****

**Statistics for Machine Learning**

Personal Notes



Department of Computer Science and Engineering

**Green University of Bangladesh**

Md. Nazmul Hossain

ID: 223002089

Semester: 6th

# **TITLE:**

* Statistics for Machine Learning

# **OVERVIEW:**

# **OBJECTIVES:**

* Learn Statistics
* Learn Different type of statistics
* Descriptive vs Inferential statistics
* Measures of Central Tendency (Mean, Median, Mode)
* Measures of Variability (Range, Variance, Standard Deviation)
* Quantile/IQR
* Covariance – Co-relation
* Probability
* Data Distribution
* Transformation
* Central Limit Theorem
* Confidence Interval
* Hypothesis

# **CHAPTER 1:**

**Introduction to statistics:**

What is statistics, uses in real life, application in machine learning, example

# **CHAPTER 2:**

**Different type of statistics:**

Descriptive statistics vs Inferential statistics, topics of statistics

# **CHAPTER 3:**

**Descriptive Statistics:**

# **CHAPTER 4:**

**Probability:**

# **CHAPTER 4:**

**Inferential Statistics:**

# **CHAPTER 5:**

**Conclusion:**

# **TEXT BOOKS:**

# **REFERENCES:**

**CHAPTER 1:**

**Introduction to Statistics**

**Introduction:**

Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting data. It encompasses a wide range of techniques for summarizing data, making inferences, and drawing conclusions.

Statistical methods help quantify uncertainty and variability in data, allowing researchers and analysts to make data-driven decisions with confidence.

**Applications of Statistics in Machine Learning:**

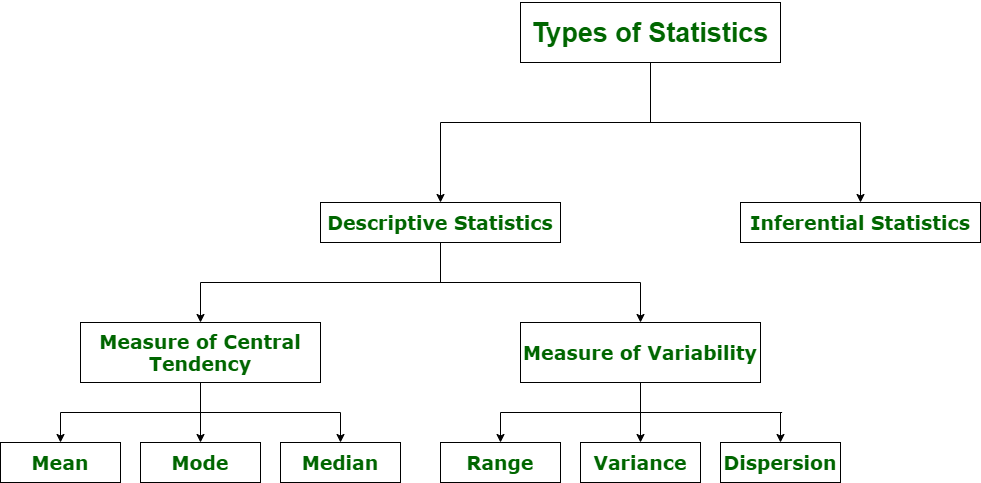
Statistics is a key component of machine learning, with broad applicability in various fields.

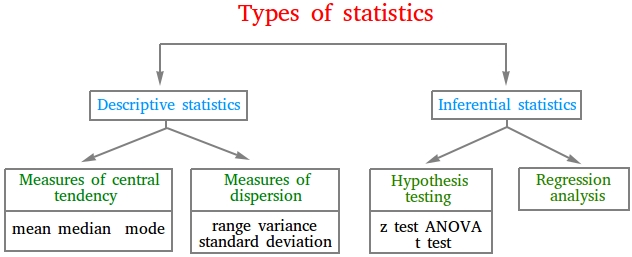
* Feature engineering relies heavily on statistics to convert geometric features into meaningful predictors for[machine learning algorithms.](https://www.geeksforgeeks.org/machine-learning-algorithms/)
* In image processing tasks like object recognition and segmentation, statistics accurately reflect the shape and structure of objects in images.
* Anomaly detection and quality control benefit from statistics by identifying deviations from norms, aiding in the detection of defects in manufacturing processes.
* Environmental observation and geospatial mapping leverage statistical analysis to monitor land cover patterns and ecological trends effectively.

Overall, statistics plays a crucial role in machine learning, driving insights and advancements across diverse industries and applications.

# **CHAPTER 2:**

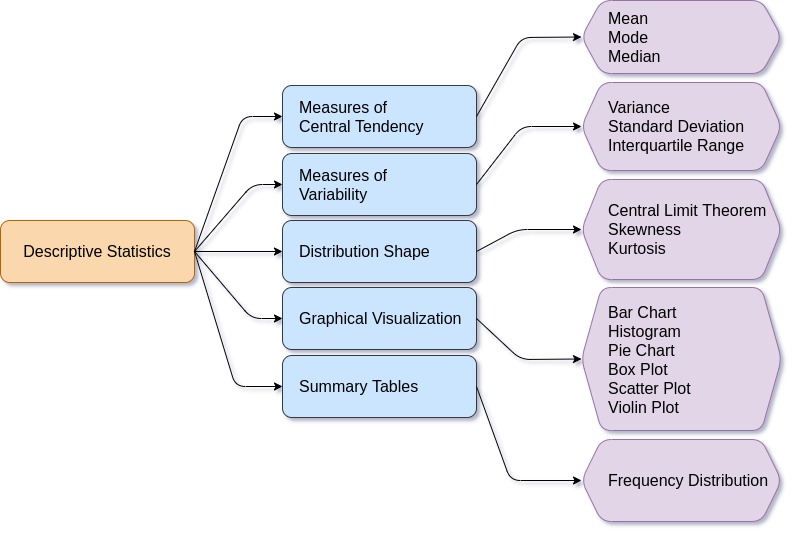
**Different type of statistics:**





# **CHAPTER 3:**

**Descriptive Statistics**

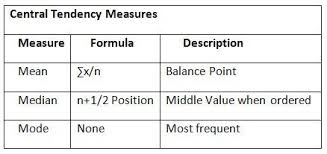


**Descriptive Statistics:**

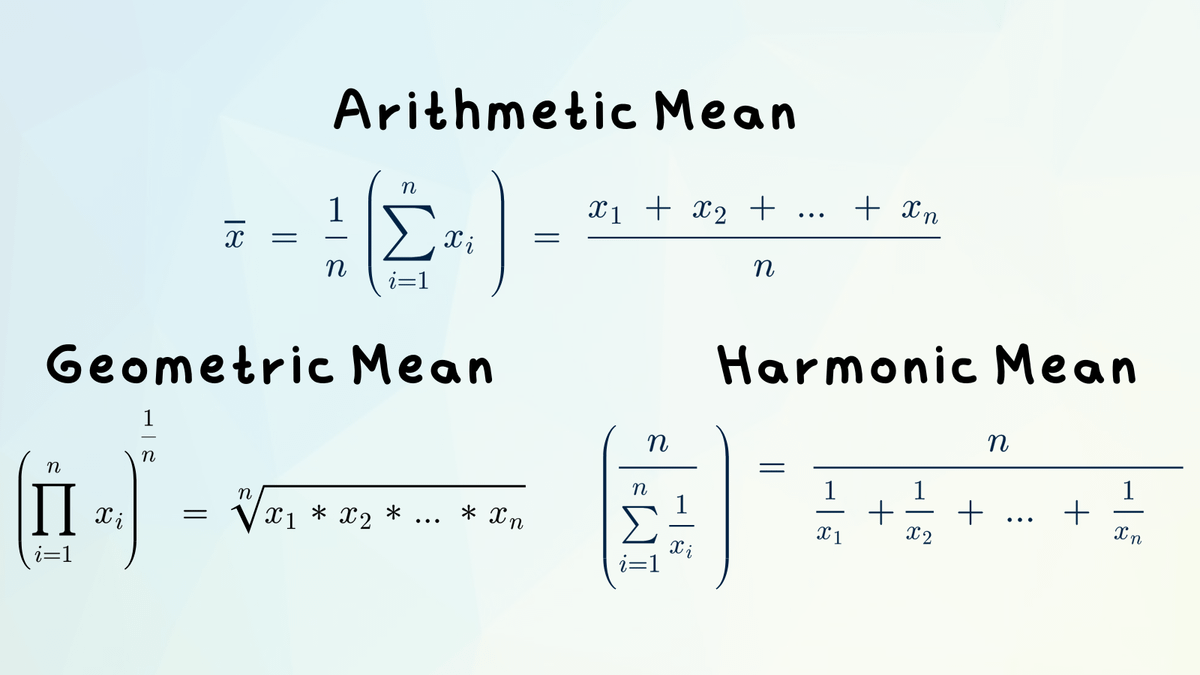
1. Measures of Central Tendency
   1. Mean, Median, Mode
2. Measures of Variability
   1. Range, Variance, Standard Deviation, IQR

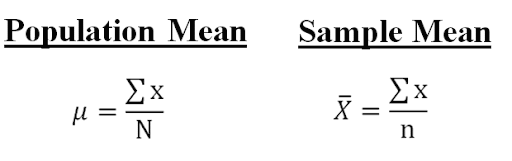
**Central Tendency:**

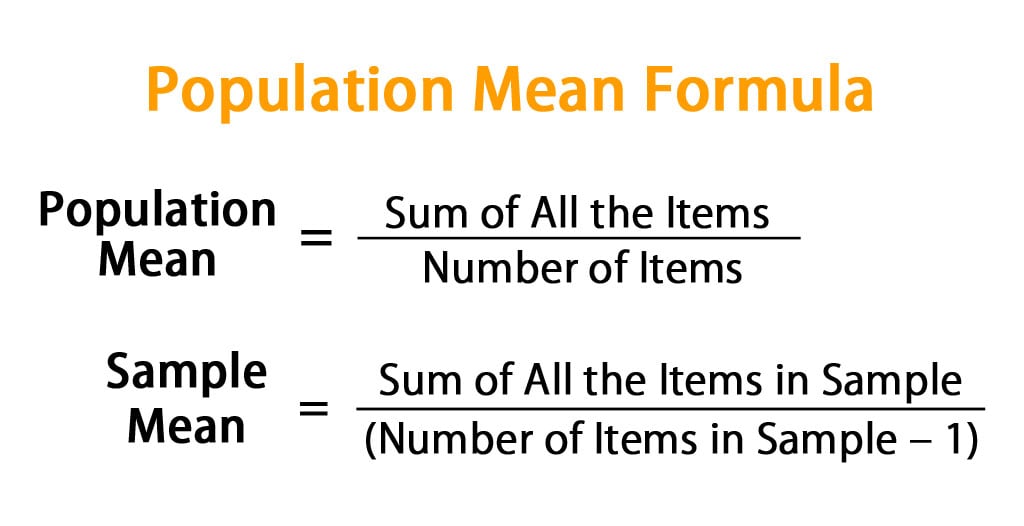
In [statistics](https://en.wikipedia.org/wiki/Statistics), a central tendency (or measure of central tendency) is a central or typical value for a [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution).



