In statistics, variables are classified into two main types: quantitative variables and qualitative variables.

1. **Quantitative Variables (Numeric Variables**): Quantitative variables are numerical measurements or quantities that represent quantities or amounts. They can be further categorized as either discrete or continuous variables.
   * **Discrete variables**: Discrete variables have a countable number of possible values and are typically whole numbers. Examples include the number of siblings a person has, the number of cars sold in a month, or the number of goals scored in a soccer match.
   * **+-Continuous variables:** Continuous variables can take on any value within a specific range or interval. They are typically measured on a continuous scale and can include decimal or fractional values. Examples include height, weight, temperature, or time taken to complete a task.

Quantitative variables allow for mathematical operations such as addition, subtraction, multiplication, and division. Summary statistics like mean, median, standard deviation, and correlation are often used to analyze and describe quantitative data.

1. **Qualitative Variables (Categorical Variables):** Qualitative variables, also known as categorical variables or attributes, describe characteristics or qualities of individuals or objects being studied. They do not have a numerical value and can be further classified as nominal or ordinal variables.
   * **Nominal variables:** Nominal variables represent categories or groups with no inherent order or ranking. **Examples** include gender (male, female), marital status (single, married, divorced), or eye color (blue, green, brown).
   * **Ordinal variables**: Ordinal variables have categories that can be ordered or ranked in a meaningful way. However, the intervals between the categories may not be equal. **Examples** include education level (high school, bachelor's, master's, Ph.D.), survey ratings (strongly agree, agree, neutral, disagree, strongly disagree), or customer satisfaction levels (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied).

**Qualitative variables are typically analysed using methods such as frequency distributions, bar charts, pie charts, or contingency tables. Measures like mode and percentage are commonly used to summarize and interpret qualitative data**.

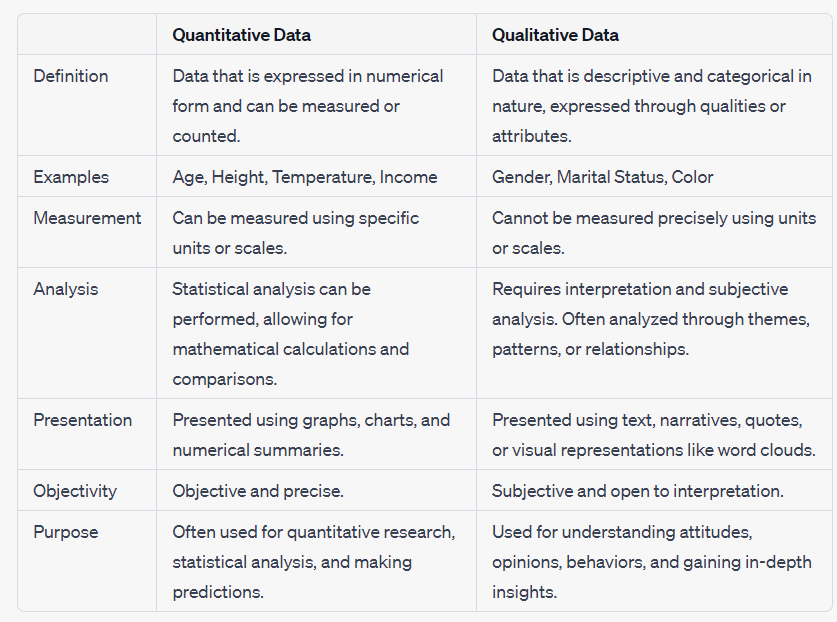
Both quantitative and qualitative variables have their respective roles in statistical analysis. Quantitative variables provide numerical measurements that allow for precise calculations, while qualitative variables provide information about categories and characteristics that help understand the composition or distribution of a population or sample. The choice of variable type depends on the nature of the data being collected and the research objectives

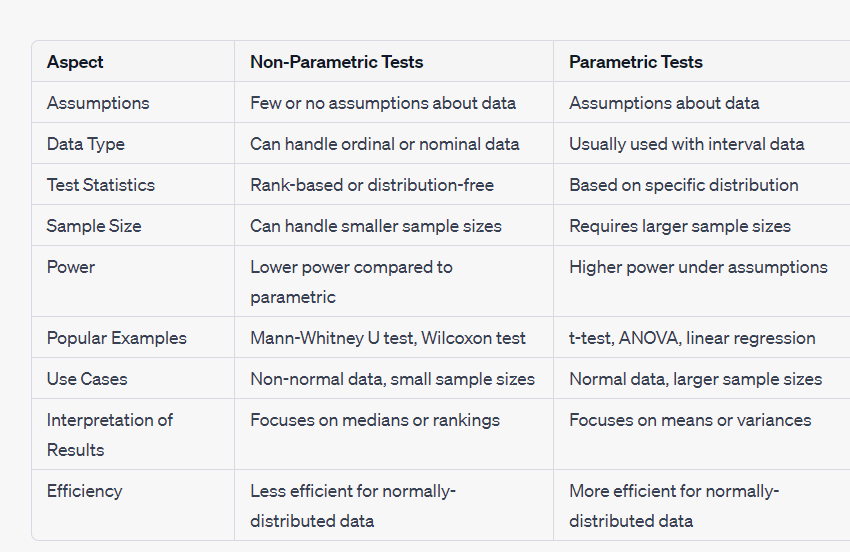
**In statistics, variables are often classified into different measurement scales, which provide information about the nature and properties of the data. The four commonly recognized measurement scales are nominal, ordinal, interval, and ratio scales. Let's go through each of them:**

1. **Nominal Scale**: A nominal scale is the lowest level of measurement and represents qualitative or categorical data. Variables at the nominal scale have categories or labels that are mutually exclusive and have no inherent order or numerical value. **Examples** include gender (male/female), eye color (blue/green/brown), or car models (Toyota/Honda/Ford).
2. **Ordinal Scale:** An ordinal scale represents data that can be ranked or ordered, but the intervals between categories may not be equal or measurable. While the categories have a relative order, the differences between them may not have a consistent interpretation. Examples include rating scales (excellent/good/fair/poor**), educational levels (elementary/middle/high school/college),** or survey responses with options like strongly agree/agree/neutral/disagree/strongly disagree.
3. **Interval Scale:** An interval scale represents data that can be ordered, and the intervals between values are equal and meaningful. However, there is no absolute zero point. Variables at the interval scale allow for meaningful calculations of differences between values but not ratios. Examples include temperature measured in Celsius or Fahrenheit, years (AD), or calendar **months**.
4. **Ratio** Scale: A ratio scale is the highest level of measurement and possesses all the properties of the previous three scales. It has a natural zero point and allows for meaningful ratios between values. **Variables at the ratio scale include quantities that have a true zero point, such as height, weight, time, or income**.

The choice of measurement scale affects the statistical techniques and operations that can be applied to the data. Nominal and ordinal scales often require non-parametric statistical tests, while interval and ratio scales allow for more advanced parametric statistical methods.

Understanding the measurement scale of a variable is crucial for selecting appropriate statistical analyses, summarizing data, and interpreting the results accurately.





**Bins in statistics refer to the intervals or ranges into which the data is divided in order to create a frequency distribution or histogram**.

**An outlier is a data point that significantly deviates from the other values in a dataset. It is an observation that lies an abnormal distance away from other observations. Outliers can be caused by various factors, such as measurement errors, experimental errors, or genuinely unusual or extreme values.**

**Outliers can have a significant impact on statistical analyses and calculations because they can skew the results. They can affect measures of central tendency (such as the mean), as well as measures of dispersion (such as the range or standard deviation)**

In statistics, sampling techniques are methods used to select a subset of individuals or observations from a larger population. These techniques help researchers gather data efficiently and make inferences about the population based on the sample. Here are some commonly used sampling techniques:

1. **Simple Random Sampling:** Simple random sampling involves selecting individuals from the population in a completely random manner, where each individual has an equal chance of being selected. This can be done by assigning a unique identifier to each individual in the population and using a random number generator to select the desired sample size.
2. **Stratified Sampling**: Stratified sampling involves dividing the population into distinct subgroups or strata based on certain **characteristics (e.g., age, gender, geographic location)** and then randomly selecting individuals from each stratum in proportion to their representation in the population. This technique ensures that the sample represents the diversity within the population.
3. **Systematic Sampling:** Systematic sampling involves selecting individuals from the population at fixed intervals. The starting point is chosen randomly, and then individuals are selected based on a predetermined pattern (e.g., every 10th person). This technique provides a relatively simple sampling method and is useful when the population is ordered in some way.
4. **Convenience Sampling:** Convenience sampling involves selecting individuals who are readily available and accessible to the researcher. While this method is convenient, it may introduce bias because the sample may not be representative of the entire population. Convenience sampling is often used in exploratory or pilot studies but should be interpreted with caution. Suppose a researcher is studying the opinions of university students about a new campus policy. To collect data, the researcher might stand in the university cafeteria and approach the first 50 students who pass by during lunchtime to participate in the study. In this case, the researcher is using convenience sampling by selecting participants based on their availability and proximity.2) Let's say a high school teacher wants to gather student feedback on a new teaching method. Instead of using a random selection process, the teacher decides to use convenience sampling by asking the first five students who arrive early for class to participate in a brief survey about their experience with the teaching method. The teacher selects these students simply because they were readily available and happened to be present before the class started.
5. **Multistage Sampling**: Multistage sampling involves multiple stages of sampling, combining different sampling techniques. It is often used in complex surveys or studies with large populations. For example, a researcher might use stratified sampling to divide the population into strata, then use systematic sampling within each stratum to select clusters, and finally use simple random sampling to select individuals within each selected cluster. To obtain a representative sample for studying the voting preferences of residents in a small town with three neighborhoods (A, B, and C), a researcher follows these steps:

Randomly select one neighborhood (e.g., Neighborhood A) out of the three.

