide the number of data points into four more or less equal parts, or quarters. Quantiles are the 0th, 25th, 50th, 75th, and 100th percentile values or the 0th, 25th, 50th, 75th, and 100th percentile values.

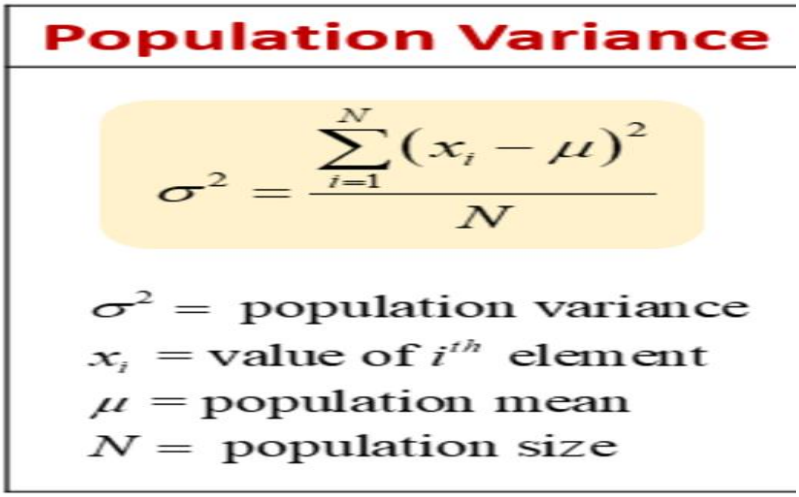
**Interquartile Range (IQR)**— The difference between the third and first quartiles is defined by the interquartile range. The partitioned values that divide the entire series into four equal parts are known as quartiles. So, there are three quartiles. The first quartile, known as the lower quartile, is denoted by Q1, the second quartile by Q2, and the third quartile by Q3, known as the upper quartile. As a result, the interquartile range equals the upper quartile minus the lower quartile.

**IQR = Upper Quartile – Lower Quartile**

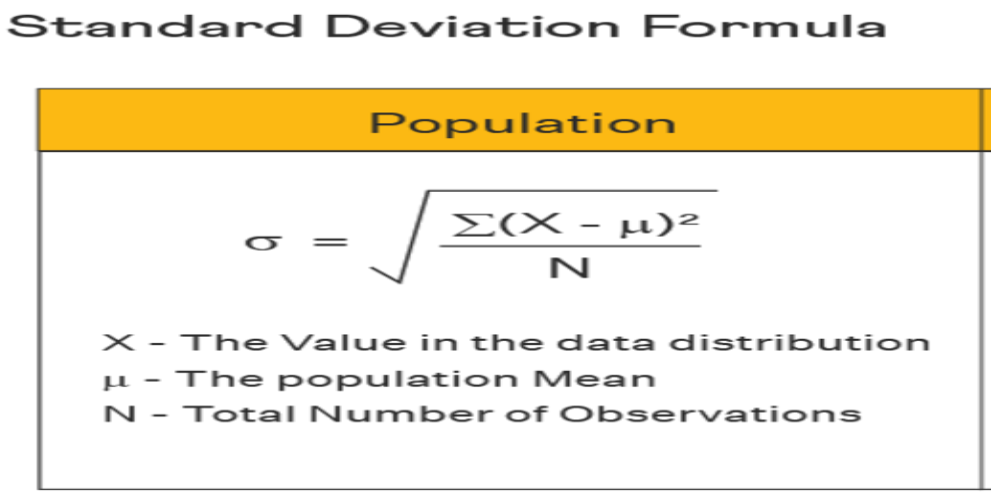
**= Q3 − Q1**




· **Variance**: The dispersion of a data collection is measured by variance. It is defined technically as the average of squared deviations from the mean.




· **Standard Deviation**: The standard deviation is a measure of data dispersion WITHIN a single sample selected from the study population. The square root of the variance is used to compute it. It simply indicates how distant the individual values in a sample are from the mean. To put it another way, how dispersed is the data from the sample? As a result, it is a sample statistic.




## Relationship Between Variables

· **Causality**: The term “causation” refers to a relationship between two events in which one is influenced by the other. There is causality in statistics when the value of one event, or variable, grows or decreases as a result of other events.

