Statistics

# Types of Data

## Categorical

Represents groups and categories

## Numerical

* Discrete – Finite Values. Eg. SAT Score, countable objects
* Continuous – Infinite, uncountable. Eg. Weight, Height, etc

# Level of Measurement

## Qualitative

Categorical values

* Nominal – Unordered Categories
* Ordinal – Ordered categories

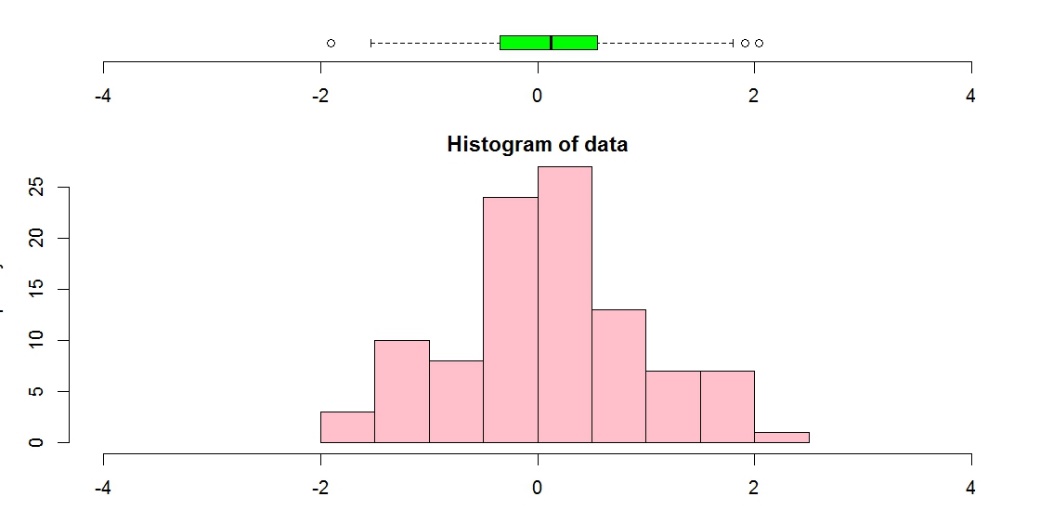
## Quantitive

Numrerical values

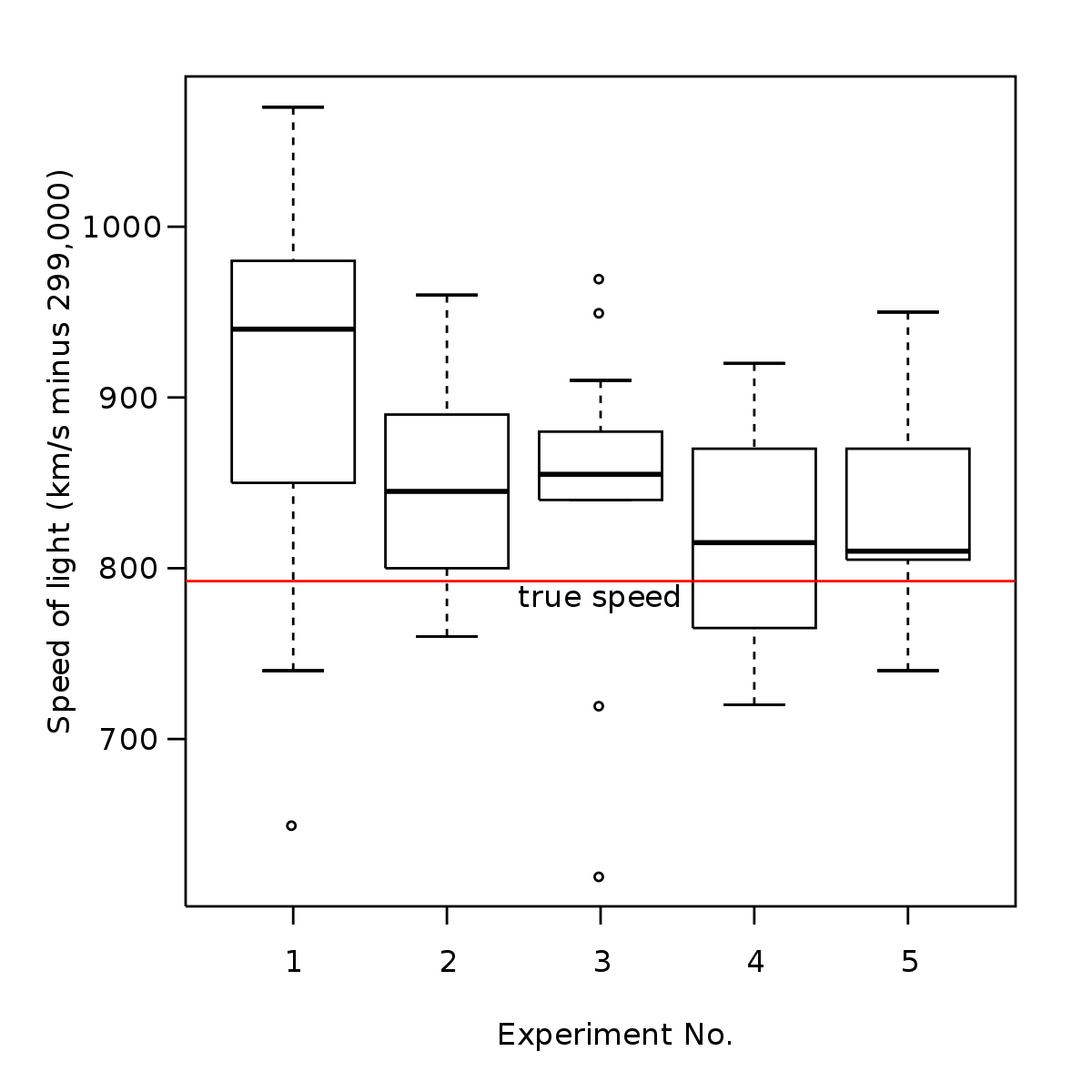
* Interval – Doesn’t have true zero(degree Celsius/farhenheit)
* Ratio – Has a true zero(length, etc)

# Visual Representation of data

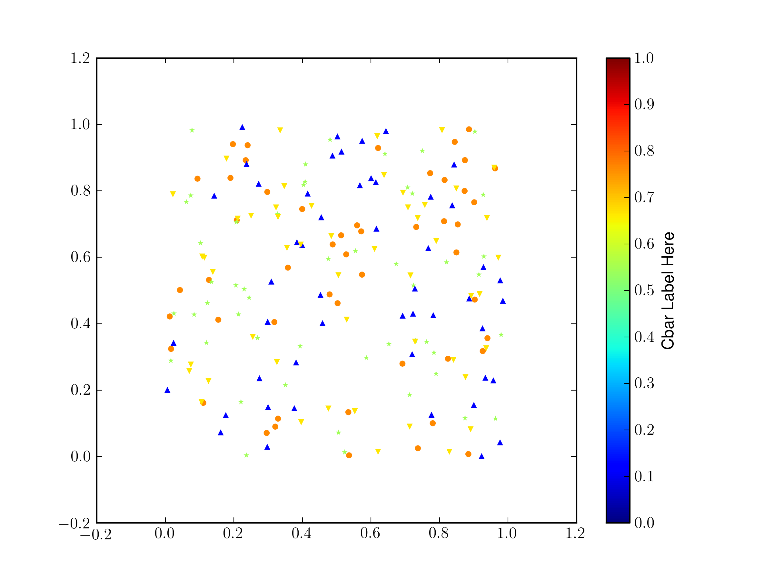
1. **Numerical Data:**
   * **Histogram:** Used to display the distribution of continuous data. It shows the frequency of data within specified intervals (bins).



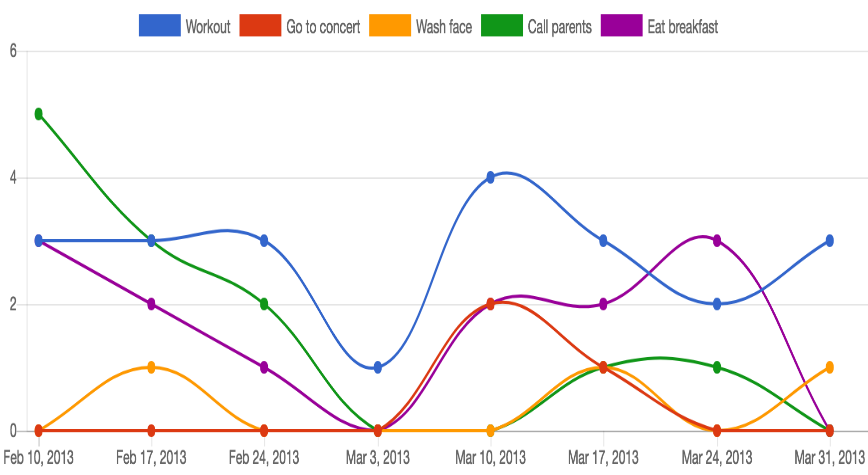
* + **Box Plot (Box-and-Whisker Plot):** Represents the distribution of data and shows the median, quartiles, and outliers.



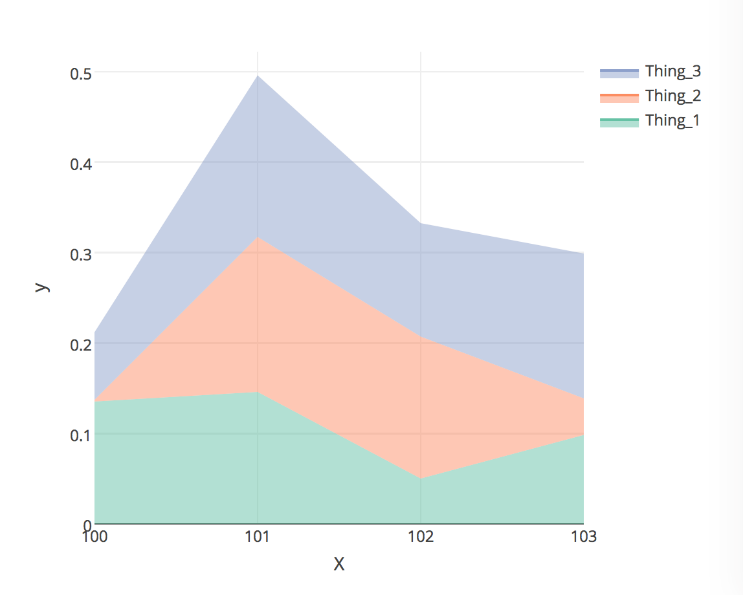
* + **Scatter Plot:** Used to visualize the relationship between two continuous variables.



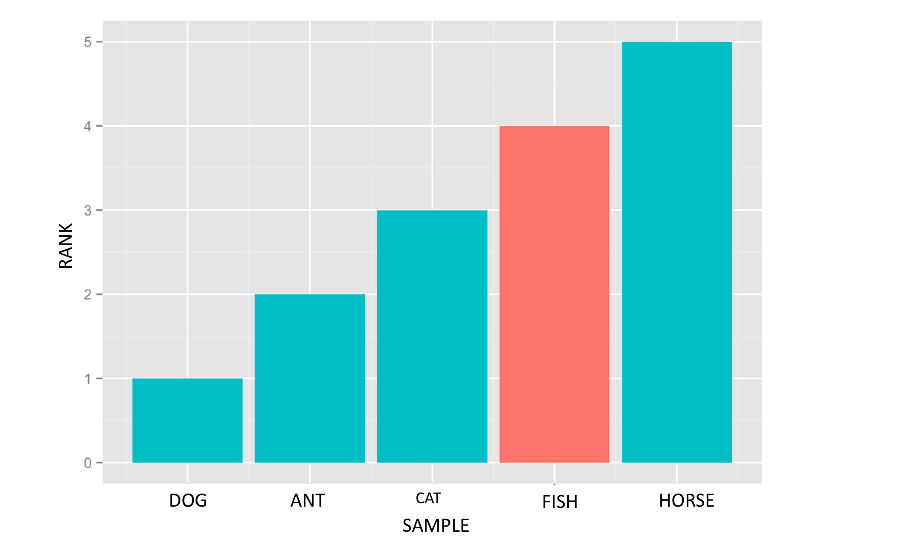
* + **Line Chart:** Suitable for showing trends and patterns in data over time or continuous variable values.



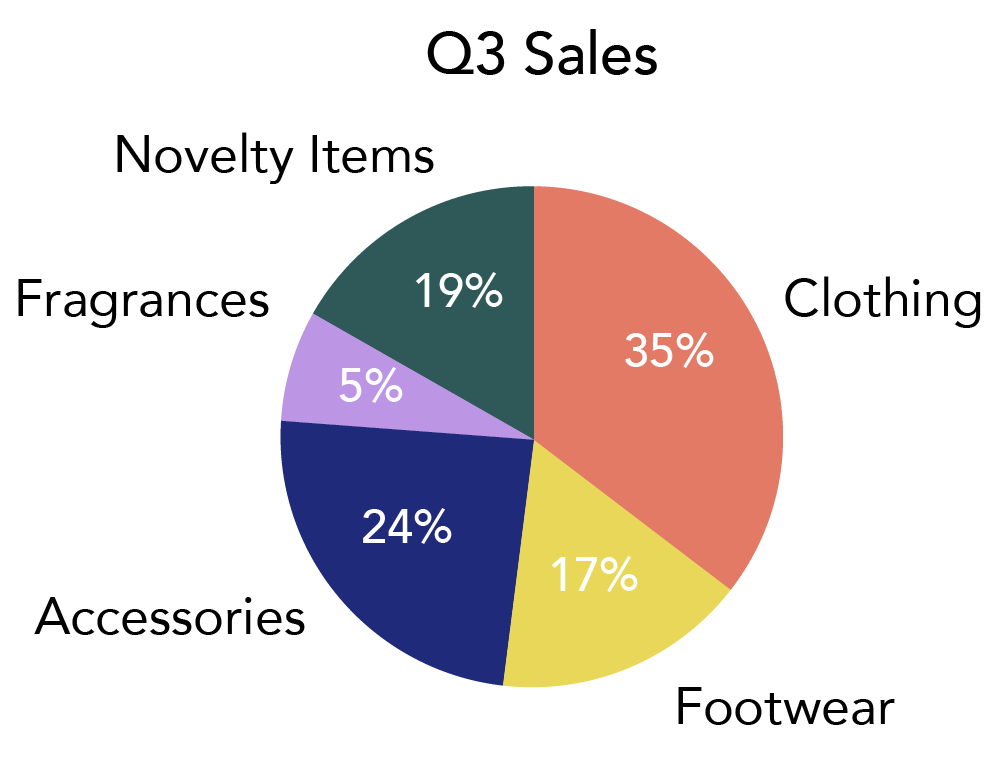
* + **Area Chart:** Similar to a line chart, but the area below the line is filled, making it suitable for representing cumulative data.



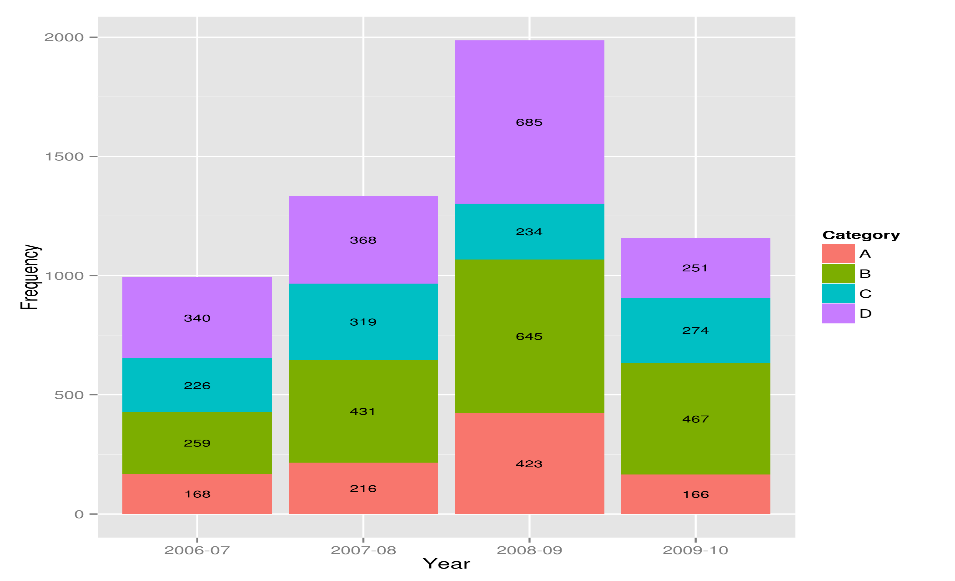
1. **Categorical Data:**
   * **Bar Chart:** Used to display the frequency of categorical data. The length of the bars represents the count or proportion of each category.



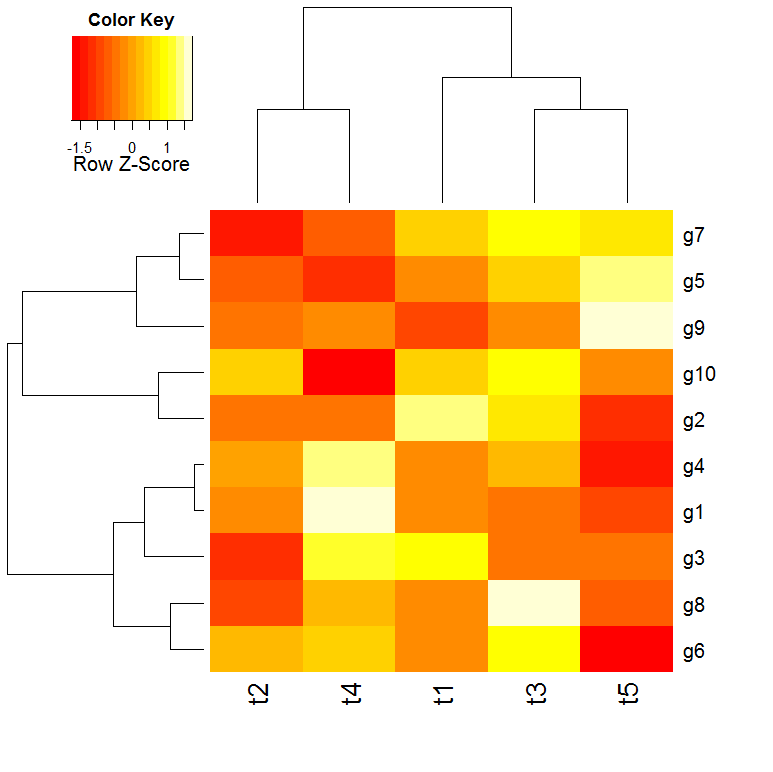
* + **Pie Chart:** Represents the proportion of each category in a whole, where the whole is divided into slices (sectors) based on the category's relative size.



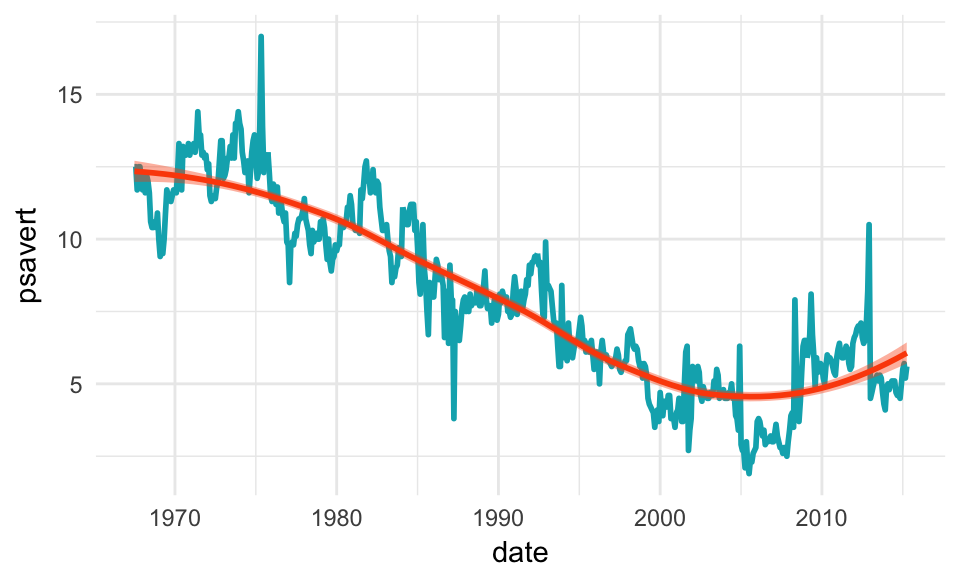
* + **Stacked Bar Chart:** Combines multiple bar charts to show the composition of each category by subcategories.