Randomly choose a specific number of households within the selected neighborhood (e.g., four households).

Interview all eligible voters residing in each selected household.

The choice of sampling technique depends on various factors, including the research objective, available resources, population characteristics, and practical considerations. The goal is to select a sample that is representative of the population and allows for valid inferences and generalizations.

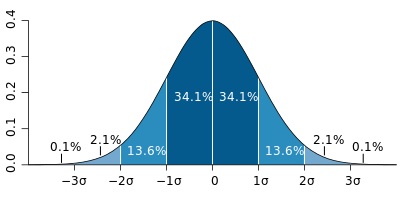
…………………………………………………………………………………………………………………………………………………

**The variance is a statistical measure that quantifies the spread or dispersion of a set of data points around their mean. It measures how far each data point is from the mean and gives an indication of the variability within the data set.**

**Dispersion refers to the extent or degree of spread or variability within a set of data points. In one line, dispersion can be defined as the range of the data, which is the difference between the maximum and minimum values:**

**Dispersion = max(data) - min(data)**

In this expression, **data** represents the set of data points for which we want to calculate the dispersion. By finding the maximum and minimum values in the data set and taking their difference, we obtain a measure of how spreadc out the data points are.



In this image, the curve **represents a bell curve**. The highest point of the curve corresponds to the mean, which is the center of the distribution. The curve is symmetrical, with data points evenly distributed around the mean. The standard deviation determines the spread of the curve, with a larger standard deviation resulting in a wider curve and a smaller standard deviation resulting in a narrower curve. The area under the curve represents the probability of observing a data point within a specific range.

**Quartiles** are statistical measures that divide a dataset into four equal parts. The four quartiles, denoted as **Q1, Q2 (also known as the median), Q3, and Q4**, provide information about the distribution of values within the dataset. Here's a brief explanation of each quartile:

1. **Q1 (First Quartile):** Q1 represents the lower quartile and marks the boundary below which the lowest 25% of the data lies**. It splits the dataset into the bottom 25% and the top 75% of the values.**
2. **Q2 (Second Quartile):** Q2 is **the median** of the dataset and divides the data into two halves**. It represents the value below which 50% of the data falls and above which the remaining 50% falls.**
3. **Q3 (Third Quartile):** Q3 is the upper quartile and indicates the boundary below which the lowest 75% of the data lies. It splits the dataset into the bottom 75% and the top 25% of the values.
4. **Q4 (Fourth Quartile): Q4 is an alternative term used to refer to the maximum value in the dataset or the upper limit of the dataset. It is not a traditional quartile but may be used in some contexts**.

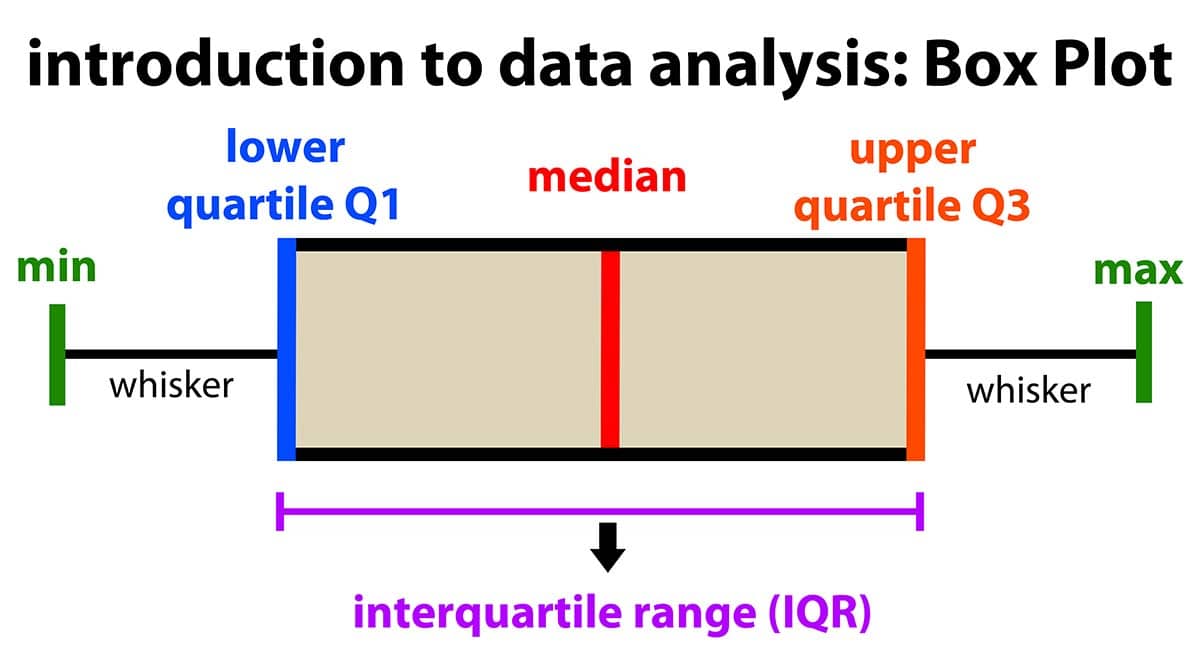
These quartiles are particularly useful for analyzing and understanding the spread and central tendency of a dataset.

**A box plot, also known as a box-and-whisker plot, is a graphical representation of the distribution of a dataset. It provides a visual summary of key statistical measures, including the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum values. Here's how a box plot is constructed:**

1. **Minimum and Maximum:** The minimum and maximum values of the dataset are identified. These values represent the lowest and highest data points, respectively.
2. **Quartiles (Q1, Q2, Q3**): The quartiles divide the dataset into four equal parts. Q1 represents the lower quartile (25th percentile), Q2 is the median (50th percentile), and Q3 is the upper quartile (75th percentile). These quartiles provide insights into the spread and central tendency of the data.
3. **Box:** A box is drawn to represent the interquartile range (IQR), which is the range between Q1 and Q3. It spans the middle 50% of the dataset. The box is typically drawn from Q1 to Q3, with a line indicating the median (Q2) inside the box.
4. **Whiskers:** Whiskers extend from the box to the minimum and maximum values, or to a predefined range based on certain criteria. They represent the overall range of the dataset, excluding any outliers.
5. **Outliers:** Outliers, if present, are individual data points that fall outside the whiskers. These points are plotted as individual dots or circles to indicate their deviation from the central data distribution.

The box plot provides a concise and visual summary of the dataset, allowing for comparisons between different groups or distributions. It helps identify skewness, variability, and the presence of outliers.

Box plots are used to visualize and analyze the distribution of a dataset, highlighting key statistical measures such as quartiles, median, minimum, and maximum values. They also help identify outliers and provide a concise summary of the data's spread and central tendency.



**The normal distribution, also known as the Gaussian distribution,** is a continuous probability distribution that is widely used in statistics. It is **often referred to as the "bell curve" due to its** characteristic shape.

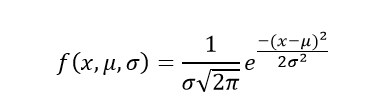
In a normal distribution, the data is symmetrically distributed around the mean, **with the highest frequency of values occurring at the mean**. The shape of the distribution is determined by two parameters: the mean (μ) and the standard deviation (σ). **The mean represents the center of the distribution, while the standard deviation determines the spread or dispersion of the data points.**

In conclusion, the normal distribution is a fundamental concept in statistics, representing a symmetric bell-shaped pattern of data. It is defined by its mean and standard deviation, and it is widely used for analyzing and modeling various real-world phenomena. The properties and characteristics of the normal distribution make it a valuable tool for statistical inference and decision-making processes.

**In probability theory and statistics, the Normal Distribution, also called the Gaussian Distribution, is the most significant continuous probability distribution. Sometimes it is also called a bell curve.** A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics.