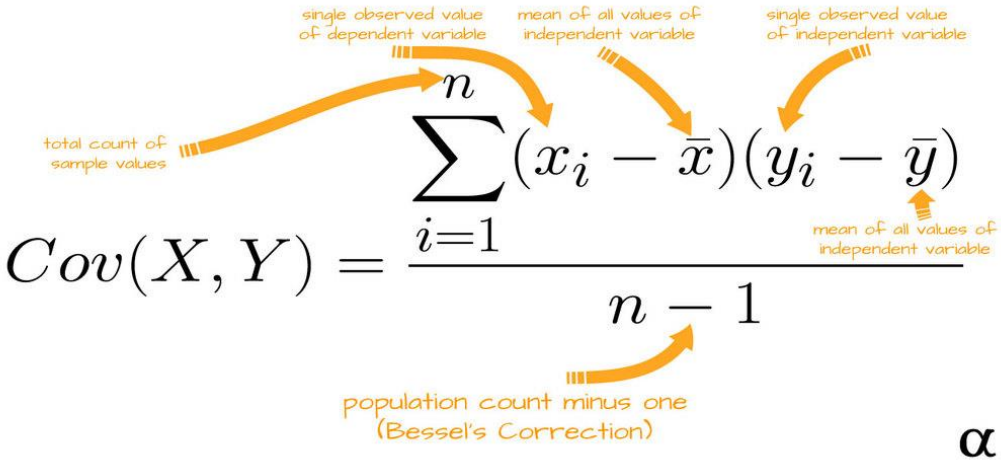
Each of the events we just observed may be thought of as a variable, and as the number of hours worked grows, so does the amount of money earned. On the other hand, if you work fewer hours, you will earn less money.

· **Covariance**: Covariance is a measure of the relationship between two random variables in mathematics and statistics. The statistic assesses how much – and how far – the variables change in tandem. To put it another way, it’s a measure of the variance between two variables. The metric, on the other hand, does not consider the interdependence of factors. Any positive or negative value can be used for the variance.

## *The following is how the values are interpreted:*

· Positive covariance: When two variables move in the same direction, this is called positive covariance.

· Negative covariance indicates that two variables are moving in opposite directions.




**Correlation**: Correlation is a statistical method for determining whether or not two quantitative or categorical variables are related. To put it another way, it's a measure of how things are connected. Correlation analysis is the study of how variables are connected.

## *Ø Here are a few examples of data with a high correlation:*

1) Your calorie consumption and weight.

2) The amount of time you spend studying and your grade point average

*Ø Here are some examples of data with poor (or no) correlation:*

1) **The expense of vehicle washes and the time it takes to get a Coke at the station.**
2) **The crime rate and house price**

Correlations are useful because they allow you to forecast future behaviour by determining what relationship variables exist. In the social sciences, such as government and healthcare, knowing what the future holds is critical. Budgets and company plans are also based on these facts.

$$Correlation = \frac{Cov\,(x,y)}{\sigma x * \sigma y}$$

## Probability

In a Random Experiment, the probability is a measure of the likelihood that an event will occur. The number of favorable outcomes in an experiment with n outcomes is denoted by x. The following is the formula for calculating the probability of an event.

Probability (Event) = Favourable Outcomes/Total Outcomes = x/n

Let's look at a simple application to better understand probability. If we need to know if it's raining or not. There are two possible answers to this question: "Yes" or "No." It is possible that it will rain or not rain. In this case,

we can make use of probability. The concept of probability is used to forecast the outcomes of coin tosses, dice rolls, and card draws from a deck of playing cards.

## Probability Distributions

### *Probability Distribution Functions*

**1) Probability Mass Function (PMF)**: The probability distribution of a discrete random variable is described by the PMF, which is a statistical term.

**2) Probability Density Function (PDF)**: The probability distribution of a continuous random variable is described by the word PDF, which is a statistical term.

**3) Cumulative Density Function (CDF)**: The cumulative distribution function can be used to describe the continuous or discrete distribution of random variables.

### *Continuous Probability Distribution*

**1) Uniform Distribution**: Uniform distribution is a sort of probability distribution in statistics in which all events are equally likely. Because the chances of drawing a heart, a club, a diamond, or a spade are equal, a deck of cards contains uniform distributions.

3) **Normal/Gaussian Distribution**: The normal distribution, also known as the Gaussian distribution, is a symmetric probability distribution

centred on the mean, indicating that data around the mean occur more frequently than data far from it. The normal distribution will show as a bell curve on a graph.

