



Green University

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:

- This Note is for Statistics for Machine Learning.

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 5:

Inferential Statistics:

CHAPTER 6:

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.

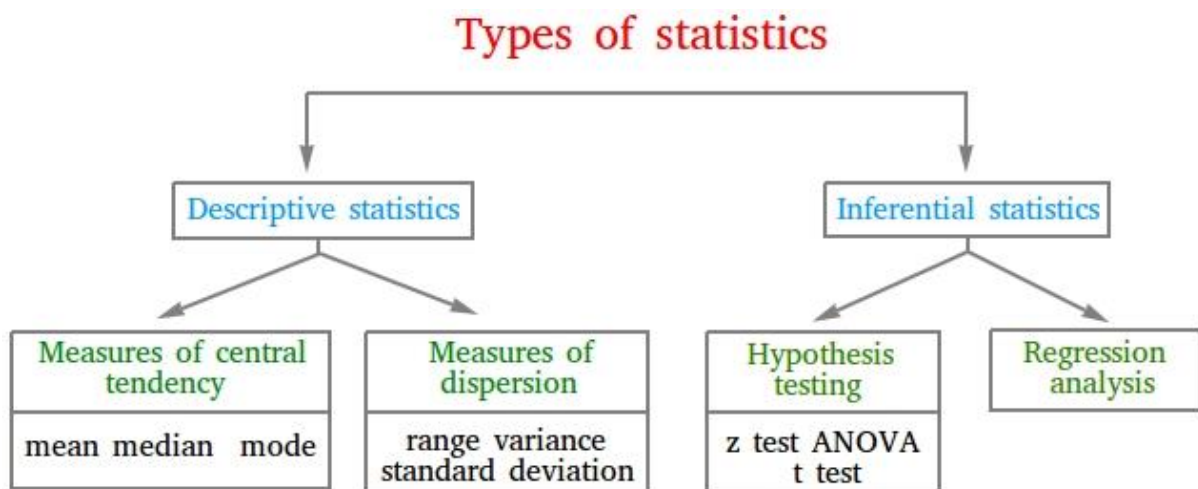
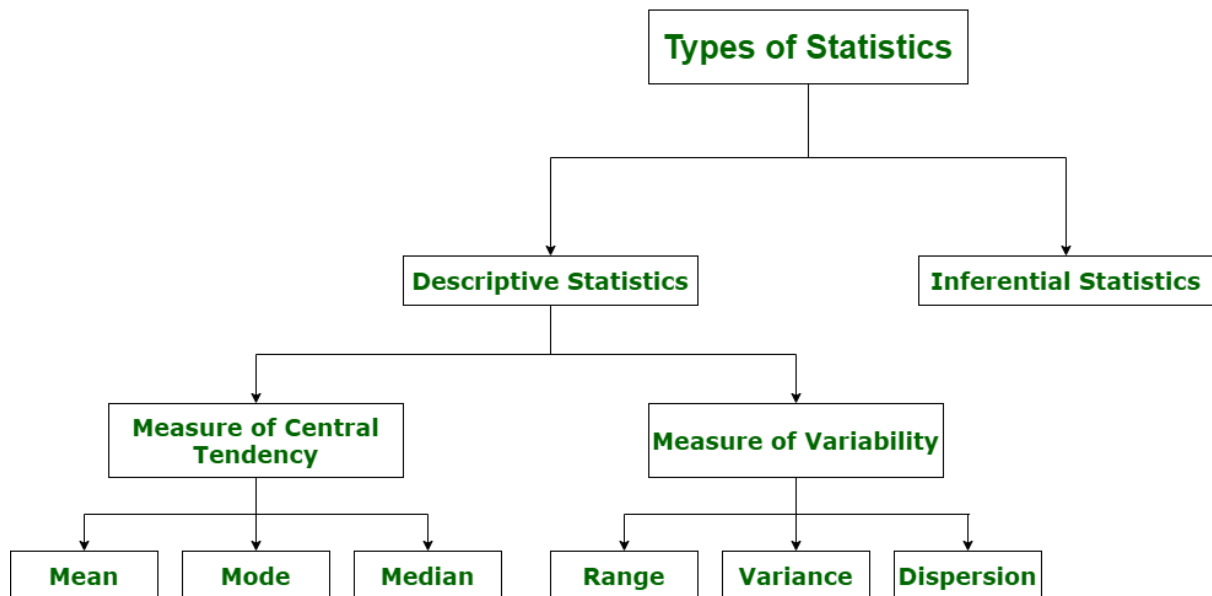
Applications of Statistics in Machine Learning:

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

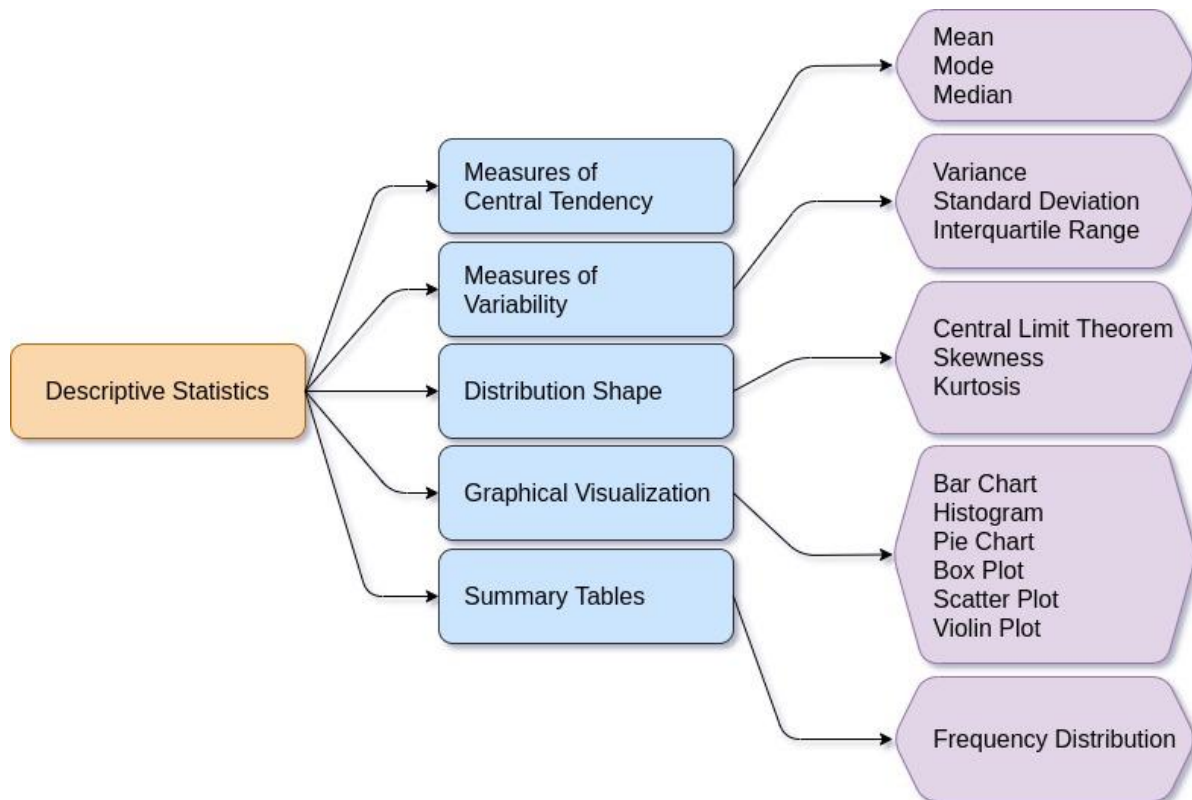
- Feature engineering relies heavily on statistics to convert geometric features into meaningful predictors for machine learning algorithms.
- In image processing tasks like object recognition and segmentation, statistics accurately reflect the shape and structure of objects in images.
- Anomaly detection 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:

- I. Measures of Central Tendency
 - a. Mean, Median, Mode
- II. Measures of Variability
 - a. Range, Variance, Standard Deviation, IQR

Central Tendency:

In statistics, a central tendency (or measure of central tendency) is a central or typical value for a probability distribution.

| Central Tendency Measures | | |
|---------------------------|------------------|---------------------------|
| Measure | Formula | Description |
| Mean | $\sum x/n$ | Balance Point |
| Median | $n+1/2$ Position | Middle Value when ordered |
| Mode | None | Most frequent |

Mean:

Arithmetic Mean

$$\bar{x} = \frac{1}{n} \left(\sum_{i=1}^n x_i \right) = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Geometric Mean

$$\left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}} = \sqrt[n]{x_1 * x_2 * \dots * x_n}$$

Harmonic Mean

$$\left(\frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \right) = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

Population Mean

$$\mu = \frac{\sum x}{N}$$

Sample Mean

$$\bar{X} = \frac{\sum x}{n}$$

Population Mean Formula

$$\text{Population Mean} = \frac{\text{Sum of All the Items}}{\text{Number of Items}}$$

$$\text{Sample Mean} = \frac{\text{Sum of All the Items in Sample}}{(\text{Number of Items in Sample} - 1)}$$

Median:

$$\text{Median} = \begin{cases} X\left[\frac{n+1}{2}\right] & \text{if } n \text{ is odd} \\ \frac{X\left[\frac{n}{2}\right] + X\left[\frac{n}{2} + 1\right]}{2} & \text{if } n \text{ is even} \end{cases}$$

where

- **n** is number of observations in a data set
- **X** is the ordered/sorted list of values in the data set

Mode:

Mode Formula



$$\text{Mode} = L + h \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)}$$

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

- a) Range \rightarrow Max-Min
- b) Variance \rightarrow Spread from mean
- c) Standard Deviation \rightarrow Square root of Variance

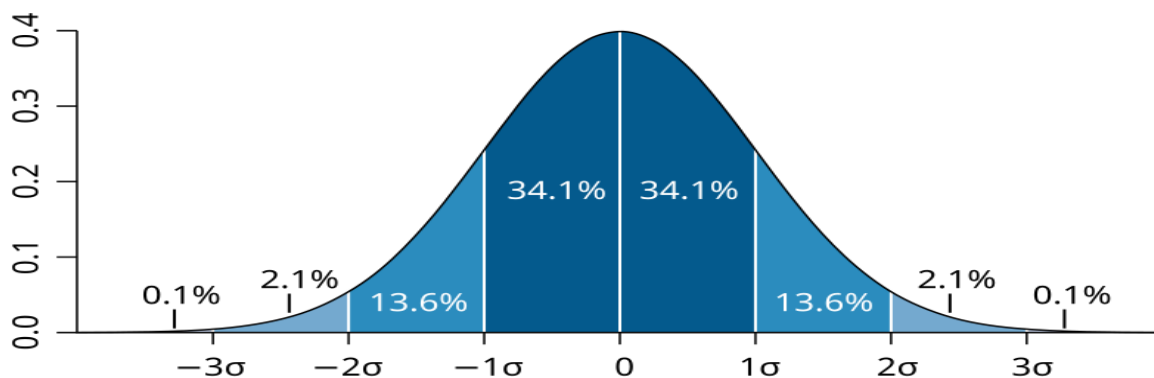
| DISPERSION MEASURES | |
|---|---|
| economipedia. | |
| Variance | Standard deviation |
| $\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2}{N}$ <p> $X \rightarrow$ Variable for which variance is to be calculated $x_i \rightarrow$ Observation number i of the variable X. i can take values between 1 and n $N \rightarrow$ Number of observations $\bar{X} \rightarrow$ It is the mean of the variable X </p> | $\sigma = \sqrt{\frac{\sum_1^N (x_i - \bar{X})^2}{N}}$ |
| Statistical range | Coefficient of variation |
| $R = \text{Max}_x - \text{Min}_x$ <p> $R \rightarrow$ It is the range $\text{Max}_x \rightarrow$ It is the maximum value of the sample or population $\text{Min}_x \rightarrow$ It is the minimum value of the sample or population $X \rightarrow$ Variable for which statistical range is to be calculated </p> | $CV = \frac{\sigma_x}{ \bar{X} }$ <p> $X \rightarrow$ Variable for which coefficient of variation is to be calculated $\sigma_x \rightarrow$ Standard deviation of variable X $\bar{X} \rightarrow$ It is the mean of the variable X in absolute value with </p> |

Variance:

| Population | Sample |
|---|---|
| $\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$ <p> μ - Population Average x_i - Individual Population Value n - Total Number of Population σ^2 - Variance of Population </p> | $S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ <p> \bar{x} - Sample Average x_i - Individual Population Value n - Total Number of Sample S^2 - Variance of Sample </p> |

Standard Deviation:



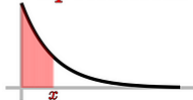


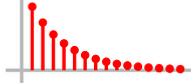
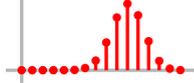
| Population | Sample |
|--|---|
| $\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$ <p> X - The Value in the data distribution μ - The population Mean N - Total Number of Observations </p> | $s = \sqrt{\frac{\sum (X - \bar{x})^2}{n - 1}}$ <p> X - The Value in the data distribution \bar{x} - The Sample Mean n - Total Number of Observations </p> |



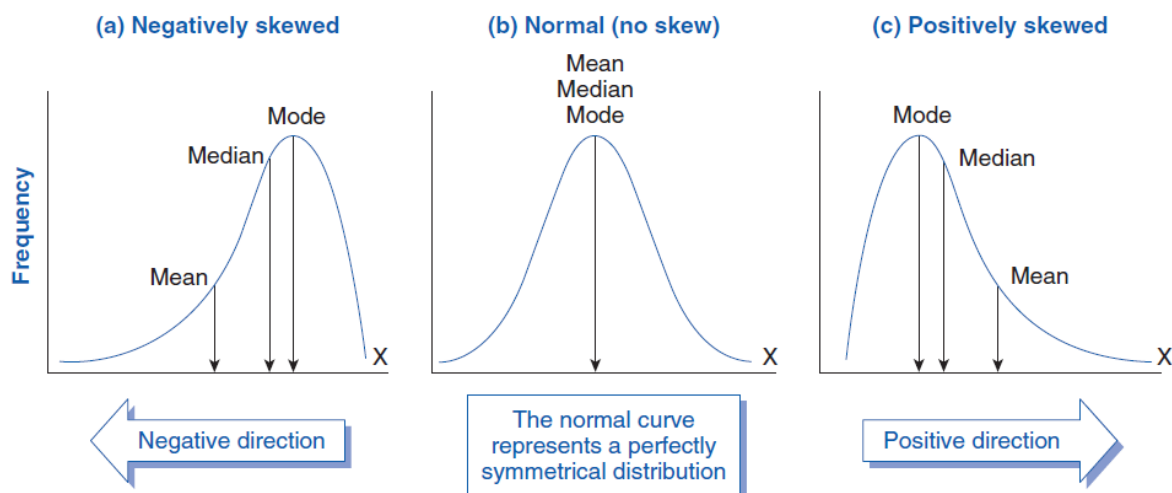
68-95-99.7 Rule

Data Distribution:

- Normal Distribution
- Binomial Distribution
- Poisson Distribution
- Exponential Distribution
- Uniform Distribution
- Student's Distribution

| Probability Distributions | | | | |  |
|---------------------------|--|---|---|--|---|
| Continuous | Uniform  $\mu = \frac{a+b}{2} \quad \sigma = \sqrt{\frac{(b-a)^2}{12}}$ $P(X < x) = \frac{x-a}{b-a}$ | Exponential  $\mu = \frac{1}{\gamma} \quad \sigma = \frac{1}{\gamma}$ $P(X < x) = 1 - e^{-\gamma x}$ | Normal  $z = \frac{x - \mu}{\sigma}$ $P(X < x) \Rightarrow \text{Use Z-Chart}$ | Key γ = rate parameter z = z-score p = probability of success n = # of trials N = population size K = # of success states | |
| | Discrete | Binomial  $\mu = n \cdot p \quad \sigma = \sqrt{n \cdot p \cdot (1-p)}$ $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ | Geometric  $\mu = \frac{1}{p} \quad \sigma = \frac{\sqrt{1-p}}{p}$ $P(X = x) = (1-p)^{x-1} p$ | Hypergeometric  $\mu = n \frac{K}{N} \quad \sigma = \sqrt{n \frac{K(N-K)(N-n)}{N^2(N-1)}}$ $P(X = x) = \frac{\binom{K}{x} \binom{N-K}{n-x}}{\binom{N}{n}}$ | |

Skewed Distribution:



Quantile/IQR:

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

- a) Quartile $\rightarrow Q_1, Q_2, Q_3$
- b) Decile $\rightarrow D_1-D_9$
- c) Percentile $\rightarrow P_1-P_{99}$
- d) Inter Quartile Range (IQR) $\rightarrow Q_1 - Q_3$

UNGROUPED DATA

Quartiles, Deciles, & Percentiles

$$Q_k = \frac{k}{4} (n+1)$$

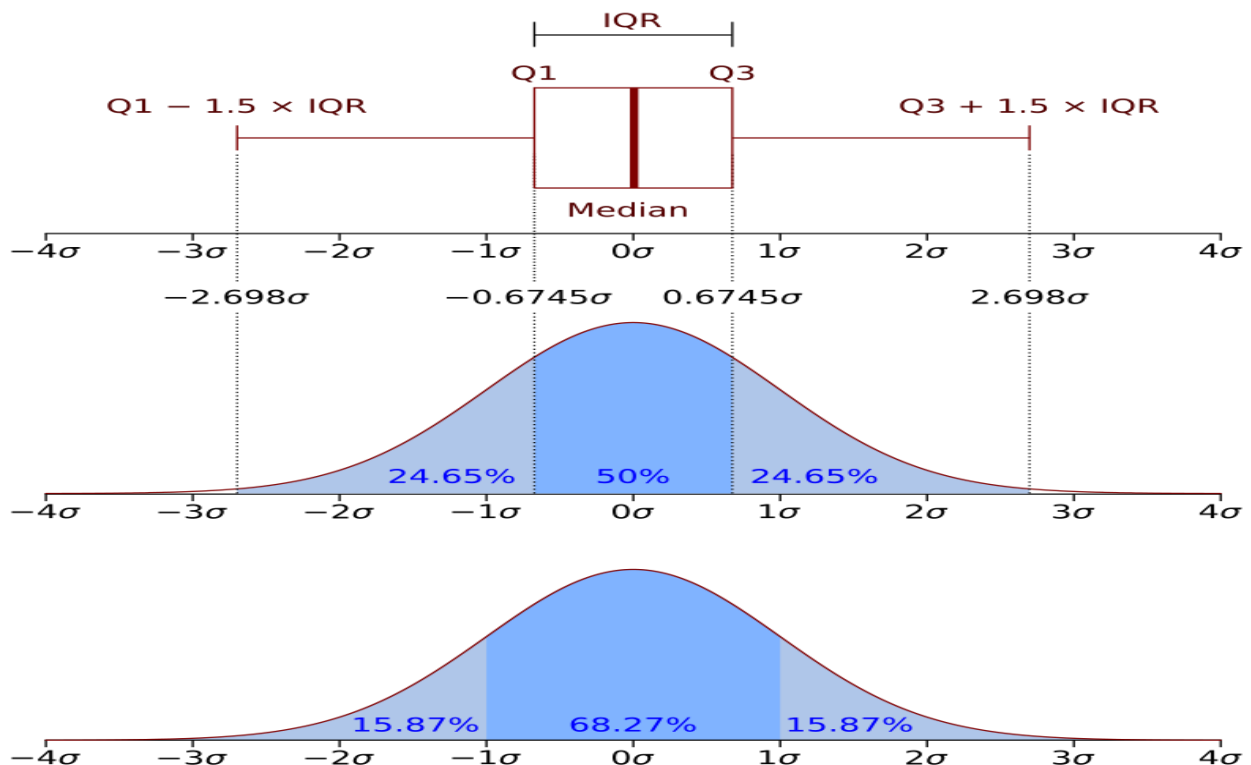
$$D_k = \frac{k}{10} (n+1)$$

$$P_k = \frac{k}{100} (n+1)$$

2
4
↑
 D_2
6
↑
 Q_1
6
8
↑
 D_5
10
11
↑
 Q_3
18
↑
 P_{82}
20

Quick and Simple!
(Filipino)

IQR ($Q_3 - Q_1$)



Covariance & Correlation:

Covariance

Population Covariance Formula

$$\text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample Covariance

$$\text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Correlation

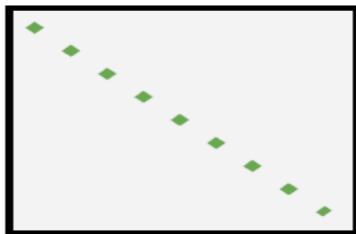
$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$$

Where

$$\sigma_x = \frac{\sqrt{\sum (x_i - \bar{x})^2}}{\sqrt{n-1}}$$

$$\sigma_y = \frac{\sqrt{\sum (y_i - \bar{y})^2}}{\sqrt{n-1}}$$

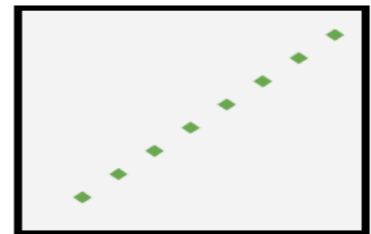
Covariance:



Large Negative Covariance



Nearly Zero Covariance



Large Positive Covariance

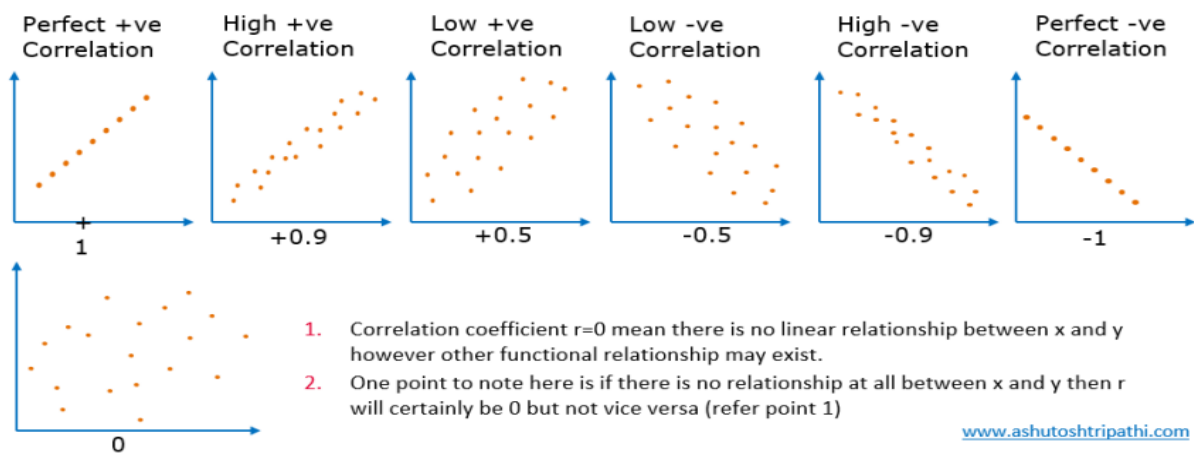
Correlation:

Correlation coefficient r is number between -1 to +1 and tells us how well a regression line fits the data and defined by

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where,

- s_{xy} is the covariance between x and y
- s_x and s_y are the standard deviations of x and y respectively.



CHAPTER 4: **Probability**

Probability:

A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable 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 (average), standard deviation, skewness, and kurtosis.

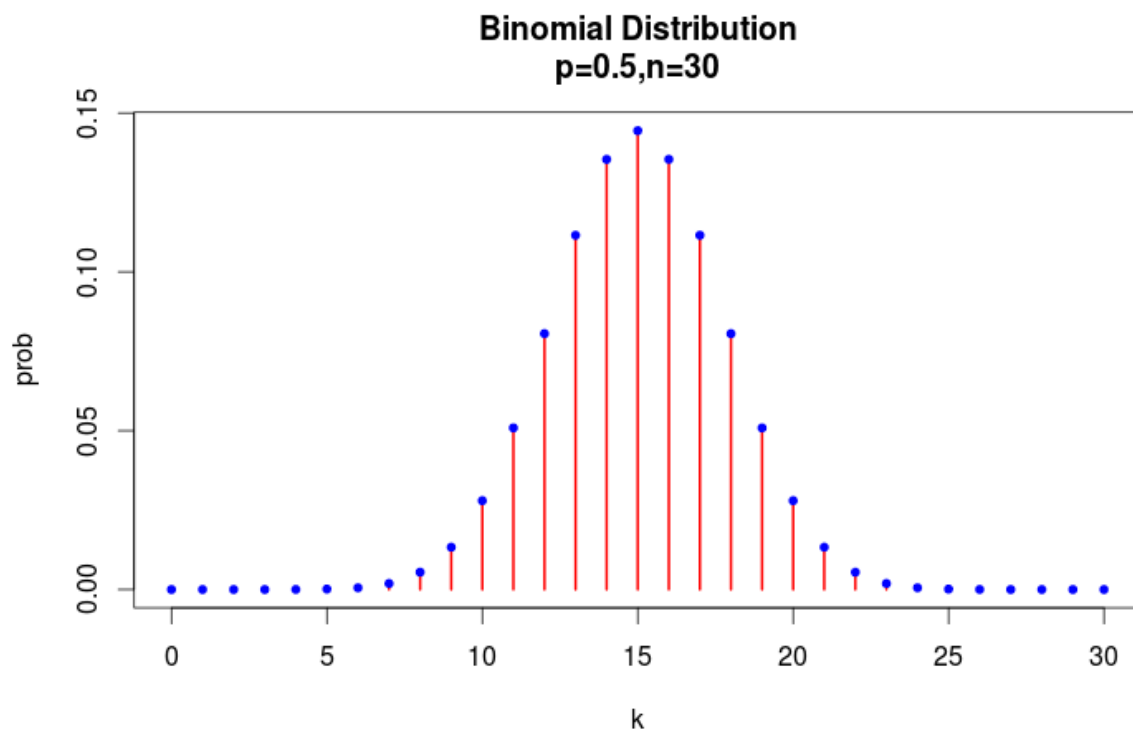
What is Random Variable in Statistics?

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
- Continuous probability distributions

Discrete probability distributions



Continuous probability distributions

Continuous Probability Distributions

- Random variable X can assume infinitely many real values
- Cumulative distribution function