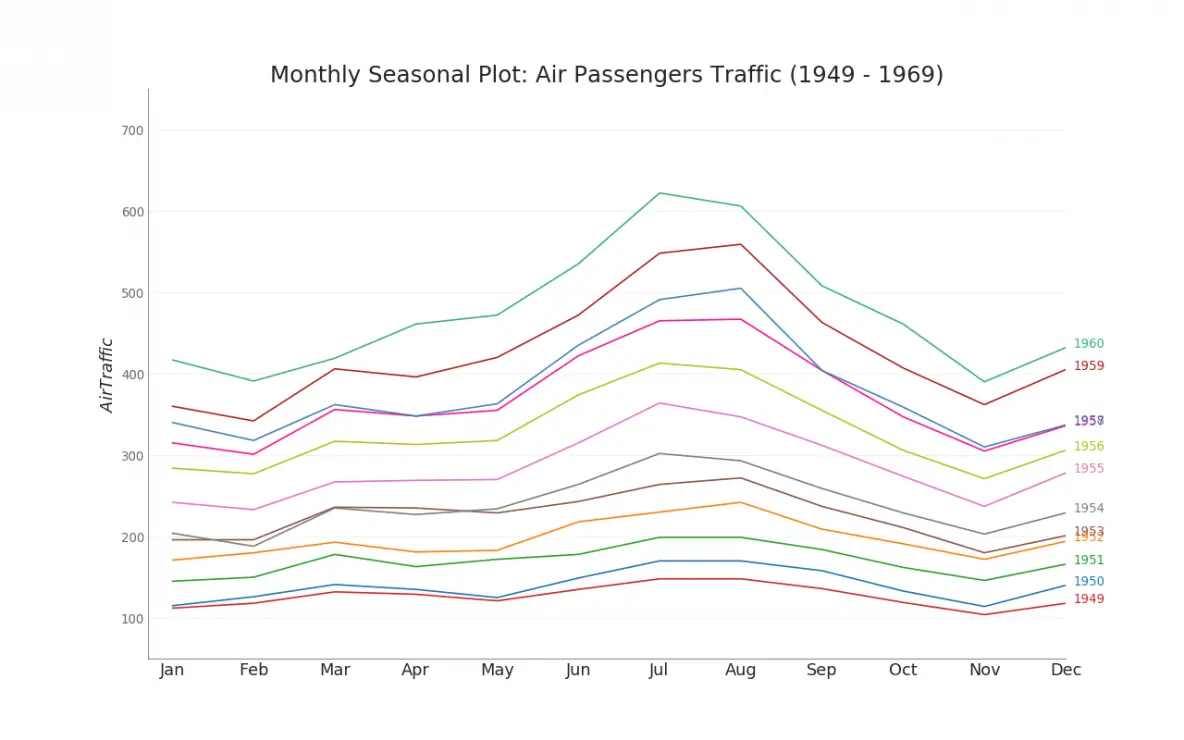
* + **Heatmap:** A matrix of colors used to represent the relationship between two categorical variables. Intensity of colors indicates the strength of the relationship.



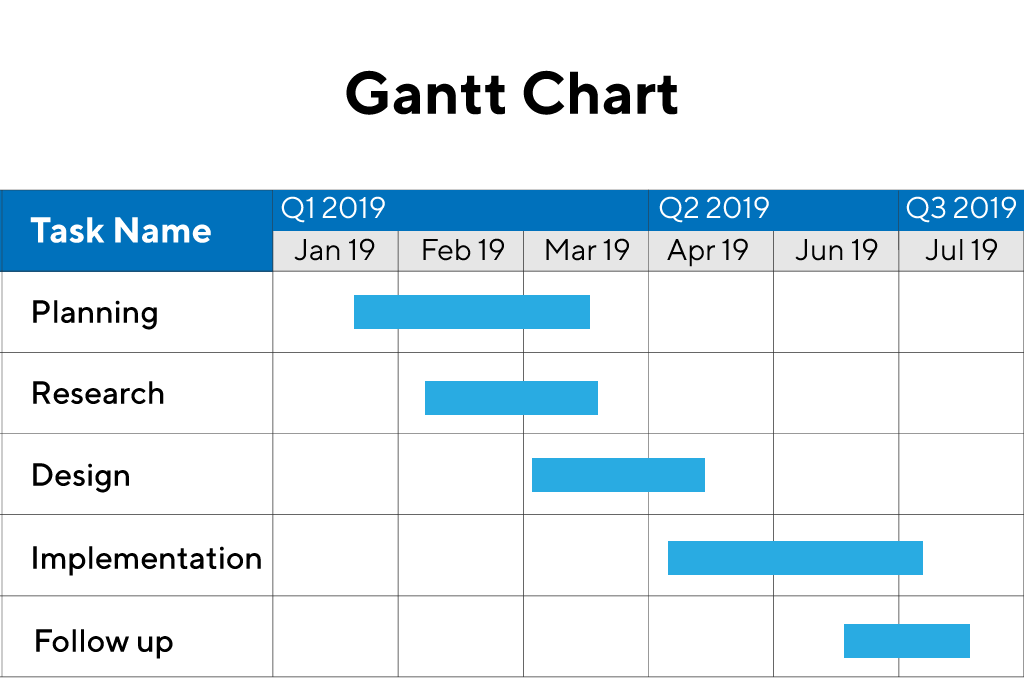
1. **Time Series Data:**
   * **Time Series Plot:** Plots data points against time, helping to identify trends, seasonality, and patterns over time.



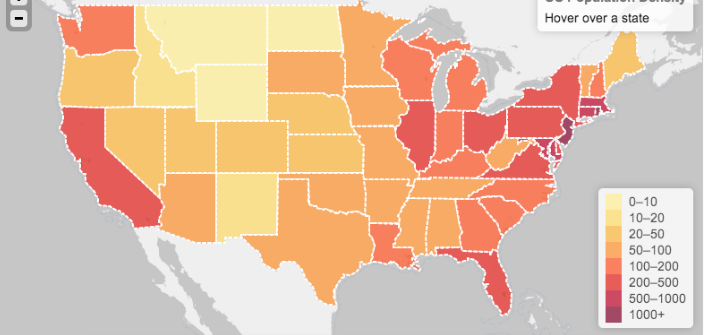
* + **Seasonal Plot:** Focuses on seasonal patterns within a time series to observe recurring patterns.



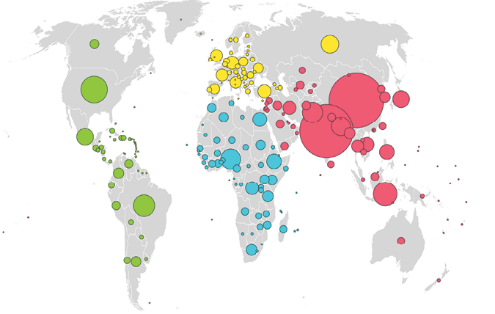
* + **Gantt Chart:** Represents activities over time, commonly used in project management to show schedules and progress.



1. **Geospatial Data:**
   * **Choropleth Map:** Uses color variations to represent data on a geographical map, where regions are shaded based on the value of the variable being visualized.



* + **Bubble Map:** Similar to a choropleth map, but data is represented using bubbles of different sizes and colors, usually over specific locations.



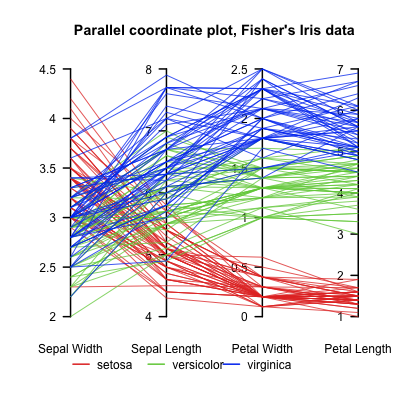
1. **Hierarchical Data:**
   * **Tree Map:** Displays hierarchical data as nested rectangles, where the size and color of each rectangle represent different metrics.



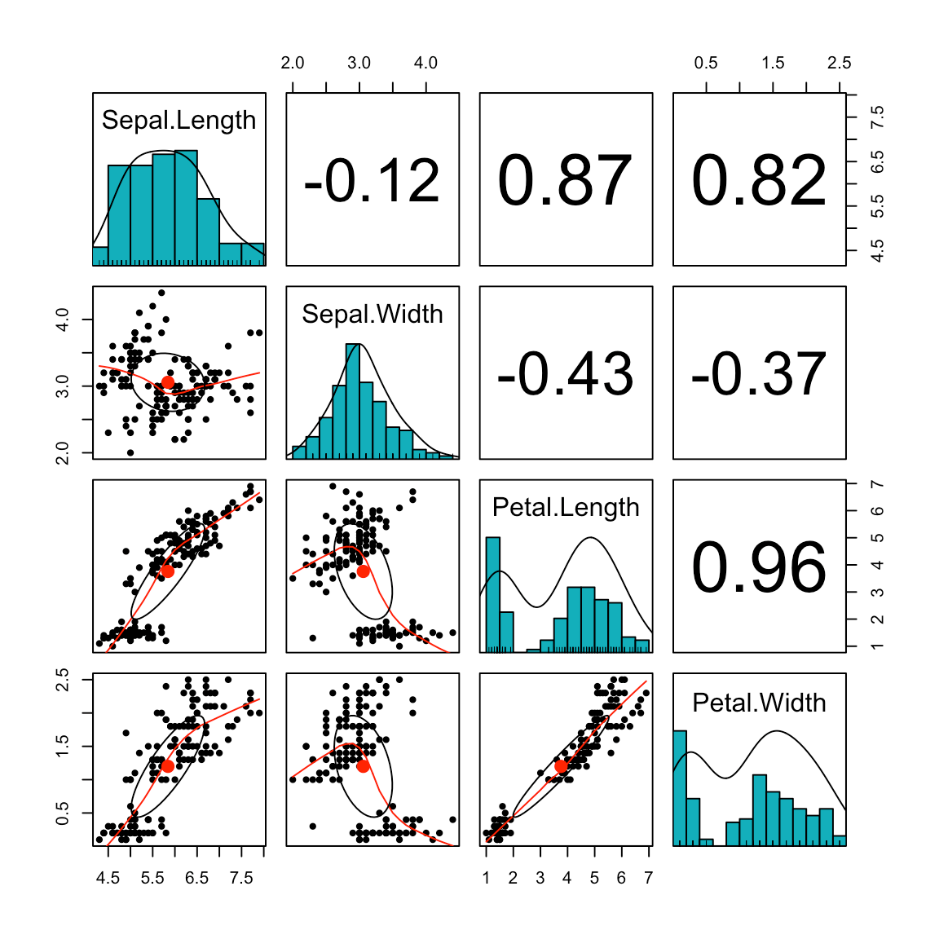
* + **Sunburst Chart:** A radial version of the tree map, displaying hierarchical data with nested rings.



1. **Multivariate Data:**
   * **Parallel Coordinates:** Suitable for visualizing multivariate data. Each variable is represented as a vertical axis, and lines connect data points based on their values.



* + **Scatterplot Matrix (SPLOM):** Combines multiple scatter plots to visualize relationships between multiple variables simultaneously.



# Part 1 – Probability

Probability is a measure of the likelihood of an event occurring. It quantifies uncertainty and is expressed as a value between 0 and 1, where 0 represents impossibility (event will not occur) and 1 represents certainty (event will definitely occur).

**Sample Space:** The set of all possible outcomes of an experiment is called the sample space and is denoted by "S."

**Event:** An event is a subset of the sample space, representing a particular outcome or a combination of outcomes.

**Probability of an Event (P):** The probability of an event "E," denoted by P(E), is the ratio of the number of favorable outcomes to the total number of possible outcomes in the sample space.

P(E) = Number of favorable outcomes / Total number of outcomes in S

**Basic Properties of Probability:**

* **0 <= P(E) <= 1**: Probability of any event lies between 0 and 1, inclusive.
* **P(S) = 1**: The probability of the entire sample space is 1.
* **P(∅) = 0**: The probability of an empty event (impossible event) is 0.

**Complementary Event:** The complement of an event "E," denoted by E', is the event that "E" does not occur. The probability of the complement of an event is given by:

P(E') = 1 - P(E)

**Union of Events (OR):** The probability of the union of two events "E" and "F," denoted by P(E ∪ F), is the probability that either "E" or "F" occurs.

P(E ∪ F) = P(E) + P(F) - P(E ∩ F)

**Intersection of Events (AND):** The probability of the intersection of two events "E" and "F," denoted by P(E ∩ F), is the probability that both "E" and "F" occur.

**Independent Events:** Two events "E" and "F" are independent if the occurrence of one event does not affect the occurrence of the other. For independent events:

P(E ∩ F) = P(E) \* P(F)

**Conditional Probability:** The conditional probability of an event "E" given that event "F" has occurred, denoted by P(E|F), is the probability of "E" under the condition that "F" has already occurred.

P(E|F) = P(E ∩ F) / P(F), where P(F) > 0

**Bayes' Theorem:** Bayes' theorem relates conditional probabilities. It allows us to update the probability of an event based on new information.

P(E|F) = (P(F|E) \* P(E)) / P(F), where P(F) > 0

**Random Variables:** A random variable is a function that maps outcomes of an experiment to numerical values. It can be either discrete (takes on distinct values) or continuous (takes on any value within a range).

**Probability Distribution:** The probability distribution of a random variable provides the probabilities of all possible values that the random variable can take.

**Expected Value (Mean):** The expected value of a random variable, denoted by E(X), is the average value it is expected to take over the long run.

E(X) = Σ [x \* P(X=x)] (for discrete) E(X) = ∫[x \* f(x)] dx (for continuous)

In general mathematics, the term "mean" refers to the arithmetic average of a set of numbers. It is calculated by adding all the numbers in the dataset together and then dividing that sum by the total number of data points. The mean is commonly used as a measure of central tendency, indicating the typical or average value in the dataset.

For example, given the set of numbers: 5, 7, 9, 11, and 13, the mean would be calculated as follows:

Mean = (5 + 7 + 9 + 11 + 13) / 5 = 45 / 5 = 9

In probability and statistics, the term "mean" is used to refer specifically to the expected value of a random variable. A random variable is a variable that can take on different values with certain probabilities. The mean of a random variable is the average value that one would expect to obtain if the random variable were repeatedly measured an infinite number of times.

The mean of a random variable is calculated by summing the products of each possible value of the variable and its corresponding probability. It provides valuable information about the long-term behavior of the random variable.

For example, consider a random variable X representing the outcome of rolling a fair six-sided die. The possible values of X are 1, 2, 3, 4, 5, and 6, each with a probability of 1/6. The mean of this random variable would be calculated as:

Mean (μ) = (1 \* 1/6) + (2 \* 1/6) + (3 \* 1/6) + (4 \* 1/6) + (5 \* 1/6) + (6 \* 1/6) = 3.5

**Variance:** Variance measures the spread of the random variable's values around its mean.

Var(X) = E[(X - E(X))^2]

In probability and statistics, variance is a measure of how much the values of a random variable vary or spread out from its mean. It quantifies the dispersion of the data points around the expected value (mean) of the random variable. A high variance indicates that the values are more spread out, while a low variance indicates that the values are closer to the mean.

The variance (σ^2) of a discrete random variable X is calculated using the following formula:

σ^2 = Σ [ (x - μ)^2 \* P(x) ]

Where:

* σ^2 is the variance of the random variable X.
* Σ denotes the summation symbol, which means you sum up the expression for each possible value of X.
* x represents each possible value of the random variable X.
* μ is the mean (expected value) of the random variable X.
* P(x) is the probability of the random variable taking the value x.

To calculate the variance, follow these steps:

1. Find the mean (μ) of the random variable X.
2. For each possible value x of X, calculate the squared difference between x and the mean (x - μ)^2.
3. Multiply each squared difference by its corresponding probability P(x).
4. Sum up all these products to get the variance.

Let's illustrate this with an example:

Consider a random variable X representing the outcome of rolling a fair six-sided die. The possible values of X are 1, 2, 3, 4, 5, and 6, each with a probability of 1/6. We already calculated the mean (μ) in the previous response as 3.5.

Now, to calculate the variance (σ^2):

σ^2 = [(1 - 3.5)^2 \* 1/6] + [(2 - 3.5)^2 \* 1/6] + [(3 - 3.5)^2 \* 1/6] + [(4 - 3.5)^2 \* 1/6] + [(5 - 3.5)^2 \* 1/6] + [(6 - 3.5)^2 \* 1/6]

σ^2 = 7.5 / 6 ≈ 1.25

**Standard Deviation:** The standard deviation is the square root of the variance and represents the average deviation from the mean.

SD(X) = √Var(X)

## Probability Frequency Distribution

A probability frequency distribution is a tabular representation of the probabilities associated with the various outcomes of a random experiment or process. It shows the likelihood of each outcome occurring and is particularly useful when dealing with discrete random variables. Each outcome is associated with its corresponding probability, and the sum of all probabilities in the distribution is equal to 1.

To create a probability frequency distribution, follow these steps:

1. Identify the possible outcomes: Determine all the possible outcomes of the random variable in question. These outcomes should be mutually exclusive and exhaustive, meaning that they cover all possible cases.
2. Assign probabilities: For each outcome, calculate or obtain the probability of that particular event occurring. Ensure that the probabilities are non-negative and add up to 1.
3. Construct the distribution: Create a table with two columns - one for the outcomes and the other for their corresponding probabilities.
4. Display the results: List each outcome along with its associated probability in the table.
5. Verify the sum: Make sure that the probabilities add up to 1 to ensure that you have accounted for all possible outcomes.

Here's a simple example to illustrate a probability frequency distribution:

Suppose we have a fair six-sided die, and we want to find the probability of each face showing up when the die is rolled.

1. Identify the possible outcomes: The possible outcomes of rolling the die are the numbers 1, 2, 3, 4, 5, and 6.
2. Assign probabilities: Since the die is fair, each outcome has an equal chance of occurring. So, the probability of each outcome is 1/6.
3. Construct the distribution:

| **Outcome (X)** | **Probability (P(X))** |
| --- | --- |
| 1 | 1/6 |
| 2 | 1/6 |
| 3 | 1/6 |
| 4 | 1/6 |
| 5 | 1/6 |
| 6 | 1/6 |

1. Display the results: The table shows the probability distribution of rolling a fair six-sided die.
2. Verify the sum: The sum of all the probabilities is 1, indicating that we have accounted for all possible outcomes.

Probability frequency distributions provide a clear and concise way to understand the likelihood of different events occurring in a random experiment, making it easier to analyze and interpret the data.

# Part 2 – Combinatorics

## Permutations

In mathematics, a permutation refers to an arrangement of objects in a specific order. It is a way of selecting and arranging a subset of items from a larger set without repetition and considering the order of arrangement.

### Without Repetition

The number of permutations of "r" objects selected from a total of "n" objects is denoted as "nPr" or "P(n, r)" and is calculated using the permutation formula:

nPr = n! / (n - r)!

* n is the total number of objects in the set.
* r is the number of objects selected or to be arranged.
* "!" denotes the factorial, which is the product of all positive integers from 1 to that number.

Examples:

You have a group of 4 friends (A, B, C, and D), and you want to arrange them in a straight line for a photograph. How many different ways can they stand?

In this example, we are selecting and arranging all 4 friends (n = 4) in a row (r = 4).

Number of permutations (4 friends in a row) = 4P4 = 4! / (4 - 4)! = 4! / 0! = (4 × 3 × 2 × 1) / 1 = 24

So, there are 24 different ways the 4 friends can stand in a row for the photograph.

Example 2: You have a set of 6 books (A, B, C, D, E, and F), and you want to arrange 3 of them on a bookshelf. How many different ways can you do this?

In this example, we are selecting and arranging 3 books (r = 3) from a set of 6 books (n = 6).

Number of permutations (3 books from 6) = 6P3 = 6! / (6 - 3)! = 6! / 3! = (6 × 5 × 4) / (3 × 2 × 1) = 120 / 6 = 20

So, there are 20 different ways to arrange 3 books from the set of 6 on the bookshelf.

Example 3: You have a bag with 5 different colored marbles (Red, Blue, Green, Yellow, and Orange). You want to pick 2 marbles randomly from the bag. How many different ways can you do this?

In this example, we are selecting and arranging 2 marbles (r = 2) from a set of 5 marbles (n = 5).

Number of permutations (2 marbles from 5) = 5P2 = 5! / (5 - 2)! = 5! / 3! = (5 × 4) / (2 × 1) = 20 / 2 = 10

So, there are 10 different ways to pick 2 marbles from the set of 5 marbles.

### With Repetition

In permutations with repetition, elements from the set can be used more than once in the arrangement. This means that if the set has duplicate elements, they can be included multiple times in different positions of the permutation.

The formula for permutations with repetition is n^r, where "n" is the number of distinct elements in the set, and "r" is the number of elements to be arranged.

For example, let's consider the set {A, B, C}, and we want to arrange 3 elements from this set with repetition allowed. The possible permutations are:

AAA, AAB, AAC, ABA, ABB, ABC, ACA, ACB, ACC, BAA, BAB, BAC, BBA, BBB, BBC, BCA, BCB, BCC, CAA, CAB, CAC, CBA, CBB, CBC, CCA, CCB, CCC

Here, each element can be repeated three times, and there are 27 possible permutations (3^3) in total.

P(n, p) = n^p

## Combination

n mathematics, a combination refers to the selection of items from a set without regard to the order in which they are arranged. Unlike permutations, combinations do not consider the order of selection, which means that the same group of elements in a different order is not counted separately.

The number of combinations of "r" elements selected from a total of "n" elements is denoted as "nCr" or "C(n, r)" and is calculated using the combination formula:

nCr = n! / (r! \* (n - r)!)

Where:

* n is the total number of elements in the set.
* r is the number of elements to be selected or chosen.
* "!" denotes the factorial, which is the product of all positive integers from 1 to that number.

The factorial of a non-negative integer "n" is denoted as "n!" and is calculated as follows:

n! = n × (n - 1) × (n - 2) × ... × 2 × 1

For example, let's say we have a group of 5 friends (A, B, C, D, and E), and we want to select 3 of them to form a committee:

n = 5 (total number of friends) r = 3 (number of friends to be selected)

Number of combinations (3 friends from 5) = 5C3 = 5! / (3! \* (5 - 3)!) = 5! / (3! \* 2!) = (5 × 4 × 3) / (3 × 2 × 1) = 10

### Combination with separate sample set

When dealing with separate sample sets, the concept of combinations can be extended to calculate the number of ways to choose elements from each sample set without considering the order of selection within each set. This is useful in various statistical and probability scenarios, especially when comparing two or more groups of elements.

Suppose you have "n1" elements in the first sample set and "n2" elements in the second sample set. You want to choose "r1" elements from the first sample set and "r2" elements from the second sample set.

The number of combinations when selecting "r1" elements from the first sample set and "r2" elements from the second sample set is calculated using the combination formula:

C(n1, r1) \* C(n2, r2)

Where:

* C(n1, r1) represents the number of combinations for selecting "r1" elements from the first sample set with "n1" elements.
* C(n2, r2) represents the number of combinations for selecting "r2" elements from the second sample set with "n2" elements.

For example, let's say you have two sample sets, A and B, with 4 elements each. You want to choose 2 elements from set A and 3 elements from set B:

n1 = 4 (number of elements in set A) r1 = 2 (number of elements to choose from set A)

n2 = 4 (number of elements in set B) r2 = 3 (number of elements to choose from set B)

Number of combinations = C(4, 2) \* C(4, 3) = 6 \* 4 = 24

So, there are 24 different ways to choose 2 elements from set A and 3 elements from set B.