**Normal Distribution Formula**

The probability density function of normal or gaussian distribution is given by;



Where,

* x is the variable
* μ is the mean
* σ is the standard deviation
* **Approximately 68% of the data falls within one standard deviation of the mean. (i.e., Between Mean- one Standard Deviation and Mean + one standard deviation)**
* **Approximately 95% of the data falls within two standard deviations of the mean. (i.e., Between Mean- two Standard Deviation and Mean + two standard deviations)**
* **Approximately 99.7% of the data fall within three standard deviations of the mean. (i.e., Between Mean- three Standard Deviation and Mean + three standard deviations)**

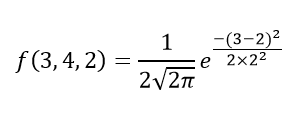
**Question 1: Calculate the probability density function of normal distribution using the following data. x = 3, μ = 4 and σ = 2.**

Solution: Given, variable, x = 3

Mean = 4 and

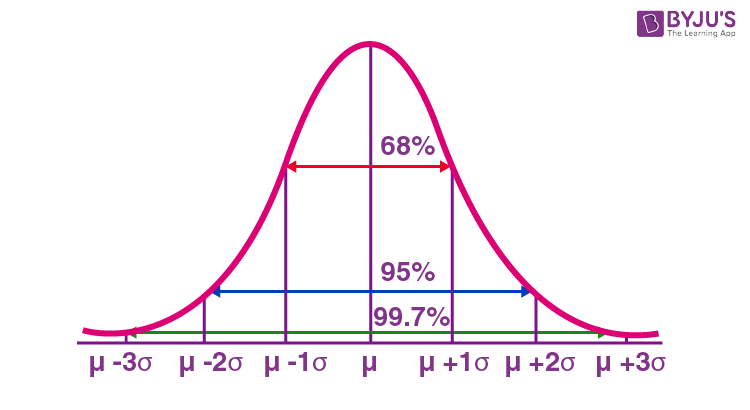
Standard deviation = 2

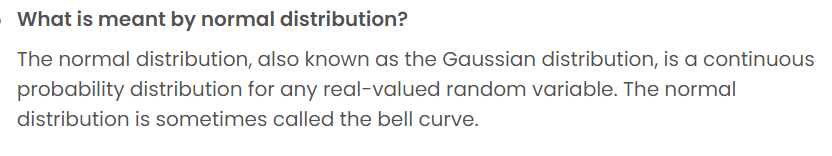
By the formula of **the probability density of normal distribution**, we can write;



Hence, f(3,4,2) = 1.106.

**Thus, the empirical rule is also called the 68 – 95 – 99.7 rule.**





Normal Distribution Properties

Some of the important properties of the normal distribution are listed below:

* **In a normal distribution, the mean, median and mode are equal.(i.e., Mean = Median= Mode).**
* **The total area under the curve should be equal to 1.**
* **The normally distributed curve should be symmetric at the centre**.
* **The normal distribution should be defined by the mean and standard deviation.**
* The normal distribution curve must have only one peak. (i.e., Unimodal)
* The curve approaches the x-axis, but it never touches, and it extends farther away from the mean.

**Applications**

The normal distributions are closely associated with many things such as:

* Marks scored on the test
* Heights of different persons
* Size of objects produced by the machine
* Blood pressure and so on.

f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}

**The standard normal distribution, also known as the z-distribution, is a specific form of the normal distribution. It is a special case where the mean (μ) is 0 and the standard deviation (σ) is 1.**

T**he standard normal distribution is often used as a reference distribution because it simplifies calculations and allows for easy comparison of values across different normal distributions. By standardizing values to the z-score, we can determine how many standard deviations a particular value is away from the mean of the standard normal distribution.**

**Also, the standard normal distribution is centred at zero, and the standard deviation gives the degree to which a given measurement deviates from the mean.**

Standard Normal Distribution Uses

* **The standard normal distribution is a tool to translate a normal distribution into numbers. We may use it to get more information about the data set than was initially known**.
* Standard normal distribution allows us to quickly estimate the probability of specific values befalling in our distribution or compare data sets with varying means and standard deviations.
* Also, the z-score of the standard normal distribution is interpreted as the number of standard deviations a data point falls above or below the mean.

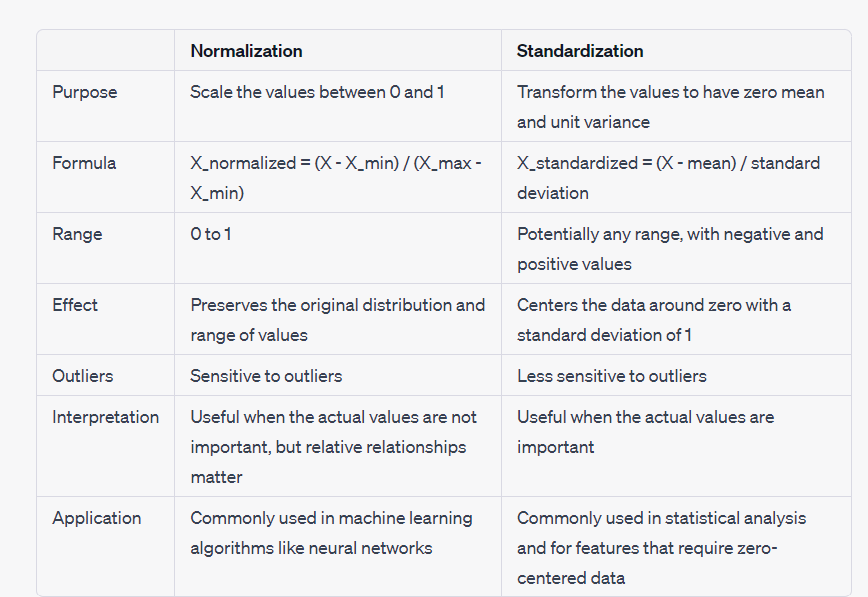
Characteristics of Standard Normal Distribution

A z-score of a standard normal distribution is a standard score that indicates how many standard deviations are away from the mean an individual value (x) lies:

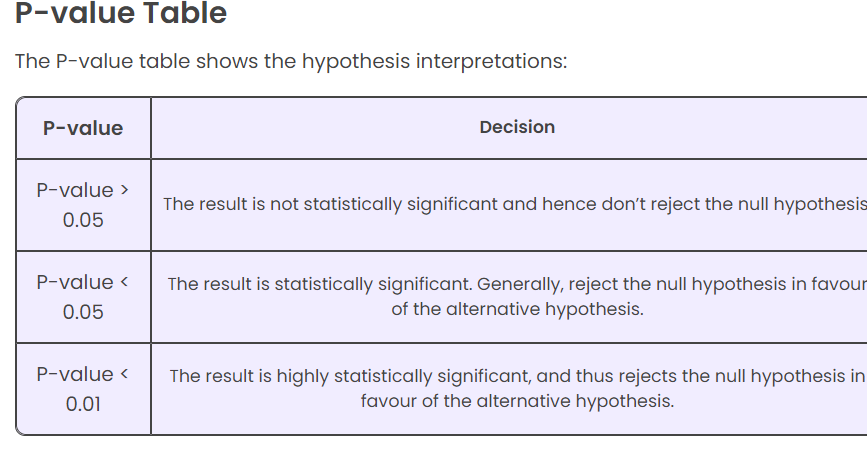
* When z-score is positive, the x-value is greater than the mean
* When z-score is negative, the x-value is less than the mean
* When z-score is equal to 0, the x-value is equal to the mean

The empirical rule, or the 68-95-99.7 rule of standard normal distribution, tells us where most values lie in the given normal distribution. Thus, for the standard normal distribution, 68% of the observations lie within 1 standard deviation of the mean; 95% lie within two standard deviations of the mean; 99.7% lie within 3 standard deviations of the mean.

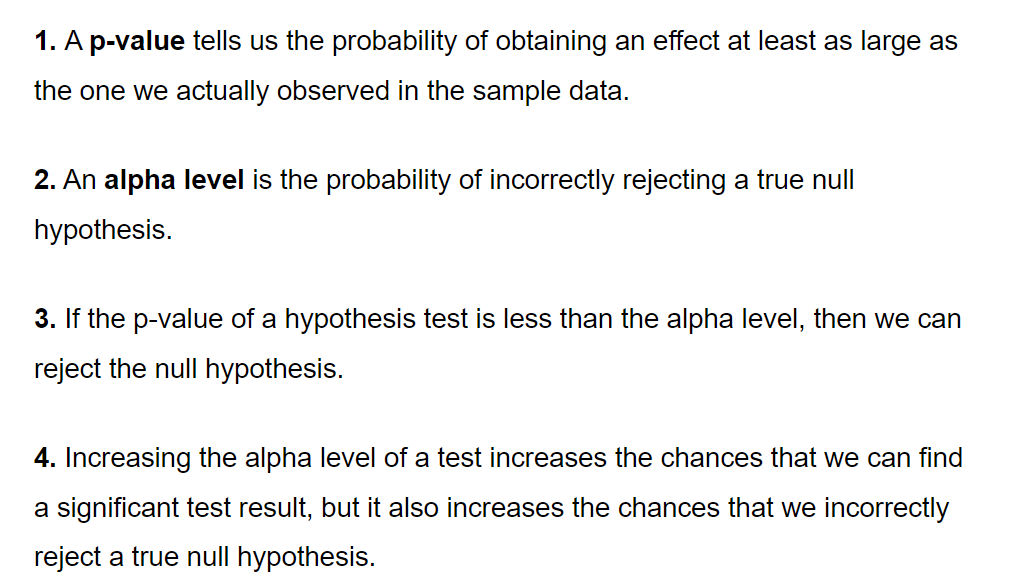
**The standard normal distribution is applied in statistics for hypothesis testing, constructing confidence intervals, standardizing variables, probability calculations, and data transformation.**

****

**In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical**[**hypothesis test**](https://www.investopedia.com/terms/h/hypothesistesting.asp)**, assuming that the**[**null hypothesis**](https://www.investopedia.com/terms/n/null_hypothesis.asp)**is correct**

* **A p-value is a statistical measurement used to validate a hypothesis against observed data.**
* **A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.**
* **The lower the p-value, the greater the statistical significance of the observed difference.**
* 

**The P-value is known as the probability** value. It is defined as the probability of getting a result that is either the same or more extreme than the actual observations. The P-value is known as the level of marginal significance within the hypothesis testing that represents the probability of occurrence of the given event**.. If the P-value is small, then there is stronger evidence in favour of the alternative hypothesi**s.



