Statistics

1. Introduction to statistics

- What is statistics?
- Types of statistics - Descriptive and Inferential
- Types of data
- Population and Sample
- Sampling Techniques
- Statistical Data Analysis Steps

2. Descriptive Statistics

- Measures of central tendency - Mean, Median and Mode
- Measures of dispersion - Range, Variance, Standard Deviation, Percentilesand Quartiles
- Frequency
- Graphical Representations - Boxplots, Histograms, Scatterplots
- Outliers and understanding their impact
- Correlation and Covariance

DataCrats

3. Probability

- Basic Probability concepts: sample space, event etc
- Types of events
  - Disjoint or Non-Disjoint event
  - Independent or dependent event
- Conditional Probability
- Bayes Theorem
- Probability Distributions
  - Random Variables and its Types (Discrete & Continuous)
  - PMF and PDF
- Discrete Distributions
  - Binomial
  - Bernoulli
- Continuous Distributions
  - Uniform
  - Normal
    - Standard Normal Distribution

- Standardization
- Normalization
- Empirical Rule

4. Inferential Statistics

- Relationship with Descriptive statistics
- Point and Interval estimation
- Confidence Interval (Z / T distribution)
- Hypothesis testing
    - Types of hypotheses: Null and Alternate
    - Level of significance and p value
    - Type of errors
    - One tailed or two tailed test
    - Types of tests in statistics (z test, t test, ANOVA, Chi square etc)

DataCrats

What is statistics?

Statistics is a branch of mathematics that involves collecting, analysing, interpretingand drawing conclusions from information/data. It provides methods for making inferences about the characteristics of a population based on a limited set of observations or data points.

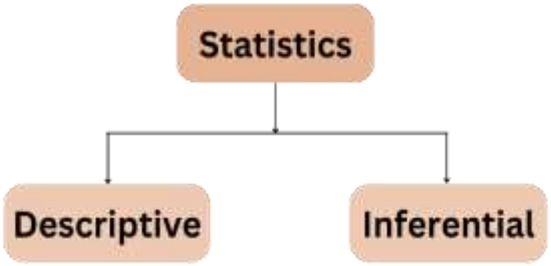
Note: Data(plural) are measurements or observations, A datum (singular) is asingular measurement.

To be more specific, here are some claims that we have heard on several occasions.

- 4 out of 5 dentists recommend Dentine.
- Almost 85% of lung cancers in men and 45% in women are tobacco related.
- There is an 40% chance that in a room full of 30 people that at least twopeople will share the same birthday
- The average score of the students in the math test was 75 out of 100.
- The average monthly sales of the company increased by 10% compared tothe previous year.
- The minimum number of participants required for the workshop is 60.

Types of statistics

Statistics can be broadly categorized into two main types:



- **Descriptive Statistics**

  Descriptive Statistics is a branch of statistics that deals with the collection, presentation, and interpretation of data. The primary goal of this statistics is to summarize and describe the main features of a dataset. This involves organizing and simplifying large amounts of data in a meaningful way to make it more understandable.

  DataCrafts

  These statistics are the foundation for more advanced statistical analysis and are essential for making informed decisions based on data. It consists of methods for organizing and summarizing information.

  Key features to describe about data:

  - What is the centre of the data? (location)
  - How much does the data vary? (scale)
  - What is the shape of the data? (shape)

  These can be described by summary statistics.

  It includes the construction of graphs, charts, and tables, and the calculation of various descriptive measures such as averages, measures of variation, percentiles.

- **Inferential Statistics**

Inferential Statistics involves drawing conclusions or inferences about a population based on information obtained from a sample of population. The key idea is to use the information obtained from a representative sample to make generalizations and predictions about the entire population.

Inferential statistics often involves the use of probability theory and statistical methods to make probabilistic statements about population parameters. It helps researchers make decisions, formulate policies and draw conclusions in situations where it may be impractical or impossible to study an entire population.

Key concepts related to inferential statistics in the context of data analysis:

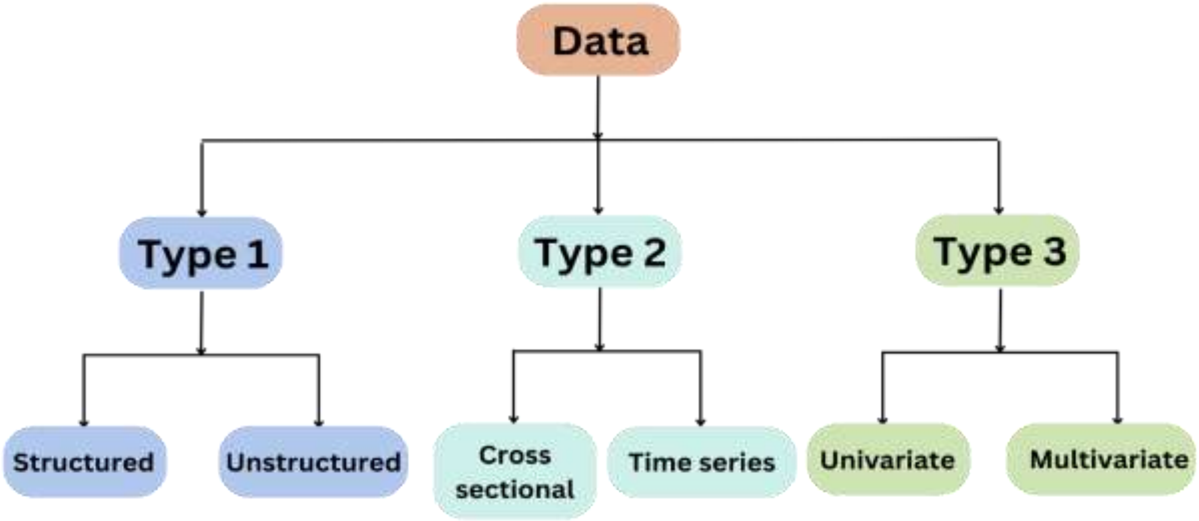
- Testing whether the average exam scores of two groups are significantly different.
- Estimating the average height of a population with a 95% confidence interval.
- Analysing the relationship between hours of study and exam scores.

Descriptive and Inferential statistics are interrelated. It is almost always necessary to use methods of descriptive statistics to organize and summarize the information obtained from a sample before methods of inferential statistics can be used to make more thorough analysis of subject under investigation.

DataCrafts

Types of data

Data can be broadly divided into different types based on its nature and characteristics.



- **Type 1:**
  - **Structured:** Structured data refers to data that is organized in rows and columns and well-defined manner.

    Examples: Tables in a relational database, spreadsheet

DataCrafts

  - **Unstructured:** It refers to data that lacks a predefined data model or simply not structured.

    Examples: Text documents (emails, articles, social media posts), Multimedia content (images. Videos., audio recordings), web pages and other free form text

- **Type 2:**
  - **Cross-sectional:** This data is collected at a single point in time, or over a very short period, and it involves observations of multiple subjects or entities.

    Examples: Survey data collected from individuals in a city at a specific date, Marks obtained in a test

  - **Time series:** Time series data involves observations taken over a sequence of time intervals.

    Examples: Monthly sales data for a product over several years, daily stock prices for a particular company over a month

- **Type 3:**
  - **Univariate:** It involves data consisting of a single variable.
  - **Multivariate:** It involves data consisting of two or more variables.

After analysing the type of data, identifying the type of variables is necessary and it also comes under the types of data.

- **Nominal**: Nominal data represents categories or labels with no inherent order or ranking.

  Examples: Gender, colours

- **Ordinal**: Ordinal data represents categories with a clear order or ranking, but the intervals between the categories are not uniform or meaningful.

  Examples: Education levels, customer satisfaction ratings, no. of cars owned by a household

- **Categorical**: Categorical data represents categories and can be either nominal or ordinal.

  Examples: Types of cars, product categories

DataCrats

- **Numerical**: Numerical data includes both discrete and continuous data and represents measurable quantities.

  Examples: Temperature, income, age

- **Interval**: Interval data has meaningful interval between values, but there is no true zero point.

  Examples: Temperature measured in Celsius or Fahrenheit, IQ scores,

- **Ratio**: Ratio data has meaningful interval between values and it has a true zero point, indicating the absence of the attribute being measured.

  Examples: Height, income, age, weight

Population and Sample

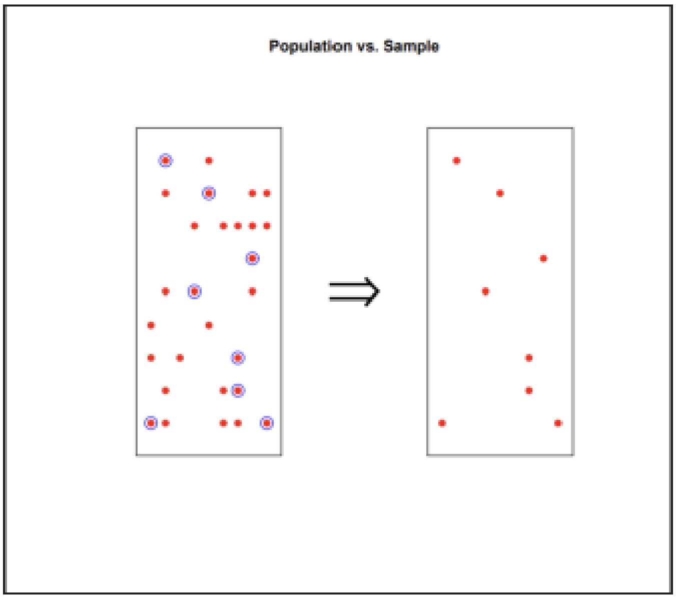
- **Population:** The population is the entire group of individuals, objects or observations. It includes all possible members that meet the criteria for inclusion in the study.
  Example**:** All people in India, All customers of Netflix,

- **Sample:** A sample is a subset of the population, selected for study or analysis. It is a representative group that is used to draw inferences about thelarger population.
  Example: 10k people from India, 300 customers of Netflix

Population

Sample

DataCrats



In statistical analysis, the goal is often to make inferences about a population based on observations from a sample. Various sampling techniques and statistical methods are employed to ensure that the sample is a fair and accurate representation of the population of interest.