In situations with more than two separate sample sets, you can extend the same principle by multiplying the combinations of each set. For example, if you have three sets with n1, n2, and n3 elements, and you want to choose r1, r2, and r3 elements from each set, the total number of combinations would be:

C(n1, r1) \* C(n2, r2) \* C(n3, r3)

Combinations with separate sample sets allow you to explore various possibilities when dealing with multiple groups of elements and can be applied in various statistical and experimental settings.

### Combination with Repetition

Combination with repetition, also known as combinations with replacement, refers to the selection of items from a set where repeated elements are allowed in the selection. In other words, combination with repetition allows you to choose elements from the set multiple times, and the order of selection is not considered.

The number of combinations with repetition of "r" elements selected from a total of "n" elements is denoted as "n + r - 1 choose r" or "(n + r - 1)Cr" and is calculated using the combination with repetition formula:

(n + r - 1)Cr = (n + r - 1)! / (r! \* (n - 1)!)

Where:

* n is the total number of distinct elements in the set.
* r is the number of elements to be chosen or selected.
* "!" denotes the factorial, which is the product of all positive integers from 1 to that number.

For example, let's consider a set of 3 different colors (Red, Green, Blue), and we want to select 4 colors from this set with repetition allowed:

n = 3 (total number of distinct colors) r = 4 (number of colors to be selected)

Number of combinations with repetition (4 colors from 3) = (3 + 4 - 1)C4 = (6)! / (4! \* 2!) = (6 × 5 × 4 × 3 × 2 × 1) / ((4 × 3 × 2 × 1) \* (2 × 1)) = 15

So, there are 15 different ways to select 4 colors from the set of 3 colors with repetition allowed.

Combination with repetition is commonly used in scenarios where you need to count the number of ways to form a group or choose elements from a set when duplication is allowed. It is relevant in various real-world problems, such as counting the number of possible combinations in experiments with replacement, or when choosing items from a menu with duplicate options.

# Part 3 – Sets

In mathematics, a set is a collection of distinct elements or objects grouped together as a single entity. The elements within a set can be anything: numbers, letters, colors, people, or any other objects that share common characteristics. Sets are a fundamental concept in set theory, a branch of mathematics that deals with the study of collections of objects and their properties.

Notation: Sets are usually denoted by capital letters, and their elements are listed within curly braces {}. For example:

* A = {1, 2, 3, 4, 5} represents a set A with elements 1, 2, 3, 4, and 5.
* B = {apple, banana, orange} represents a set B with elements apple, banana, and orange.

Properties of Sets:

1. Distinct Elements: A set contains only distinct or unique elements. If an element is repeated in the set, it is still considered only once in the set.
2. No Order: The elements in a set have no specific order, meaning that the elements are considered without any arrangement or sequence.
3. Cardinality: The cardinality of a set refers to the number of elements it contains. It is denoted by |A|, where A is the set. For example, if A = {1, 2, 3, 4, 5}, then |A| = 5.

Basic Set Operations:

1. Union (⋃): The union of two sets A and B, denoted by A ⋃ B, is a set that contains all elements that are in A, in B, or in both sets.
2. Intersection (⋂): The intersection of two sets A and B, denoted by A ⋂ B, is a set that contains all elements that are common to both sets.
3. Difference (−): The difference of two sets A and B, denoted by A - B, is a set that contains all elements that are in A but not in B.
4. Complement (′): The complement of a set A with respect to a universal set U, denoted by A', is a set that contains all elements in U that are not in A.
5. Subset (⊆): If all elements of set A are also elements of set B, then A is a subset of B, denoted as A ⊆ B.
6. Proper Subset (⊂): If A is a subset of B, and there exists at least one element in B that is not in A, then A is a proper subset of B, denoted as A ⊂ B.
7. Empty set : No elements in set.

Sets play a crucial role in various mathematical concepts, including logic, algebra, calculus, and probability. They provide a foundation for defining relationships between elements and are essential in the development of more complex mathematical structures.

## Independent and Dependent Events

1. Independent Events: Two events are considered independent if the occurrence of one event does not affect the probability of the occurrence of the other event. In other words, the outcome of one event has no influence on the outcome of the other event.

Mathematically, two events A and B are independent if and only if:

P(A and B) = P(A) \* P(B)

Where:

* P(A and B) is the probability that both events A and B occur.
* P(A) is the probability of event A occurring.
* P(B) is the probability of event B occurring.

For example, if you toss a fair coin twice, the outcome of the first toss (heads or tails) has no impact on the outcome of the second toss. The probability of getting heads on the first toss and tails on the second toss is 0.5 \* 0.5 = 0.25, and this is also the probability of getting tails on the first toss and heads on the second toss.

1. Dependent Events: Two events are considered dependent if the occurrence of one event affects the probability of the occurrence of the other event. In this case, the outcome of one event influences the outcome of the other event.

Mathematically, two events A and B are dependent if and only if:

P(A and B) = P(A) \* P(B|A)

P(B|A) = P(A and B)/ P(A)

Where:

* P(A and B) is the probability that both events A and B occur.
* P(A) is the probability of event A occurring.
* P(B|A) is the conditional probability of event B occurring given that event A has occurred.

For example, if you draw a card from a standard deck without replacement, the probability of drawing a spade on the first draw is 13/52. However, if you don't replace the card back into the deck and draw again, the probability of drawing a spade on the second draw, given that a spade was drawn on the first draw, becomes 12/51. The second draw is dependent on the outcome of the first draw.

Understanding the independence or dependence of events is crucial in probability calculations and real-world applications. It helps in making predictions, assessing risks, and analyzing various scenarios in different fields, such as statistics, finance, and engineering.

## Law of Total Probability

The Law of Total Probability is a fundamental principle in probability theory that allows us to calculate the probability of an event by considering all possible outcomes based on different conditions or partitions. It is particularly useful when dealing with events that can be broken down into mutually exclusive and exhaustive cases.

The Law of Total Probability states that for any event "A" and a partition of the sample space into mutually exclusive events {B₁, B₂, ..., Bₙ}, the probability of event "A" can be expressed as the sum of the probabilities of "A" given each partition multiplied by the probability of each partition:

P(A) = Σ [P(A | Bᵢ) \* P(Bᵢ)]

Where:

* P(A) is the probability of event "A."
* P(A | Bᵢ) is the conditional probability of event "A" given that the partition "Bᵢ" occurred.
* P(Bᵢ) is the probability of the partition "Bᵢ" occurring.
* Σ represents the summation sign, indicating that we sum up the products for all partitions {B₁, B₂, ..., Bₙ}.

The partitions {B₁, B₂, ..., Bₙ} should satisfy two conditions:

1. They are mutually exclusive, meaning that they have no common elements and cannot occur simultaneously.
2. They are exhaustive, meaning that their union covers the entire sample space, and there are no other possible outcomes.

## Additive and Multiplicative Law

The Additive and Multiplicative Laws are fundamental principles in probability theory and combinatorics that help calculate probabilities and counts of outcomes in various situations. Let's explain each law:

1. Additive Law (Union): The Additive Law, also known as the Union Rule, is used to calculate the probability of the union of two or more mutually exclusive events. Mutually exclusive events are events that cannot occur simultaneously.

For two events A and B, the Additive Law states that the probability of either event A or event B occurring is given by the sum of their individual probabilities minus the probability of their intersection:

P(A ∪ B) = P(A) + P(B) - P(A ∩ B)

Where:

* P(A ∪ B) represents the probability of either event A or event B occurring (the union of A and B).
* P(A) and P(B) are the probabilities of events A and B, respectively.
* P(A ∩ B) is the probability of both events A and B occurring (the intersection of A and B).

1. Multiplicative Law (Intersection): The Multiplicative Law, also known as the Intersection Rule, is used to calculate the probability of the intersection of two or more independent events. Independent events are events where the occurrence of one event does not affect the occurrence of the other event.

For two independent events A and B, the Multiplicative Law states that the probability of both event A and event B occurring is the product of their individual probabilities:

P(A ∩ B) = P(A) × P(B)

Where:

* P(A ∩ B) represents the probability of both event A and event B occurring (the intersection of A and B).
* P(A) and P(B) are the probabilities of events A and B, respectively.

It's important to note that the Additive Law is used for mutually exclusive events (e.g., rolling either a 1 or 2 on a fair six-sided die), while the Multiplicative Law is used for independent events (e.g., flipping a coin and rolling a die, where the outcomes are unrelated).

## Bayes Theorem

Bayes' theorem, also known as Bayes' law or Bayes' rule, is a fundamental concept in probability theory and statistics. It allows us to update the probability of an event based on new evidence or information. Bayes' theorem is named after the Reverend Thomas Bayes, who first formulated it.

Suppose we have two events, A and B, and we want to find the probability of event A occurring given that event B has occurred. In mathematical notation, we denote this conditional probability as P(A|B), which reads as "the probability of A given B."

Bayes' theorem states:

P(A|B) = (P(B|A) × P(A)) / P(B)

Where:

* P(A|B) is the conditional probability of event A occurring given that event B has occurred.
* P(B|A) is the conditional probability of event B occurring given that event A has occurred.
* P(A) is the prior probability of event A occurring (i.e., the probability of A before considering any new evidence).
* P(B) is the prior probability of event B occurring (i.e., the probability of B before considering any new evidence).

In words, Bayes' theorem tells us that the probability of A given B is equal to the likelihood of B given A (P(B|A)) times the prior probability of A (P(A)), divided by the prior probability of B (P(B)). It helps us incorporate new information (P(B|A)) into our initial beliefs (P(A)) to update the probability of the event of interest (P(A|B)).

# Part 4 – Probability Distribution

## Mean, Variance, Std. Deviation

In statistics, the mean, variance, and standard deviation are measures of central tendency and variability used to describe the distribution of data. These measures can be calculated for both samples and populations.

* Sample: A sample is a subset of data taken from a larger group or population. It is used to make inferences about the entire population.
* Population: The population refers to the entire group of individuals, items, or events that we are interested in studying. It includes all possible elements with the characteristics of interest.

1. Population Mean (μ): The population mean represents the average value of all the individuals in the entire population. It is denoted by the Greek letter "μ."

Population Mean Formula: μ = (Σx) / N

Where: μ = Population mean Σx = Sum of all the individual values in the population N = Total number of individuals in the population

1. Sample Mean (x̄): The sample mean represents the average value of a subset of individuals (the sample) taken from the entire population. It is denoted by the symbol "x̄."

Sample Mean Formula: x̄ = (Σx) / n

Where: x̄ = Sample mean Σx = Sum of all the individual values in the sample n = Sample size (number of individuals in the sample)

1. Population Variance (σ^2): The population variance measures the average squared deviation of each individual value from the population mean. It is denoted by "σ^2."

Population Variance Formula: σ^2 = Σ[(x - μ)^2] / N

Where: σ^2 = Population variance Σ = Summation symbol (x - μ)^2 = Squared difference between each individual value (x) and the population mean (μ) N = Total number of individuals in the population

1. Sample Variance (s^2): The sample variance measures the average squared deviation of each individual value from the sample mean. It is denoted by "s^2."

Sample Variance Formula: s^2 = Σ[(x - x̄)^2] / (n - 1)

Where: s^2 = Sample variance Σ = Summation symbol (x - x̄)^2 = Squared difference between each individual value (x) and the sample mean (x̄) n = Sample size (number of individuals in the sample)

Note that in the formula for the sample variance, we divide by (n - 1) instead of n to account for the fact that the sample mean is used to estimate the population mean, introducing some bias in the calculation.

1. Population Standard Deviation (σ): The population standard deviation is the square root of the population variance. It is denoted by "σ."

Population Standard Deviation Formula: σ = √σ^2

1. Sample Standard Deviation (s): The sample standard deviation is the square root of the sample variance. It is denoted by "s."

Sample Standard Deviation Formula: s = √s^2

The sample mean, sample variance, and sample standard deviation are used when we have data from a subset (sample) of the entire population, and we want to estimate the corresponding population parameters (mean, variance, and standard deviation). If we have data from the entire population, we use the population mean, population variance, and population standard deviation directly.

## Probability Function

### Discrete Probability Function

A discrete probability function, also known as a probability mass function (PMF), is used to describe the probabilities of all possible outcomes of a discrete random variable. A discrete random variable takes on specific, distinct values with gaps between them. The probability mass function gives the probability of each value occurring.

For a discrete random variable X, the probability mass function is denoted as P(X = x), where "x" represents a specific value of the random variable X. The function must satisfy the following properties:

1. Non-negativity: P(X = x) ≥ 0 for all x.
2. Summation: The sum of the probabilities over all possible values of X must equal 1: Σ P(X = x) = 1.

For example, let's consider rolling a fair six-sided die. The discrete probability function for this situation would be: P(X = 1) = 1/6 P(X = 2) = 1/6 P(X = 3) = 1/6 P(X = 4) = 1/6 P(X = 5) = 1/6 P(X = 6) = 1/6

**Continuous Probability Function:** A continuous probability function, also known as a probability density function (PDF), is used to describe the probabilities of different intervals or ranges of values for a continuous random variable. A continuous random variable can take on any value within a given range, and there are no gaps between the possible values.

For a continuous random variable X, the probability density function is denoted as f(x), where "x" represents a specific value of the random variable X. The function must satisfy the following properties:

1. Non-negativity: f(x) ≥ 0 for all x.
2. Total area under the curve: The integral of the PDF over the entire range of X must equal 1: ∫ f(x) dx = 1.

Since the probability of a single point for a continuous random variable is infinitesimally small, we calculate probabilities for intervals. The probability of X falling within a certain interval [a, b] is given by the integral of the PDF over that interval: P(a ≤ X ≤ b) = ∫[a, b] f(x) dx.

An example of a continuous probability function is the standard normal distribution, which is often denoted as f(x) = (1/√(2π)) \* e^(-x^2/2). The probability of X falling within a certain range can be calculated using integration.

These probability functions are essential tools in probability theory and are used to analyze and predict outcomes in a wide range of fields, including statistics, economics, physics, and more.

## Discrete Distributions

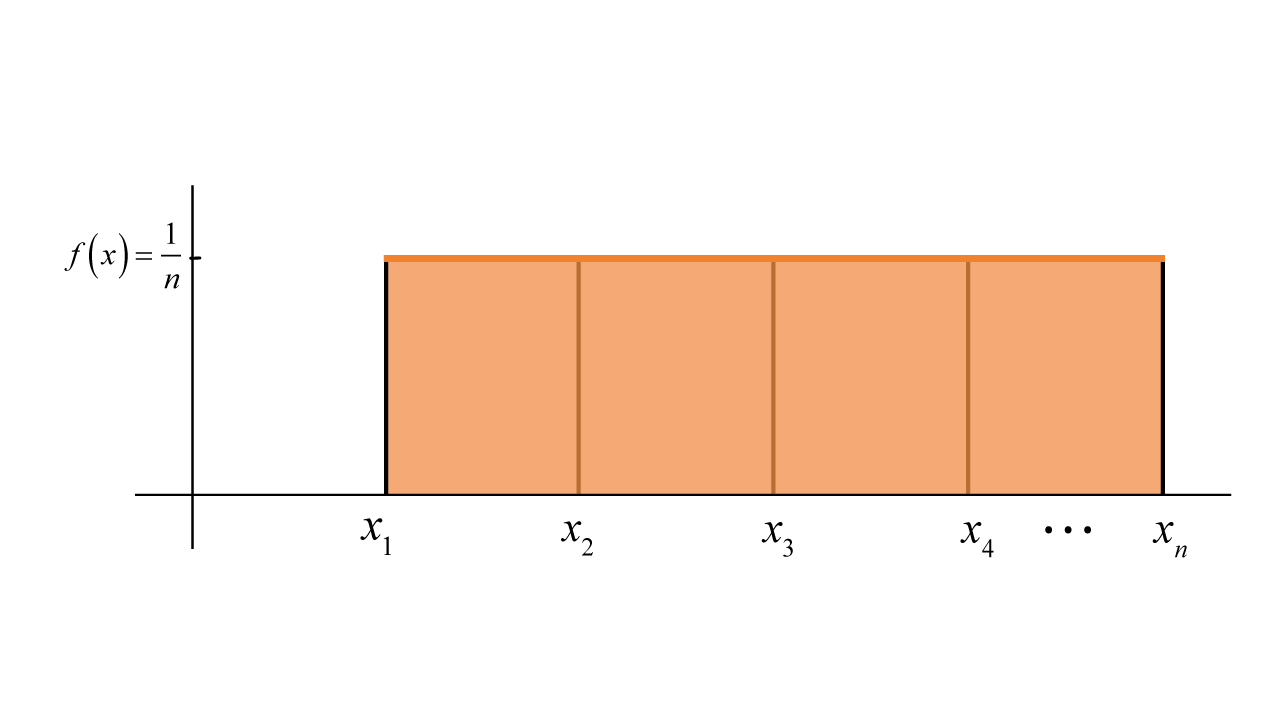
Sure, let's delve deeper into four common discrete distributions: the uniform distribution, Bernoulli distribution, binomial distribution, and Poisson distribution.

### Uniform Distribution

The uniform distribution is a simple discrete distribution where all possible outcomes are equally likely. In other words, each value has the same probability of occurrence. The uniform distribution is typically denoted as U(a, b), where "a" is the minimum value and "b" is the maximum value of the range.

Probability Mass Function (PMF) for the uniform distribution: P(X = x) = 1 / (b - a + 1), for x ∈ {a, a + 1, a + 2, ..., b}

Example: Rolling a fair six-sided die, where each outcome from 1 to 6 is equally likely.



### Bernoulli Distribution

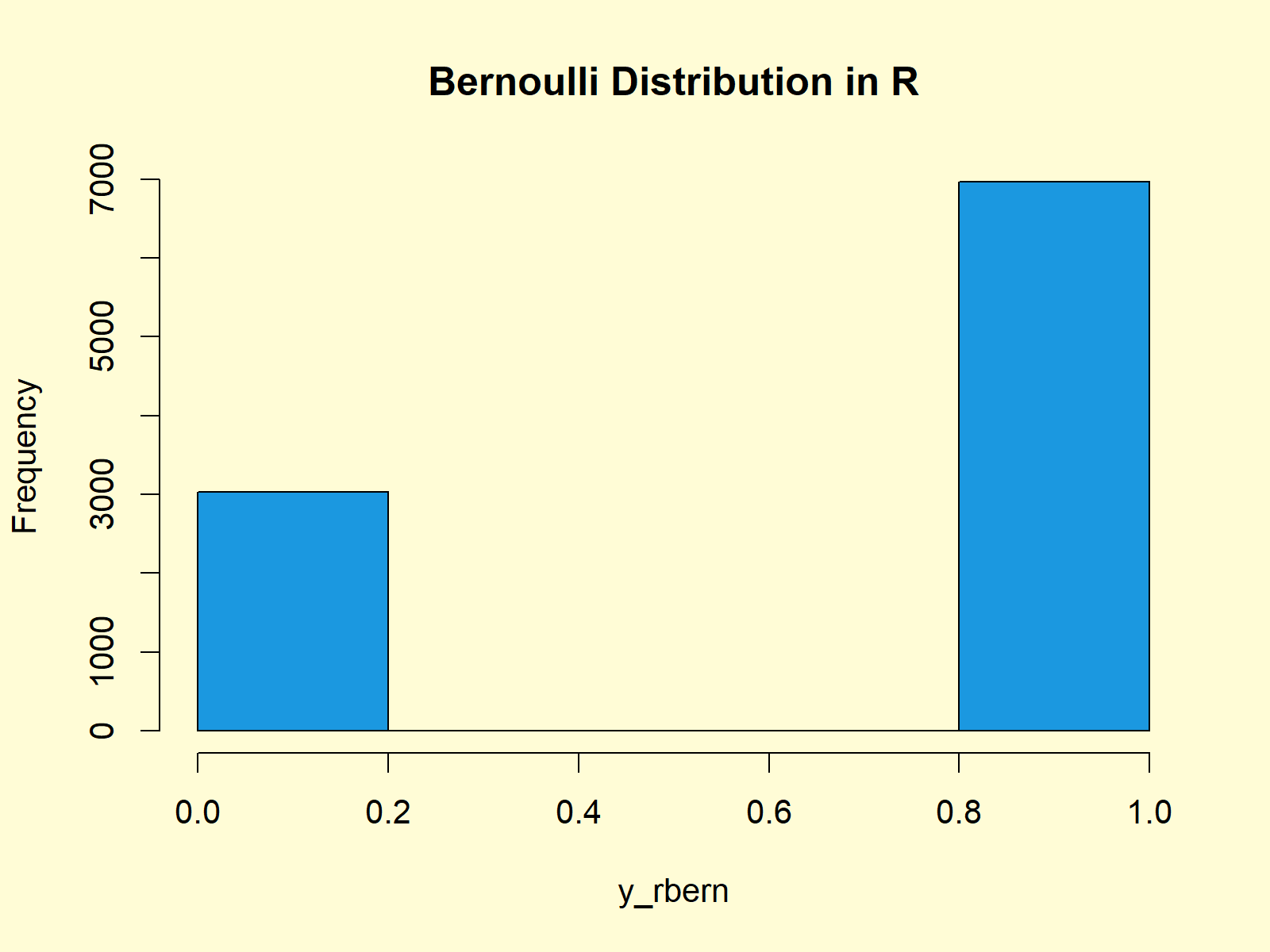
The Bernoulli distribution models a random experiment with two possible outcomes: success (usually denoted as 1) with probability "p" and failure (usually denoted as 0) with probability "1 - p."

Probability Mass Function (PMF) for the Bernoulli distribution: P(X = x) = p^x \* (1 - p)^(1 - x), for x ∈ {0, 1}

Mean E(x) = p

Var(x) = p \* (1-p)

Example: Tossing a coin, where "Heads" (success) occurs with probability p and "Tails" (failure) occurs with probability 1 - p.