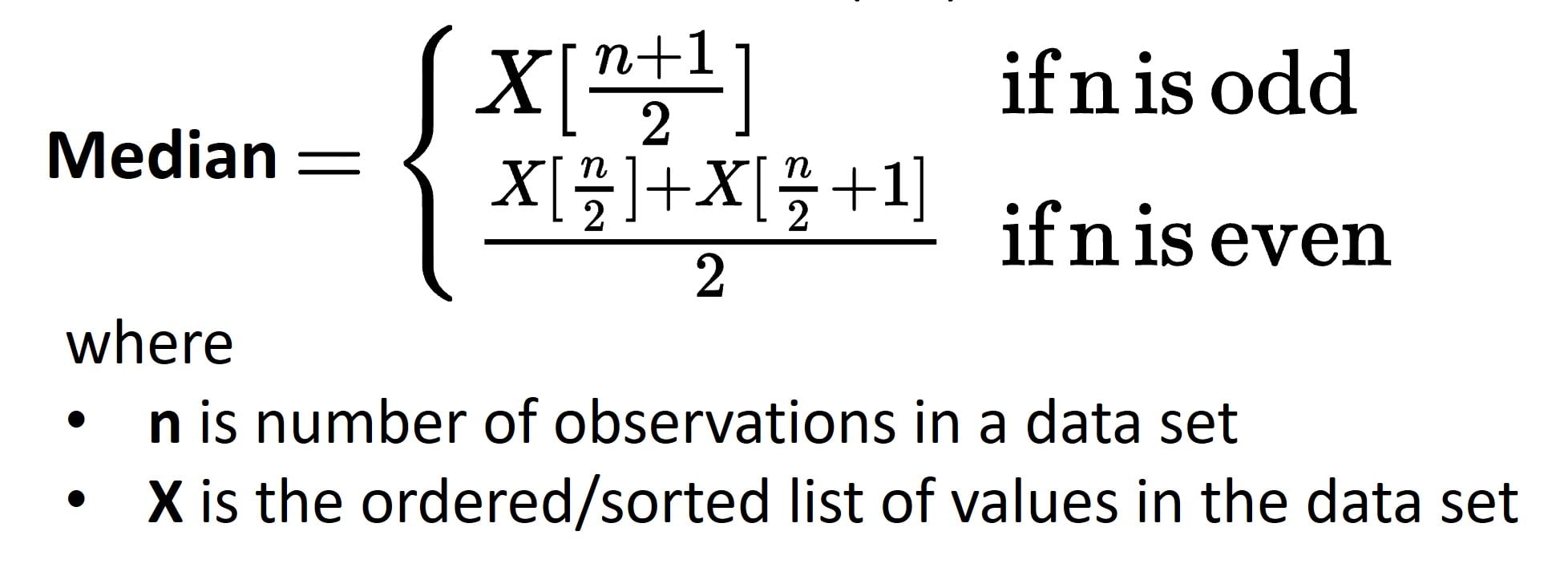
**Mean:**



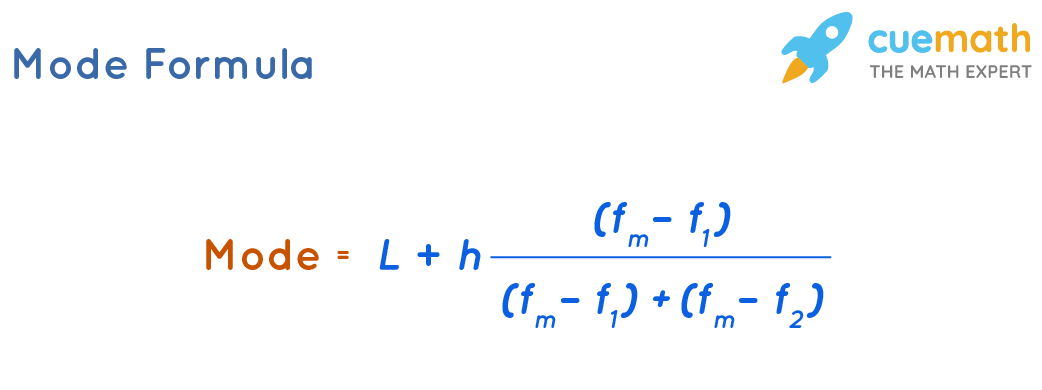




**Median:**

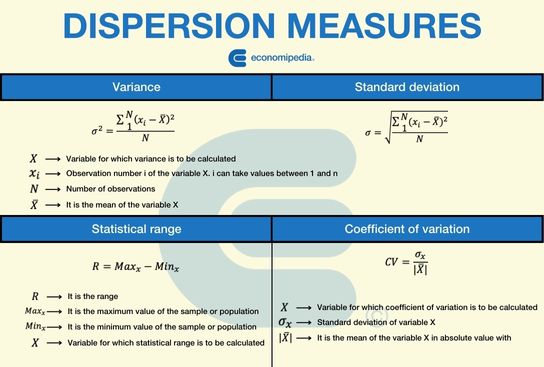


**Mode:**

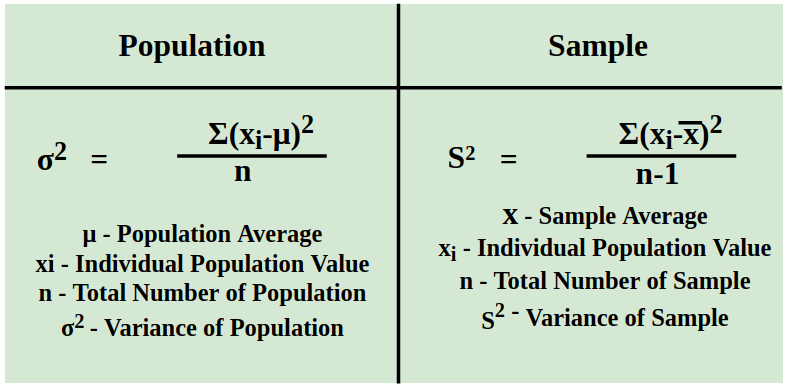


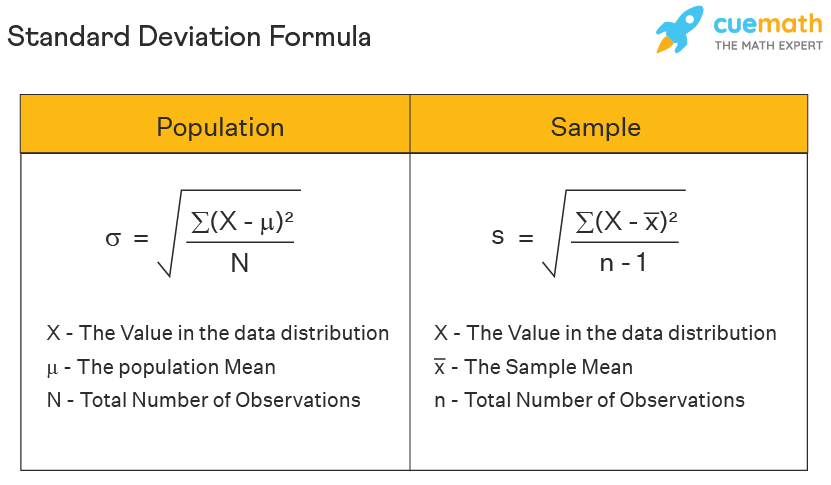
**Measures of Dispersions:** It measures the Spread of Data from Mean

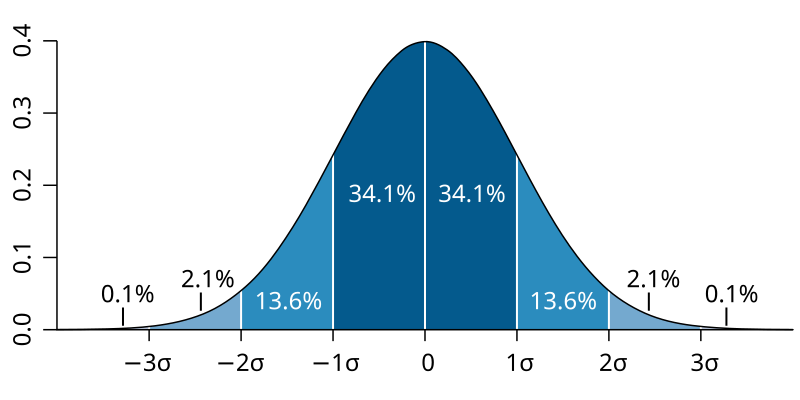
1. Range 🡪 Max-Min
2. Variance 🡪 Spread from mean
3. Standard Deviation 🡪 Square root of Variance