$$F(x) = \int_{-\infty}^x f(v) dv$$

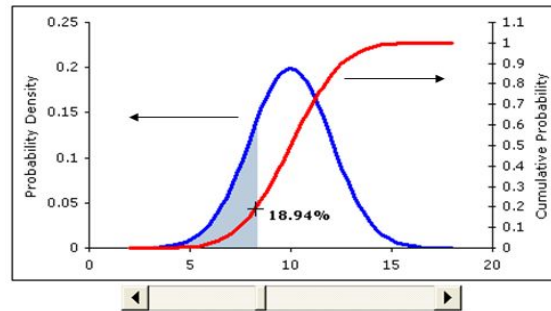
- Probability density function

$$f(x) = \frac{dF(x)}{dx}$$

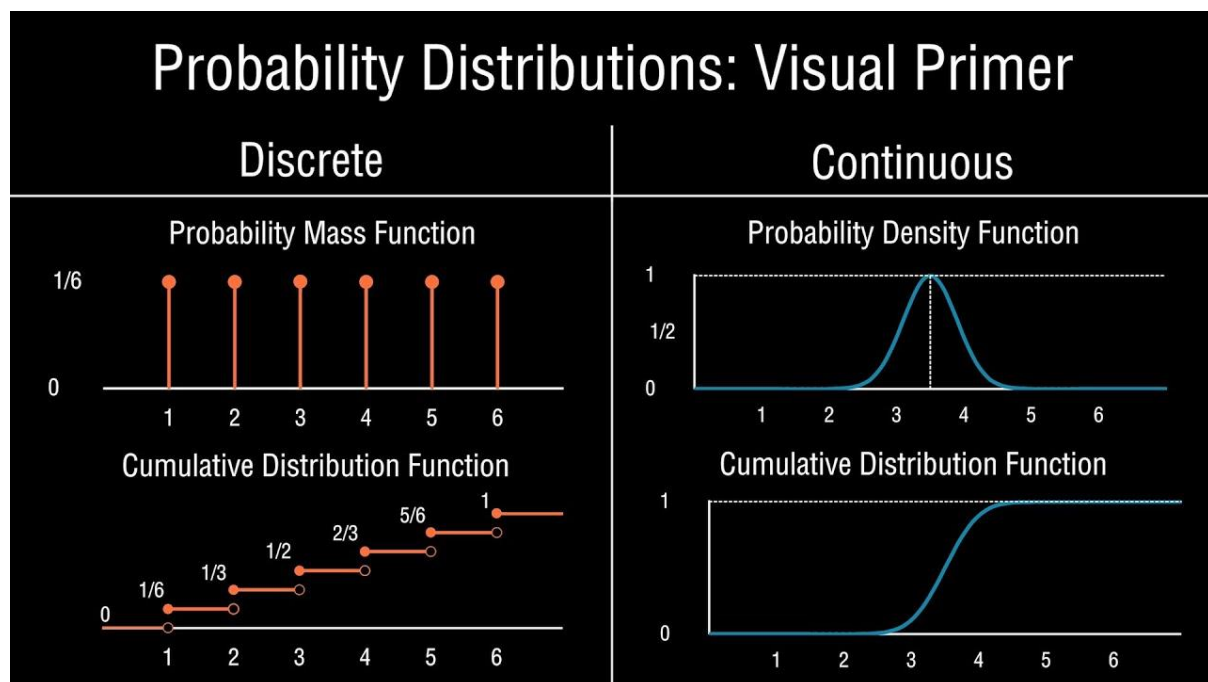
- Properties

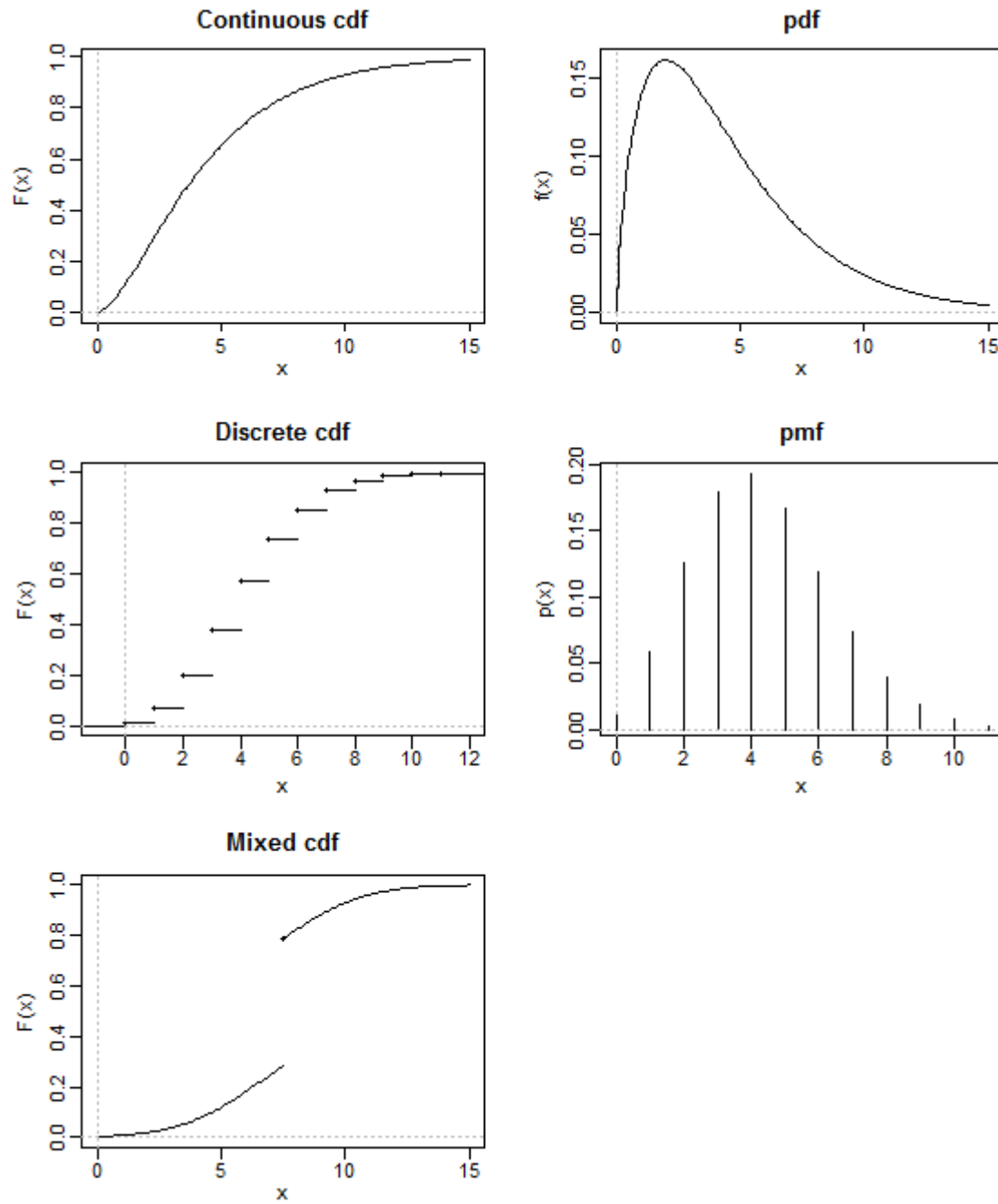
$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(v) dv$$

$$\int_{-\infty}^{\infty} f(v) dv = 1$$



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.

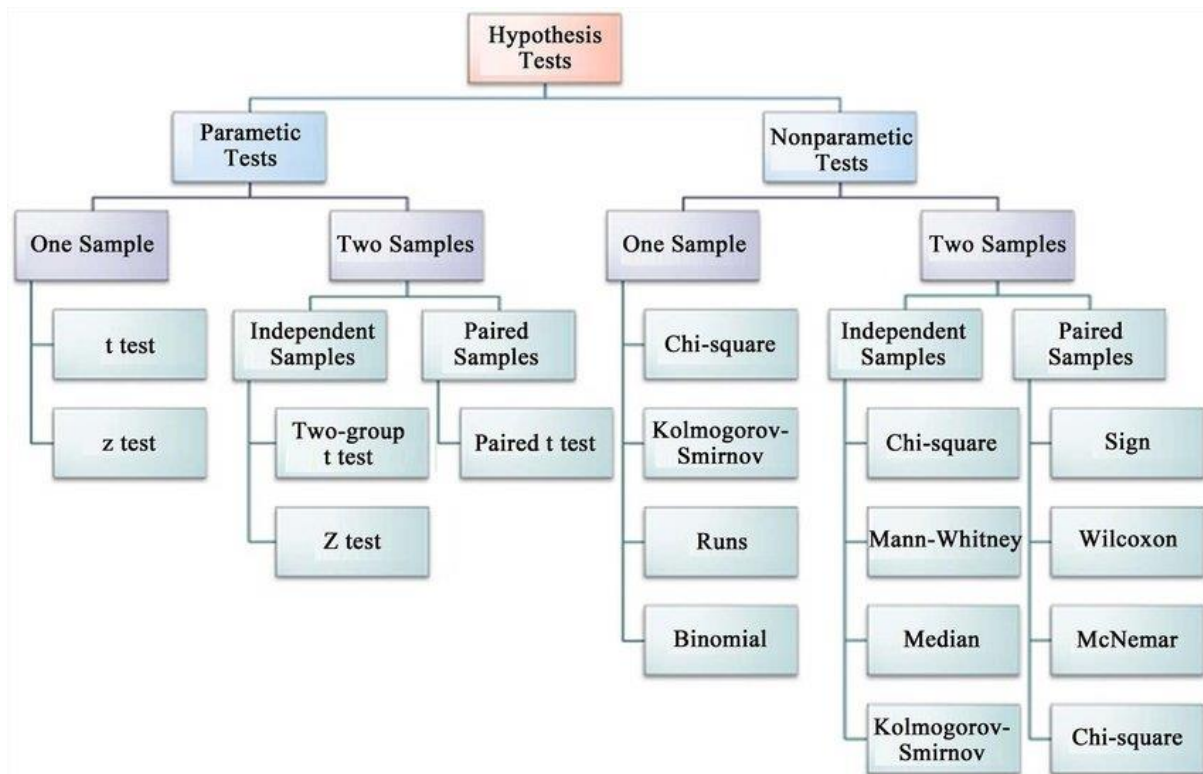
When to use which test:

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.