**The relationship between the p-value and the alpha value is straightforward: if the p-value is smaller than the alpha value, the result is considered statistically significant, and the null hypothesis is rejected in favor of the alternative hypothesis. Conversely, if the p-value is larger than the alpha value, the result is not considered statistically significant, and there is insufficient evidence to reject the null hypothesis**.

……………………………………………………………………………………………………………………………….

**Hypothesis testing** is a statistical procedure used to make inferences and draw conclusions about a population based on sample data. **It involves formulating a null hypothesis (H0) and an alternative hypothesis (H1 or Ha), and then collecting and analyzing data to determine whether there is sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis**.

You decide to get tested for COVID-19 based on mild symptoms. There are two errors that could potentially occur:

* **Type I error (false positive):** the test result says you have coronavirus, but you actually don’t.
* **Type II error (false negative):** the test result says you don’t have coronavirus, but you actually do.

**The probability of making a Type I error is the**[**significance level**](https://www.scribbr.com/statistics/statistical-significance/#significance-level)**, or alpha (α), while the probability of making a Type II error is beta (β). These risks can be minimized through careful planning in your study design.**

**Type I Error**

A type I error appears when the [null hypothesis](https://byjus.com/maths/null-hypothesis/) (H0) of an experiment is true, but still, it is rejected.A type I error is often called a false positive (an event that shows that a given condition is present when it is absent). **The type I error significance level or rate level is the probability of refusing the null hypothesis given that it is true. It is represented by Greek letter α (alpha) and is also known as alpha level. Usually, the significance level or the probability of type i error is set to 0.05 (5%), assuming that it is satisfactory to have a 5% probability of inaccurately rejecting the null hypothesis**

**Type II Error**

A type II error appears when the null hypothesis is false but mistakenly fails to be refused. A type II error is also known as false negative (where a real hit was rejected by the test and is observed as a miss), in an experiment checking for a condition with a final outcome of true or false.

A type II error is assigned when a true [alternative hypothesis](https://byjus.com/maths/alternative-hypothesis/) is not acknowledged.

The rate level of the type II error is represented by the Greek letter β (beta) and linked to the power of a test (which equals 1−β).

| **Aspect** | **One-Tailed Test** | **Two-Tailed Test** |
| --- | --- | --- |
| Hypotheses | One-sided hypothesis | Two-sided hypothesis |
| Direction | Tests for a specific direction of effect | Tests for any direction of effect |
| Critical Region | Only one tail of the distribution | Both tails of the distribution |
| Rejection Region | One-sided rejection region (either upper or lower) | Two-sided rejection region (both upper and lower) |
| Type 1 Error | **α (alpha) divided by 2** | α (alpha) |
| Example | Testing if a new drug increases blood pressure | Testing if a new drug has any effect on blood pressure |
| Calculation | Critical value for the chosen tail only | Critical values for both tails of the distribution |
| P-value | P-value calculated for the chosen tail only | P-value calculated for both tails of the distribution |
| Statistical Power | Higher power due to focusing on one **direction** | Lower power due to dividing α between two directions |

**What Is a Confidence Interval?**

* **A confidence interval displays the probability that a parameter will fall between a pair of values around the mean.**
* Confidence intervals measure the degree of uncertainty or certainty in a sampling method.
* **They are also used in hypothesis testing and regression analysis.**
* Statisticians often use p-values in conjunction with confidence intervals to gauge **statistical significance.**
* **They are most often constructed using confidence levels of 95% or 99%.**

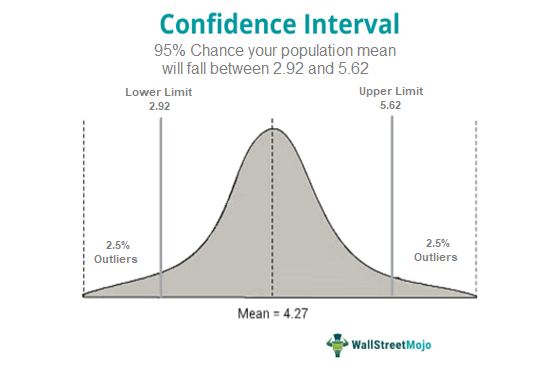
**What is the Definition of Point Estimation?**

Point estimators **are defined as functions that can be used to find the approximate value of a particular point from a given population parameter. The sample data of a population is used to find a point estimate or a statistic that can act as the best estimate of an unknown parameter that is given for a population**.

Let's say you want to estimate the average height of students in a university. You collect the heights of a sample of 50 students and calculate the sample mean. The sample mean, in this case, would be your point estimate for the population parameter, which is the average height

**Comparison Chart**

| **BASIS FOR COMPARISON** | **SAMPLE MEAN** | **POPULATION MEAN** |
| --- | --- | --- |
| Meaning | Sample mean is the arithmetic mean of random sample values drawn from the population. | Population mean represents the actual mean of the whole population. |
| Symbol | x̄ (pronounced as x bar) | μ (Greek term mu) |
| Calculation | Easy | Difficult |
| Accuracy | Low | High |
| Standard deviation | When calculated using sample mean, is denoted by (s). | When calculated using population mean, is denoted by (σ). |

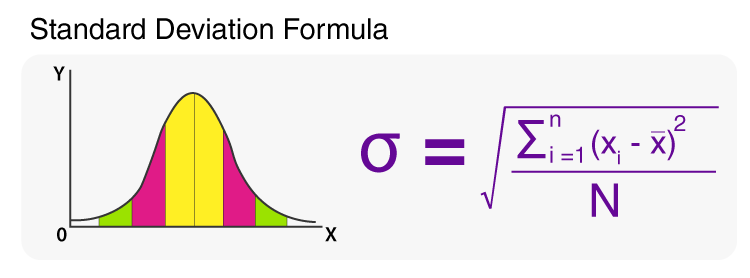


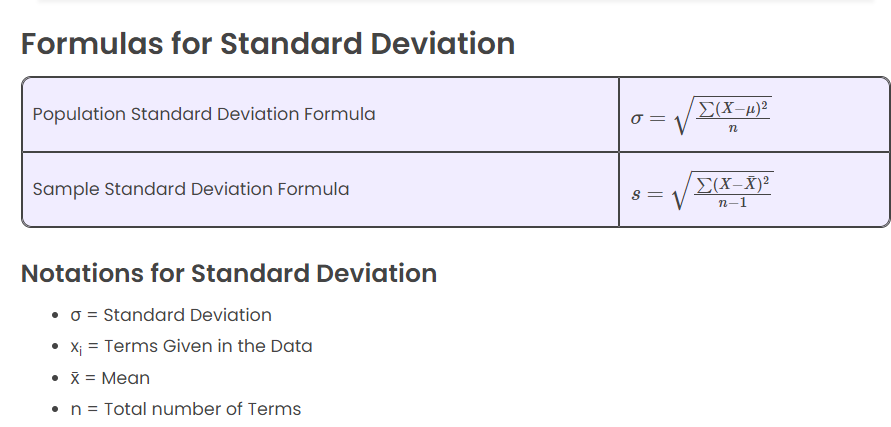
**Confidence**, in statistics, is another way to describe probability. For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.

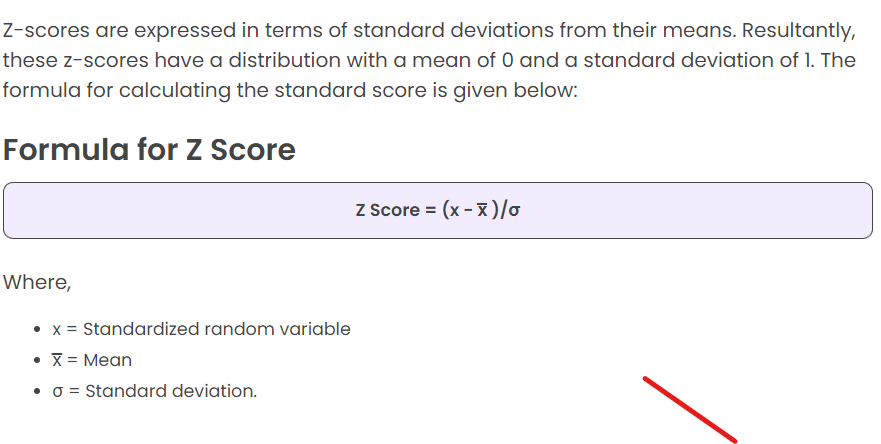
**Standard deviation formula** is used to find the values of a particular data that is dispersed. In simple words, the standard deviation is defined as the deviation of the values or data from an average mean. **Lower standard deviation concludes that the values are very close to their average. Whereas higher values mean the values are far from the mean value.** It should be noted that the [standard deviation](https://byjus.com/maths/standard-deviation/) value can never be negative.