**Variance:**



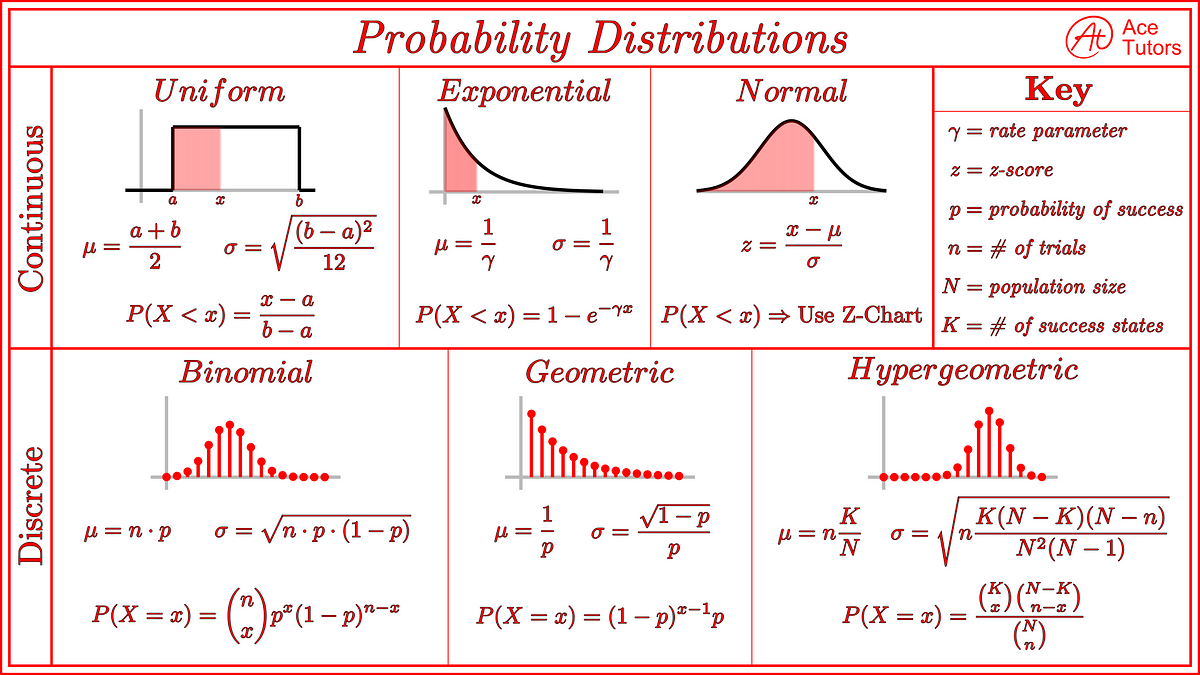
**Standard Deviation:**



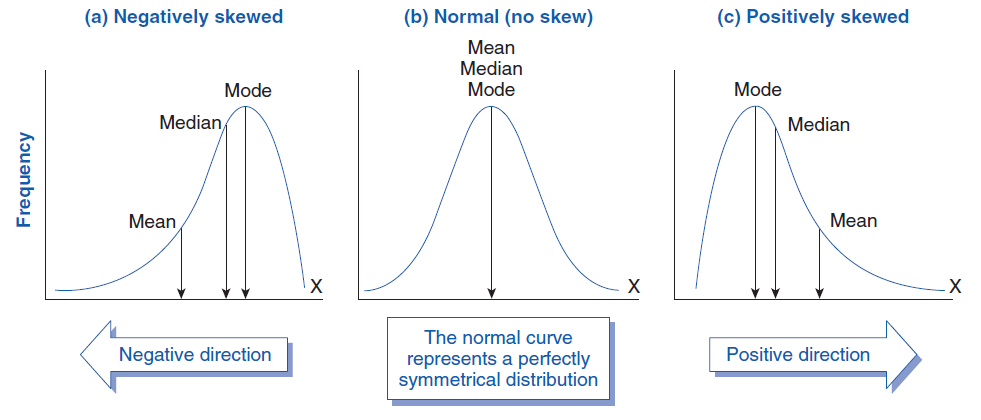
**68-95-99.7 Rule**

**Data Distribution:**

1. Normal Distribution
2. Binomial Distribution
3. Poisson Distribution
4. Exponential Distribution
5. Uniform Distribution
6. Student’s Distribution



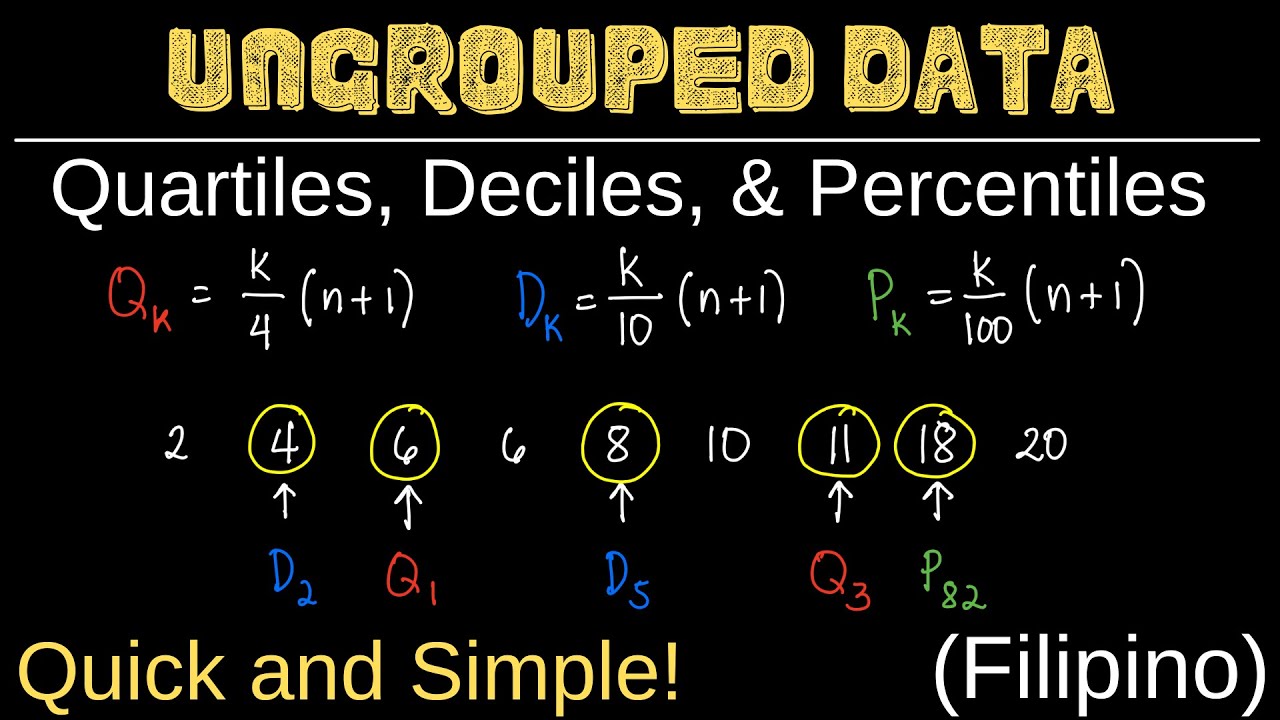
**Skewed Distribution:**



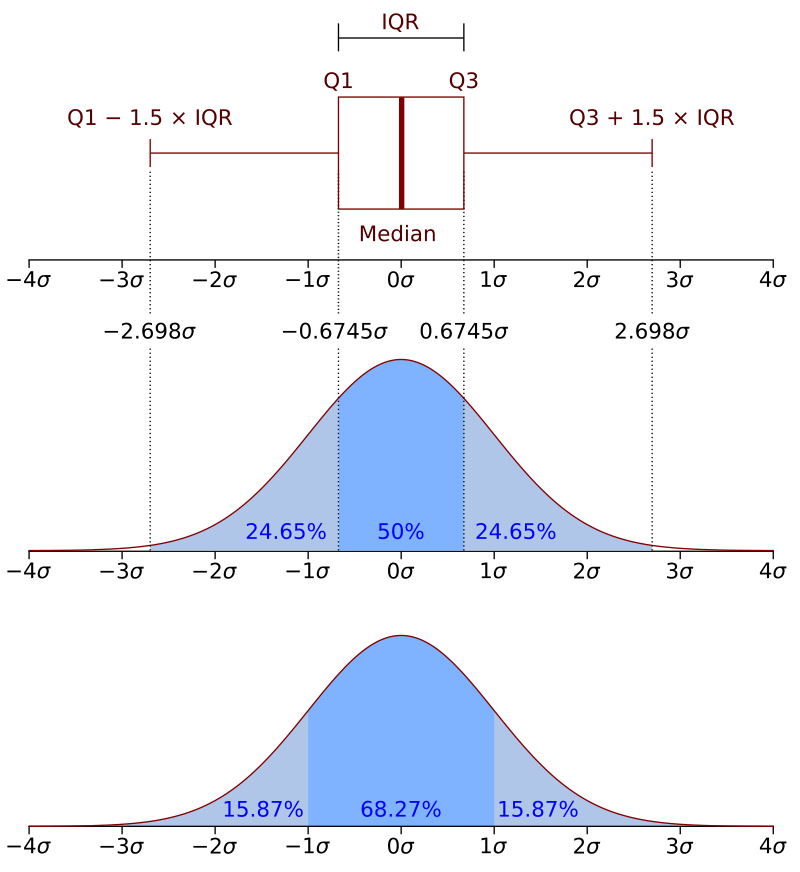
**Quantile/IQR:**

It divides a dataset into equal subset for describe that data and specific portion