### Binomial Distribution

The binomial distribution models the number of successes in a fixed number "n" of independent Bernoulli trials, each with a probability "p" of success.

Probability Mass Function (PMF) for the binomial distribution:

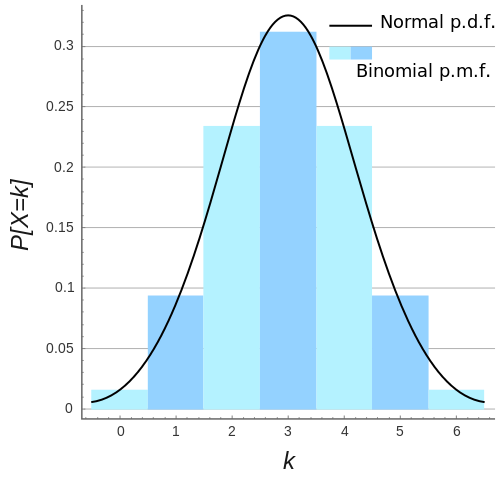
P(X = x) = C(n, x) \* p^x \* (1 - p)^(n - x), for x ∈ {0, 1, 2, ..., n}

Where C(n, x) is the binomial coefficient, given by C(n, x) = n! / (x! \* (n - x)!).

E(X) = n \* p

Var(X) = n \* p \* (1-p)

Example: The number of "Heads" obtained when flipping a coin 10 times, where the probability of getting "Heads" in a single toss is p.



### Poisson Distribution

The Poisson distribution models the number of events that occur in a fixed interval of time or space when the events occur at a constant average rate "λ."

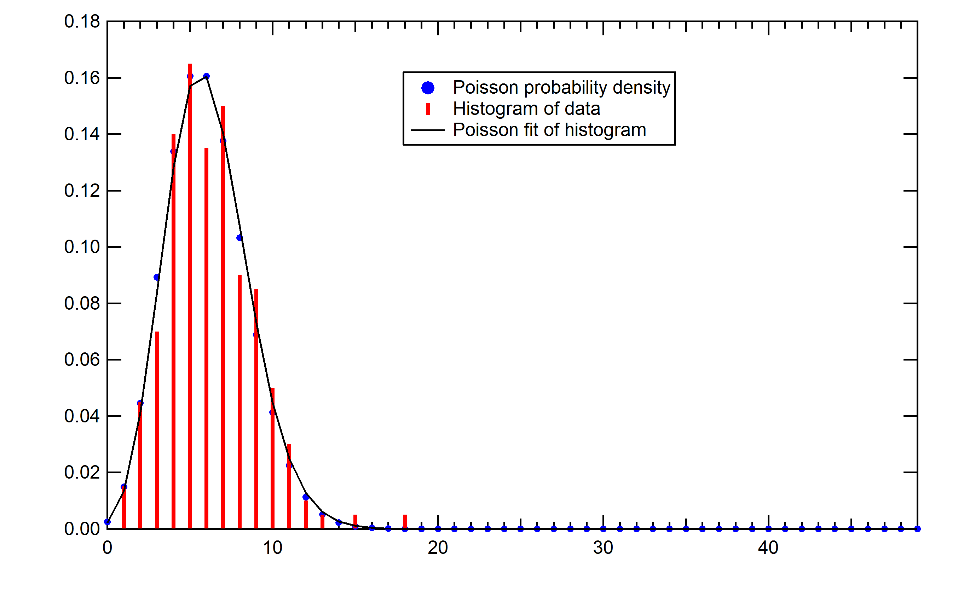
Probability Mass Function (PMF) for the Poisson distribution:

P(X = x) = (e^(-λ) \* λ^x) / x!, for x ∈ {0, 1, 2, ...}

E(X) = λ

Var(X) = λ

Example: The number of customers arriving at a store in an hour, given that the average arrival rate is λ customers per hour.



These discrete distributions are fundamental in probability theory and statistics, and they find applications in various fields such as biology, economics, engineering, and more, for modeling and analyzing random phenomena with discrete outcomes.

## Continuous Distributions

Variables that can take infinitely many consecutive values. Graph is a smooth curve.

Given a continuous random variable X with a probability density function (PDF) denoted as f(x), the likelihood of an interval [a, b] is given by integrating the PDF over that interval. Mathematically, the likelihood P(a ≤ X ≤ b) is expressed as follows:

P(a ≤ X ≤ b) = ∫[a, b] f(x) dx

Here, ∫[a, b] represents the definite integral of f(x) with respect to x, taken over the interval [a, b]. The result of this integration represents the probability that X falls within the interval [a, b].

It's important to note that the likelihood of an interval will be between 0 and 1, as it represents the probability of the event occurring within that range.

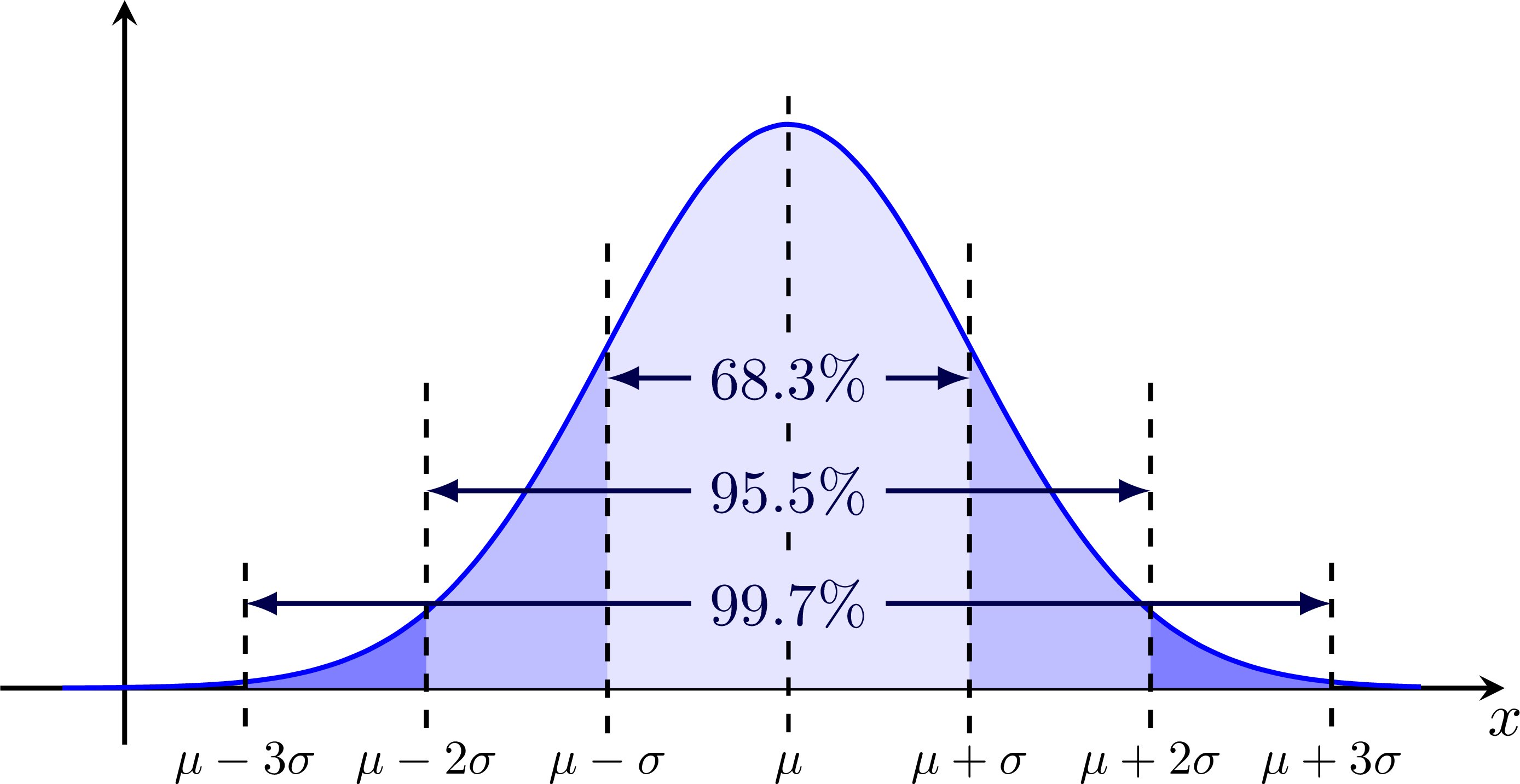
### Normal Distribution (Gaussian Distribution)

PDF: f(x) = (1 / (σ \* √(2π))) \* e^(-(x - μ)^2 / (2σ^2))

Mean (μ): μ

Variance (σ^2): σ^2

Explanation: The normal distribution is symmetric and bell-shaped, characterized by the mean (μ) and variance (σ^2), and widely used due to the central limit theorem.



### Standard Normal Distribution

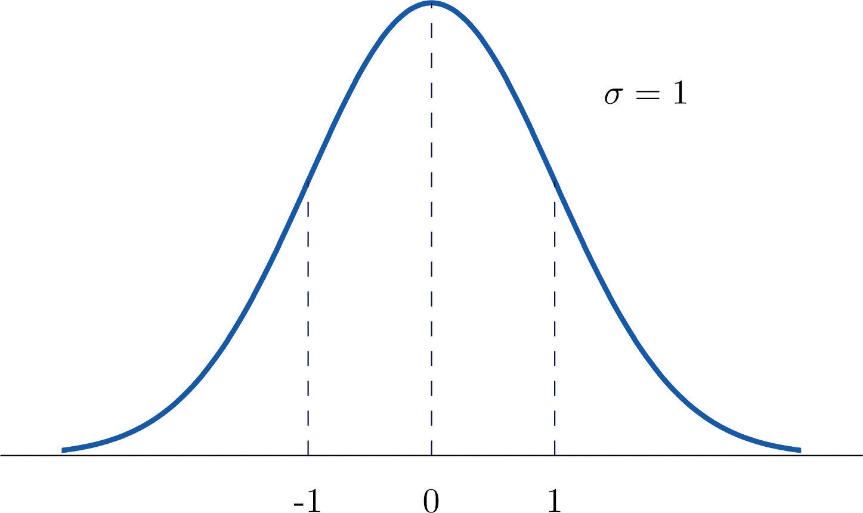
PDF: f(z) = (1 / √(2π)) \* e^(-z^2 / 2)

Mean (μ): μ = 0

Variance (σ^2): σ^2 = 1

Explanation: The standard normal distribution is a special case of the normal distribution with mean 0 and variance 1, often used for standardizing data.

To convert normal distribution to standard normal distribution:  
z = (x - μ)/ σ



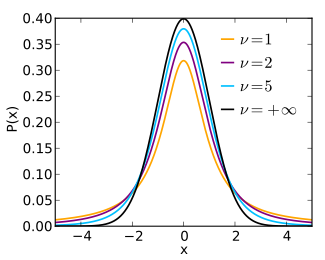
### Student's t-Distribution (t-Distribution)

PDF: Formula involving degrees of freedom (df) determines the shape.

Mean (μ): For df > 1, the mean is 0.

Variance (σ^2): For df > 2, the variance is df / (df - 2).

Explanation: The t-distribution is used for small sample sizes and when the population standard deviation is unknown in statistical tests.



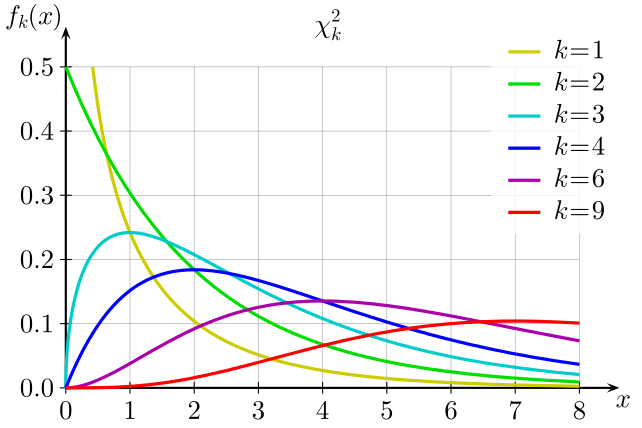
### Chi-Squared DistributionPDF

Formula involving degrees of freedom (df) determines the shape.

Mean (μ): μ = df

Variance (σ^2): σ^2 = 2 \* df

Explanation: The chi-squared distribution is used in tests involving the sum of squares of independent standard normal random variables.



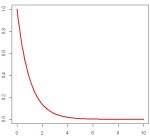
### Exponential Distribution

PDF: f(x) = λ \* e^(-λx) for x ≥ 0

Mean (μ): μ = 1 / λ

Variance (σ^2): σ^2 = 1 / λ^2

Explanation: The exponential distribution models time between events in a Poisson process, often used in survival analysis and reliability engineering.



### Logistic Distribution

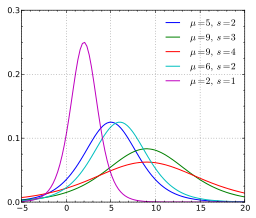
PDF: f(x) = (e^(-x)) / (1 + e^(-x))^2

Mean (μ): μ

Variance (σ^2): σ^2 = (π^2 / 3) , if s = 0 ;

Else = (s^2 \* π^2 )/ 3),

Explanation: The logistic distribution is an S-shaped distribution used for modeling growth processes and in logistic regression.It shows how continuous variable can affect probability of binary outcome.



# Part 5 – Descriptive Statistics

## Measures of Central Tendency

**Mean:** The arithmetic average of all the values in the dataset. It is calculated by summing up all the values and dividing by the number of data points. The mean is sensitive to extreme values, and it represents the "center" of the data distribution.

**Median**: The middle value of the dataset when arranged in ascending or descending order. It separates the lower and upper halves of the data and is less affected by extreme values. The median is a robust measure of central tendency.

**Mode:** The value that appears most frequently in the dataset. A dataset can have one mode (unimodal) or multiple modes (bimodal, trimodal, etc.). It is particularly useful for categorical data.

## Measures of Variability (Dispersion)

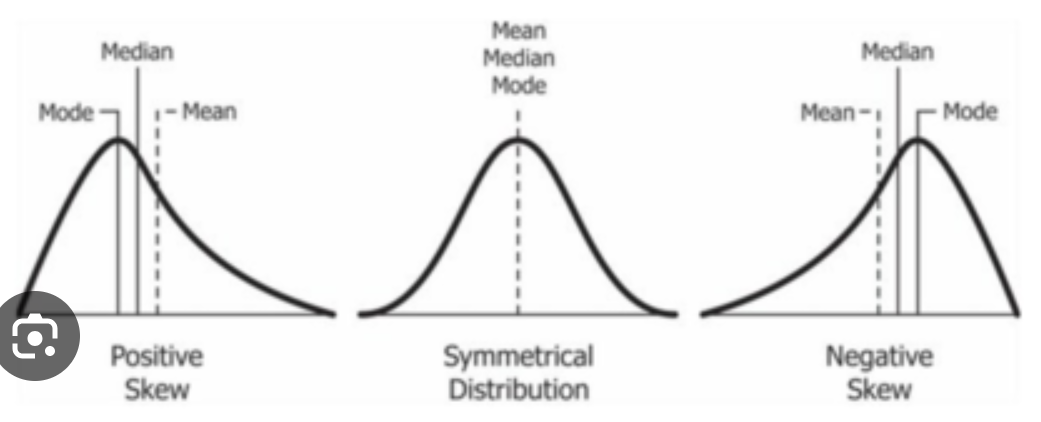
**Range**: The difference between the maximum and minimum values in the dataset, providing a measure of the spread of the data. It is easy to calculate but is sensitive to outliers.

**Variance**: The average of the squared differences between each data point and the mean. It quantifies how much the data points deviate from the mean. Larger variance indicates more spread-out data points.

**Standard Deviation:** The square root of the variance, providing a more interpretable measure of data dispersion. It is widely used because it is in the same unit as the original data.

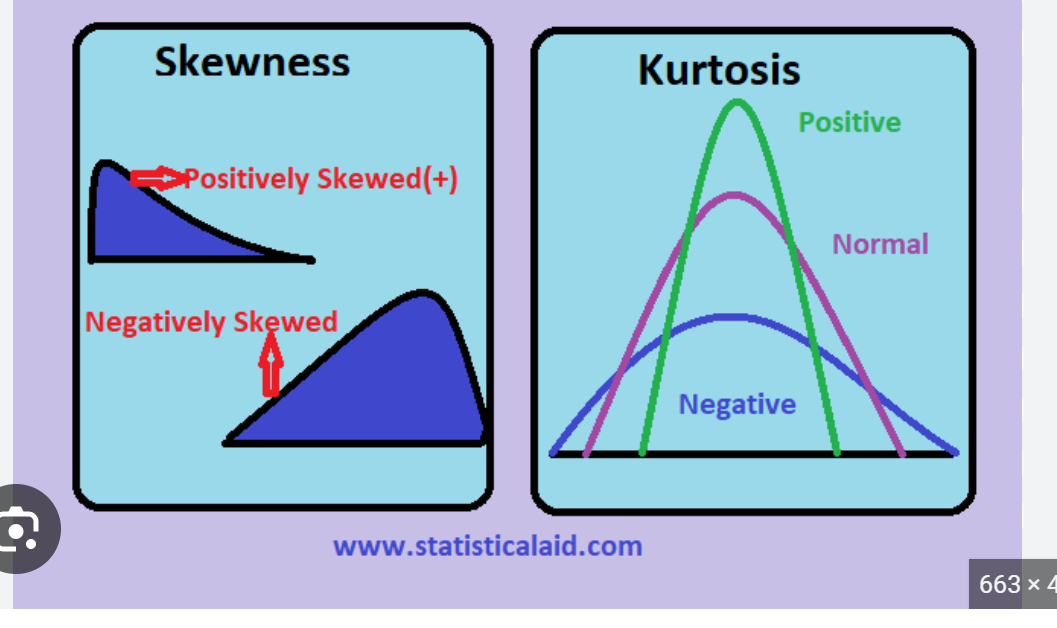
Measures of Shape and Distribution:

**Skewness:** A measure of the asymmetry of the data distribution. Positive skewness indicates a longer right tail, meaning the data is skewed to the right. Negative skewness indicates a longer left tail, and symmetric data has skewness close to 0.



Kurtosis: A measure of the "tailedness" of the data distribution. High kurtosis indicates heavier tails compared to a normal distribution, while low kurtosis indicates lighter tails.

* Kurtosis greater than 3: Leptokurtic
* Kurtosis less than 3: Platykurtic
* Kurtosis equal to 3: Mesokurtic

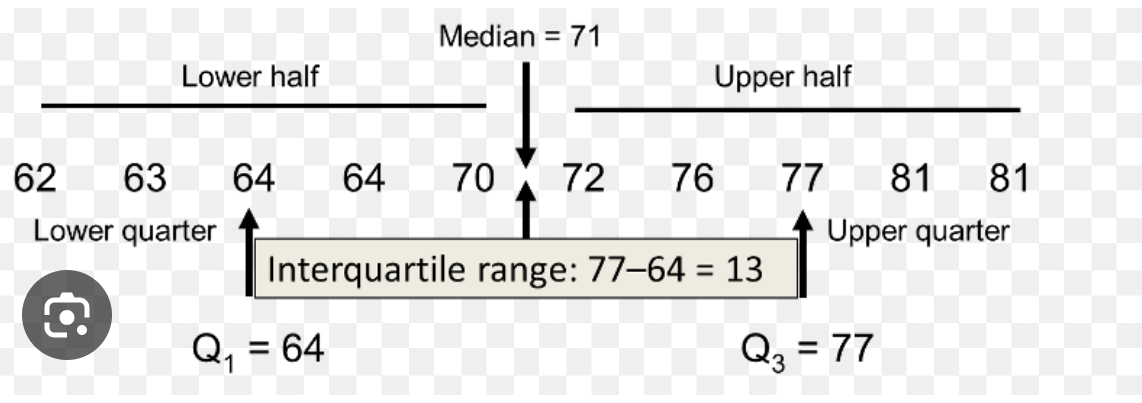


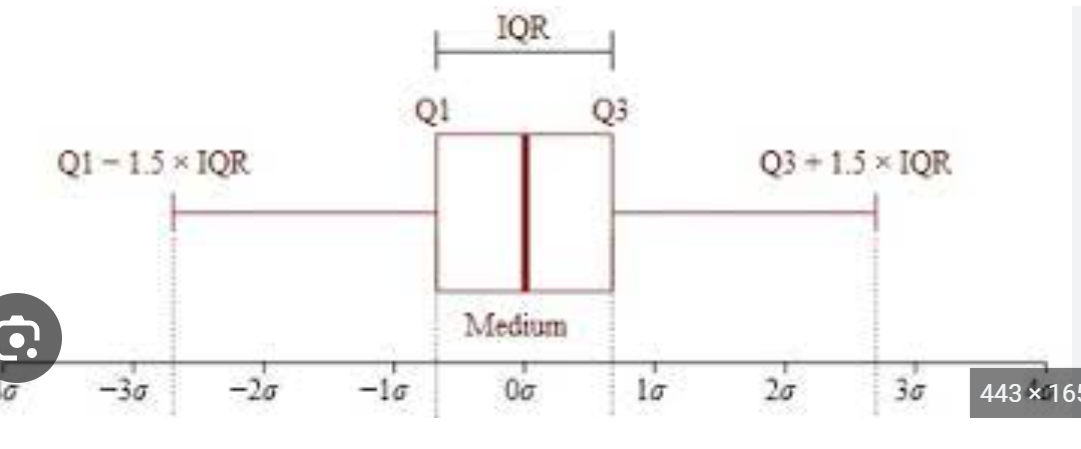
## Percentiles

Percentiles divide the data into 100 equal parts. The kth percentile represents the value below which k% of the data falls. For example, the 25th percentile (Q1) is the value below which 25% of the data falls, and the 75th percentile (Q3) is the value below which 75% of the data falls.

Interquartile Range (IQR):

The IQR is the range of the middle 50% of the data, which is the difference between the third quartile (Q3) and the first quartile (Q1). It is a measure of spread that is not affected by extreme values and is useful for identifying outliers.