**Standard Deviation is of two types:**

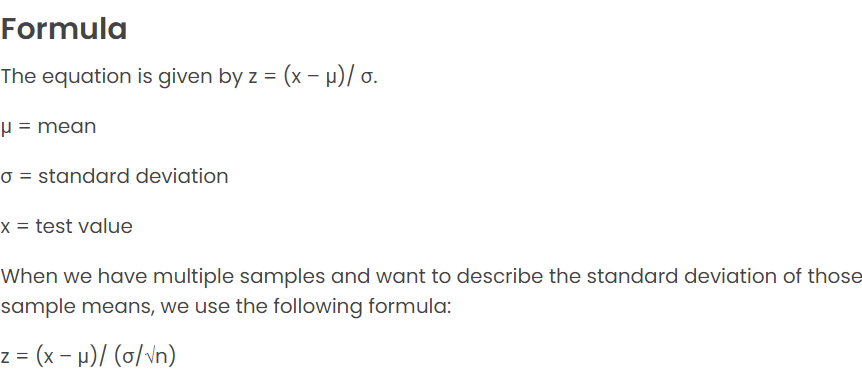
1. Population Standard Deviation
2. Sample Standard Deviation

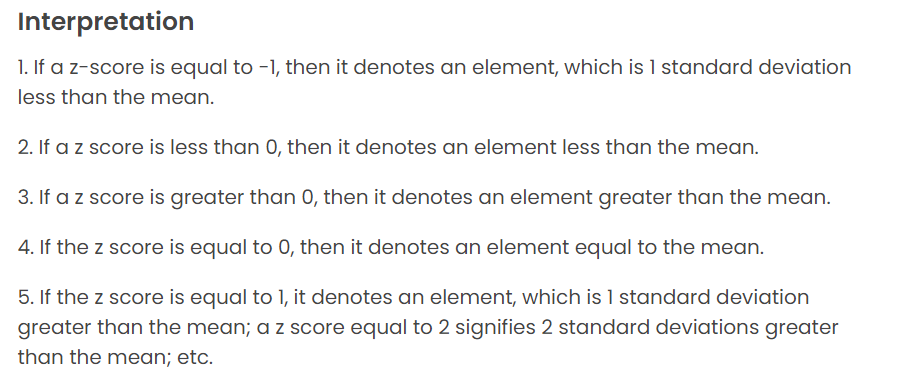






**A measure of how many standard deviations below or above the population mean a raw score is called z score**. It will be positive if the value lies above the mean and negative if it lies below the mean. It is also known as **standard score**. It indicates how many standard deviations an entity is, from the mean**. In order to use a z-score, the mean μ and also the population standard deviation σ should be known**. A z score helps to calculate the probability of a score occurring within a standard normal distribution. It also enables us to compare two scores that are from different samples.



****

**Z-score is used in a medical field to find how a certain new born baby’s weight compares to the mean weight of all babies. It is used to find how a certain shoe size compares to the mean population size.**

The t-test is any statistical hypothesis test in which the test statistic follows a Student’s t-distribution under the null hypothesis. It can be used to determine if two sets of data are significantly different from each other, and is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known

A **chi-squared test** (symbolically represented as **χ2**) is basically a data analysis on the basis of observations of a random set of variables. Usually, it is a comparison of two statistical data sets. This test was introduced by **Karl Pearson** in 1900 for [categorical data analysis and distribution](https://byjus.com/maths/categorical-data/). So it was mentioned as **Pearson’s chi-squared test**.

The chi-square test is used to estimate how likely the observations that are made would be, by considering the assumption of the [null hypothesis](https://byjus.com/maths/null-hypothesis/) as true.

A hypothesis is a consideration that a given condition or statement might be true, which we can test afterwards. Chi-squared tests are usually created from a sum of squared falsities or errors over the sample variance.

When we consider, the null speculation is true, the sampling distribution of the test statistic is called as **chi-squared distribution**. **The chi-squared test helps to determine whether there is a notable difference between the normal frequencies and the observed frequencies in one or more classes or categories. It gives the probability of independent variables.**

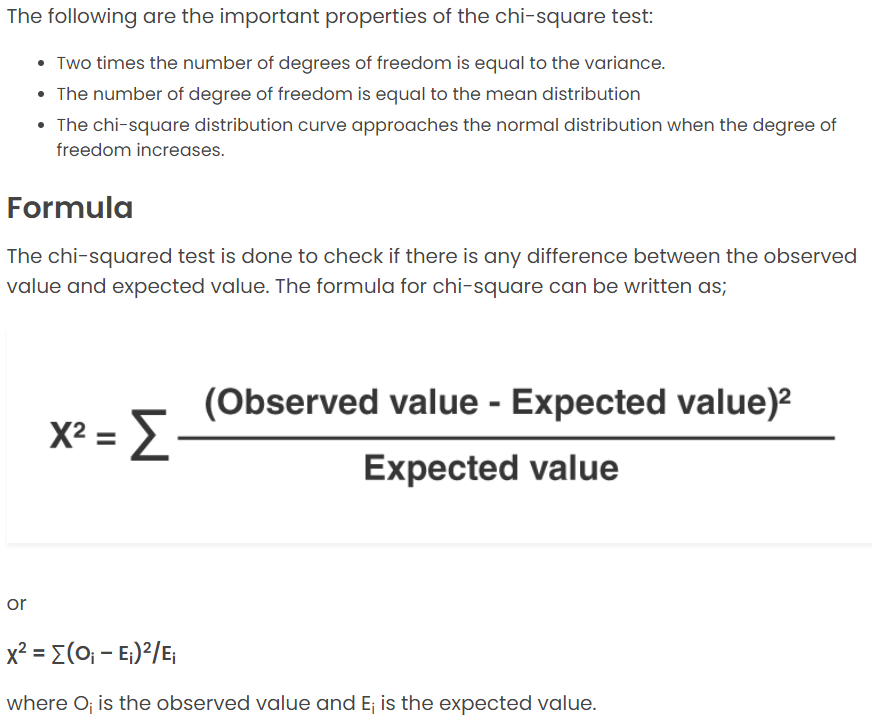
**Note**: **Chi-squared test is applicable only for categorical data, such as men and women falling under the categories of Gender, Age, Height, etc.**

**What is a chi-square test used for?**

**The chi-squared test is done to check if there is any difference between the observed value and the expected value.**

**What is a good chi-square value?**

**A good chi-square value is assumed to be 5. As we know, for the chi-square approach to be valid, the expected frequency should be at least 5.**



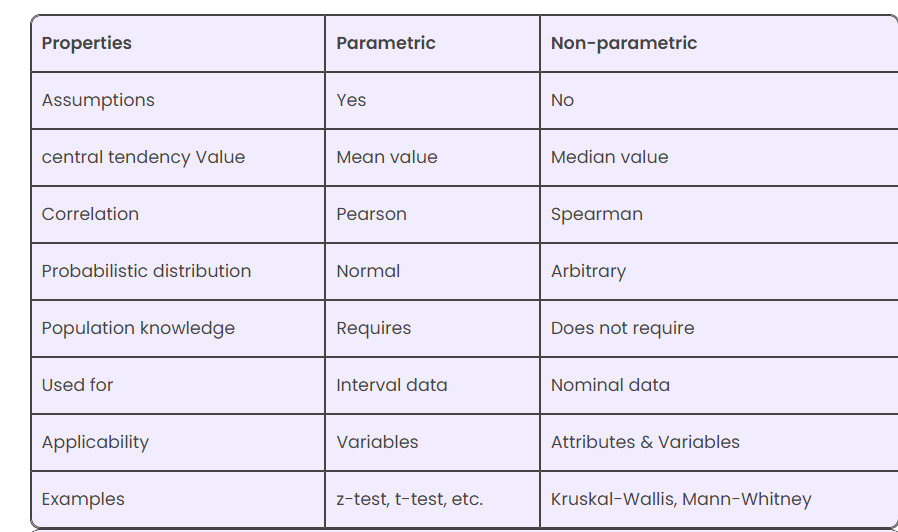
**Parametric Test Definition**

In Statistics, a parametric test is a kind of hypothesis test which gives generalizations for generating records regarding the mean of the primary/original population. The t-test is carried out based on the students’ t-statistic, which is often used in that value.