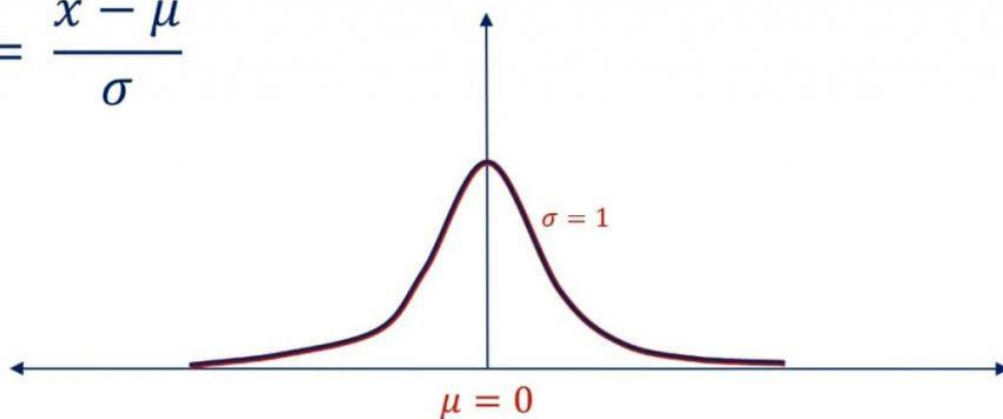


A Standard Normal Variate (SNV):

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.

STANDARDIZATION

$$Z = \frac{x - \mu}{\sigma}$$



$$Z \sim N(0,1)$$

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

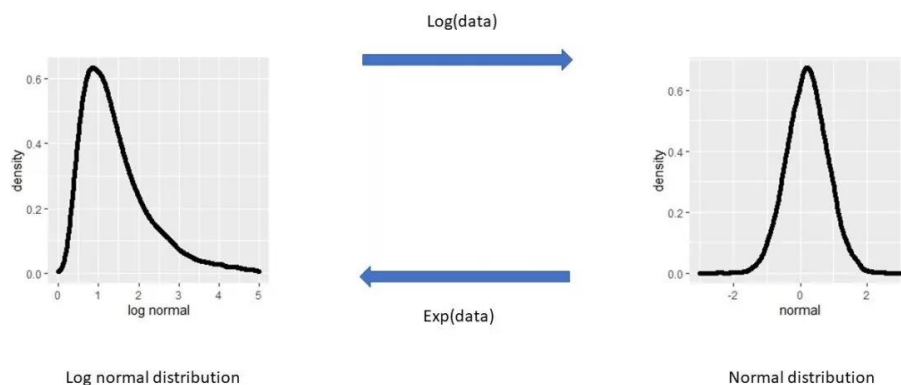
Box-Cox transformation: Transforms non-normal dependent variables into a normal shape

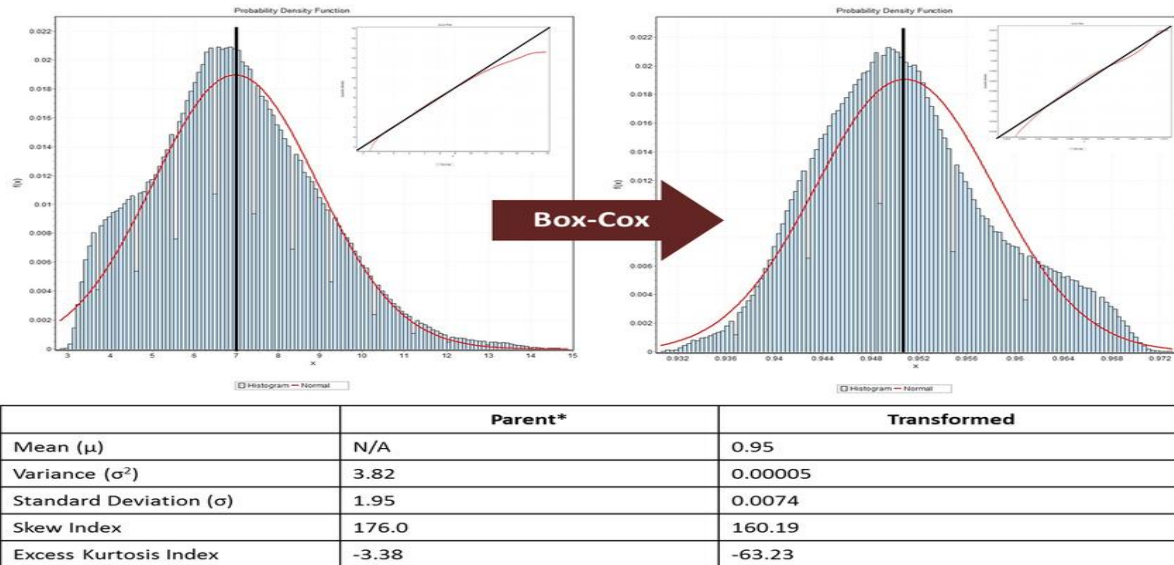
Reciprocal transformation: Can help scale down large values or spread out small values in a right-skewed distribution

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:**



The Box-Cox Transformation

$$y = (x^\lambda - 1) / \lambda, \quad \text{if } \lambda \neq 0$$

$$y = \log(x), \quad \text{if } \lambda = 0$$

x = untransformed variable
 y = transformed variable
 λ = Box-Cox parameter

Yeo-Johnson transformation

$$y_i^{(\lambda)} = \begin{cases} ((y_i + 1)^\lambda - 1) / \lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \ln(y_i + 1) & \text{if } \lambda = 0, y \geq 0 \\ -((-y_i + 1)^{(2-\lambda)} - 1) / (2 - \lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\ln(-y_i + 1) & \text{if } \lambda = 2, y < 0 \end{cases}$$

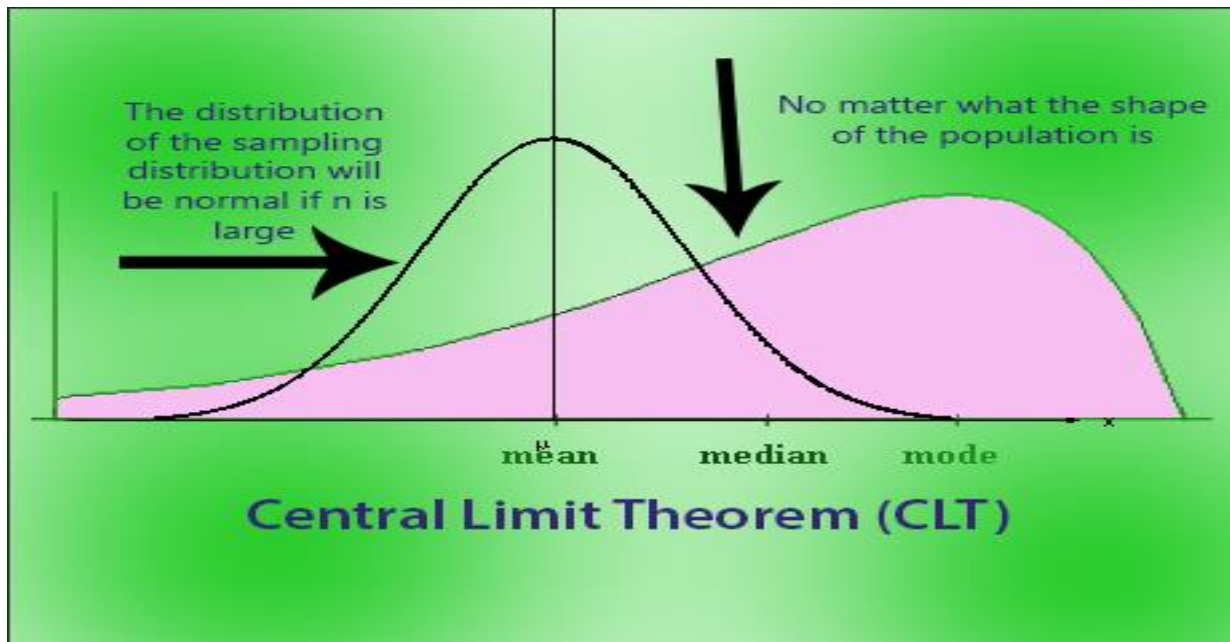
Reciprocal transformation:

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.