## Correlation and Covariance

**Covariance:** Covariance measures the degree to which two variables change together. It indicates the extent of the linear relationship between the variables. If the covariance is positive, it means that when one variable increases, the other tends to increase as well. Conversely, if the covariance is negative, it indicates that when one variable increases, the other tends to decrease. A covariance of zero suggests that there is no linear relationship between the variables.

Mathematically, the covariance (cov) between two random variables X and Y with n data points can be calculated as:

cov(X, Y) = Σ [(X\_i - X̄) \* (Y\_i - Ȳ)] / n

Where:

* X̄ and Ȳ are the means of X and Y, respectively.
* (X\_i - X̄) and (Y\_i - Ȳ) are the deviations from the mean for each data point.

The main limitation of covariance is that its magnitude is not standardized, making it difficult to interpret the strength of the relationship. Therefore, it is often more informative to use correlation.

**Correlation:** Correlation is a standardized measure of the linear relationship between two variables. It scales the covariance by the product of the standard deviations of the two variables, resulting in a value between -1 and 1. A correlation of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship (although there could still be other types of relationships).

The Pearson correlation coefficient (r) is the most commonly used measure of correlation:

r = cov(X, Y) / (σ\_X \* σ\_Y)

Where:

* cov(X, Y) is the covariance between X and Y.
* σ\_X and σ\_Y are the standard deviations of X and Y, respectively.

Correlation is a powerful tool to assess the strength and direction of the relationship between two variables, and it allows for easy comparison between different pairs of variables.

# Part 6 – Inferential Statistics

Inferential statistics is a branch of statistics that deals with drawing conclusions and making inferences about a population based on data collected from a sample. The primary goal of inferential statistics is to use sample data to make predictions or generalizations about a larger population.

## Estimation

Estimation involves using sample data to estimate unknown population parameters. Point estimates provide a single value as the best guess for the population parameter, while interval estimates give a range of values within which the parameter is likely to lie.

## Frequency Distribution

Frequency distribution, also known as frequency table, is a statistical representation of the number of occurrences of different values or categories in a dataset.

For qualitative data (categorical), the frequency distribution involves counting the number of occurrences of each category or class.

For example, let's say you have a dataset of favorite colors of individuals:

| **Favorite Color** | **Frequency** |
| --- | --- |
| Blue | 12 |
| Red | 8 |
| Green | 5 |
| Yellow | 3 |

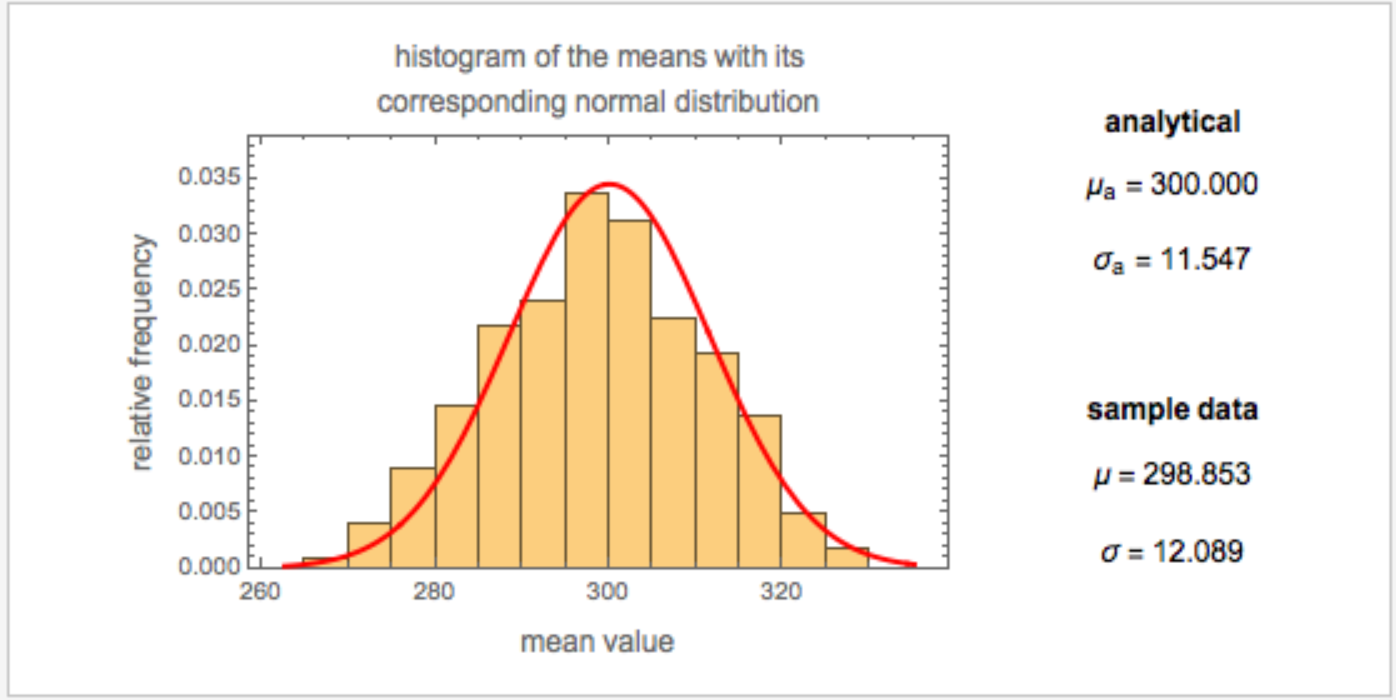
For quantitative data (numeric), the frequency distribution involves grouping the data into intervals or bins and counting the number of data points that fall within each interval.

Here's an example of a frequency distribution for quantitative data (exam scores):

| **Score Range** | **Frequency** |
| --- | --- |
| 60 - 69 | 8 |
| 70 – 79 | 12 |
| 80 – 89 | 15 |
| 90 – 99 | 10 |

## Central Limit Theorem

### Sample Mean Distribution



The distribution of sample means, also known as the sampling distribution of the mean, is a theoretical probability distribution that represents all possible sample means that could be obtained from a given population when drawing samples of a fixed size.

Here's how the distribution of sample means is formed:

1. Select Samples: Take multiple random samples (with replacement) from the population, each of the same fixed sample size.
2. Calculate Sample Mean: For each sample, calculate the mean of the data points in that sample.
3. Record Sample Means: Create a list or a histogram of all the sample means obtained from different samples.

The key idea is that as you take more and more samples from the population and calculate their means, the distribution of those sample means will start to approximate a normal distribution, regardless of the shape of the original population distribution. This phenomenon is described by the Central Limit Theorem.

The Central Limit Theorem states that if you have a large enough sample size (typically n ≥ 30), the distribution of sample means will be approximately normally distributed, even if the population from which the samples are drawn is not normally distributed.

The mean of the sampling distribution of the sample means will be equal to the population mean, and the standard deviation (standard error) will be equal to the population standard deviation divided by the square root of the sample size.

#### Standard Error of the Mean (SE or SEM)

For a sample mean (x̄) with sample size (n) drawn from a population with standard deviation (σ):

SE = σ / √n

If the population standard deviation (σ) is unknown and estimated using the sample standard deviation (s), the formula becomes:

SE = s / √n

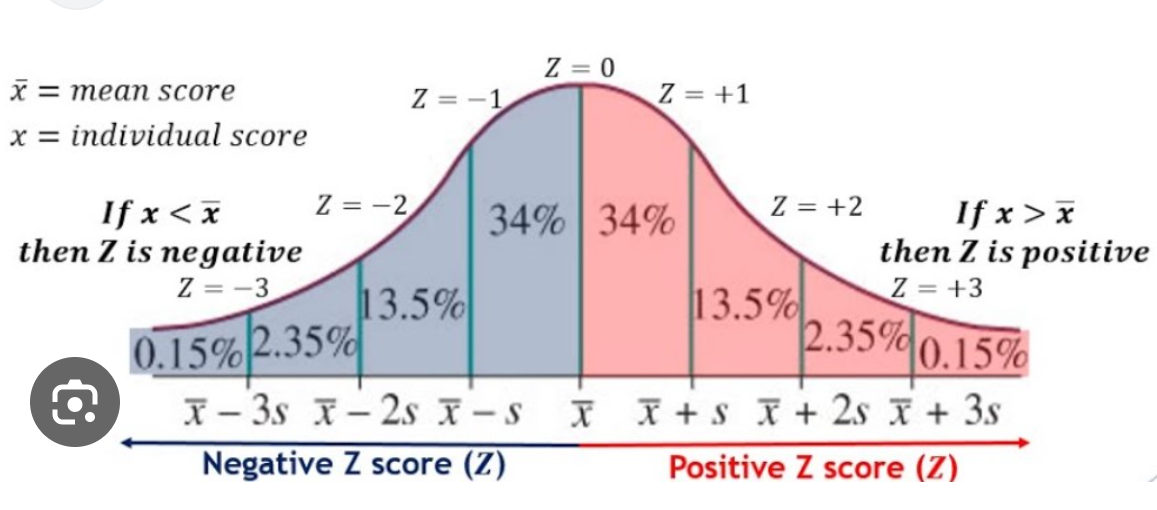
### Properties of a Normal Curve

The normal curve, also known as the Gaussian distribution, is a specific type of probability distribution that is characterized by several important properties. It is a continuous and symmetric probability distribution with a bell-shaped curve. Here are the key properties of a normal curve:

1. Symmetry: The normal curve is perfectly symmetrical around its mean (μ). This means that the probability of obtaining a value to the left of the mean is the same as the probability of obtaining a value to the right of the mean.
2. Unimodality: The normal curve is unimodal, which means it has a single peak. The peak corresponds to the mean of the distribution.
3. Asymptotic: The tails of the normal curve extend infinitely in both directions along the x-axis. However, the curve gets increasingly closer to the x-axis but never touches it.
4. Mean, Median, and Mode: The mean, median, and mode of a normal distribution are all equal and located at the center of the distribution.
5. Empirical Rule: The normal curve follows the empirical rule, also known as the 68-95-99.7 rule. Approximately 68% of the data falls within one standard deviation (σ) of the mean, about 95% falls within two standard deviations, and nearly 99.7% falls within three standard deviations.
6. Standard Deviation: The spread or dispersion of the normal curve is determined by the standard deviation (σ). A larger standard deviation results in a wider and flatter curve, while a smaller standard deviation leads to a narrower and taller curve.
7. Area Under the Curve: The total area under the normal curve is equal to 1, representing the entire probability space. This means that the probability of obtaining any value within the range of the distribution is 1 or 100%.
8. Inflection Points: The points of inflection are located one standard deviation away from the mean. At these points, the curve changes from curving upwards to curving downwards or vice versa.

## Z Score

The z-score, also known as the standard score, is a statistical measure that quantifies how many standard deviations a data point is away from the mean of its distribution. It is used to standardize and compare data points from different distributions, allowing us to understand how each data point relates to the overall distribution.



The formula for calculating the z-score for a data point (x) in a distribution with mean (μ) and standard deviation (σ) is:

z = (x - μ) / σ

Where:

* z is the z-score of the data point.
* x is the individual data point whose z-score is being calculated.
* μ is the mean of the distribution.
* σ is the standard deviation of the distribution.

Key points about z-scores:

1. Interpretation: A positive z-score indicates that the data point is above the mean, while a negative z-score indicates it is below the mean. The magnitude of the z-score indicates how many standard deviations the data point is from the mean.
2. Standardization: By converting data to z-scores, different datasets with different units and scales can be standardized and compared on a common scale.
3. Mean and Z-Score: The z-score of the mean of a dataset is always 0 because (x - μ) = 0 when x equals the mean.
4. Standard Deviation and Z-Score: The z-score of a data point that is equal to the mean plus or minus one standard deviation is approximately ±1, respectively. This follows from the z-score formula.
5. Outliers: Z-scores can be used to identify outliers in a dataset. Data points with extreme z-scores (far from 0) may be considered outliers.
6. Probability and Z-Score: Z-scores are used to find probabilities associated with specific values in a normal distribution. By converting data points to z-scores, we can use standard normal distribution tables or statistical software to find probabilities for specific z-score values.

Example:

Let's calculate the z-scores for a simple dataset to illustrate how z-score standardization works. Consider the following dataset of exam scores:

Dataset: [75, 85, 90, 70, 80]

Step 1: Calculate the Mean (μ) μ = (75 + 85 + 90 + 70 + 80) / 5 = 80

Step 2: Calculate the Standard Deviation (σ) To calculate the standard deviation, first, find the squared differences between each data point and the mean, then compute the variance and finally take the square root to get the standard deviation.

Variance = [(75 - 80)² + (85 - 80)² + (90 - 80)² + (70 - 80)² + (80 - 80)²] / 5 Variance = [25 + 25 + 100 + 100 + 0] / 5 = 50

Standard Deviation (σ) = √50 ≈ 7.07

Step 3: Calculate Z-Scores Now, we can calculate the z-scores for each data point using the formula:

z = (x - μ) / σ

For each data point (x), calculate the z-score:

For x = 75: z = (75 - 80) / 7.07 ≈ -0.71

For x = 85: z = (85 - 80) / 7.07 ≈ 0.71

For x = 90: z = (90 - 80) / 7.07 ≈ 1.41

For x = 70: z = (70 - 80) / 7.07 ≈ -1.41

For x = 80: z = (80 - 80) / 7.07 ≈ 0

The resulting z-scores are approximately -0.71, 0.71, 1.41, -1.41, and 0, respectively.

Interpretation:

1. A z-score of -0.71 for the score 75 means that this data point is about 0.71 standard deviations below the mean.
2. A z-score of 0.71 for the score 85 means that this data point is about 0.71 standard deviations above the mean.
3. A z-score of 1.41 for the score 90 means that this data point is about 1.41 standard deviations above the mean.
4. A z-score of -1.41 for the score 70 means that this data point is about 1.41 standard deviations below the mean.
5. A z-score of 0 for the score 80 means that this data point is exactly at the mean.

These z-scores provide a standardized representation of the original data and allow for comparisons across different datasets or variables.

### Z Table

The z-table, also known as the standard normal table or the standard normal distribution table, comes into play when we want to find the probability associated with specific z-scores in a standard normal distribution. The standard normal distribution is a special case of the normal distribution with a mean (μ) of 0 and a standard deviation (σ) of 1.

The z-table provides the cumulative probabilities for various z-scores. It is a precomputed table that contains the area under the standard normal curve from the left tail (mean) up to a given z-score. Each entry in the table corresponds to a z-score, and the value in the table represents the probability (or percentage) of getting a z-score less than or equal to that value.

Using the z-table, you can find the probabilities associated with different z-scores in a standard normal distribution. These probabilities are often used in statistical calculations and hypothesis testing.

**For Example,**

z-score = 1.96, and you want to find the probability of getting a z-score less than or equal to 1.96 in the standard normal distribution.

Step 1: Locate the z-score in the z-table. The z-table typically provides values for z-scores with two decimal places. In our case, we have a z-score of 1.96.

Step 2: Find the row and column corresponding to the z-score in the z-table. In the z-table, find the row that starts with "1.9". Then, find the column that corresponds to the hundredths place "0.06". The intersection of this row and column will give us the probability associated with the z-score.

The probability of obtaining a z-score less than or equal to 1.96 in a standard normal distribution is approximately 0.975. In percentage terms, it's 97.5%.

Similarly, you can use the z-table to find probabilities for negative z-scores, which correspond to the left tail of the standard normal distribution. For example, if you want to find the probability of getting a z-score less than or equal to -1.96, you would look up the value in the z-table for z = -1.96, which is also approximately 0.025. You can also find value for positive Z score and subtract it from 1.(1 – 0.975) = 0.025.

Keep in mind that the z-table is specific to the standard normal distribution (mean = 0, standard deviation = 1). To use the z-table for non-standard normal distributions, you would first need to standardize the data by converting it to z-scores using the formula (z = (x - μ) / σ), and then use the z-table to find probabilities for the standard normal distribution.

### Area between two points on a table

Let's walk through an example of finding the area between two points on a standard normal distribution using z-scores.

Example: Suppose we have a standard normal distribution (mean = 0, standard deviation = 1), and we want to find the area between two z-scores: z1 = -1.5 and z2 = 1.5.

Step 1: Identify the two z-scores. z1 = -1.5 z2 = 1.5

Step 2: Use the z-table to find the probabilities. Using the z-table, we can find the probabilities associated with z1 and z2.

From the z-table, we find:

Probability at z1 (P(z < -1.5)) ≈ 0.0668 Probability at z2 (P(z < 1.5)) ≈ 0.9332

Step 3: Calculate the area between the two z-scores. To find the area between z1 and z2, subtract the probability associated with z1 from the probability associated with z2:

Area between z1 and z2 = Probability at z2 - Probability at z1 Area between z1 and z2 ≈ 0.9332 - 0.0668 Area between z1 and z2 ≈ 0.8664

Interpretation: The area between z1 = -1.5 and z2 = 1.5 on the standard normal distribution is approximately 0.8664. In percentage terms, it's 86.64%.

This means that about 86.64% of the data falls between z1 = -1.5 and z2 = 1.5 on the standard normal distribution.

The process of finding the area between two z-scores on a standard normal distribution using the z-table allows us to determine the proportion of data within a specific range of z-scores, which is useful in various statistical analyses and probability calculations.

### Finding Z score from probability

Here are the steps to find the z-score from a given probability using the z-table:

Step 1: Identify the probability (area) you want to find the z-score for. Let's say the given probability is P.

Step 2: Locate the probability in the z-table. Search for the closest value to the given probability in the z-table. The z-table typically provides probabilities with four decimal places, so you may need to round the probability to the nearest value in the table.

Step 3: Read the z-score from the z-table. Once you've located the closest value to the given probability in the z-table, read the corresponding z-score value. This z-score is the value you are looking for.

For example, let's say you want to find the z-score corresponding to a probability of P = 0.8508.

Step 1: Given probability P = 0.8508.

Step 2: Locate the probability in the z-table. In the z-table, the closest value to 0.8508 is 0.8515, which corresponds to a z-score of approximately 1.04(Row 1.00, colum 0.04).

Step 3: Read the z-score from the z-table. The z-score corresponding to a probability of 0.8508 is approximately 1.04.

Keep in mind that z-scores are standardized values, and they are specific to the standard normal distribution (mean = 0, standard deviation = 1). If you are working with a non-standard normal distribution, you may need to convert the given probability to a z-score using the appropriate mean and standard deviation for that distribution.

### Probability of Distribution of Sample Means

Mean of Sample = Population Mean

Standard error:

For a sample mean (x̄) with sample size (n) drawn from a population with standard deviation (σ):

SE = σ / √n

If the population standard deviation (σ) is unknown and estimated using the sample standard deviation (s), the formula becomes:

SE = s / √n

To calculate the z-score using sample means, you need to have the following information:

1. Sample Mean (x̄): The mean of the sample you are working with.
2. Population Mean (μ): The mean of the population from which the sample is drawn.
3. Population Standard Deviation (σ): The standard deviation of the population.

The formula to calculate the z-score using sample means is:

z = (x̄ - μ) / (σ / √n)

Where:

* z is the z-score.
* x̄ is the sample mean.
* μ is the population mean.
* σ is the population standard deviation.
* n is the sample size.

The z-score represents the number of standard deviations that the sample mean is away from the population mean. It standardizes the sample mean, allowing for comparison with the population mean on a common scale.

Here's an example to illustrate how to calculate the z-score using sample means:

Example: Suppose you have a sample of 50 students' exam scores. The sample mean is 75, and the population mean (μ) of all students in the school is 70. The population standard deviation (σ) is 8.

Sample Mean (x̄) = 75 Population Mean (μ) = 70 Population Standard Deviation (σ) = 8 Sample Size (n) = 50

Now, let's calculate the z-score:

z = (x̄ - μ) / (σ / √n) z = (75 - 70) / (8 / √50) z = 5 / (8 / 7.07) z ≈ 5 / 1.13 z ≈ 4.425

The z-score is approximately 4.425.

Interpretation: The sample mean of 75 is approximately 4.425 standard deviations above the population mean of 70.

The z-score helps to quantify the difference between the sample mean and the population mean in terms of standard deviations, providing a standardized measure of how representative the sample mean is of the population mean.

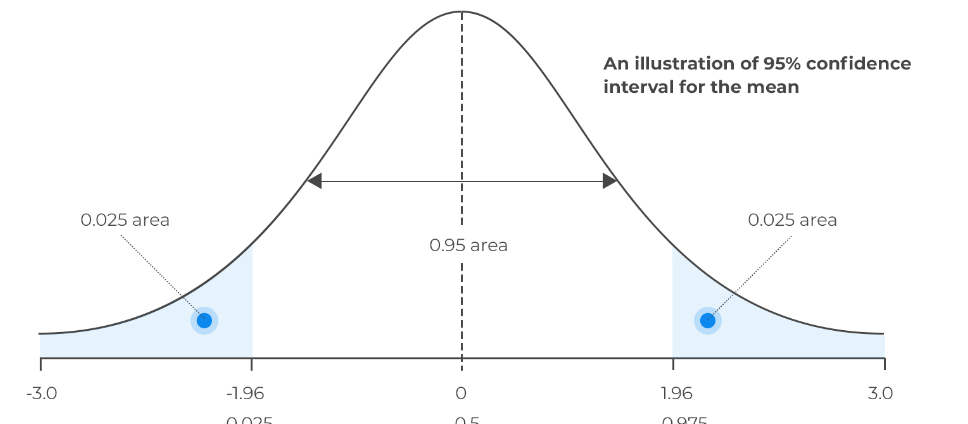
## Confidence Interval

A confidence interval is a range of values that is constructed around a sample statistic (such as the sample mean or sample proportion) to provide an estimate of the likely range of the corresponding population parameter (such as the population mean or population proportion). It is a fundamental concept in inferential statistics and is used to make inferences about the population based on sample data.