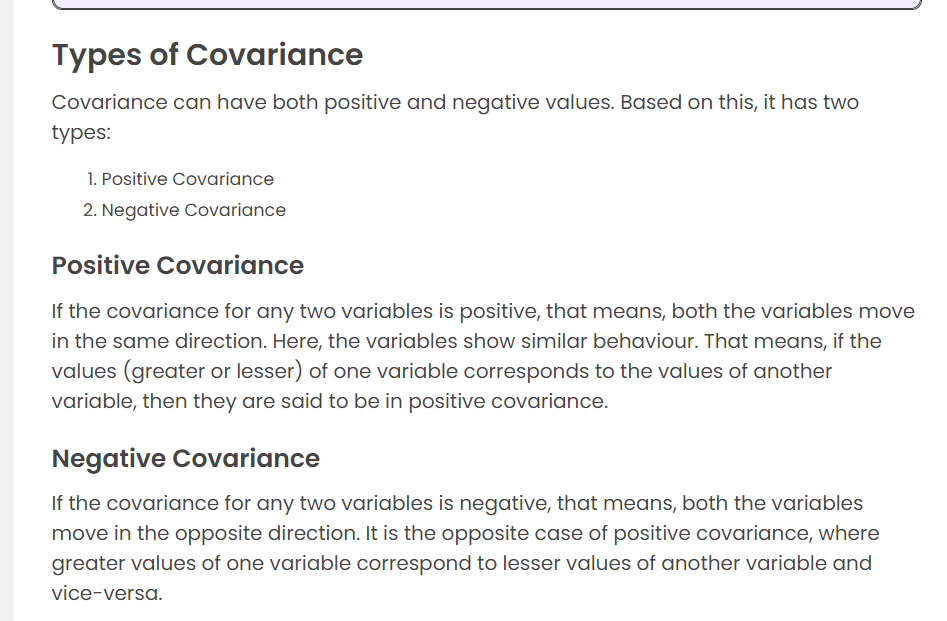
The t-statistic test holds on the underlying hypothesis, which includes the normal distribution of a variable. In this case, the mean is known, or it is considered to be known. For finding the sample from the population, population variance is identified. It is hypothesized that the variables of concern in the population are estimated on an interval scale.

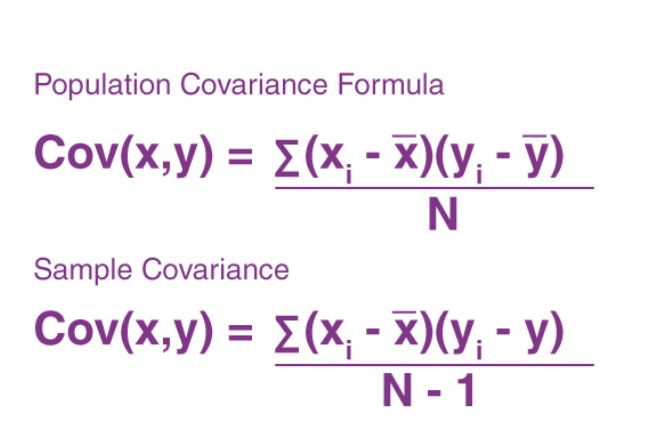
**Non-Parametric Test Definition**

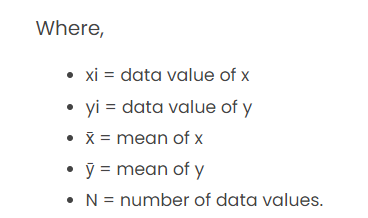
The non-parametric test does not require any population distribution, which is meant by distinct parameters. It is also a kind of hypothesis test, which is not based on the underlying hypothesis. **In the case of the non-parametric test, the test is based on the differences in the median**. So **this kind of test is also called a distribution-free test.** The test variables are determined on the nominal or ordinal level. If the independent variables are non-metric, the non-parametric test is usually performed.

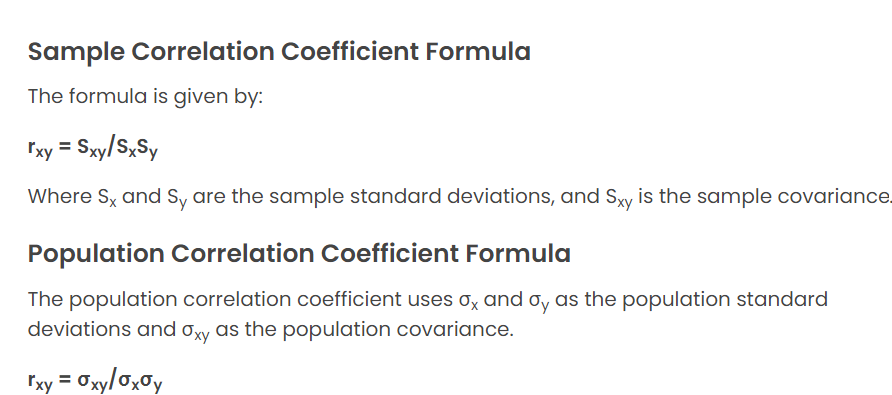
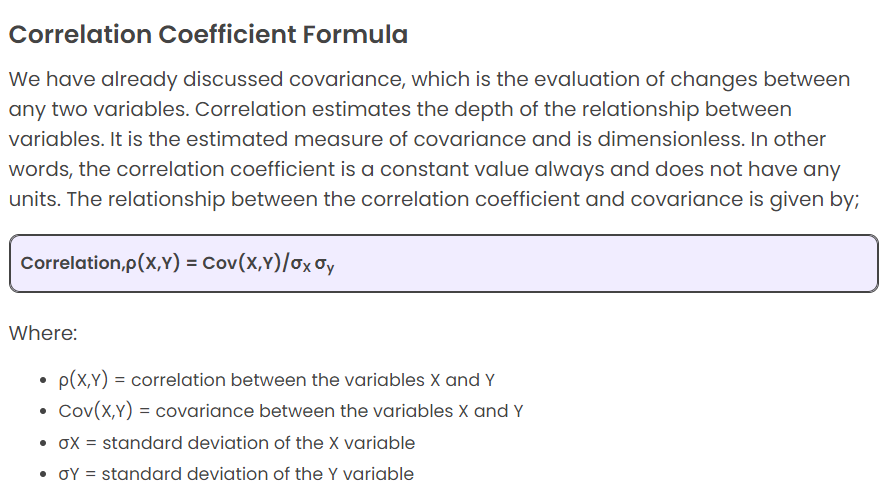


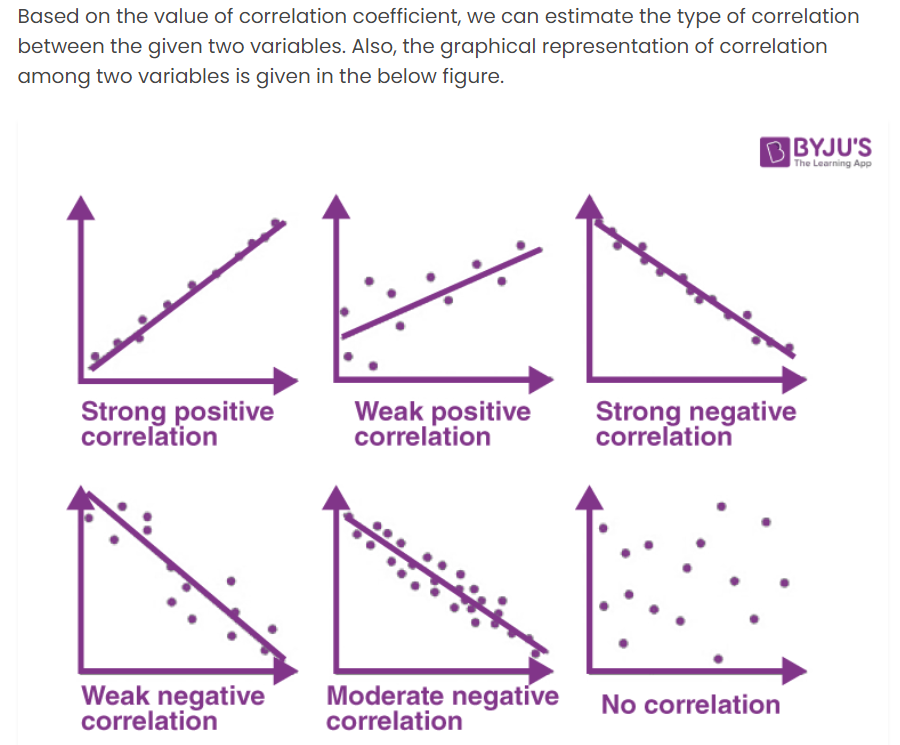
**Covariance** is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, **it defines the changes between the two variables, such that change in one variable is equal to change in another variable**. This is the property of a function of maintaining its form when the variables are linearly transformed. Covariance is measured in units, which are calculated by multiplying the units of the two variables.