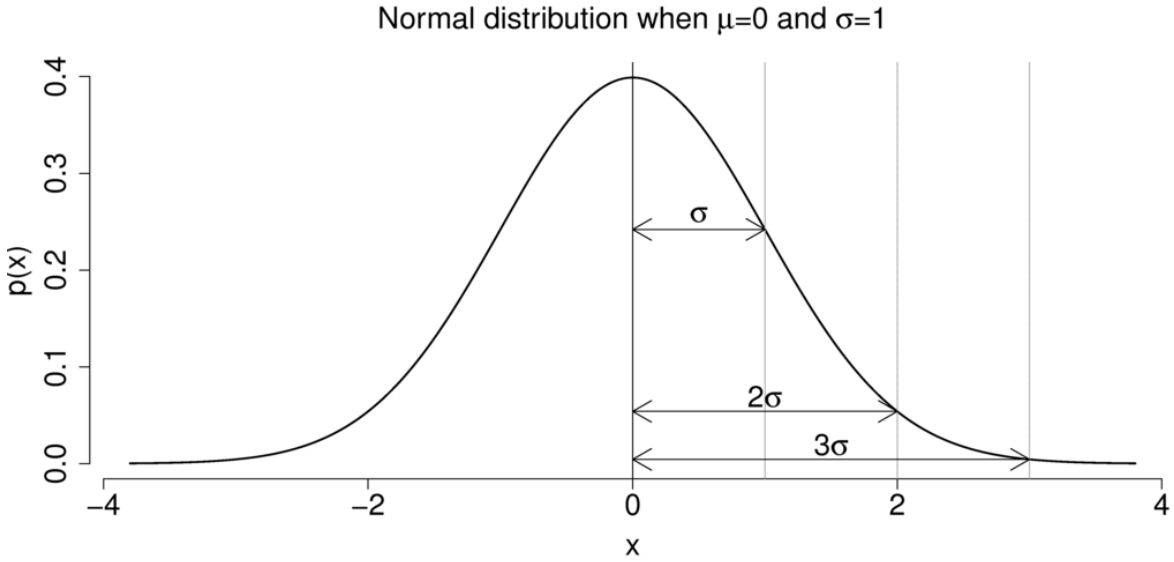
## *Points to remember: –*

· A probability bell curve is referred to as a normal distribution.

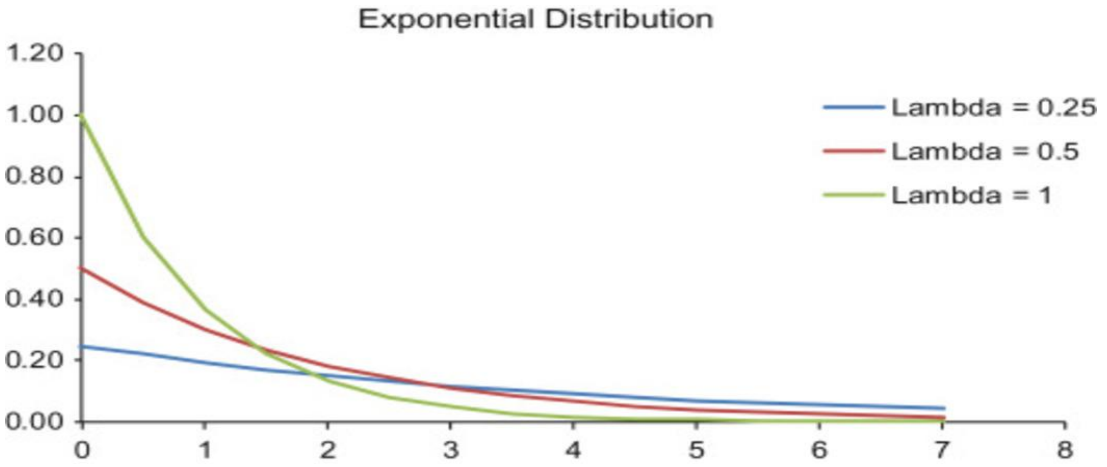
· The mean of a normal distribution is 0 and the standard deviation is 1. It has a kurtosis of 3 and zero skew.

· Although all symmetrical distributions are normal, not all normal distributions are symmetrical.

· Most pricing distributions aren't totally typical.




**3) Exponential Distribution**: The exponential distribution is a continuous distribution used to estimate the time it will take for an event to occur. For example, in physics, it is frequently used to calculate radioactive decay, in engineering, it is frequently used to calculate the time required to receive a defective part on an assembly line, and in finance, it is frequently used to calculate the likelihood of a portfolio of financial assets defaulting. It can also be used to estimate the likelihood of a certain number of defaults occurring within a certain time frame.




4) **Chi-Square Distribution**: A continuous distribution with degrees of freedom is called a chi-square distribution. It's used to describe a sum of squared random variable's distribution. It's also used to determine whether a data distribution's goodness of fit is good, whether data series are independent, and to estimate confidence intervals around variance and standard deviation for a random variable from a normal distribution. Furthermore, the chi-square distribution is a subset of the gamma distribution.