The confidence interval is characterized by two key components:

1. Point Estimate: The point estimate is the sample statistic itself, which serves as the best estimate of the population parameter. For example, the sample mean (x̄) is often used as the point estimate for the population mean (μ).
2. Margin of Error: The margin of error is the range of values added to and subtracted from the point estimate to create the interval. It represents the uncertainty or variability associated with the estimate. The margin of error is typically determined by the standard error of the sample statistic and the desired level of confidence.

The confidence level, often denoted as (1 - α), where α is the significance level (typically 0.05), represents the probability that the confidence interval will contain the true population parameter. Commonly used confidence levels include 90%( 1 - α = 0.9; α = 0.1), 95%, and 99%.



The formula to construct a confidence interval for a sample mean is:

Confidence Interval for Population Mean (μ) = x̄ ± (Z \* (σ / √n))

**(Z \* (σ / √n)) is Margin of Error(E)**

If population standard deviation not present, use sample deviation(s)

Where:

* x̄ is the sample mean.
* Z is the critical value from the standard normal distribution corresponding to the desired confidence level(α/2). Find Z score for probability value α/2.
* σ is the population standard deviation (or sample standard deviation if the population standard deviation is unknown).
* n is the sample size.

Example: Suppose we have a sample of 100 students and we want to estimate the average height of all students in the school. The sample mean height is 165 cm, and the population standard deviation is known to be 8 cm.

Sample Mean (x̄) = 165 cm Population Standard Deviation (σ) = 8 cm Sample Size (n) = 100 Desired Confidence Level = 95% (α = 0.05)

Step 1: Find the critical value (Z) for the desired confidence level. For a 95% confidence level, the critical value (Z) from the standard normal distribution is approximately 1.96. This value corresponds to a two-tailed test with α = 0.05.

Step 2: Calculate the margin of error (ME). The margin of error is given by: ME = Z \* (σ / √n) ME = 1.96 \* (8 / √100) ≈ 1.568

Step 3: Construct the confidence interval. The confidence interval is given by: CI = x̄ ± ME CI = 165 ± 1.568

Step 4: Interpret the confidence interval. The 95% confidence interval for the average height of all students in the school is approximately (163.432, 166.568) cm.

Interpretation: We are 95% confident that the true average height of all students in the school lies between 163.432 cm and 166.568 cm. This means that if we were to repeat the sampling and construct a confidence interval for the population mean in this manner, about 95% of the intervals would contain the true average height.

## Hypothesis Testing

Hypothesis testing is a statistical method used to make inferences about a population based on sample data. It involves formulating two competing hypotheses, the null hypothesis (H0) and the alternative hypothesis (Ha), and then using sample data to determine which hypothesis is more likely or supported by the evidence.

The main steps in hypothesis testing are as follows:

1. Formulate the Null Hypothesis (H0): The null hypothesis is a statement of no effect, no difference, or no change in the population. It represents the status quo or the commonly accepted belief. It is denoted as H0.

For example, if we want to test whether a new drug has a significant effect on reducing blood pressure, the null hypothesis might state that the drug has no effect on blood pressure (i.e., there is no difference in blood pressure between the treated and control groups).

1. Formulate the Alternative Hypothesis (Ha): The alternative hypothesis is the statement that contradicts the null hypothesis. It represents what the researcher wants to find evidence for or what they believe to be true. It is denoted as Ha.

In the above example, the alternative hypothesis (Ha) would state that the new drug has a significant effect on reducing blood pressure (i.e., there is a difference in blood pressure between the treated and control groups).

1. Set the Significance Level (α): The significance level (α) is the probability of rejecting the null hypothesis when it is true. It is the threshold for determining whether the evidence against the null hypothesis is strong enough to reject it.
2. Commonly used significance levels are 0.05 (5%) and 0.01 (1%). A significance level of 0.05 means that there is a 5% chance of making a Type I error, which is the probability of rejecting the null hypothesis when it is true.
3. Collect and Analyze the Data: Collect a sample from the population of interest and analyze the data to obtain the sample statistic (e.g., sample mean, sample proportion) that will be used in the hypothesis test.
4. Perform the Hypothesis Test: Using the sample data and the chosen test statistic, perform the hypothesis test to calculate the test statistic value and the corresponding p-value. The p-value represents the probability of observing the sample data or more extreme data if the null hypothesis is true.
5. Make a Decision: Compare the p-value to the significance level (α). If the p-value is less than or equal to α, you reject the null hypothesis in favor of the alternative hypothesis. If the p-value is greater than α, you fail to reject the null hypothesis.
6. Draw Conclusion: Based on the decision made in Step 6, draw a conclusion about the population. If the null hypothesis is rejected, it suggests evidence in favor of the alternative hypothesis. If the null hypothesis is not rejected, there is not enough evidence to support the alternative hypothesis.

### Z and T Test

Both the Z-test and the T-test are statistical methods used for hypothesis testing, but they are applied under different circumstances depending on the information available about the population and sample data.

#### Z-Test

The Z-test is used when you have a large sample size (typically n ≥ 30) or when you know the population standard deviation (σ). It is appropriate for hypothesis testing when the population standard deviation is known or can be assumed. The Z-test is based on the standard normal distribution (Z-distribution).

The formula for the Z-test statistic is:

Z = (x̄ - μ) / (σ / √n)

where:

* Z is the Z-test statistic.
* x̄ is the sample mean.
* μ is the population mean (specified in the null hypothesis).
* σ is the population standard deviation.
* n is the sample size.

The Z-test compares the sample mean to the population mean under the null hypothesis and determines whether the difference is statistically significant.

In both the Z-test and T-test, the test statistic is compared to a critical value from the respective distribution (Z-distribution or t-distribution) based on the desired significance level (α) to determine whether to reject or fail to reject the null hypothesis. The smaller the p-value associated with the test statistic, the stronger the evidence against the null hypothesis. If the p-value is less than the chosen significance level (α), the null hypothesis is rejected in favor of the alternative hypothesis. Otherwise, the null hypothesis is not rejected.

Top of Form

**p-value:** It is a probability that quantifies the strength of evidence against the null hypothesis. In hypothesis testing, we compare the sample data to what would be expected under the assumption of the null hypothesis, and the p-value tells us the likelihood of observing such extreme results if the null hypothesis is true.

#### Two Sample Z Test

The two-sample z-test is a hypothesis test used to compare the means of two independent groups to determine if they are significantly different from each other. It is suitable when the sample sizes are large, and the population standard deviations are known or the sample sizes are large enough to assume normality.

Here are the steps to conduct a two-sample z-test:

Step 1: State the Hypotheses:

* Null Hypothesis (H0): The means of the two groups are equal (μ1 = μ2).
* Alternative Hypothesis (Ha): The means of the two groups are not equal (μ1 ≠ μ2) or one mean is greater than the other (μ1 > μ2 or μ1 < μ2).

Step 2: Set the Significance Level (α): Choose a significance level (commonly 0.05) to determine the critical region for the test.

Step 3: Calculate the Test Statistic (z): The z-test statistic is given by the formula:

z = (x̄1 - x̄2) / √(σ1² / n1 + σ2² / n2)

where x̄1 and x̄2 are the sample means of the two groups, σ1 and σ2 are the population standard deviations (or sample standard deviations if the sample size is large), and n1 and n2 are the sample sizes of the two groups.

Step 4: Determine the Critical Region or p-value:

* For a two-tailed test: Find the critical z-values (zα/2 and -zα/2) based on the chosen significance level (α/2) from the standard normal distribution table. The critical region is outside these z-values.
* For a one-tailed test: Find the critical z-value (zα or -zα) based on the chosen significance level α from the standard normal distribution table. The critical region is either to the right or left of this z-value.

Alternatively, you can find the p-value corresponding to the calculated z-value using the standard normal distribution table or statistical software.

Step 5: Make a Decision:

* If the calculated z-value falls in the critical region or the p-value is less than the chosen significance level (α), reject the null hypothesis (H0) and conclude that there is a significant difference between the means of the two groups.
* If the calculated z-value does not fall in the critical region and the p-value is greater than the chosen significance level (α), fail to reject the null hypothesis (H0) and conclude that there is no significant difference between the means of the two groups.

Remember that the two-sample z-test assumes that the sample sizes are large enough or the data is normally distributed. If these assumptions are not met, you may consider using a two-sample t-test (for small sample sizes) or non-parametric tests like the Mann-Whitney U test.

#### Types of Error

In the context of hypothesis testing, there are two types of errors that can occur: Type I error and Type II error. These errors represent the incorrect decisions made during the hypothesis testing process.

1. Type I Error (False Positive): A Type I error occurs when we reject the null hypothesis (H0) when it is actually true. In other words, we conclude that there is a significant effect or difference in the population when there is, in fact, no such effect or difference. This error is also known as a false positive or a "alpha error."

The probability of committing a Type I error is denoted by the symbol α (alpha) and is set as the significance level in hypothesis testing. For example, if the significance level is 0.05 (5%), it means there is a 5% chance of making a Type I error.

1. Type II Error (False Negative): A Type II error occurs when we fail to reject the null hypothesis (H0) when it is actually false. In other words, we fail to detect a significant effect or difference in the population when there is, in fact, a real effect or difference. This error is also known as a false negative or a "beta error."

The probability of committing a Type II error is denoted by the symbol β (beta). The complement of β, denoted as 1 - β, is called the "power" of the test, which represents the probability of correctly rejecting the null hypothesis when it is false.

The relationship between Type I and Type II errors is often inverse: as you reduce the probability of one type of error, the probability of the other type of error may increase and vice versa. It is essential to strike a balance between these error types depending on the context and consequences of the specific hypothesis test.

In summary:

* Type I Error (False Positive): Rejecting the null hypothesis when it is true. Probability denoted by α (alpha).
* Type II Error (False Negative): Failing to reject the null hypothesis when it is false. Probability denoted by β (beta), and power = 1 - β.

#### Examples

Example: Suppose a company claims that the average delivery time for their online orders is 3 days. You want to test this claim using a sample of 50 recent orders. The sample mean delivery time is 2.8 days, and the population standard deviation is 0.6 days. Conduct a Z-test to determine if there is enough evidence to support the company's claim.

Step 1: State the Hypotheses: Null Hypothesis (H0): The population mean delivery time is 3 days. (μ = μ0).

Alternative Hypothesis (Ha): The population mean delivery time is not 3 days. (μ ≠ μ0).

Step 2: Set the Significance Level (α): Let's set the significance level (α) to 0.05 (5%).

Step 3: Calculate the Z-Test Statistic: The Z-test statistic is calculated using the formula: Z = (x̄ - μ) / (σ / √n) where:

* x̄ is the sample mean (2.8 days),
* μ is the population mean under the null hypothesis (3 days),
* σ is the population standard deviation (0.6 days),
* n is the sample size (50 orders).

Z = (2.8 - 3) / (0.6 / √50) Z = -0.2 / (0.6 / 7.07) Z ≈ -0.2 / 0.085 Z ≈ -2.35

Step 4: Determine the Critical Value and P-Value: Since this is a two-tailed test, we need to find the critical value for a 95% confidence level (α/2 = 0.025). From the standard normal distribution table or using statistical software, the critical value is approximately ±1.96.

The p-value can also be determined using statistical software, which is the probability of obtaining a Z-test statistic as extreme as -2.35.

Step 5: Make a Decision: Since the absolute value of the Z-test statistic (-2.35) is greater than the critical value (1.96) and the p-value is less than the significance level (0.05), we reject the null hypothesis (H0).

Step 6: Interpret the Result: Based on the Z-test, there is enough evidence to reject the company's claim that the average delivery time is 3 days. The sample data suggests that the average delivery time is significantly different from 3 days.

###### Example 2: One Tailed Test

Suppose a company claims that the average IQ of their employees is 110. You want to test this claim using a sample of 50 randomly selected employees. The sample mean IQ is 114, and the population standard deviation is known to be 15. Conduct a one-tailed Z-test to determine if there is enough evidence to support the company's claim that the average IQ of their employees is 110.

Step 1: State the Hypotheses: Null Hypothesis (H0): The population mean IQ is 110 (μ = 110). Alternative Hypothesis (Ha): The population mean IQ is greater than 110 (μ > 110).

Step 2: Set the Significance Level (α): Let's set the significance level (α) to 0.05 (5%). Here, you use full (α) not (α/2)

Step 3: Calculate the Z-Test Statistic: The Z-test statistic for Ha: μ > 110 is calculated as follows: Z = (x̄ - μ0) / (σ / √n) Z = (114 - 110) / (15 / √50) Z = 4 / (15 / 7.07) Z ≈ 4 / 2.13 Z ≈ 1.88

Step 4: Determine the Critical Value or p-Value: Since this is a one-tailed test in the right direction (greater than), we need to find the critical value associated with a significance level of 0.05 from the standard normal distribution table. The critical value is approximately 1.645.

**Note: If the alternative hypothesis was opposite(μ < 110) the critical value will be -1.645(it will be negative). Rest logic remains same**

Alternatively, you can calculate the p-value using statistical software or Z-tables. The p-value represents the probability of obtaining a Z-test statistic as extreme as 1.88.

Step 5: Make a Decision: The Z-test statistic (1.88) is greater than the critical value (1.645), and the p-value is less than the significance level (0.05). Therefore, we reject the null hypothesis (H0) in favor of the alternative hypothesis (Ha).

Step 6: Interpret the Result: Based on the one-tailed Z-test, there is enough evidence to support the company's claim that the average IQ of their employees is greater than 110. The sample data suggests that the average IQ is significantly higher than 110.

###### Exmaple 3: One tailed left

Example: Suppose we have a population of individuals, and it is known that the average height in the population is 170 cm. We want to determine if a sample of 50 individuals from this population has a significantly lower average height. The sample mean height is 167 cm, and the population standard deviation is known to be 5 cm.

Step 1: State the Hypotheses: Null Hypothesis (H0): The population mean height is equal to or greater than 170 cm (μ ≥ 170). Alternative Hypothesis (Ha): The population mean height is less than 170 cm (μ < 170).

Step 2: Set the Significance Level (α): Let's set the significance level (α) to 0.05 (5%).

Step 3: Calculate the One-Tailed Z-Test Statistic: The one-tailed z-test statistic for a one-tailed test (less than) is calculated as follows: z = (x̄ - μ) / (σ / √n)

where:

* x̄ is the sample mean height (167 cm).
* μ is the population mean height under the null hypothesis (170 cm).
* σ is the population standard deviation (5 cm).
* n is the sample size (50).

z = (167 - 170) / (5 / √50) z = -3 / (5 / √50) z ≈ -3 / (5 / 3.162) z ≈ -3 / 1.581 z ≈ -1.896

Step 4: Determine the Critical Value or p-Value: Since this is a one-tailed test (less than), we need to find the critical value associated with a significance level of 0.05 from the standard normal distribution table in the left tail.

Alternatively, we can calculate the p-value for the z-statistic of -1.896 using statistical software or z-tables.

Step 5: Make a Decision: Using the standard normal distribution table or statistical software, we find that the critical value for a one-tailed test with a significance level of 0.05 is approximately -1.645 (in the left tail).

The p-value for a z-statistic of -1.896 is p ≈ 0.029 (this is the probability of obtaining a z-statistic as extreme as -1.896).

Since the z-statistic (-1.896) is less than the critical value (-1.645), and the p-value is less than our significance level (0.05), we reject the null hypothesis (H0).

Step 6: Interpret the Result: Based on the one-tailed z-test with a one-tailed hypothesis (less than), there is enough evidence to support the claim that the average height of the sample of individuals is significantly less than the population mean height of 170 cm. The sample data suggests that the average height is statistically lower than the population mean height.

###### Two Sample Z Test

Suppose we have data on the weights of individuals from Group A and Group B. The sample data is as follows:

Group A: 25, 28, 30, 32, 26, 29; Group B: 22, 24, 26, 28, 30

Step 1: State the Hypotheses:

* Null Hypothesis (H0): The mean weight of Group A is equal to the mean weight of Group B (μA = μB).
* Alternative Hypothesis (Ha): The mean weight of Group A is not equal to the mean weight of Group B (μA ≠ μB).

Step 2: Set the Significance Level (α): Let's set the significance level (α) to 0.05 (5%).

Step 3: Calculate the Sample Means and Standard Deviations:

Sample Mean (x̄A) for Group A = (25 + 28 + 30 + 32 + 26 + 29) / 6 ≈ 28.33

Sample Mean (x̄B) for Group B = (22 + 24 + 26 + 28 + 30) / 5 = 26

Sample Standard Deviation (sA) for Group A ≈ 2.52

Sample Standard Deviation (sB) for Group B ≈ 2.83

Step 4: Calculate the Test Statistic (z):

z = (x̄A - x̄B) / √(σA² / nA + σB² / nB)

z = (28.33 - 26) / √(2.52² / 6 + 2.83² / 5)

z ≈ 2.29

Step 5: Find the Critical Region or p-value: Since this is a two-tailed test, we need to find the critical z-values for α/2 = 0.05/2 = 0.025 from the standard normal distribution table. The critical values for a 5% significance level are approximately -1.96 and 1.96.

Step 6: Make a Decision: The calculated z-value (z ≈ 2.29) is greater than 1.96 (the critical value for α/2 = 0.025), indicating that the z-value falls in the critical region. Additionally, the p-value corresponding to z ≈ 2.29 is less than 0.05. Therefore, we reject the null hypothesis.

Step 7: Interpret the Result: Based on the analysis, we can conclude that there is a significant difference in the average weights between Group A and Group B at the 5% significance level. The average weight of one group is significantly different from the average weight of the other group.

Top of Form

#### T-Test

The T-test is used when you have a small sample size (typically n < 30) and the population standard deviation (σ) is unknown. It is also used when the population standard deviation is known, but the sample size is small.

The T-test is based on the t-distribution, which is similar to the standard normal distribution but takes into account the smaller sample size and the uncertainty in estimating the population standard deviation.

There are two main types of T-tests:

##### a. One-Sample T-Test:

The one-sample T-test compares the sample mean to a known value (often the population mean under the null hypothesis). It is used to determine whether the sample mean is significantly different from a specified value.

The formula for the one-sample T-test statistic is:

t = (x̄ - μ) / (s / √n)

where:

* t is the T-test statistic.
* x̄ is the sample mean.
* μ is the known value (e.g., population mean under the null hypothesis).
* s is the sample standard deviation.
* n is the sample size.

##### b. Independent Samples T-Test (Two-Sample T-Test):

The independent samples T-test compares the means of two independent groups to determine whether there is a statistically significant difference between them. It is used when comparing two separate samples to see if their means are significantly different.

The formula for the independent samples T-test statistic is:

t = (x̄1 - x̄2) / √((s1^2 / n1) + (s2^2 / n2))

where:

* t is the T-test statistic.
* x̄1 and x̄2 are the sample means of the two groups.
* s1 and s2 are the sample standard deviations of the two groups.
* n1 and n2 are the sample sizes of the two groups.

##### Paired T Test

The paired t-test, also known as the dependent samples t-test, is a statistical hypothesis test used to compare the means of two related groups. It is used when the data points in one group are paired or matched with the data points in the other group. Paired data occur when the measurements are made on the same individuals or subjects before and after some intervention or treatment.

The paired t-test is designed to determine whether there is a significant difference between the means of the paired measurements. It takes into account the dependency or correlation between the data points, which makes it suitable for situations where the same individuals or items are measured under two different conditions or at two different time points.

Assumptions of the Paired T-Test: Before conducting a paired t-test, it is essential to check that the following assumptions are met:

1. Paired Observations: The data should consist of paired observations, meaning each subject or item in one group is related to a corresponding subject or item in the other group.
2. Normality: The differences between the paired measurements should be approximately normally distributed. The t-test is relatively robust to deviations from normality, especially with moderate sample sizes.
3. Symmetric Distribution: The distribution of the differences should be approximately symmetric around the mean difference.
4. Interval or Ratio Data: The paired measurements should be continuous data on either an interval or ratio scale.

Formula for the Paired T-Test: The paired t-test compares the mean of the differences between the paired observations to a specified value (often assumed to be zero) under the null hypothesis. The formula for the t-test statistic in the paired t-test is as follows:

t = (x̄d - μd) / (sd / √n)

where:

* t is the t-test statistic.
* x̄d is the mean of the differences between the paired observations.
* μd is the mean difference under the null hypothesis (usually assumed to be 0).
* sd is the sample standard deviation of the differences.
* n is the number of pairs (sample size).

To calculate the sample standard deviation (sd) of the differences for the paired t-test, follow these steps:

1. Determine the differences between the paired observations: For each pair of measurements, subtract the value in the second condition (after the intervention or treatment) from the value in the first condition (before the intervention or treatment). You will end up with a set of differences.
2. Calculate the mean of the differences (x̄d): Add up all the differences and divide by the total number of pairs to find the mean of the differences.
3. Calculate the sum of squared differences (SSD): For each difference, square the value obtained in Step 1. Then, add up all the squared differences.
4. Calculate the sample variance (s^2) of the differences: Divide the sum of squared differences (SSD) from Step 3 by the degrees of freedom, which is equal to the number of pairs minus 1 (df = n - 1).

s^2 = SSD / (n - 1)

1. Calculate the sample standard deviation (sd) of the differences: Finally, take the square root of the sample variance (s^2) obtained in Step 4 to get the sample standard deviation (sd).

sd = √(s^2)

Once you have the sample standard deviation (sd) of the differences, you can use it in the formula for the paired t-test:

t = (x̄d - μd) / (sd / √n)

where:

* t is the t-test statistic.
* x̄d is the mean of the differences between the paired observations.
* μd is the mean difference under the null hypothesis (usually assumed to be 0).
* sd is the sample standard deviation of the differences.
* n is the number of pairs (sample size).

##### Degrees Of Freedom

The degrees of freedom (df) for different tests refer to the number of independent pieces of information available for estimating a parameter.