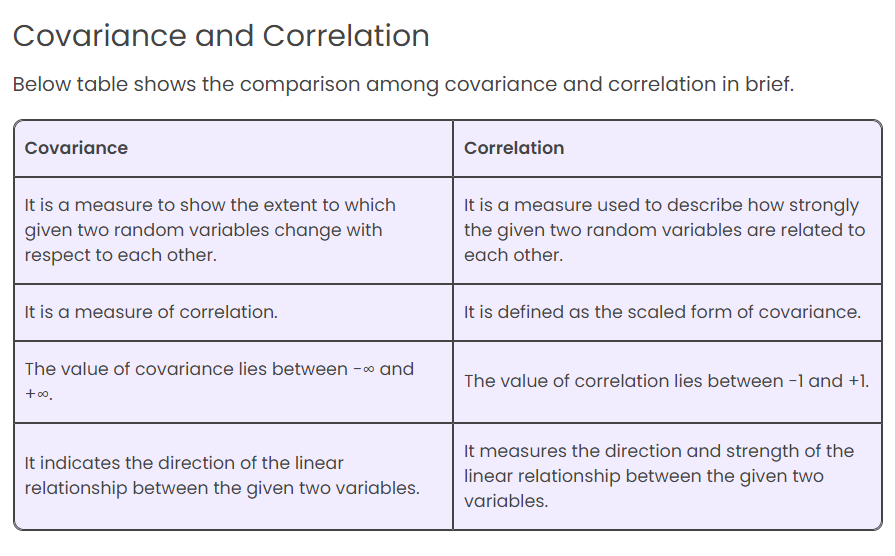


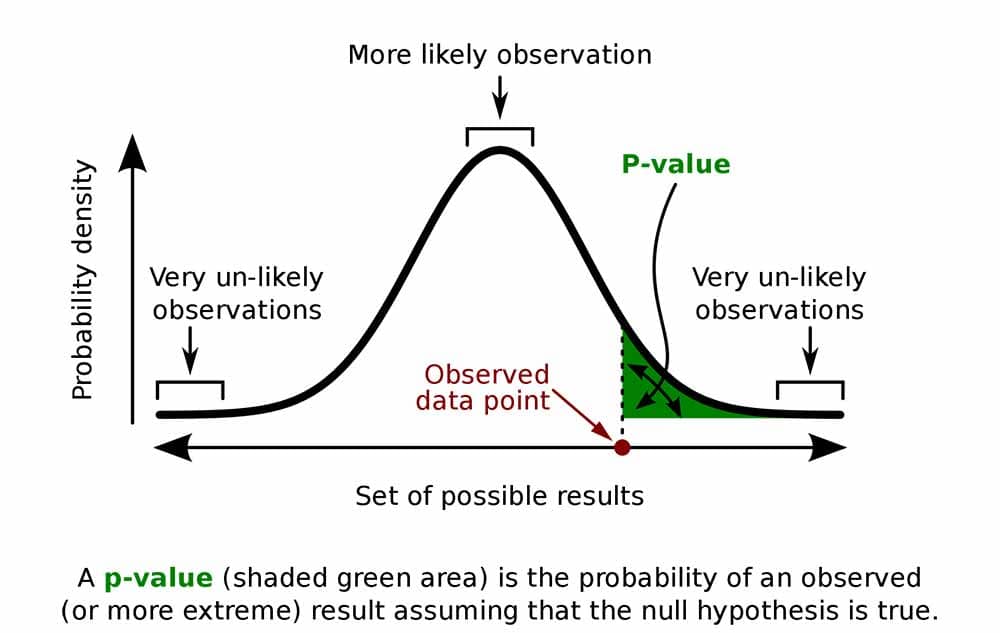


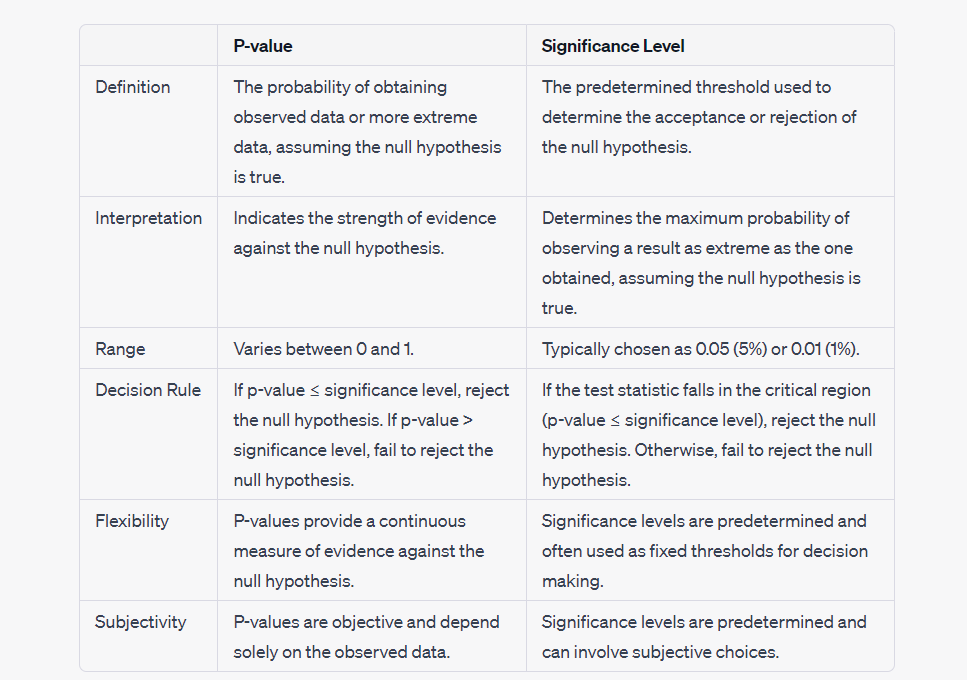




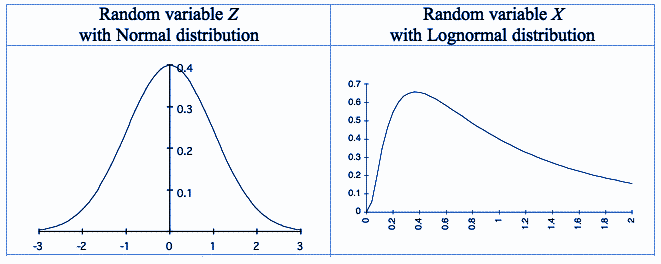








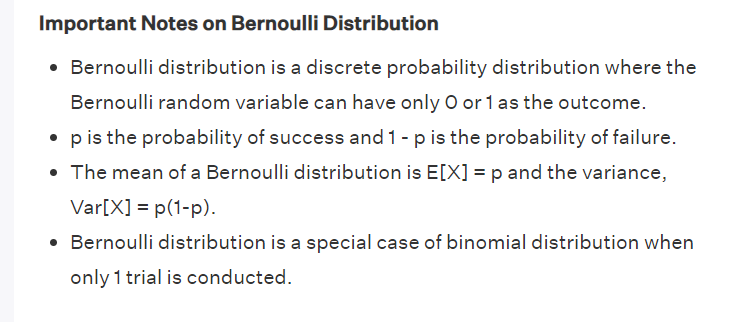
**A probability distribution of outcomes which is symmetrical or forms a bell curve is called a normal distribution. A log-normal distribution can be formed from a normal distribution using logarithmic mathematics**.

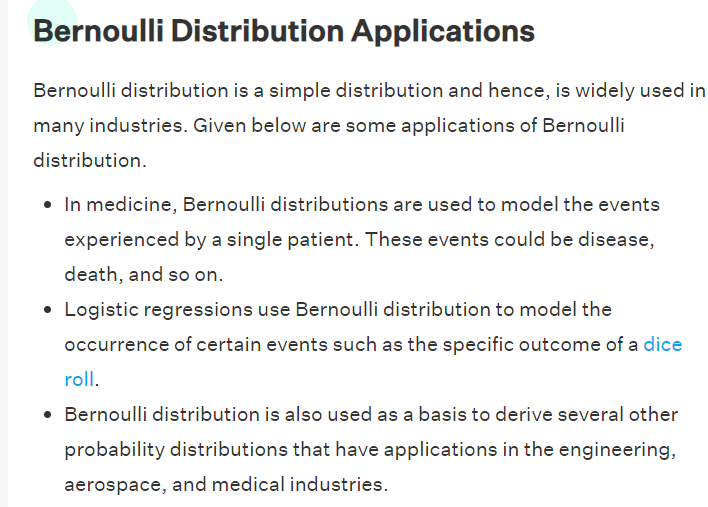
If X is a random variable and Y=ln(X) is normally distributed, then X is said to be distributed lognormally. **Similarly, if Y has a normal distribution, then the exponential function of Y will be having a lognormal distribution, i.e. X=exp(Y).** 

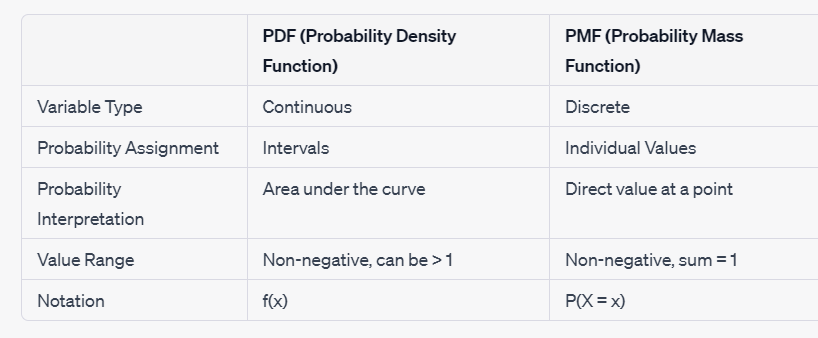
**Applications of Lognormal distribution**