1. Quartile 🡪 Q1,Q2,Q3
2. Decile 🡪 D1-D9
3. Percentile 🡪 P1-P99
4. Inter Quartile Range (IQR) 🡪 Q1 – Q3



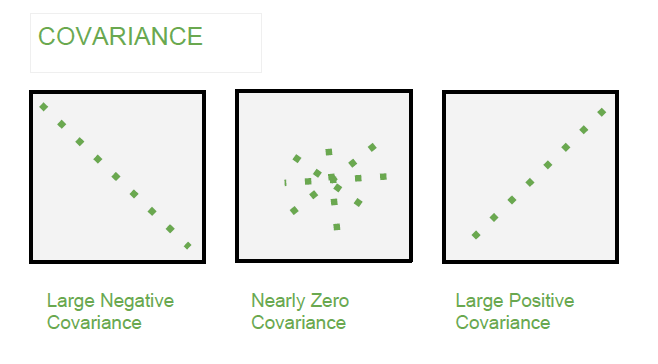
**IQR (Q3-Q1)**



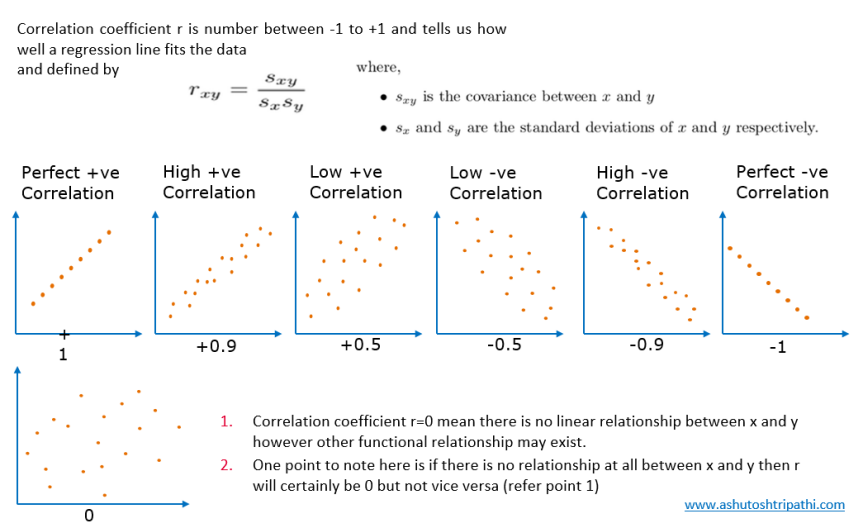
**Covariance & Correlation:**

|  |  |
| --- | --- |
| **Covarience** | **Correlation** |
| https://cdn1.byjus.com/wp-content/uploads/2019/02/Covariance-Formula.png | https://www.postnetwork.co/wp-content/uploads/2022/11/CovCorthumbnail.png |

**Covariance:**



**Correlation:**



**CHAPTER 4:**

**Probability**

**Probability:**

A probability distribution is a statistical function that describes all the possible values and likelihoods that a [random variable](https://www.investopedia.com/terms/r/random-variable.asp) can take within a given range. This range will be bounded between the minimum and maximum possible values. However, where the possible value is likely to be plotted on the probability distribution depends on several factors. These factors include the distribution’s [mean](https://www.investopedia.com/terms/m/mean.asp) (average), [standard deviation](https://www.investopedia.com/terms/s/standarddeviation.asp), [skewness](https://www.investopedia.com/terms/s/skewness.asp), and [kurtosis](https://www.investopedia.com/terms/k/kurtosis.asp).

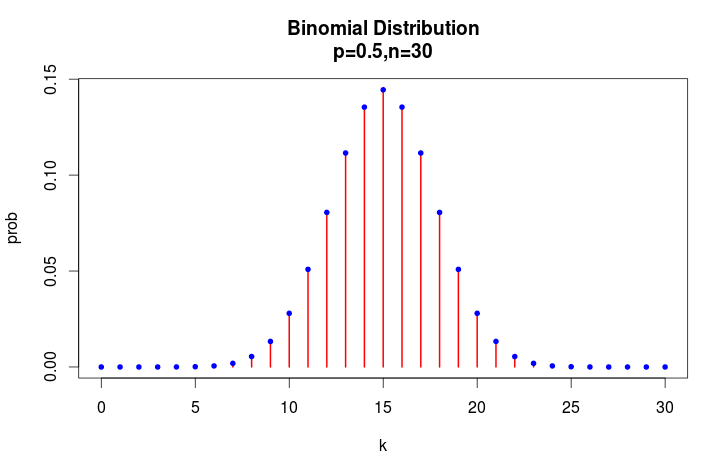
**What is Random Variable in Statistics?**

In probability, a real-valued function, defined over the sample space of a random experiment, is called a random variable. That is, the values of the random variable correspond to the outcomes of the random experiment. Random variables could be either discrete or continuous. In this article, let’s discuss the different types of random variables.

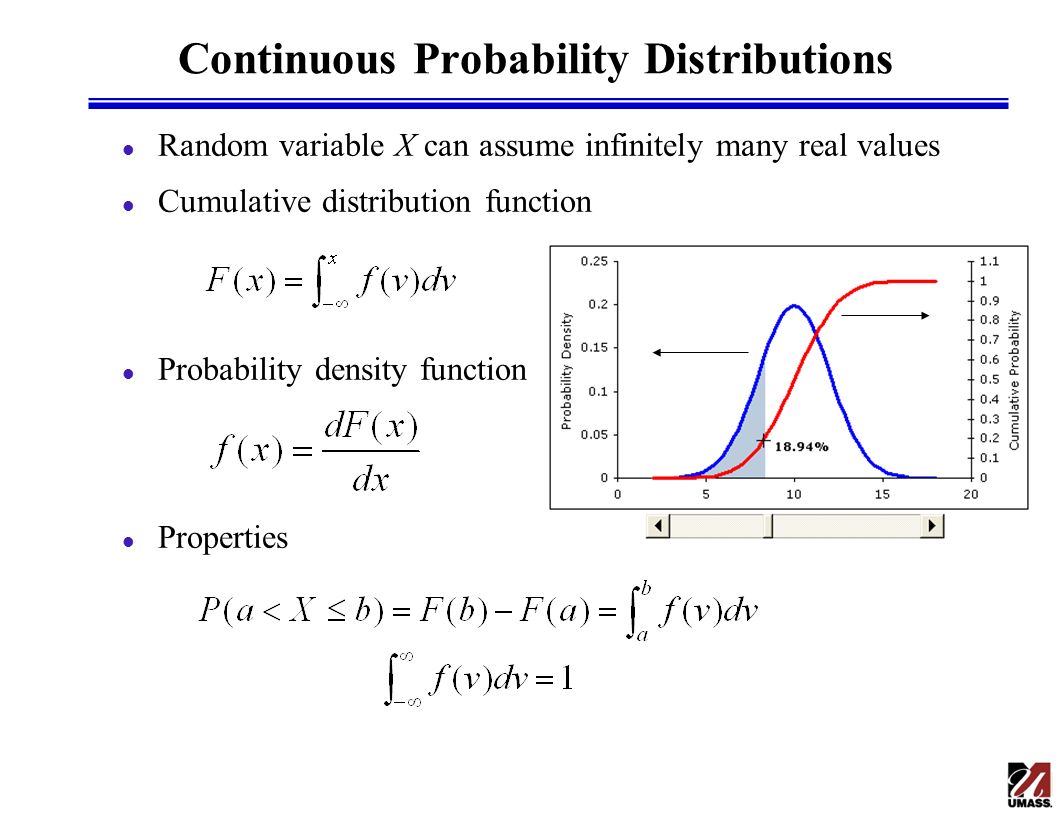
**There are two types of probability distributions:**

* [Discrete probability distributions](https://www.scribbr.com/statistics/probability-distributions/#discrete)
* [Continuous probability distributions](https://www.scribbr.com/statistics/probability-distributions/#continuous)

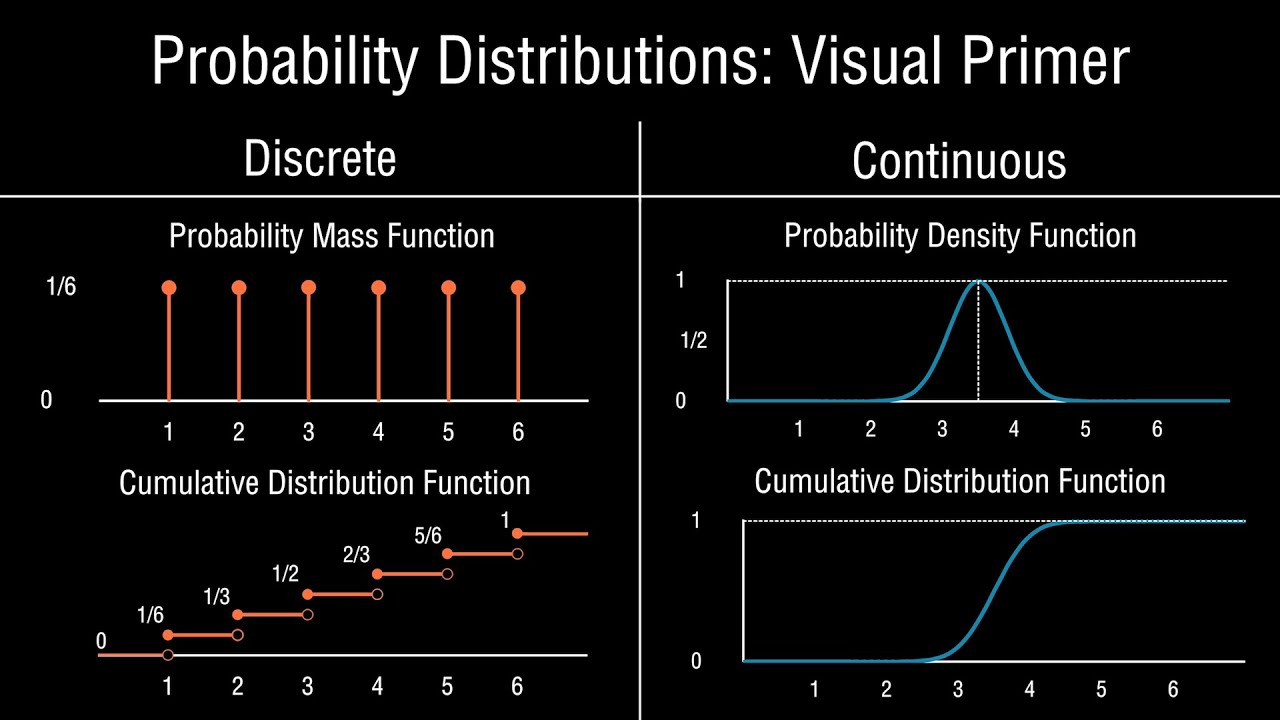
[**Discrete probability distributions**](https://www.scribbr.com/statistics/probability-distributions/#discrete)