## *Discrete Probability Distribution*

**1) Bernoulli Distribution**: A Bernoulli distribution is a discrete probability distribution for a Bernoulli trial, which is a random experiment with just two outcomes (named "Success" or "Failure" in most cases). When flipping a coin, the likelihood of getting ahead (a "success") is 0.5. "Failure" has a chance of 1 – P. (where p is the probability of success, which also equals 0.5 for a coin toss). For n = 1, it is a particular case of the binomial distribution. In other words, it's a single-trial binomial distribution (e.g. a single coin toss).

## BERNOULLI DISTRIBUTION

- A Bernoulli trial is an experiment with only two outcomes. An r.v. $X$ has Bernoulli(p) distribution if

$$X = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1-p \end{cases}; 0 \le p \le 1$$

$$P(X = x) = p^x (1-p)^{1-x} \text{ for } x = 0,1; \text{ and } 0 < p < 1$$

11

**2) Binomial Distribution**: A discrete distribution is a binomial distribution. It's a well-known probability distribution. The model is then used to depict a variety of discrete phenomena seen in business, social science, natural science, and medical research.

Because of its relationship with a binomial distribution, the binomial distribution is commonly employed. For binomial distribution to be used, the following conditions must be met:

1. There are n identical trials in the experiment, with n being a limited number.

2. Each trial has only two possible outcomes, i.e., each trial is a Bernoulli's trial.

3. One outcome is denoted by the letter S (for success) and the other by the letter F (for failure) (for failure).

4. From trial to trial, the chance of S remains the same. The chance of success is represented by p, and the likelihood of failure is represented by q (where p+q=1).

5. Each trial is conducted independently.

6. The number of successful trials in n trials is the binomial random variable . **3) Poisson Distribution**: A Poisson distribution is a probability distribution used in statistics to show how many times an event is expected to happen over a certain amount of time. To put it another way, it's a count distribution. Poisson distributions are frequently accustomed comprehend independent events that occur at a gradual rate during a selected timeframe.

The Poisson distribution is a discrete function, which means the variable can only take values from a (possibly endless) list of possibilities. To put it another way, the variable can't take all of the possible values in any continuous range. The variable can only take the values 0, 1, 2, 3, etc., with no fractions or decimals, in the Poisson distribution (a discrete distribution).

## Hypothesis Testing and Statistical Significance in Statistics

Hypothesis testing may be a method within which an analyst verifies a hypothesis a couple of population parameters. The analyst's approach is set by the kind of the info and also the purpose of the study. the utilization of sample data to assess the plausibility of a hypothesis is thought of as hypothesis testing.

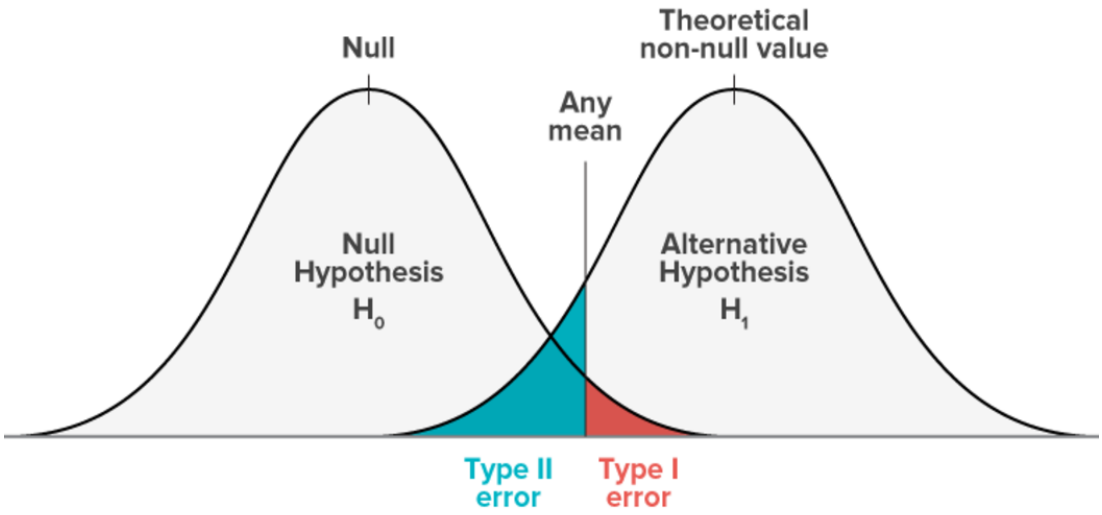
# *Null and Alternative Hypothesis*

## *Null Hypothesis (H0)*

A population parameter (such as the mean, standard deviation, and so on) is equal to a hypothesised value, according to the null hypothesis. The null hypothesis is a claim that is frequently made based on previous research or specialised expertise.