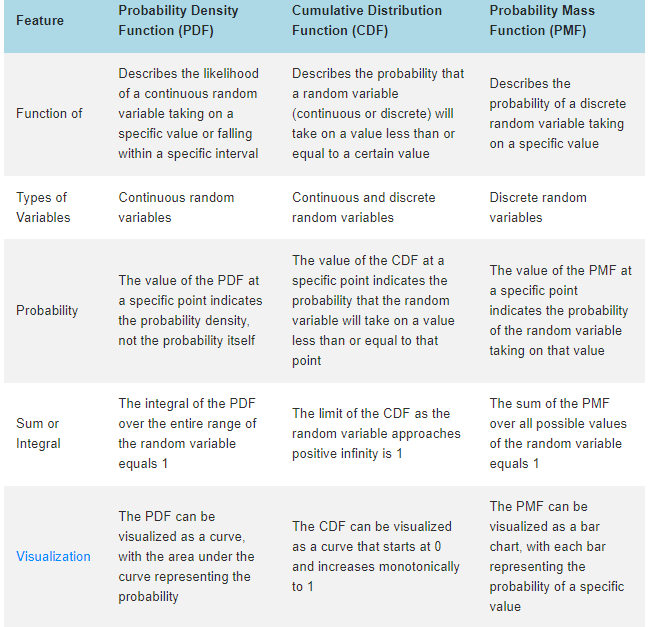
* **The length of comments posted on social media website discussion forums follows a lognormal distribution**
* **Time spent by a user on online articles (jokes, news etc.) follows a lognormal distribution**
* **In economics, to analyse the income of the population other than higher-income individuals**
* **In order to analyse the fluctuations in the stock markets**

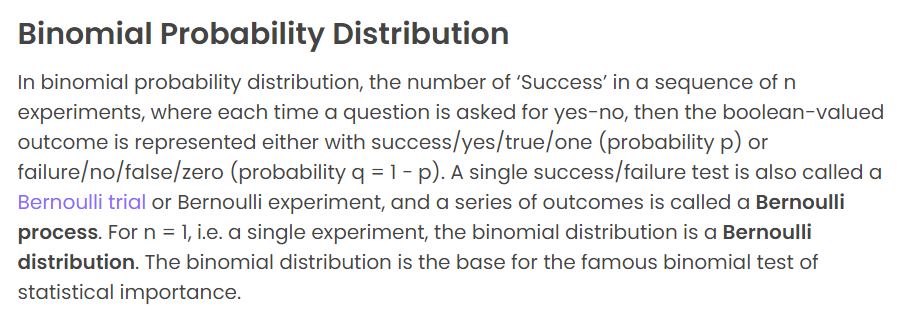
**Bernoulli distribution** is a discrete probability distribution. It describes the probability of achieving a “success” or “failure” from a Bernoulli trial. A Bernoulli trial is an event that has only two possible outcomes (success or failure). For example, will a coin land on heads (success) or tails (failure)?

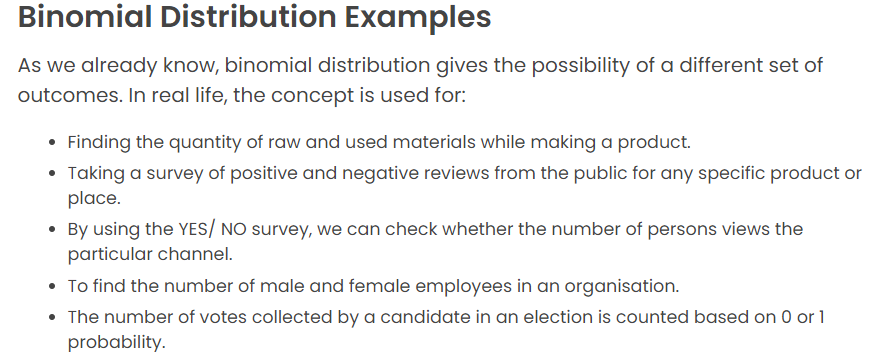


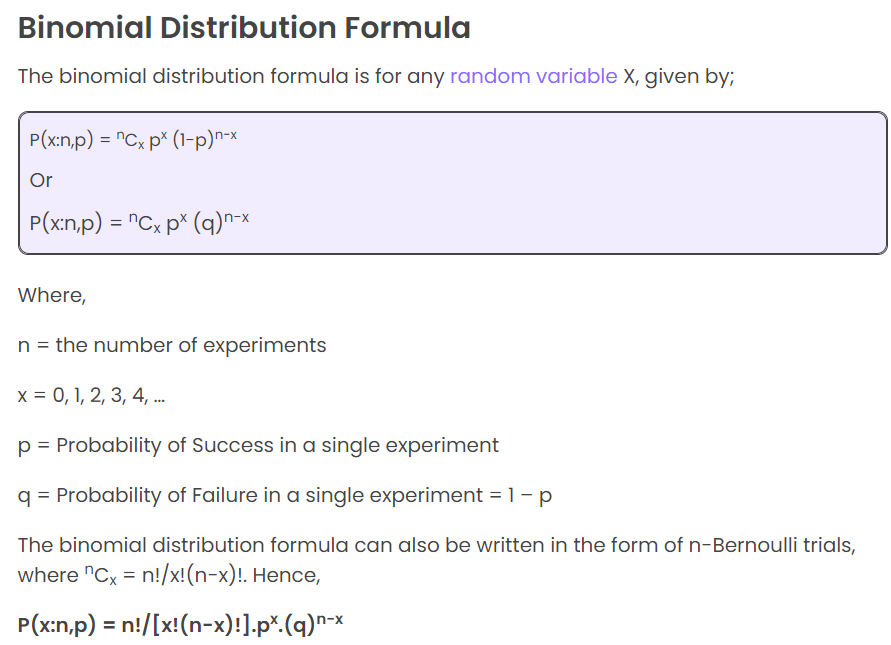


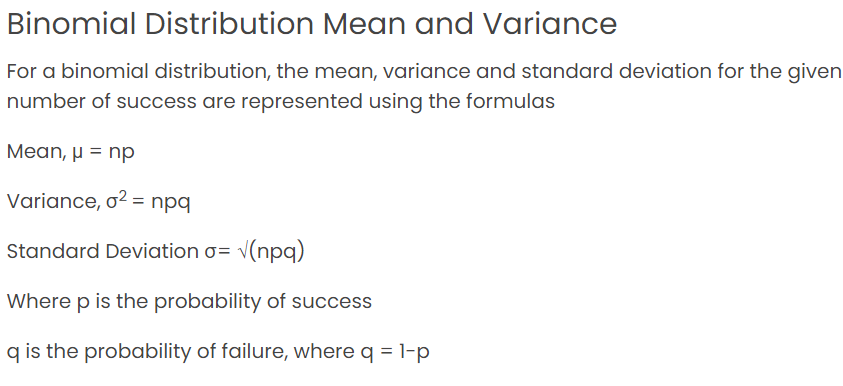


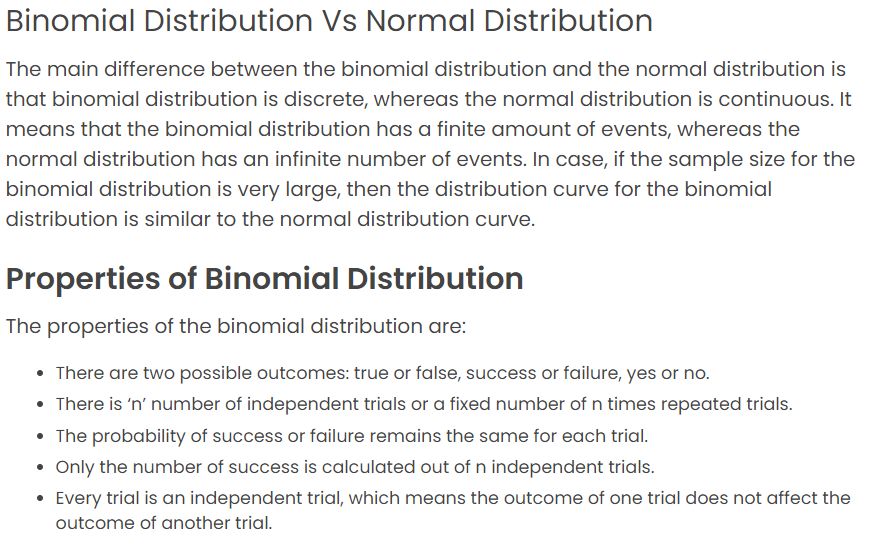












**Dispersion:** The dispersion of heights can be measured using statistics like the range, standard deviation, or variance. For example, if the range of heights is large, it indicates a wide dispersion, meaning there is a significant difference between the tallest and shortest student.

**Distribution:** The distribution of heights can be described by examining the shape and pattern of the data. For instance, if the heights follow a normal distribution, it means most students have heights around the mean, with fewer students having heights that are significantly higher or lower.

**Dispersion:** The dispersion would tell us how spread out the test scores are. For example, if the test scores range from 60 to 90, we can say that there is a high dispersion or variability in the scores.

**Distribution: The** distribution would describe how the test scores are distributed among the students. For example, if the majority of students scored around 80, and only a few scored below 70 or above 90, we can say that the distribution is skewed towards the higher score