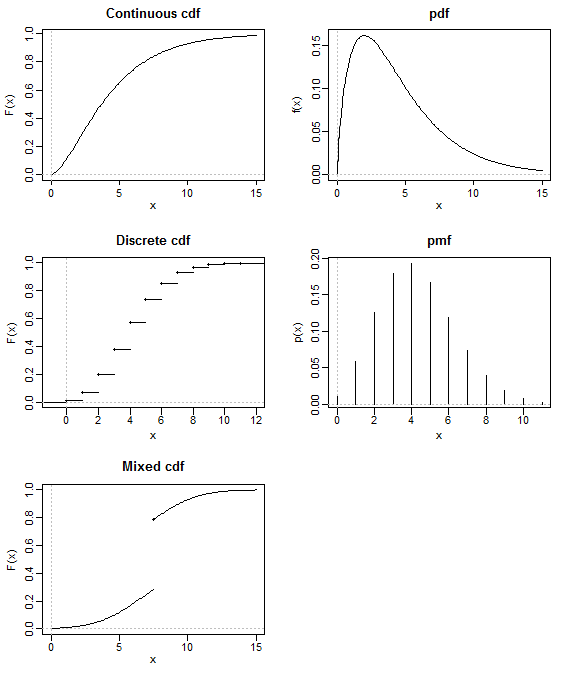


[**Continuous probability distributions**](https://www.scribbr.com/statistics/probability-distributions/#continuous)



**PDF and CDF and PMF**





**CHAPTER 5:**

**Inferential Statistics**

**Inferential Statistics**

After learning the foundational concepts of statistics, such as relationships in data (covariance and correlation), and probability distributions, it’s time to move into Inferential Statistics—a critical branch of statistics used extensively. While descriptive statistics focuses on summarizing and visualizing data, inferential statistics allows us to make predictions, draw conclusions, and generalize insights from a sample to an entire population.

**Parametric and Non-Parametric:**

A parametric test is a statistical test that makes assumptions about the population distribution, typically assuming the data follows a normal distribution, while a non-parametric test does not make such assumptions and is considered "distribution-free," meaning it can be used when the data distribution is unknown or does not meet the requirements for a parametric test; non-parametric tests often rely on ranking data instead of raw values.

**Assumptions about distribution:**

Parametric tests assume the data follows a specific distribution (like normal), while non-parametric tests do not make such assumptions.

**Data types:**

Parametric tests are usually best suited for continuous data, while non-parametric tests can handle ordinal data (ranked data) as well.

**Sensitivity to outliers:**

Parametric tests can be heavily influenced by outliers, whereas non-parametric tests are often more robust to outliers.

**Examples of parametric tests:**

T-test, ANOVA (Analysis of Variance), and Z-test.

**Examples of non-parametric tests:**

Mann-Whitney U test, Wilcoxon signed-rank test, and Kruskal-Wallis test.

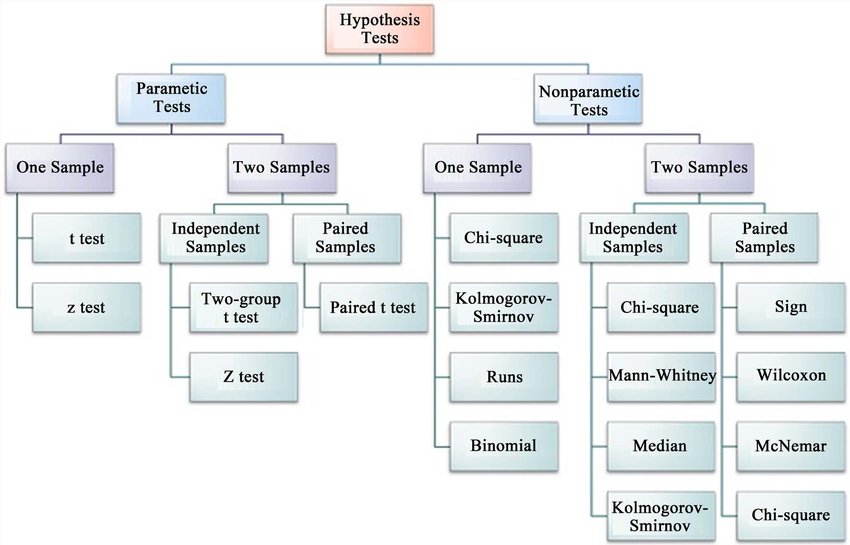
**When to use which test:**

**Use a parametric test:**

When you are confident that your data is normally distributed and you want to analyze the mean.

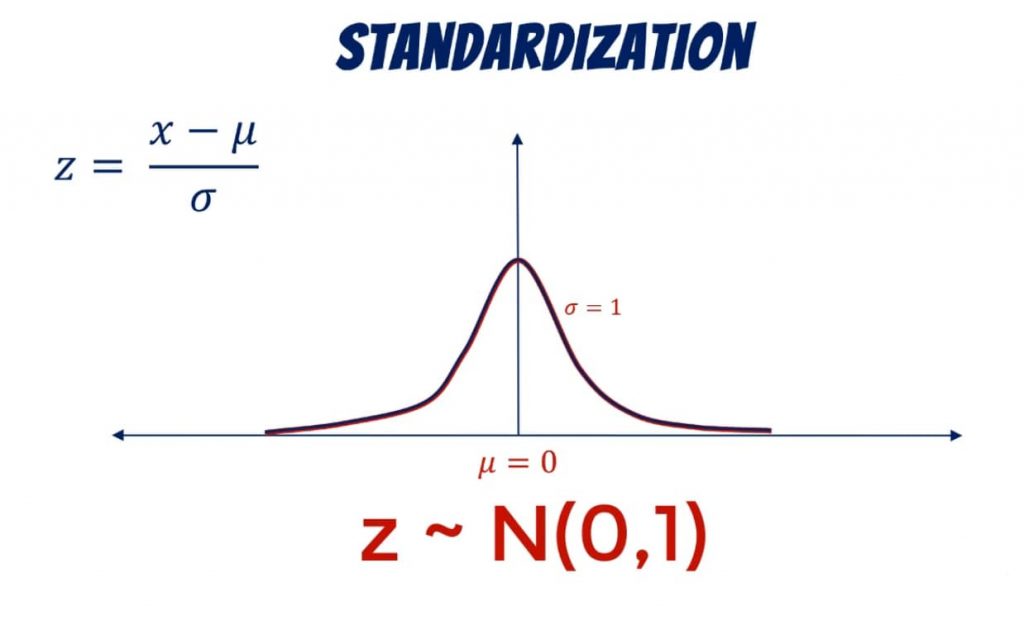
**Use a non-parametric test:**

When the data distribution is unknown, skewed, or has many outliers, or when you are analyzing ordinal data.



**A Standard Normal Variate (SNV):**

A standard normal variate (SNV) is a random variable that follows a normal distribution with a mean of 0 and a standard deviation of 1. It's also a normalization method that transforms data to have these characteristics.



**Transformation:**

In statistics, data transformation is the process of applying a mathematical function to each data point in a set. This is done to improve the data's appearance, interpretability, or to make it more suitable for statistical analysis.

**Examples of data transformations**

**Log transformation**: Used to transform right-skewed data, but only works for positive non-zero data

**Rank transformation**: Replaces data points with their rank when sorted, usually when the value is less important than its order

**Box-Cox transformation**: Transforms non-normal dependent variables into a normal shape

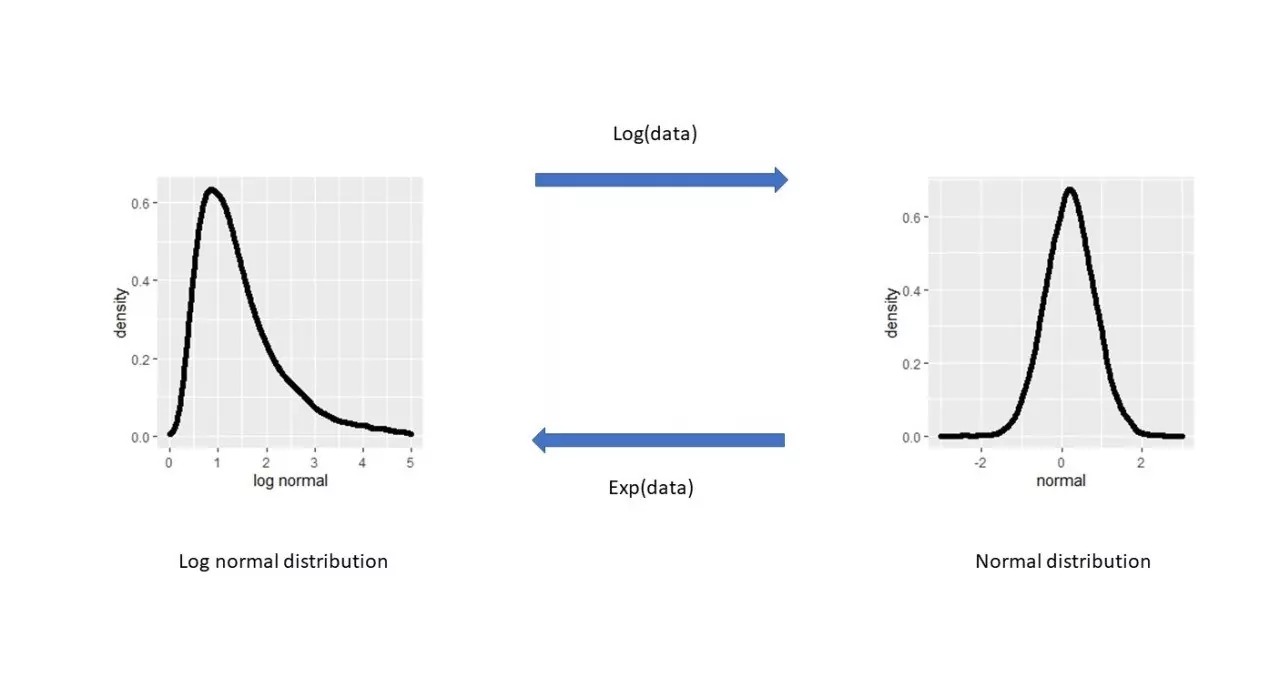
**Reciprocal transformation**: Can help scale down large values or spread out small values in a right-skewed distribution