1. Independent Samples T-Test (Two-Sample T-Test): Degrees of Freedom (df) = (n1 - 1) + (n2 - 1) where n1 and n2 are the sample sizes of the two groups being compared.
2. Paired Samples T-Test: Degrees of Freedom (df) = n - 1 where n is the number of pairs (sample size).
3. One-Sample T-Test: Degrees of Freedom (df) = n - 1 where n is the sample size.
4. Analysis of Variance (ANOVA): Degrees of Freedom (df) for between-group variation = k - 1 where k is the number of groups being compared. Degrees of Freedom (df) for within-group variation = N - k where N is the total sample size.
5. Chi-Squared Test of Independence (Contingency Table Test): Degrees of Freedom (df) = (Number of rows - 1) × (Number of columns - 1)

##### Examples

###### One Sample Test(One Tailed)

Suppose we have a population of students, and it is known that the average exam score in the population is 75. We want to know if a sample of 30 students from this population has a significantly higher average exam score. The sample mean exam score is 78.5, and the sample standard deviation is 5.

Step 1: State the Hypotheses: Null Hypothesis (H0): The population mean exam score is equal to 75 (μ = 75). Alternative Hypothesis (Ha): The population mean exam score is greater than 75 (μ > 75).

Step 2: Set the Significance Level (α): Let's set the significance level (α) to 0.05 (5%).

Step 3: Calculate the One-Sample T-Test Statistic: The one-sample t-test statistic is calculated as follows: t = (x̄ - μ) / (s / √n)

where:

* x̄ is the sample mean exam score (78.5).
* μ is the population mean exam score under the null hypothesis (75).
* s is the sample standard deviation (5).
* n is the sample size (30).

t = (78.5 - 75) / (5 / √30) t = 3.5 / (5 / √30) t ≈ 3.5 / 0.9129 t ≈ 3.83

Step 4: Determine the Critical Value or p-Value: Since this is a one-tailed test in the right direction (greater than), we need to find the critical value associated with a significance level of 0.05 from the t-distribution table. Alternatively, we can calculate the p-value.

Step 5: Make a Decision: Using a t-distribution table or statistical software, we find that the critical value for a one-tailed test with 29 degrees of freedom (df = n - 1)(column) and a significance level of 0.05(row) is approximately 1.699.

The p-value for a t-statistic of 3.83 and 29 degrees of freedom is very small, let's say p < 0.001 (it is much smaller than our chosen significance level).

Since the t-statistic (3.83) is greater than the critical value (1.699) and the p-value is less than our significance level (0.05), we reject the null hypothesis (H0).

Step 6: Interpret the Result: Based on the one-sample t-test, there is enough evidence to support the claim that the average exam score of the sample is significantly higher than 75. The sample data suggests that the average exam score is statistically greater than the population mean of 75.

Top of Form

Bottom of Form

###### One Sample Test – Two tailed

Suppose we have a population of fruits, and it is known that the average weight of fruits in the population is 150 grams. We want to know if a sample of 20 fruits from this population has a significantly different average weight. The sample mean weight of the fruits is 160 grams, and the sample standard deviation is 10 grams.

Step 1: State the Hypotheses: Null Hypothesis (H0): The population mean weight of fruits is equal to 150 grams (μ = 150). Alternative Hypothesis (Ha): The population mean weight of fruits is not equal to 150 grams (μ ≠ 150).

Step 2: Set the Significance Level (α): Let's set the significance level (α) to 0.05 (5%).

Step 3: Calculate the One-Sample T-Test Statistic: The one-sample t-test statistic is calculated as follows: t = (x̄ - μ) / (s / √n)

where:

* x̄ is the sample mean weight of fruits (160 grams).
* μ is the population mean weight of fruits under the null hypothesis (150 grams).
* s is the sample standard deviation (10 grams).
* n is the sample size (20).

t = (160 - 150) / (10 / √20) t = 10 / (10 / √20) t ≈ 10 / 2.236 t ≈ 4.472

Step 4: Determine the Critical Value or p-Value: Since this is a two-tailed test, we need to find the critical values associated with a significance level of 0.05/2 = 0.025 from the t-distribution table. Alternatively, we can calculate the p-value.

Step 5: Make a Decision: Using a t-distribution table or statistical software, we find that the critical values for a two-tailed test with 19 degrees of freedom (df = n - 1) and a significance level of 0.025 are approximately ±2.093.

The p-value for a t-statistic of 4.472 and 19 degrees of freedom is very small, let's say p < 0.001 (it is much smaller than our chosen significance level).

Since the absolute value of the t-statistic (4.472) is greater than the critical values (±2.093), and the p-value is less than our significance level (0.05), we reject the null hypothesis (H0).

Step 6: Interpret the Result: Based on the one-sample t-test, there is enough evidence to support the claim that the average weight of the sample of fruits is significantly different from 150 grams. The sample data suggests that the average weight of the fruits is statistically different from the population mean of 150 grams.

###### Independent T Test One Tailed

Suppose we have two groups of students: Group A and Group B. We want to determine if Group A's average exam score is significantly lower than Group B's average exam score. We have the following information:

Group A: Sample size (n1) = 30, Sample mean exam score (x̄1) = 75, Sample standard deviation (s1) = 8 Group B: Sample size (n2) = 35, Sample mean exam score (x̄2) = 80, Sample standard deviation (s2) = 10

Step 1: State the Hypotheses: Null Hypothesis (H0): The population mean exam score of Group A is equal to or greater than the population mean exam score of Group B (μ1 ≥ μ2). Alternative Hypothesis (Ha): The population mean exam score of Group A is less than the population mean exam score of Group B (μ1 < μ2).

Step 2: Set the Significance Level (α): Let's set the significance level (α) to 0.05 (5%).

Step 3: Calculate the Independent Samples T-Test Statistic: The independent samples t-test statistic for a one-tailed test (less than) is calculated as follows: t = (x̄1 - x̄2) / √((s1^2 / n1) + (s2^2 / n2))

t = (75 - 80) / √((8^2 / 30) + (10^2 / 35)) t = -5 / √((64/30) + (100/35)) t = -5 / √(2.1333 + 2.8571) t = -5 / √4.9904 t ≈ -5 / 2.237 t ≈ -2.233

Step 4: Determine the Critical Value or p-Value: Since this is a one-tailed test (less than) with 30 + 35 - 2 = 63 degrees of freedom (df = n1 + n2 - 2), we need to find the critical value associated with a significance level of 0.05 from the t-distribution table in the left tail.

Alternatively, we can calculate the p-value for the t-statistic of -2.233 using statistical software or t-tables.

Step 5: Make a Decision: Using a t-distribution table or statistical software, we find that the critical value for a one-tailed test with 63 degrees of freedom and a significance level of 0.05 is approximately -1.667 (in the left tail).

The p-value for a t-statistic of -2.233 and 63 degrees of freedom is p ≈ 0.014 (this is the probability of obtaining a t-statistic as extreme as -2.233).

Since the t-statistic (-2.233) is less than the critical value (-1.667), and the p-value is less than our significance level (0.05), we reject the null hypothesis (H0).

Step 6: Interpret the Result: Based on the independent samples t-test with a one-tailed hypothesis (less than), there is enough evidence to support the claim that the average exam score of Group A is significantly lower than the average exam score of Group B. The sample data suggests that Group A's average exam score is statistically less than Group B's average exam score.

###### Independent Sample Two Tailed

Suppose we have two groups of students: Group X and Group Y. We want to determine if there is a significant difference in the average exam scores between these two groups. We have the following information:

Group X: Sample size (n1) = 40, Sample mean exam score (x̄1) = 78, Sample standard deviation (s1) = 7 Group Y: Sample size (n2) = 45, Sample mean exam score (x̄2) = 82, Sample standard deviation (s2) = 9

Step 1: State the Hypotheses: Null Hypothesis (H0): The population mean exam score of Group X is equal to the population mean exam score of Group Y (μ1 = μ2). Alternative Hypothesis (Ha): The population mean exam score of Group X is different from the population mean exam score of Group Y (μ1 ≠ μ2).

Step 2: Set the Significance Level (α): Let's set the significance level (α) to 0.05 (5%).

Step 3: Calculate the Independent Samples T-Test Statistic: The independent samples t-test statistic for a two-tailed test is calculated as follows: t = (x̄1 - x̄2) / √((s1^2 / n1) + (s2^2 / n2))

t = (78 - 82) / √((7^2 / 40) + (9^2 / 45)) t = -4 / √((49/40) + (81/45)) t = -4 / √(1.225 + 1.8) t = -4 / √3.025 t ≈ -4 / 1.74 t ≈ -2.299

Step 4: Determine the Critical Value or p-Value: Since this is a two-tailed test, we need to find the critical values associated with a significance level of 0.05/2 = 0.025 from the t-distribution table. Alternatively, we can calculate the p-value for the absolute value of the t-statistic (|t|) of 2.299 using statistical software or t-tables.

Step 5: Make a Decision: Using a t-distribution table or statistical software, we find that the critical values for a two-tailed test with 40 + 45 - 2 = 83 degrees of freedom (df = n1 + n2 - 2) and a significance level of 0.025 are approximately ±1.987.

The p-value for an absolute t-statistic of 2.299 and 83 degrees of freedom is p ≈ 0.023 (this is the probability of obtaining a t-statistic as extreme as 2.299 in either tail).

Since the absolute value of the t-statistic (|t| = 2.299) is greater than the critical values (±1.987), and the p-value is less than our significance level (0.05), we reject the null hypothesis (H0).

Step 6: Interpret the Result: Based on the independent samples t-test with a two-tailed hypothesis, there is enough evidence to support the claim that there is a significant difference in the average exam scores between Group X and Group Y. The sample data suggests that the population mean exam scores of the two groups are different from each other.

###### Paired T Test Two Tailed

Suppose we have a group of 15 students who participated in a training program. We have their scores on a test before and after the training. We want to determine if there is a significant improvement in the students' test scores after the training. The data is as follows:

Before Training: 78, 85, 72, 90, 88, 81, 75, 80, 85, 82, 79, 76, 86, 89, 83 After Training: 80, 88, 75, 92, 89, 83, 78, 83, 87, 84, 80, 77, 88, 91, 85

Step 1: State the Hypotheses: Null Hypothesis (H0): There is no significant difference in the students' test scores before and after the training program (μd = 0). Alternative Hypothesis (Ha): There is a significant difference in the students' test scores before and after the training program (μd ≠ 0).

Step 2: Set the Significance Level (α): Let's set the significance level (α) to 0.05 (5%).

Step 3: Calculate the Paired T-Test Statistic: The paired t-test statistic is calculated as follows:

1. Calculate the differences between the paired scores: Differences = After Training - Before Training

Differences = (80 - 78), (88 - 85), (75 - 72), (92 - 90), (89 - 88), (83 - 81), (78 - 75), (83 - 80), (87 - 85), (84 - 82), (80 - 79), (77 - 76), (88 - 86), (91 - 89), (85 - 83)

Differences = 2, 3, 3, 2, 1, 2, 3, 3, 2, 2, 1, 1, 2, 2, 2

1. Calculate the mean of the differences (x̄d): x̄d = (ΣDifferences) / n

x̄d = (2 + 3 + 3 + 2 + 1 + 2 + 3 + 3 + 2 + 2 + 1 + 1 + 2 + 2 + 2) / 15 x̄d = 33 / 15 x̄d = 2.2

1. Calculate the sample standard deviation of the differences (sd): sd = √((Σ(Differences - x̄d)^2) / (n - 1))

sd = √((Σ(2 - 2.2)^2 + (3 - 2.2)^2 + ... + (2 - 2.2)^2) / (15 - 1))

sd = √((0.04 + 0.64 + ... + 0.04) / 14)

sd = √(0.8 / 14)

sd ≈ √0.0571

sd ≈ 0.2392

1. Calculate the t-test statistic: t = x̄d / (sd / √n)

t = 2.2 / (0.2392 / √15)

t ≈ 2.2 / (0.2392 / 3.8729)

t ≈ 2.2 / 0.0619

t ≈ 35.5682

Step 4: Determine the Critical Value or p-Value: Since this is a two-tailed test, we need to find the critical values associated with a significance level of 0.05/2 = 0.025 from the t-distribution table. Alternatively, we can calculate the p-value for the t-statistic of 35.5682 using statistical software or t-tables.

Step 5: Make a Decision: Using a t-distribution table or statistical software, we find that the critical values for a two-tailed test with 14 degrees of freedom (df = n - 1) and a significance level of 0.025 are approximately ±2.145.

The p-value for a t-statistic of 35.5682 and 14 degrees of freedom is p < 0.001 (this is the probability of obtaining a t-statistic as extreme as 35.5682).

Since the t-statistic (35.5682) is greater than the critical values (±2.145), and the p-value is less than our significance level (0.05), we reject the null hypothesis (H0).

Step 6: Interpret the Result: Based on the paired t-test, there is enough evidence to support the claim that there is a significant improvement in the students' test scores after the training program. The sample data suggests that the students' test scores significantly increased after undergoing the training.

###### Paired T Test One tailed

Suppose we have a group of 20 students who participated in a tutoring program. We have their test scores before and after the program. We want to determine if there is a significant improvement in the students' test scores after the tutoring program. The data is as follows:

Before Tutoring: 78, 85, 72, 90, 88, 81, 75, 80, 85, 82, 79, 76, 86, 89, 83, 77, 80, 74, 78, 81 After Tutoring: 85, 88, 75, 92, 89, 83, 78, 83, 87, 84, 80, 77, 88, 91, 85, 82, 86, 80, 82, 84

Step 1: State the Hypotheses: Null Hypothesis (H0): There is no significant improvement in the students' test scores after the tutoring program (μd ≤ 0). Alternative Hypothesis (Ha): There is a significant improvement in the students' test scores after the tutoring program (μd > 0).

Step 2: Set the Significance Level (α): Let's set the significance level (α) to 0.05 (5%).

Step 3: Calculate the Differences: Calculate the differences between the paired scores (After Tutoring - Before Tutoring) for each student:

Differences = 85 - 78, 88 - 85, 75 - 72, 92 - 90, 89 - 88, 83 - 81, 78 - 75, 83 - 80, 87 - 85, 84 - 82, 80 - 79, 77 - 76, 88 - 86, 91 - 89, 85 - 83, 82 - 77, 86 - 80, 80 - 74, 82 - 78, 84 - 81

Differences = 7, 3, 3, 2, 1, 2, 3, 3, 2, 2, 1, 1, 2, 2, 2, 5, 6, 6, 4, 3

Step 4: Calculate the Mean and Standard Deviation of Differences: Calculate the mean (x̄d) and sample standard deviation (sd) of the differences:

x̄d = (ΣDifferences) / n x̄d = (7 + 3 + 3 + 2 + 1 + 2 + 3 + 3 + 2 + 2 + 1 + 1 + 2 + 2 + 2 + 5 + 6 + 6 + 4 + 3) / 20 x̄d = 65 / 20 x̄d = 3.25

sd = √((Σ(Differences - x̄d)^2) / (n - 1)) sd = √((4.6875 + 0.0625 + 0.0625 + 0.5625 + 4.5625 + 0.5625 + 0.0625 + 0.0625 + 0.5625 + 0.5625 + 2.25 + 2.25 + 0.5625 + 0.5625 + 0.5625 + 2.25 + 4.6875 + 4.6875 + 0.0625 + 0.5625) / 19) sd = √(32.2632 / 19) sd ≈ √1.6975 sd ≈ 1.3014

Step 5: Calculate the t-Test Statistic: Calculate the t-test statistic using the formula:

t = x̄d / (sd / √n) t = 3.25 / (1.3014 / √20) t ≈ 3.25 / (1.3014 / 4.4721) t ≈ 3.25 / 0.2911 t ≈ 11.159

Step 6: Determine the Critical Value or p-Value: Since this is a one-tailed test (greater than), we need to find the critical value associated with a significance level of 0.05 from the t-distribution table in the right tail. Alternatively, we can calculate the p-value for the t-statistic of 11.159 using statistical software or t-tables.

Step 7: Make a Decision: Using a t-distribution table or statistical software, we find that the critical value for a one-tailed test with 19 degrees of freedom (df = n - 1) and a significance level of 0.05 is approximately 1.729.

The p-value for a t-statistic of 11.159 and 19 degrees of freedom is p < 0.001 (this is the probability of obtaining a t-statistic as extreme as 11.159 in the right tail).

Since the t-statistic (11.159) is greater than the critical value (1.729), and the p-value is less than our significance level (0.05), we reject the null hypothesis (H0).

Step 8: Interpret the Result: Based on the paired t-test with a one-tailed hypothesis (greater than), there is enough evidence to support the claim that there is a significant improvement in the students' test scores after the tutoring program. The sample data suggests that the students' test scores significantly increased after participating in the tutoring program.

##### Confidence Intervals

1. Confidence Interval for the Population Mean (μ) with Known Population Standard Deviation (Z-Test): The confidence interval for the population mean (μ) when the population standard deviation (σ) is known is given by:

Confidence Interval = x̄ ± Z \* (σ / √n)

where:

* x̄ is the sample mean,
* Z is the critical value from the standard normal distribution corresponding to the desired confidence level,
* σ is the population standard deviation, and
* n is the sample size.

1. Confidence Interval for the Population Mean (μ) with Unknown Population Standard Deviation (T-Test): The confidence interval for the population mean (μ) when the population standard deviation (σ) is unknown is given by:

Confidence Interval = x̄ ± t \* (s / √n)

where:

* x̄ is the sample mean,
* t is the critical value from the t-distribution corresponding to the desired confidence level and degrees of freedom (df),
* s is the sample standard deviation, and
* n is the sample size.

Confidence intervals are commonly used in hypothesis testing to estimate the range of values within which the true population parameter is likely to lie. Here is a list of confidence intervals for some common hypothesis tests:

1. Confidence Interval for the Population Mean (μ) with Known Population Standard Deviation (Z-Test): The confidence interval for the population mean (μ) when the population standard deviation (σ) is known is given by:

Confidence Interval = x̄ ± Z \* (σ / √n)

where:

* x̄ is the sample mean,
* Z is the critical value from the standard normal distribution corresponding to the desired confidence level,
* σ is the population standard deviation, and
* n is the sample size.

1. Confidence Interval for the Population Mean (μ) with Unknown Population Standard Deviation (T-Test): The confidence interval for the population mean (μ) when the population standard deviation (σ) is unknown is given by:

Confidence Interval = x̄ ± t \* (s / √n)

where:

* x̄ is the sample mean,
* t is the critical value from the t-distribution corresponding to the desired confidence level and degrees of freedom (df),
* s is the sample standard deviation, and
* n is the sample size.

1. Confidence Interval for the Difference in Population Means (μ1 - μ2) in Independent Samples T-Test: The confidence interval for the difference in population means (μ1 - μ2) when comparing two independent samples is given by:

Confidence Interval = (x̄1 - x̄2) ± t \* √((s1^2 / n1) + (s2^2 / n2))

where:

* x̄1 and x̄2 are the sample means of the two groups,
* t is the critical value from the t-distribution corresponding to the desired confidence level and degrees of freedom (df),
* s1 and s2 are the sample standard deviations of the two groups, and
* n1 and n2 are the sample sizes of the two groups.

If population variance is known, formula will be:

Confidence Interval = (x̄1 - x̄2) ± z \* √(( σ 1^2 / n1) + (σ 2^2 / n2))

If population variance is unknown and assumed to be equal

Confidence Interval = (x̄1 - x̄2) ± t \* √((sp^2 / n1) + (sp^2 / n2))

Where

sp = ((n1 - 1) \* s1^2 + (n2-1) \* s2^2 )/n1+n2-2

1. Confidence Interval for the Difference in Population Means (μd) in Paired T-Test: The confidence interval for the difference in population means (μd) when comparing paired samples (dependent samples) is given by:

Confidence Interval = x̄d ± t \* (sd / √n)

where:

* x̄d is the sample mean of the differences between paired samples,
* t is the critical value from the t-distribution corresponding to the desired confidence level and degrees of freedom (df),
* sd is the sample standard deviation of the differences, and
* n is the number of paired samples.

### Chi Square Test

The chi-square test is a statistical test used to determine if there is a significant association between two categorical variables. It is based on the chi-square (χ²) statistic, which measures the difference between the observed and expected frequencies in a contingency table.

The test is commonly used to answer questions like:

* Is there a relationship between two categorical variables?
* Are the observed frequencies significantly different from the expected frequencies?

There are two main types of chi-square tests:

#### Chi-Square Test for Independence

This test is used when you want to determine if there is a significant association between two categorical variables. It is appropriate when both variables have two or more categories and the data are presented in a contingency table.

The hypothesis for the chi-square test for independence are as follows: Null Hypothesis (H0): There is no association between the two categorical variables. Alternative Hypothesis (Ha): There is an association between the two categorical variables.

The test involves the following steps:

* Create a contingency table that shows the observed frequencies for each combination of categories of the two variables.
* Calculate the expected frequencies assuming no association between the variables (i.e., assuming the null hypothesis is true).
* Calculate the chi-square statistic using the formula: χ² = Σ((Oij - Eij)² / Eij), where Oij is the observed frequency and Eij is the expected frequency for each cell in the contingency table.
* Determine the degrees of freedom (df) for the test, which is equal to (r - 1) \* (c - 1), where r is the number of rows and c is the number of columns in the contingency table.
* Find the critical value of the chi-square statistic at a chosen significance level (α) and df from the chi-square distribution table or using statistical software.
* Compare the calculated chi-square statistic with the critical value. If the calculated chi-square value is greater than the critical value, you reject the null hypothesis and conclude that there is a significant association between the two variables.

#### Chi-Square Goodness-of-Fit Test

This test is used to determine if a sample of data comes from a specific theoretical distribution. It is appropriate when you want to compare the observed frequencies of a single categorical variable with the expected frequencies based on a hypothesized distribution.

The hypothesis for the chi-square goodness-of-fit test are as follows: Null Hypothesis (H0): The observed frequencies follow the expected distribution. Alternative Hypothesis (Ha): The observed frequencies do not follow the expected distribution.

The test involves the following steps:

* Define the expected distribution based on theoretical probabilities or observed frequencies from a reference population.
* Create a contingency table that shows the observed frequencies for each category of the variable.
* Calculate the expected frequencies based on the hypothesized distribution.
* Calculate the chi-square statistic using the formula: χ² = Σ((Oi - Ei)² / Ei), where Oi is the observed frequency and Ei is the expected frequency for each category.
* Determine the degrees of freedom (df), which is equal to (k - 1), where k is the number of categories in the variable.
* Find the critical value of the chi-square statistic at a chosen significance level (α) and df from the chi-square distribution table or using statistical software.
* Compare the calculated chi-square statistic with the critical value. If the calculated chi-square value is greater than the critical value, you reject the null hypothesis and conclude that the observed frequencies do not follow the expected distribution.