## *Alternative hypothesis (H1)*

The alternative hypothesis says that a population parameter is less, more, or different than the null hypothesis's hypothesised value. The alternative hypothesis is what you believe or want to prove to be correct.



# *Type 1 and Type 2 error*

## *Type 1 error:*

A type 1 error, often referred to as a false positive, happens when a researcher rejects a real null hypothesis incorrectly. this suggests you're claiming your findings are noteworthy after they actually happened by coincidence.

Your alpha level (), which is that the p-value below which you reject the null hypothesis, represents the likelihood of constructing a sort I error. When rejecting the null hypothesis, a p-value of 0.05 suggests that you simply are willing to tolerate a 5% probability of being mistaken.

By setting p to a lesser value, you'll lessen your chances of constructing a kind I error.

## *Type 2 error:*

A type II error commonly said as a false negative happens when a researcher fails to reject a null hypothesis that's actually true. during this case, a researcher finds that there's no significant influence when, in fact, there is.

Beta () is that the probability of creating a sort II error, and it's proportional to the statistical test's power (power = 1- ). By ensuring that your test has enough power, you'll reduce your chances of constructing a sort II error.

This can be accomplished by ensuring that your sample size is sufficient to spot a practical difference when one exists.

| | Reject $H_0$ | Fail to Reject $H_0$ |
|---|---|---|
| $H_0$ Is True | Type I Error<br>$\alpha$<br>(FP) | Correct<br>$1 - \alpha$<br>(TN) |
| $H_0$ Is False | Correct<br>$1 - \beta$ ("Statistic Power")<br>(TP) | Type II Error<br>$\beta$<br>(FN) |

## *Interpretation*

**P-value**: The p-value in statistics is that the likelihood of getting outcomes a minimum of as extreme because the observed results of a statistical hypothesis test, given the null hypothesis is valid. The p-value, instead of rejection points, is employed to work out the smallest amount level of significance at which the null hypothesis is rejected. A lower p-value indicates that the choice hypothesis has more evidence supporting it.

**Critical Value**: it is a point on the test distribution that is compared to the test statistic to see if the null hypothesis should be rejected. You can declare statistical significance and reject the null hypothesis if the absolute value of your test statistic is larger than the crucial value.

**Significance Level and Rejection Region**: The probability that an event (such as a statistical test) occurred by chance is the significance level of the occurrence. We call an occurrence significant if the level is very low, i.e., the possibility of it happening by chance is very minimal. The rejection region depends on the importance level. the importance level is denoted by α and is that the probability of rejecting the null hypothesis if it's true.

**Z-Test**: The z-test may be a hypothesis test within which the z-statistic is distributed normally. The z-test is best utilized for samples with quite 30 because, in line with the central limit theorem, samples with over 30 samples are assumed to be approximately regularly distributed.

The null and alternative hypotheses, also because the alpha and z-score, should all be reported when doing a z-test. The test statistic should next be calculated, followed by the results and conclusion. A z-statistic, also called a z-score, could be a number that indicates what number of standard deviations a score produced from a z-test is above or below the mean population.

**T-Test**: A t-test is an inferential statistic that's won't see if there's a major difference within the means of two groups that are related in how. It's most ordinarily employed when data sets, like those obtained by flipping a coin 100 times, are expected to follow a traditional distribution and have unknown variances. A t-test could be a hypothesis-testing technique that will be accustomed to assess an assumption that's applicable to a population.

**ANOVA (Analysis of Variance):** ANOVA is the way to find out if experimental results are significant. **One-way ANOVA** compares two means from two independent groups using only one independent variable. **Two-way ANOVA** is the extension of one-way ANOVA using two independent variables to calculate the main effect and interaction effect.

**Chi-Square Test**: it is a test that assesses how well a model matches actual data. A chi-square statistic requires data that is random, raw, mutually exclusive, collected from independent variables, and drawn from a

large enough sample. The outcomes of a fair coin flip, for example, meet these conditions.

In hypothesis testing, chi-square tests are frequently utilised. Given the size of the sample and the number of variables in the relationship, the chi-square statistic examines the size of any disparities between the expected and actual results.

$$X^2 = \sum \frac{(\text{Observed value} - \text{Expected value})^2}{\text{Expected value}}$$

## Central Limit Theorem (CLT)

Instead of analyzing 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.

**End Notes:**

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