**Data aggregation**: Combines data at different levels to create new features or attributes

**Data discretization**: Converts data into discrete buckets or intervals, also known as binning

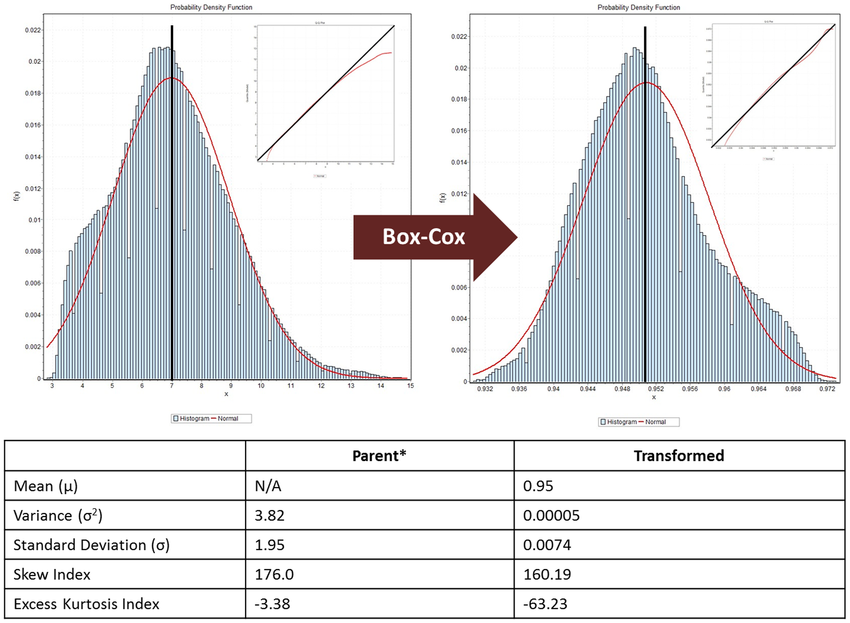
**Frequency transformation**: Measures the time between occurrences of events, such as transactions, claims, or alerts.

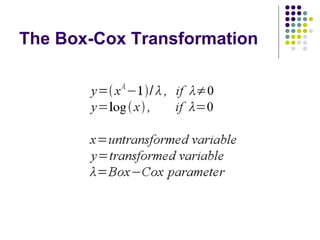
**Log transformation**:



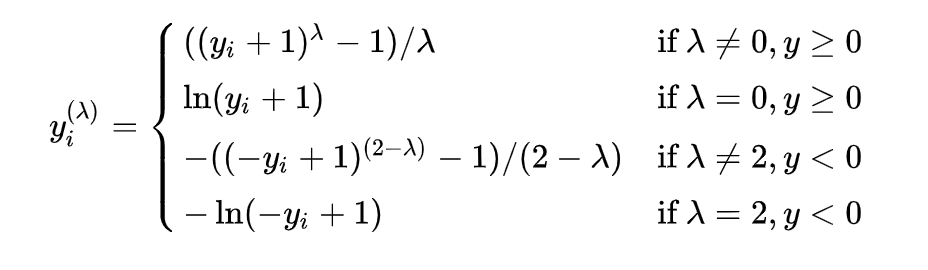
**Rank transformation**:

**Box-Cox transformation**:





**Yeo-Johnson transformation**



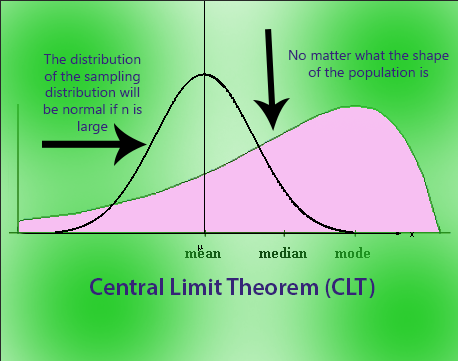
**Reciprocal transformation**:

**Data aggregation**:

**Frequency transformation**:

**Central Limit Theorem (CLT):**

The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution. A sufficiently large sample size can predict the characteristics of a population more accurately.

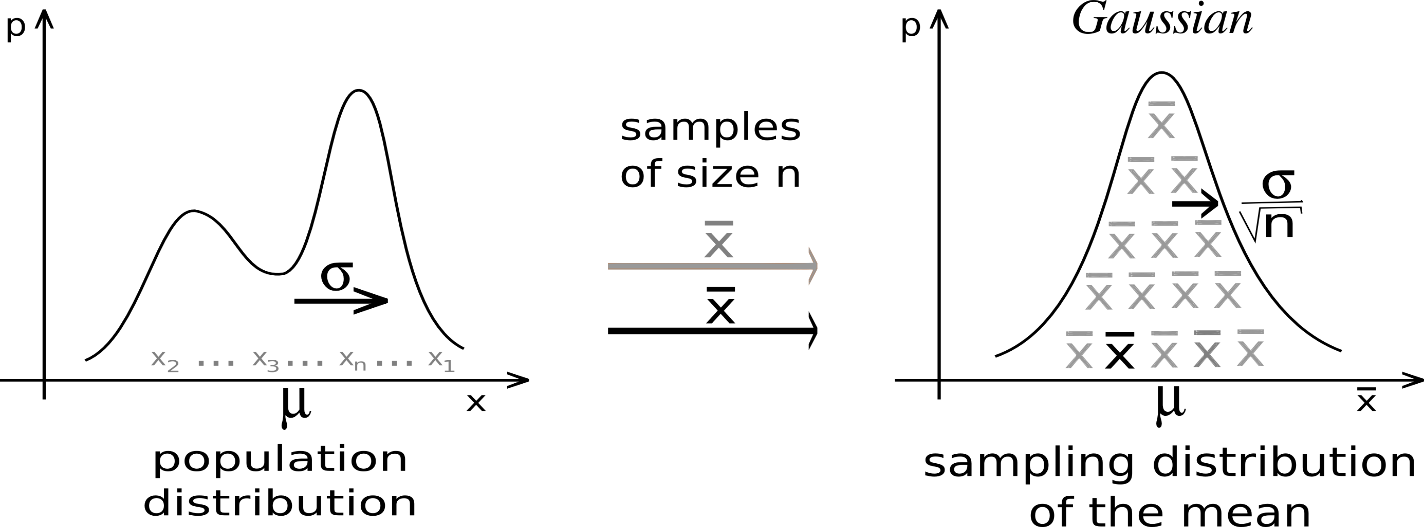


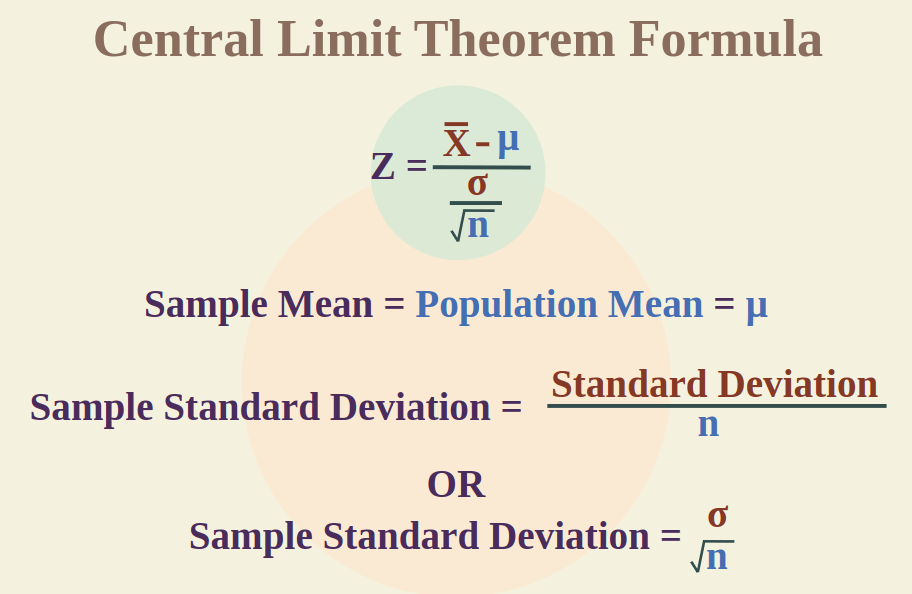
**Key points about the CLT assumptions:**