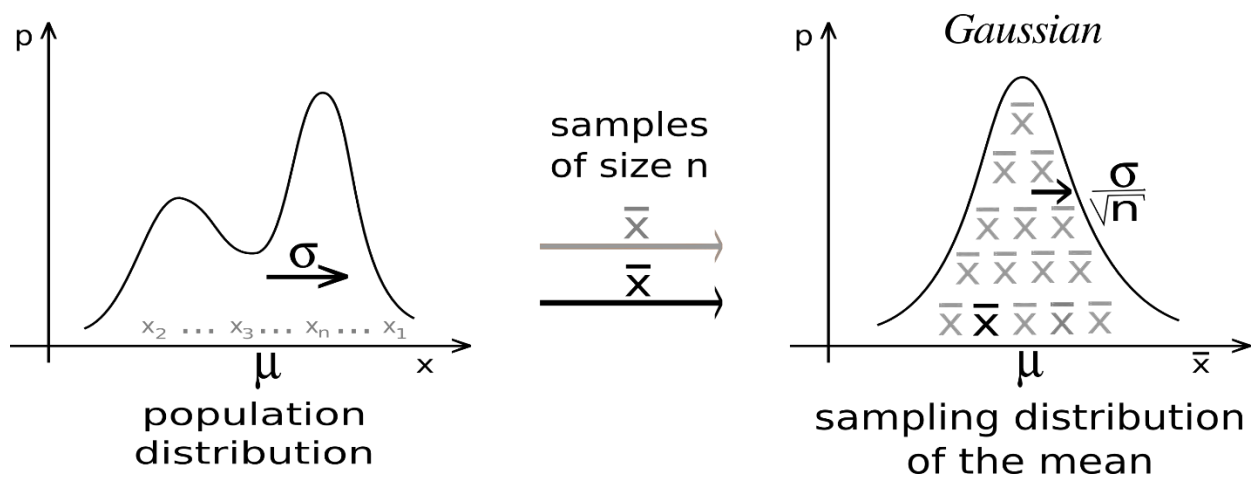


Key points about the CLT assumptions:

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.



Central Limit Theorem Formula

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Sample Mean = Population Mean = μ

Sample Standard Deviation = $\frac{\text{Standard Deviation}}{\sqrt{n}}$

OR

Sample Standard Deviation = $\frac{\sigma}{\sqrt{n}}$

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 (e.g., sample mean).

Point Estimate:

In statistics, point estimation involves the use of sample data to calculate a single value (known as a point estimate since it identifies a point in some parameter space) which is to serve as a "best guess" or "best estimate" of an unknown population parameter (for example, the population mean). More formally, it is the application of a point estimator to the data to obtain a point estimate.

Parameter:

A numerical value representing a characteristic or trait of a whole population.



Statistics:

A numerical value derived from a sample of data taken from a population.



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.

$$\mu = \bar{x} \pm Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

**Confidence Interval
Formula**



$$\text{Confidence interval} = \bar{x} \pm \bar{z}_c \left(\frac{s}{\sqrt{n}} \right)$$

CI → Confidence interval for infinite population

\bar{x} → Sample mean

z_c → Z value for confidence level

s → Sample standard deviation

n → Number of elements in a sample

$$CI = \bar{x} \pm z \frac{s}{\sqrt{n}}$$

CI = confidence interval

\bar{x} = sample mean

z = confidence level value

s = sample standard deviation

n = sample size

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 $z = (x - \mu) / \sigma$ equals open paren x minus mu close paren / sigma

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-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.

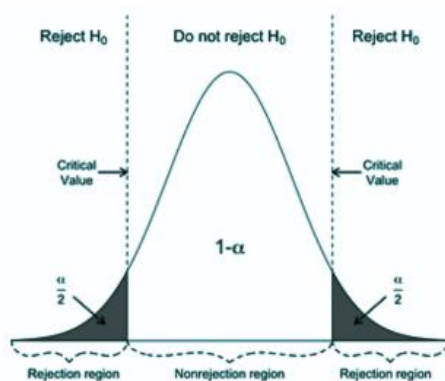
Assumption:

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.

Z-Test



Step 1: State hypotheses and identify the claim.

Step 2: Find critical value/s.

Step 3: Compute test value by using Z-Test.

Step 4: Make decision to reject or to not reject the null.

Z-TEST

✚ Formula to find the value of Z (z-test) is:

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

✚ \bar{x} = mean of sample

✚ μ_0 = mean of population

✚ σ = standard deviation of population

✚ n = no. of observations

Z- Test Formula

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \Bigg| \quad Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

T-Procedure:

One-Sample T-Test

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

\bar{x} = observed mean of the sample

μ = assumed mean

s = standard deviation

n = sample size

Two-Sample T-Test

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

\bar{x}_1 = observed mean of 1st sample

\bar{x}_2 = observed mean of 2nd sample

s_1 = standard deviation of 1st sample

s_2 = standard deviation of 2nd sample

n_1 = sample size of 1st sample

n_2 = sample size of 2nd sample

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

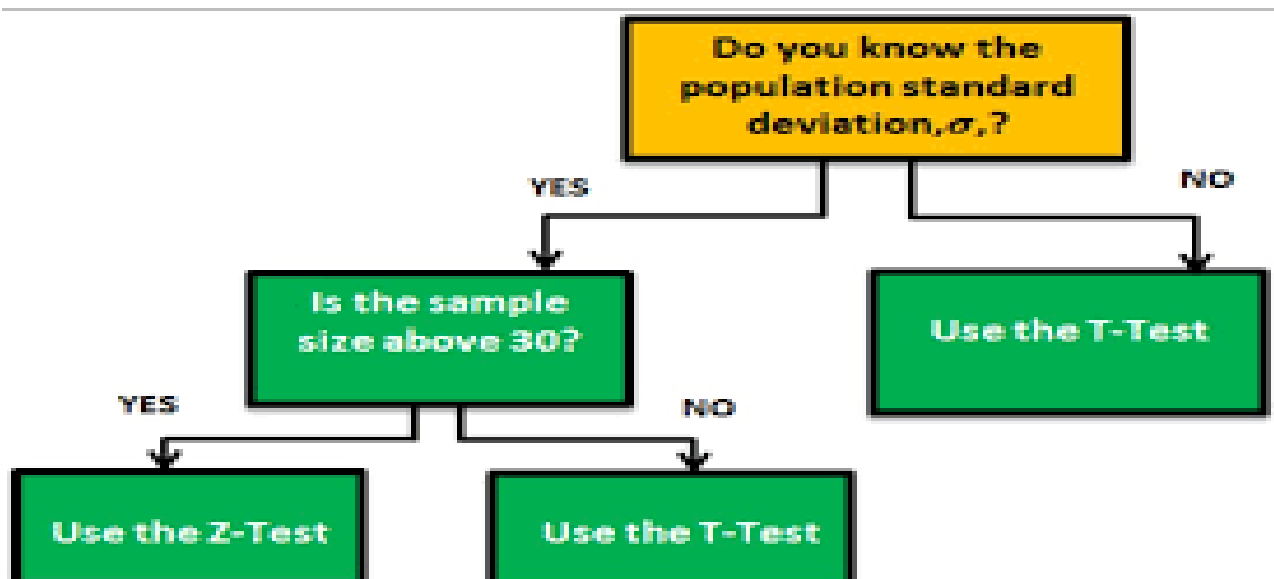
Types of t-tests

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

$$\text{Lower bound: } \bar{X}_D - t_{\alpha/2} * \frac{S_D}{\sqrt{n}}$$
$$\text{Upper bound: } \bar{X}_D + t_{\alpha/2} * \frac{S_D}{\sqrt{n}}$$



$$CI = \bar{x} \pm Z * \frac{s}{\sqrt{n}}$$

$$CI = \bar{x} \pm t * \frac{s}{\sqrt{n}}$$

$CI = \bar{x} \pm z \cdot \frac{s}{\sqrt{n}}$

Labels and connections:

- Mean value**: points to \bar{x}
- Lower/Upper limit**: points to the \pm symbol
- z-value for the confidence level**: points to z
- Standard deviation**: points to s
- Sample size**: points to n in the denominator

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 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 H_0

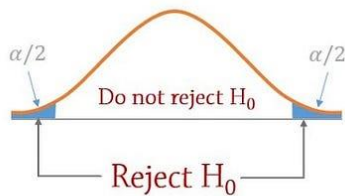
Alternative Hypothesis H_1 or H_A :

Hypothesis Testing

Two-tailed

$$H_0: \mu = 23$$

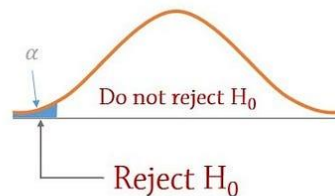
$$H_1: \mu \neq 23$$



Left-tailed

$$H_0: \mu \geq 23$$

$$H_1: \mu < 23$$

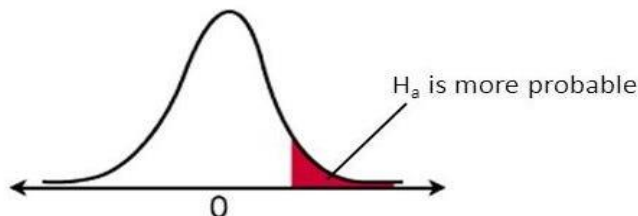
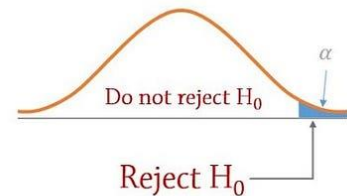


One-tailed

Right-tailed

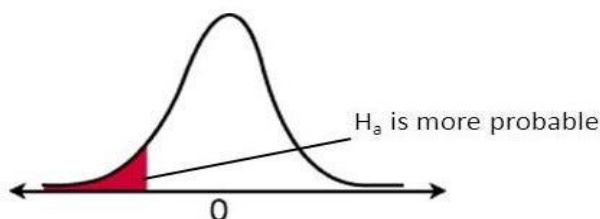
$$H_0: \mu \leq 23$$

$$H_1: \mu > 23$$



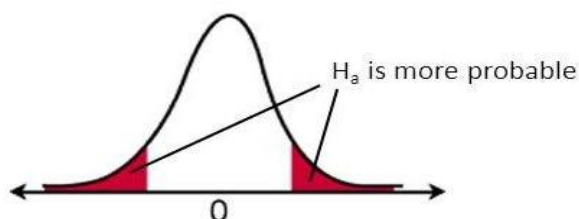
Right-tail test

$$H_a: \mu > \text{value}$$



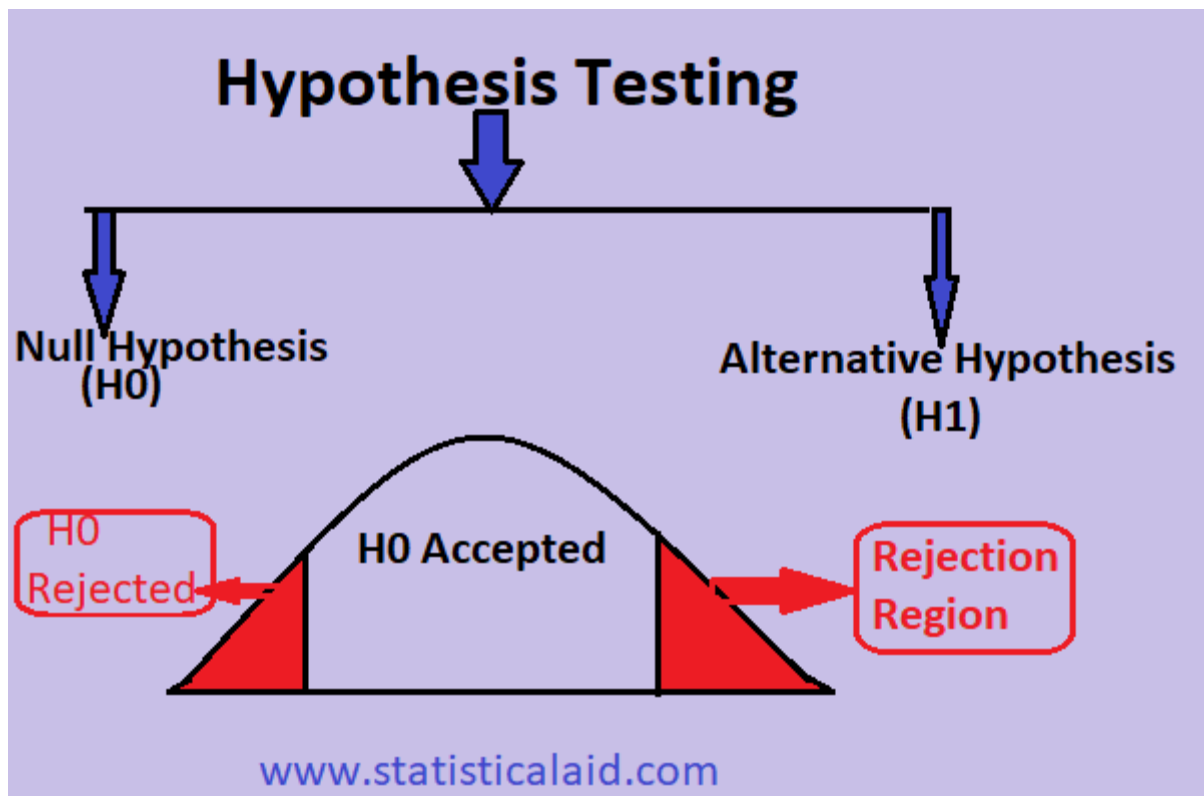
Left-tail test

$$H_a: \mu < \text{value}$$



Two-tail test

$$H_a: \mu \neq \text{value}$$



CHAPTER 6:
Conclusion