It's important to note that the chi-square test has certain assumptions, including that the data must be categorical, the observations are independent, and the expected frequencies in each cell of the contingency table are not too small (usually a rule of thumb is that no expected frequency should be less than 5). Violation of these assumptions can affect the validity of the test results.

Chi-square tests are widely used in various fields, including social sciences, healthcare, market research, and more, to analyze categorical data and detect relationships or differences between variables.

#### Examples

##### Test for independence

Suppose we conducted a survey in a school to collect data on students' favorite subjects and their gender. The data is as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Subject | : Math | Science | English | History | Art |
| Male | 30 | 25 | 20 | 15 | 10 |
| Female | 20 | 30 | 25 | 15 | 10 |

Step 1: State the Hypotheses: Null Hypothesis (H0): There is no association between students' favorite subjects and their gender. Alternative Hypothesis (Ha): There is an association between students' favorite subjects and their gender.

Step 2: Set the Significance Level (α): Let's set the significance level (α) to 0.05 (5%).

Step 3: Create the Contingency Table: Construct a contingency table with the observed frequencies for each combination of favorite subjects and gender.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Subject: | Math | Science | English | History | Art | Row Totals |
| Male | 30 | 25 | 20 | 15 | 10 | 100 |
| Female | 20 | 30 | 25 | 15 | 10 | 100 |
| Column Totals | 50 | 55 | 45 | 30 | 20 | 200 |

Step 4: Calculate the Expected Frequencies:

Assuming there is no association between favorite subjects and gender, we calculate the expected frequencies for each cell. The expected frequency for each cell is given by:

Expected Frequency (E) = (Row Total \* Column Total) / Grand Total

Expected Frequency (Male-Math) = (100 \* 50) / 200 = 25

Expected Frequency (Male-Science) = (100 \* 55) / 200 = 27.5

Expected Frequency (Male-English) = (100 \* 45) / 200 = 22.5

Expected Frequency (Male-History) = (100 \* 30) / 200 = 15

Expected Frequency (Male-Art) = (100 \* 20) / 200 = 10

Expected Frequency (Female-Math) = (100 \* 50) / 200 = 25

Expected Frequency (Female-Science) = (100 \* 55) / 200 = 27.5

Expected Frequency (Female-English) = (100 \* 45) / 200 = 22.5

Expected Frequency (Female-History) = (100 \* 30) / 200 = 15

Expected Frequency (Female-Art) = (100 \* 20) / 200 = 10

Step 5: Calculate the Chi-Square Statistic:

The chi-square statistic is calculated using the formula:

χ² = Σ((Oij - Eij)² / Eij)

where Oij is the observed frequency, and Eij is the expected frequency for each cell in the contingency table.

χ² = ((30 - 25)² / 25) + ((25 - 27.5)² / 27.5) + ((20 - 22.5)² / 22.5) + ((15 - 15)² / 15) + ((10 - 10)² / 10)

+ ((20 - 25)² / 25) + ((30 - 27.5)² / 27.5) + ((25 - 22.5)² / 22.5) + ((15 - 15)² / 15) + ((10 - 10)² / 10)

χ² ≈ 0.04 + 0.0082 + 0.139 + 0 + 0

χ² ≈ 0.1872

Step 6: Determine the Degrees of Freedom (df): The degrees of freedom (df) for the chi-square test for independence is calculated as:

df = (Number of Rows - 1) \* (Number of Columns - 1)

df = (2 - 1) \* (5 - 1)

df = 4

Step 7: Find the Critical Value: At a significance level (α) of 0.05 and degrees of freedom (df) of 4, the critical value from the chi-square distribution table is approximately 9.488.

Step 8: Make a Decision: Compare the calculated chi-square statistic (χ² ≈ 0.1872) with the critical value (9.488). Since the calculated chi-square value is much smaller than the critical value, we fail to reject the null hypothesis.

Step 9: Interpret the Result: Based on the chi-square test for independence, there is not enough evidence to conclude that there is a significant association between students' favorite subjects and their gender. The data does not support the idea that the choice of favorite subject is dependent on gender in the population from which the sample was taken.

##### Test For goodness of Fit

Suppose we collected data on the eye colors of 200 individuals in a certain population. We want to test if the observed distribution of eye colors matches the expected distribution based on the assumption that 30% have brown eyes, 45% have blue eyes, 20% have green eyes, and 5% have other colors.

Brown Eyes: 65 individuals

Blue Eyes: 85 individuals

Green Eyes: 40 individuals

Other Colors: 10 individuals

Step 1: State the Hypotheses: Null Hypothesis (H0): The observed distribution of eye colors follows the expected distribution. Alternative Hypothesis (Ha): The observed distribution of eye colors does not follow the expected distribution.

Step 2: Set the Significance Level (α): Let's set the significance level (α) to 0.05 (5%).

Step 3: Create the Contingency Table: Construct a contingency table with the observed and expected frequencies for each eye color category.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Eye Color: | Brown | Blue | Green | Other | Row Totals |
| Observed: | 65 | 85 | 40 | 10 | 200 |
| Expected: | 200\*0.3=60 | 200\*0.45=90 | 200\*0.2=40 | 200\*0.05=10 | 200 |
|  |  |  |  |  |  |

Step 4: Calculate the Chi-Square Statistic: The chi-square statistic is calculated using the formula:

χ² = Σ((O - E)² / E)

where O is the observed frequency, and E is the expected frequency for each category.

χ² = ((65 - 60)² / 60) + ((85 - 90)² / 90) + ((40 - 40)² / 40) + ((10 - 10)² / 10)

χ² = (25/60) + (25/90) + (0) + (0)

χ² ≈ 0.4167 + 0.2778 + 0 + 0

χ² ≈ 0.6944

Step 5: Determine the Degrees of Freedom (df): The degrees of freedom (df) for the chi-square goodness-of-fit test is calculated as the number of categories (k) minus 1, where k is the number of eye color categories. In this case, there are 4 categories (brown, blue, green, other).

df = k - 1 df = 4 - 1 df = 3

Step 6: Find the Critical Value: At a significance level (α) of 0.05 and degrees of freedom (df) of 3, the critical value from the chi-square distribution table is approximately 7.815.

Step 7: Make a Decision: Compare the calculated chi-square statistic (χ² ≈ 0.6944) with the critical value (7.815). Since the calculated chi-square value is much smaller than the critical value, we fail to reject the null hypothesis.

Step 8: Interpret the Result: Based on the chi-square goodness-of-fit test, there is not enough evidence to conclude that the observed distribution of eye colors significantly deviates from the expected distribution. The data supports the idea that the distribution of eye colors in the sample follows the hypothesized distribution of 30% brown eyes, 45% blue eyes, 20% green eyes, and 5% other colors.

### ANOVA

ANOVA (Analysis of Variance) is a statistical test used to compare the means of three or more groups to determine if there are any statistically significant differences among the groups. ANOVA is an extension of the t-test, which can only compare the means of two groups.

The main idea behind ANOVA is to partition the total variation in the data into different sources of variation: variation between groups and variation within groups. If the variation between groups is much larger than the variation within groups, it suggests that there are significant differences among the groups.

There are several types of ANOVA tests, each designed for specific scenarios:

1. One-Way ANOVA: One-Way ANOVA is used when you have one categorical independent variable (factor) and one continuous dependent variable. It compares the means of three or more independent groups.
2. Two-Way ANOVA: Two-Way ANOVA is used when you have two categorical independent variables (factors) and one continuous dependent variable. It examines the interaction effects between the two factors and their individual effects on the dependent variable.
3. N-Way ANOVA: N-Way ANOVA extends the concept to more than two independent variables.

The null hypothesis (H0) for ANOVA is that there are no significant differences among the means of the groups, and any observed differences are due to random chance.

The alternative hypothesis (Ha) is that at least one group's mean is different from the others, indicating that there is a significant effect of the independent variable on the dependent variable.

Assumptions of ANOVA include:

* Independence: The observations within each group are independent of each other.
* Normality: The dependent variable is approximately normally distributed within each group.
* Homogeneity of Variance: The variance of the dependent variable is approximately equal across all groups.

If the p-value obtained from ANOVA is less than the chosen significance level (usually 0.05), we reject the null hypothesis and conclude that there are significant differences among the group means. However, ANOVA itself does not tell us which specific groups differ from each other. To determine that, post hoc tests like Tukey's HSD (Honestly Significant Difference) or Bonferroni corrections can be performed.

ANOVA is widely used in various fields, including social sciences, biological sciences, and industrial processes, to compare the means of multiple groups and understand the impact of different factors on a dependent variable.

Steps:

1. Sum of Squares Between (SSB):

SSB = Σ(ni \* (Mi - Grand Mean)²)

where ni is the number of observations in the ith group, Mi is the mean of the ith group, and Grand Mean is the mean of all the data combined.

2. Sum of Squares Within (SSW):

SSW = Σ Σ(yij - Mi)²

where yij is the jth observation in the ith group, and Mi is the mean of the ith group.

3. Sum of Squares Total (SST):

SST = Σ(yij - Grand Mean)²

where yij is the jth observation in any group, and Grand Mean is the mean of all the data combined.

4. Degrees of Freedom Between (dfB) and Within (dfW):

dfB = k - 1 dfW = N - k

where k is the number of groups, and N is the total number of observations.

5. Mean Square Between (MSB) and Mean Square Within (MSW):

MSB = SSB / dfB

MSW = SSW / dfW

6. F-statistic:

F = MSB / MSW

1. Critical Value: To compare the F-statistic with the critical value, you need to determine the critical value from the F-distribution table or using statistical software at a given significance level (α) and degrees of freedom dfB and dfW.
2. p-value: The p-value is the probability of observing an F-statistic as extreme as the one calculated, assuming the null hypothesis is true. It can be obtained using statistical software or F-distribution tables.
3. Effect Size: The effect size is a measure of the practical significance of the differences among the groups. One common effect size measure for ANOVA is eta-squared (η²) or partial eta-squared (η²p), which represents the proportion of variance in the dependent variable explained by the independent variable.

Example:

Suppose we have data on the test scores of students from three coaching classes: Class A, Class B, and Class C. Each coaching class had 10 students, and their test scores (out of 100) are as follows:

Class A: 78, 85, 88, 90, 82, 92, 80, 86, 84, 88

Class B: 75, 80, 83, 72, 79, 85, 76, 82, 78, 80

Class C: 92, 90, 88, 94, 95, 96, 90, 93, 89, 91

Step 1: State the Hypotheses: Null Hypothesis (H0): There is no significant difference in the mean test scores among the three coaching classes. Alternative Hypothesis (Ha): There is a significant difference in the mean test scores among the three coaching classes.

Step 2: Set the Significance Level (α): Let's set the significance level (α) to 0.05 (5%).

Step 3: Calculate the Grand Mean and the Mean of Each Class:

Grand Mean = (78 + 85 + 88 + 90 + 82 + 92 + 80 + 86 + 84 + 88 + 75 + 80 + 83 + 72 + 79 + 85 + 76 + 82 + 78 + 80 + 92 + 90 + 88 + 94 + 95 + 96 + 90 + 93 + 89 + 91) / 30 ≈ 85.47

Mean Class A = (78 + 85 + 88 + 90 + 82 + 92 + 80 + 86 + 84 + 88) / 10 ≈ 85.3

Mean Class B = (75 + 80 + 83 + 72 + 79 + 85 + 76 + 82 + 78 + 80) / 10 ≈ 78.0

Mean Class C = (92 + 90 + 88 + 94 + 95 + 96 + 90 + 93 + 89 + 91) / 10 ≈ 91.0

Step 4: Calculate the Sum of Squares Total (SST):

SST = Σ(yij - Grand Mean)²

= (78 - 85.47)² + (85 - 85.47)² + (88 - 85.47)² + (90 - 85.47)² + (82 - 85.47)²

+ (92 - 85.47)² + (80 - 85.47)² + (86 - 85.47)² + (84 - 85.47)² + (88 - 85.47)²

+ (75 - 85.47)² + (80 - 85.47)² + (83 - 85.47)² + (72 - 85.47)² + (79 - 85.47)²

+ (85 - 85.47)² + (76 - 85.47)² + (82 - 85.47)² + (78 - 85.47)² + (80 - 85.47)²

+ (92 - 85.47)² + (90 - 85.47)² + (88 - 85.47)² + (94 - 85.47)² + (95 - 85.47)²

+ (96 - 85.47)² + (90 - 85.47)² + (93 - 85.47)² + (89 - 85.47)² + (91 - 85.47)²

≈ 477.29 + 0.2029 + 6.9601 + 17.0809 + 8.2501 + 42.7289 + 28.8409 + 0.2284 + 2.5744 + 6.9601

+ 69.0889 + 0.2029 + 3.0625 + 157.9824 + 40.5729 + 34.1449 + 68.0625 + 81.9524 + 11.5049 + 49.5625

+ 7.8121 + 42.7289 + 11.5049 + 62.4484 + 88.7329 + 122.5569 + 34.1449 + 4.6544 + 38.2049 + 33.8544

≈ 1379.13

Step 5: Calculate the Sum of Squares Between (SSB):

SSB = Σ(ni \* (Mi - Grand Mean)²)

= 10 \* (85.3 - 85.47)² + 10 \* (78.0 - 85.47)² + 10 \* (91.0 - 85.47)²

≈ 0.0291 + 54.7051 + 41.0581

≈ 95.7923

Step 6: Calculate the Sum of Squares Within (SSW):

SSW = Σ Σ(yij - Mi)²

= (78 - 85.3)² + (85 - 85.3)² + (88 - 85.3)² + (90 - 85.3)² + (82 - 85.3)² + (92 - 85.3)² + (80 - 85.3)² + (86 - 85.3)² + (84 - 85.3)² + (88 - 85.3)²

+ (75 - 78.0)² + (80 - 78.0)² + (83 - 78.0)² + (72 - 78.0)² + (79 - 78.0)² + (85 - 78.0)² + (76 - 78.0)² + (82 - 78.0)² + (78 - 78.0)² + (80 - 78.0)²

+ (92 - 91.0)² + (90 - 91.0)² + (88 - 91.0)² + (94 - 91.0)² + (95 - 91.0)² + (96 - 91.0)² + (90 - 91.0)² + (93 - 91.0)² + (89 - 91.0)² + (91 - 91.0)²

≈ 58.09 + 0.09 + 6.76 + 22.09 + 10.89 + 46.56 + 32.49 + 0.09 + 1.69 + 6.76

+ 9.0 + 4.0 + 25.0 + 36.0 + 1.0 + 49.0 + 4.0 + 4.0 + 0.0 + 4.0 + 4.0

+ 1.0 + 1.0 + 9.0 + 9.0 + 16.0 + 25.0 + 1.0 + 4.0 + 4.0 + 0.0

≈ 240.06

Step 7: Calculate the Degrees of Freedom (df):

dfB = k - 1 = 3 - 1 = 2

dfW = N - k = 30 - 3 = 27

Step 8: Calculate the Mean Square Between (MSB) and Mean Square Within (MSW):

MSB = SSB / dfB = 95.7923 / 2 ≈ 47.89615

MSW = SSW / dfW = 240.06 / 27 ≈ 8.89111

Step 9: Calculate the F-Statistic:

F = MSB / MSW ≈ 47.89615 / 8.89111 ≈ 5.38

Step 10: Find the Critical Value or p-Value:

At a significance level (α) of 0.05 and dfB of 2 and dfW of 27, the critical value from the F-distribution table is approximately 3.35. Alternatively, using statistical software, the p-value for F ≈ 5.38 and dfB = 2, dfW = 27 is less than 0.05.

Step 11: Make a Decision:

Since the calculated F-statistic (F ≈ 5.38) is greater than the critical value (3.35) and the p-value is less than 0.05, we reject the null hypothesis. We conclude that there is a significant difference in the mean test scores among the three coaching classes.

Step 12: Post hoc Tests (if needed):

If ANOVA indicates significant differences among the groups, you may perform post hoc tests (e.g., Tukey's HSD, Bonferroni correction) to determine which specific coaching classes' scores differ significantly from each other.

### Correlation Coefficient

r = Σ((xi - x̄)(yi - ȳ)) / √(Σ(xi - x̄)² \* Σ(yi - ȳ)²)

The correlation coefficient is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges between -1 and 1, where:

* 1 indicates a perfect positive linear correlation (as one variable increases, the other increases proportionally).
* -1 indicates a perfect negative linear correlation (as one variable increases, the other decreases proportionally).
* 0 indicates no linear correlation (there is no systematic relationship between the variables).

### Which Test to use

* Continuous
  + One Sample
    - Unknown population std or sample size < 30 -One Sample T –Test
    - Known pop std and size > 30 – Z – Test
  + Two Sample
    - Paired – Paired T Test
    - Independent
      * Unknown population std or sample size < 30 –Independent T –Test
      * Known pop std and size > 30 – Two sample Z – Test
  + More than two
    - ANOVA

# Python For Statistics

## Descriptive Statistics

# Using NumPy

scores = [85, 78, 92, 88, 90, 82, 75, 80, 83, 72]

mean = np.mean(scores)

median = np.median(scores)

std\_dev = np.std(scores)

variance = np.var(scores)

min\_value = np.min(scores)

max\_value = np.max(scores)

# Using pandas

data\_series = pd.Series(scores)

data\_summary = data\_series.describe()

# Printing the results

print("Mean:", mean)

print("Median:", median)

print("Standard Deviation:", std\_dev)

print("Variance:", variance)

print("Minimum Value:", min\_value)

print("Maximum Value:", max\_value)

print("\nData Summary (using pandas):\n", data\_summary)

Mean: 83.5

Median: 83.5

Standard Deviation: 6.897927951545057

Variance: 47.5

Minimum Value: 72

Maximum Value: 92

Data Summary (using pandas):

count 10.000000

mean 83.500000

std 6.897928

min 72.000000

25% 78.250000

50% 83.500000

75% 89.500000

max 92.000000

# Create a sample DataFrame

data = {

'Scores': [85, 78, 92, 88, 90, 82, 75, 80, 83, 72]

}

df = pd.DataFrame(data)

# Calculate skewness

skewness = df['Scores'].skew()

# Calculate kurtosis

kurt = df['Scores'].kurtosis()

# Print the results

print("Skewness:", skewness)

print("Kurtosis:", kurt)

Skewness: -0.059849377475064044

Kurtosis: -0.4681648313832358

#Plots

df['Scores'].plot()

df['Scores'].hist()

df['Scores'].plot.density()

## Inferential Statistics

Step 1: Install Required Libraries If you haven't already installed SciPy, you can do so using pip:

pip install scipy

Step 2: Import Libraries Import the required libraries into your Python script or Jupyter Notebook:

import numpy as np from scipy import stats

Step 3: Prepare Data Create a list, array, or pandas DataFrame containing your data. For demonstration purposes, let's assume you have two independent samples, sample1 and sample2:

sample1 = [85, 78, 92, 88, 90]

sample2 = [75, 80, 83, 72, 79]

Step 4: Perform Inferential Statistics

### Z Score

z\_scores = stats.zscore(sample1)

# Print the z-scores

print("Z-Scores:", z\_scores)

### Independent Sample t-test

# Independent Sample t-test

t\_stat, p\_value = stats.ttest\_ind(sample1, sample2)

print("Independent Sample t-test:")

print("T-statistic:", t\_stat)

print("P-value:", p\_value)

#### Confidence Interval

# Perform two-sample t-test

t\_stat, p\_value = stats.ttest\_ind(sample1, sample2)

# Calculate the sample sizes

n1 = len(sample1)

n2 = len(sample2)

# Calculate the degrees of freedom

df = n1 + n2 - 2

# Calculate the standard errors

std\_err1 = np.std(sample1, ddof=1) / np.sqrt(n1)

std\_err2 = np.std(sample2, ddof=1) / np.sqrt(n2)

# Calculate the pooled standard error

pooled\_std\_err = np.sqrt((std\_err1\*\*2 + std\_err2\*\*2) / 2)

# Calculate the margin of error for the 95% confidence interval

margin\_of\_error = stats.t.ppf(0.975, df) \* pooled\_std\_err

# Calculate the mean difference

mean\_diff = np.mean(sample1) - np.mean(sample2)

# Calculate the lower and upper bounds of the confidence interval

lower\_bound = mean\_diff - margin\_of\_error

upper\_bound = mean\_diff + margin\_of\_error

# Print the results

print("Mean Difference:", mean\_diff)

print("Confidence Interval (95%):", (lower\_bound, upper\_bound))

### Paired Sample t-test

# Paired Sample t-test

t\_stat, p\_value = stats.ttest\_rel(sample1, sample2)

print("Paired Sample t-test:")

print("T-statistic:", t\_stat)

print("P-value:", p\_value)

### One-Way ANOVA

# One-Way ANOVA

f\_stat, p\_value = stats.f\_oneway(sample1, sample2)

print("One-Way ANOVA:")

print("F-statistic:", f\_stat)

print("P-value:", p\_value)

### Chi-square Test for Independence

# Chi-square Test for Independence

observed = np.array([[30, 15], [20, 25]]) # Contingency table

chi2\_stat, p\_value, dof, expected = stats.chi2\_contingency(observed)

print("Chi-square Test for Independence:")

print("Chi2-statistic:", chi2\_stat)

print("P-value:", p\_value)

print("Degrees of Freedom:", dof)

print("Expected Frequencies:\n", expected)

### Chi Square Goodness of Fit

# Specify the expected frequencies based on your hypothesis

expected\_freq = [25, 25, 20, 20]

# Perform the chi-square test

chi2\_stat, p\_value = stats.chisquare(observed\_freq, f\_exp=expected\_freq)

# Print the results

print("Chi-square Statistic:", chi2\_stat)

print("P-value:", p\_value)