**Random sampling:** Every member of the population should have an equal opportunity to be included in the sample.

**Independent samples:** One sample should not affect the probability of selecting another sample.

**Large sample size:** Generally, a sample size of 30 or more is considered sufficient for the CLT to hold.





**Parameter:**

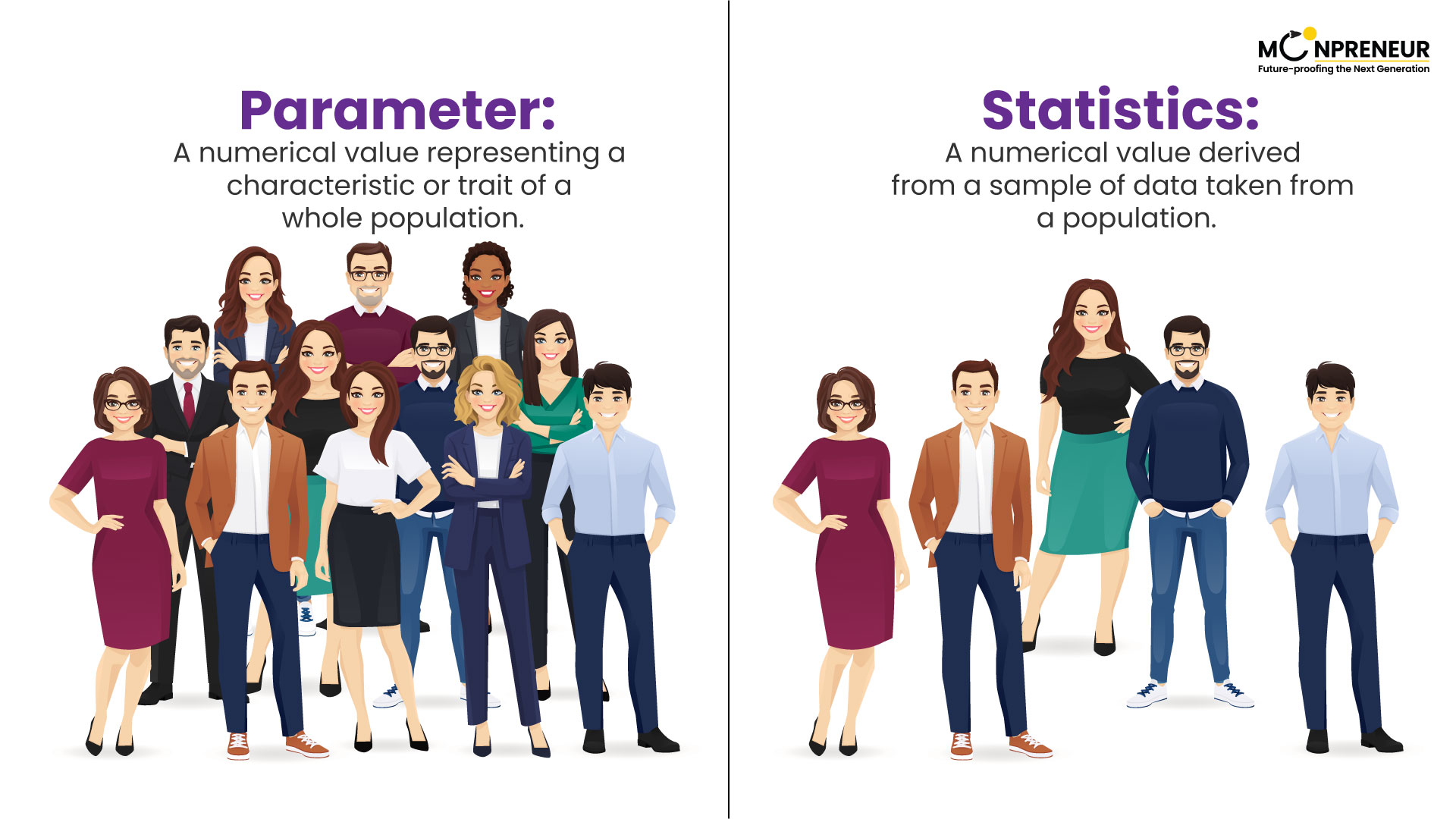
A parameter is a useful component of statistical analysis. It refers to the characteristics that are used to define a given population. It is used to describe a specific characteristic of the entire population.

**Estimate/Statistic:**

A **statistic**is a number describing a [sample](https://www.scribbr.com/methodology/population-vs-sample/) (e.g., sample mean).

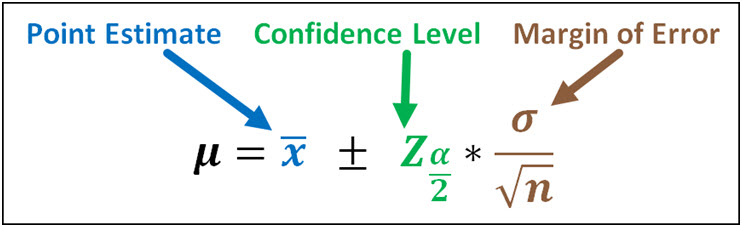
**Point Estimate:**

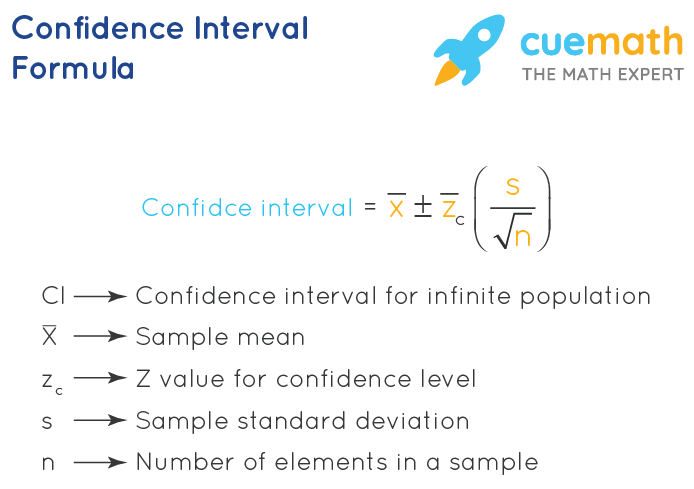
In [statistics](https://en.wikipedia.org/wiki/Statistics), point estimation involves the use of [sample](https://en.wikipedia.org/wiki/Statistical_sample) [data](https://en.wikipedia.org/wiki/Data) to calculate a single value (known as a point estimate since it identifies a [point](https://en.wikipedia.org/wiki/Point_(geometry)) in some [parameter space](https://en.wikipedia.org/wiki/Parameter_space)) which is to serve as a "best guess" or "best estimate" of an unknown population [parameter](https://en.wikipedia.org/wiki/Parameter) (for example, the [population mean](https://en.wikipedia.org/wiki/Population_mean)). More formally, it is the application of a point [estimator](https://en.wikipedia.org/wiki/Estimator) to the data to obtain a point estimate.

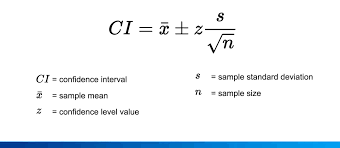


**Confidence Interval (CI):**

A confidence interval is the mean of your estimate plus and minus the variation in that estimate. This is the range of values you expect your estimate to fall between if you redo your test, within a certain level of confidence. Confidence, in statistics, is another way to describe probability.







**Z-Procedure (Sigma Known):**

Z-procedures in statistics are a set of techniques that use z-scores to compare data points, test hypotheses, and make inferences about populations.

**Z-score**

A z-score is a number that measures how many standard deviations a data point is from the mean of a group of values.

A z-score can be positive or negative, with positive values indicating the score is above the mean and negative values indicating it is below the mean.

The formula for calculating a z-score is z=(x−μ)/σz equals open paren x minus mu close paren / sigma

𝑧=(𝑥−𝜇)/𝜎

.

**Z-test**

A z-test is a statistical test that compares the mean of a sample to the mean of a population.

Z-tests are used to determine if there are statistically significant differences between two populations.

Z-tests are commonly used in fields like healthcare and data science.

Z-tests can be used when the standard deviation of the population is known, or when the sample size is large enough to use the sample variance as an estimate.

****

**Z-Test: Definition, Uses in Statistics, and Example - Investopedia**

**A z-score, or z-statistic, is a number representing how many standard deviations above or below the mean population the score deri...**

**Investopedia**

****

**Z-Test - Numeracy, Maths and Statistics - Academic Skills Kit**

**z -tests are a statistical way of testing a hypothesis, when we know the population variance σ2 . We use them when we wish to comp...**

**Numeracy, Maths and Statistics - Academic Skills Kit**

****

**z-Score: Definition, Formula, Calculation & Interpretation**

**Z-standardization is a statistical procedure used to make data points from different datasets comparable. In this procedure, each ...**

**Error! Filename not specified.**

**Datatab**

****

**Z-Test for Statistical Hypothesis Testing Explained | Built In**

**Oct 16, 2024 — A Z-test determines whether there are any statistically significant differences between the means of two populations.**

**Error! Filename not specified.**

**Built In**

****

**Z-test : Formula, Types, Examples - GeeksforGeeks**

**3 days ago — After learning about inferential statistics we now move on to a more specific technique used for making decisions based ...**

**Error! Filename not specified.**

**GeeksforGeeks**

****

**Z-Score: Meaning and Formula - Investopedia**

**The statistical formula for a value's z-score is calculated using the following formula: z = ( x - μ ) / σ Where:**

**Investopedia**

****

**Z-Test | Definition, Formula & Example - Lesson - Study.com**

**A Z-test is a statistical test that is used to determine whether or not two populations are statistically different from one anoth...**

**Error! Filename not specified.**

**Study.com**

**Z-interval**

A z-interval is a type of confidence interval that provides a range where a particular mean or proportion is expected to fall.

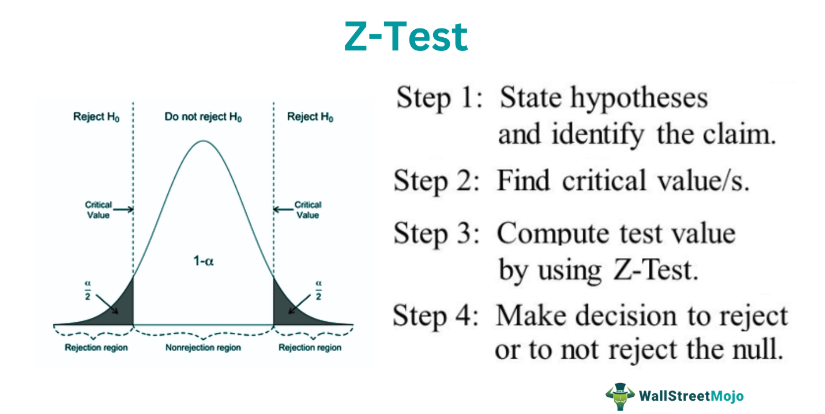
Z-intervals can be calculated from a known standard deviation.

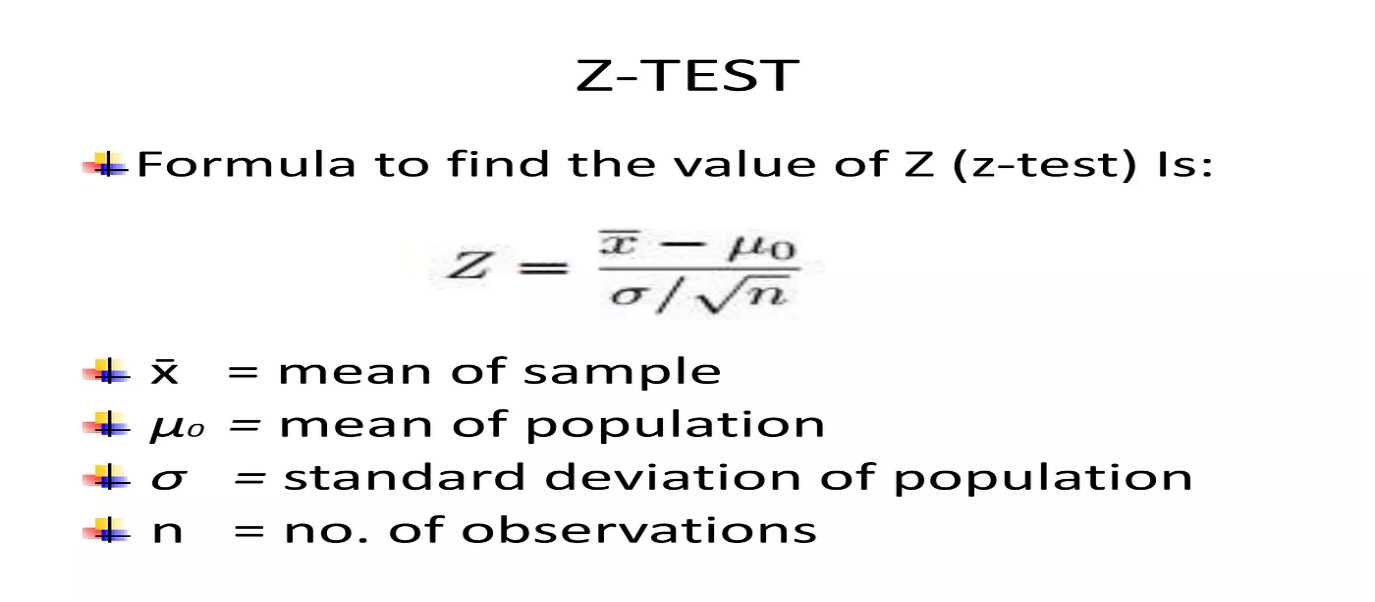
**Assumption:**

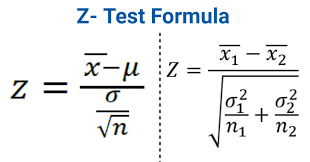
**Normal distribution:** The data is assumed to be normally distributed in the population.

**Known population standard deviation:** The standard deviation of the population must be known to use a z-test.

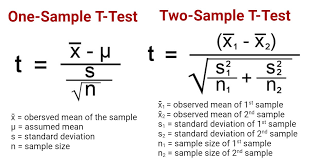
**Random sampling:** The data should be collected through a simple random sample from the population.







**T-Procedure:**



A t-test is a statistical procedure that compares the means of two or more groups to determine if there is a significant difference between them. It is also known as Student's t-test, t-statistic, or t-distribution.

**When to use a t-test**

When comparing the means of two groups, such as patients who received different treatments

When comparing the mean of a group to a known value

When comparing paired measurements

When there are a small number of sample observations (less than 30)

**How to perform a t-test?**

Define the null and alternative hypotheses

Input the mean and standard deviation values into a formula

Calculate the t-value, which measures the difference between the group means

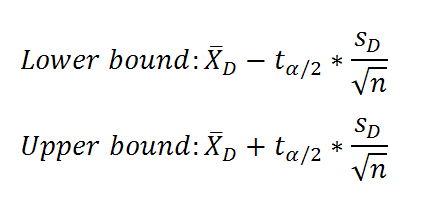
Calculate the p-value, which estimates the likelihood that the difference is due to chance

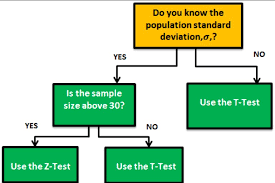
**Types of t-tests**

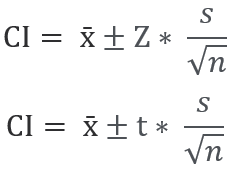
**One-sample t-test**: Compares a single group to a known value

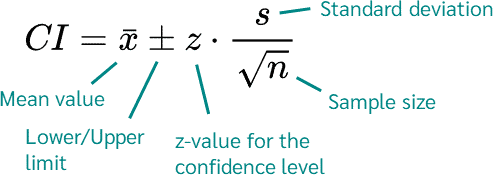
**Independent two-sample t-test**: Compares two groups to each other

**Paired t-test**: Compares paired measurements









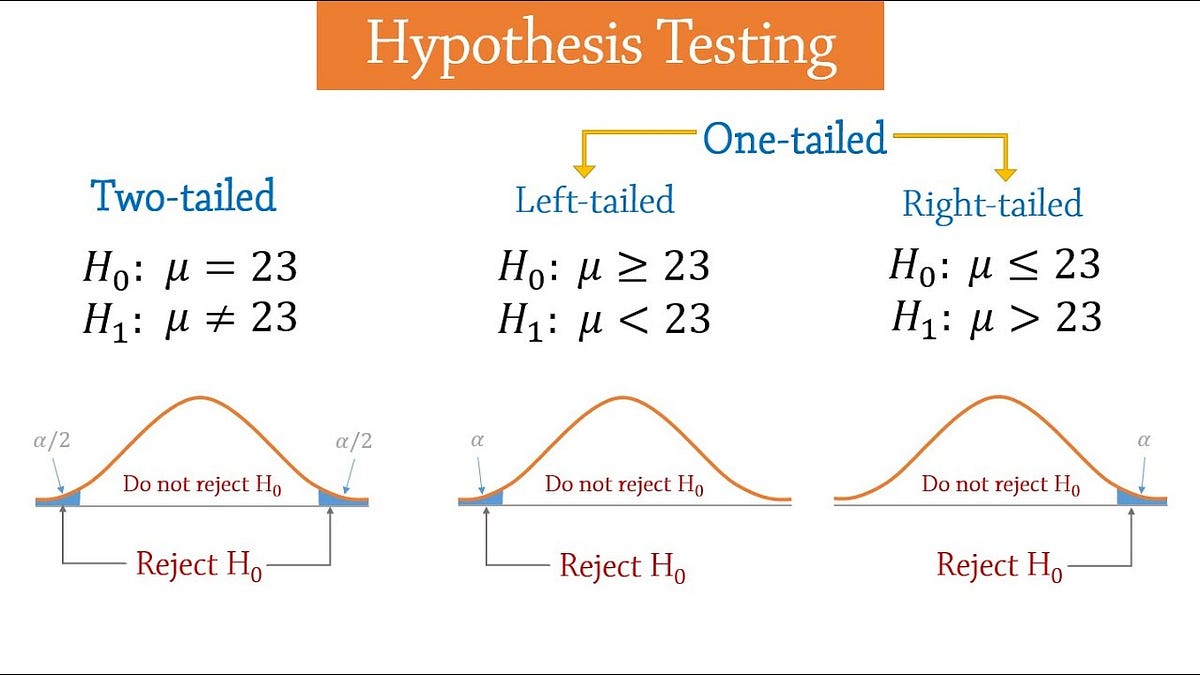
**Hypothesis:**

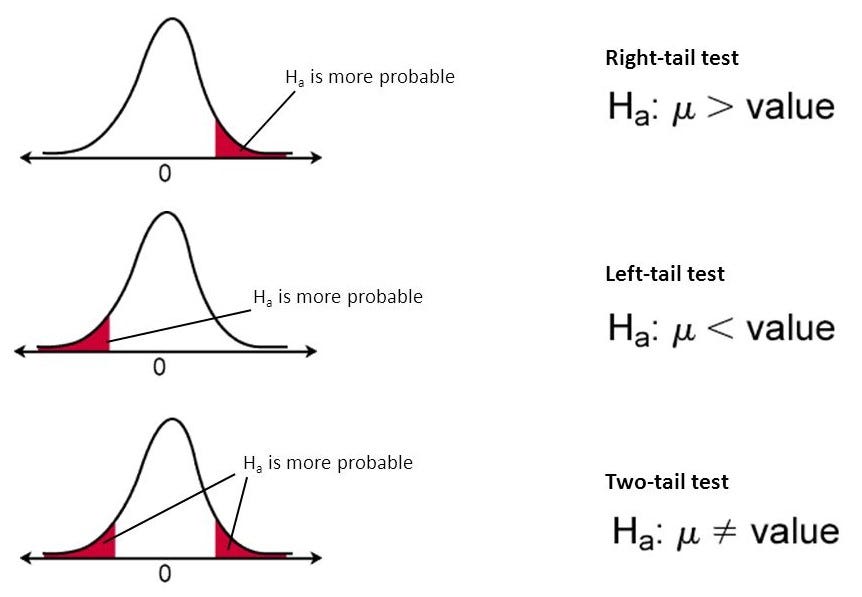
Hypothesis testing is a structured method used to determine if the findings of a study provide evidence to support a specific theory relevant to a larger population.

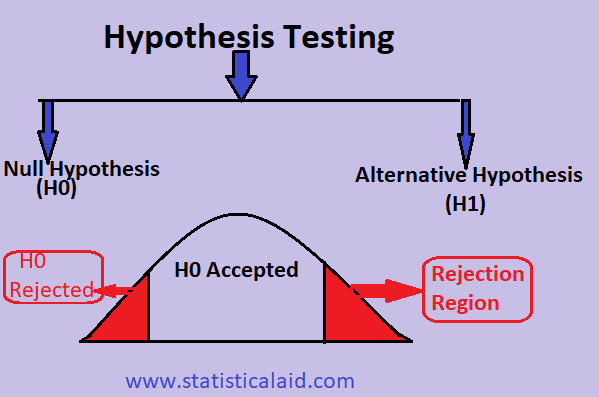
Hypothesis Testing is a type of [statistical analysis](https://www.simplilearn.com/what-is-statistical-analysis-article) in which you put your assumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables.

**Null Hypothesis H0**

**Alternative Hypothesis H1 or HA:**







# **CHAPTER 5:**

**Conclusion**