For good statistical analysis, the sample needs to be as similar as possible to the population. If they are similar enough, we say that the sample is representative of the population. The sample is used to make conclusions about the whole population. If the sample is not similar enough to the whole population, the conclusions count as useless.

The characteristics of a population are known as population parameters and characteristics that describe a sample are called sample statistics.

Why are samples used?

- To reduce cost of data collection
- When a full census cannot be taken

Sampling Techniques

Sampling techniques are methods used to select a subset of elements (a sample) from a larger population for the purpose of making inferences about that population. Here are some common sampling techniques:

- **Simple Random Sampling**: Every individual or element in the population has an equal chance of being included in the sample.

DataCrafts

- **Systematic Sampling**: A fixed interval is used to select every kth element from a list after a random starting point is chosen

  Example: Selecting every 10th person from a list of names, first 20 and last 20

- **Stratified Sampling:** Population is divided into subgroups or strata based on certain characteristics (e.g., gender, age) and then random samples are taken from each stratum.

  Example: Dividing a population of students into strata based on grade level and then randomly selecting students from each grade.

- **Clustered Sampling:** Population is divided into clusters and a random sample of clusters is selected. Then, all members within the chosen clusters are included in the sample. Unlike stratified, that selects individuals from each subgroup, it selects entire subgroups

  Example: Dividing a city into neighbourhoods, randomly selecting several neighbourhoods, and surveying all households in the chosen neighbourhoods.

The choice of sampling technique depends on various factors, including the research objectives, the nature of the population, available resources and the desired level of precision. Each sampling method has its advantages and limitations, and researchers must carefully consider the appropriateness of the technique for their specific study.

Statistical Data Analysis

Statistical data analysis involves a systematic process of inspecting, cleaning, transforming, and modelling data with the goal of discovering useful information, reaching conclusions, and support decision – making.

Here are the key steps in statistical data analysis:

1. Define the problem or research question:

   Clearly define the problem or research question you want to address. This step is crucial for guiding the entire analysis process.

2. Data Collection:

   Gather relevant data based on the research question. Data can be collected through surveys, experiments, observations or from existing datasets.

3. Data Cleaning:

   Check for errors, missing values, outliers, and inconsistencies in the data. Clean and preprocess the data to ensure its quality and reliability.

DataCrafts

4. Exploratory Data Analysis (EDA):

   Perform initial exploratory analysis to understand the characteristics of the data. This may involve summary statistics, visualizations (histograms, scatter plots), and identifying patterns or trends.

5. Data Transformation:

   If needed, transform the data to meet the assumptions of statistical methods. Common transformations include normalization, standardization, and handling categorical variables.

6. Hypothesis Formulation:

   Formulate a hypothesis based on your research question. Clearly define the null hypothesis (HO) and alternate hypothesis (H1) that you want to test.

7. Statistical Testing:

   Choose appropriate statistical tests based on the nature of your data and research question. Common tests include t-tests, chi-square tests, ANOVA, regression analysis, etc.

8. Interpretation of Results:

   Analyse the results of your statistical tests. Determine whether the evidence supports or contradicts your hypothesis. Consider the significance level and confidence intervals.

9. Draw Conclusions:

   Based on the results, draw conclusions regarding the research question. Consider the practical significance of your findings in addition to statistical significance.

10. Document the Analysis Process/ Report Making:

    Document all the steps, methods, and decisions made during the analysis. This documentation is important for transparency, reproducibility, and future reference. Prepare a report based on your conclusions, provide recommendations for future action. Discuss the implications of your findings in the context of the original research question.

Measures of Central Tendency

DataCrats

Measures of central tendency are statistical measures that describe the centre of the data. They provide a single representative value around which the entire data set tends to cluster. The three main measures of central tendency are the mean, median and mode.

1- Mean:

The mean, also known as the average, is calculated by adding up all the values in a data set and then dividing by the number of values.

$$\mu = \frac{sum\ of\ all\ values}{Number\ of\ values} = \frac{x1+x2+x3....xn}{n}$$

Example: For the dataset {2,4,6,8,10}, the mean is $\frac{2+4+6+8+10}{5}$ = 6

Properties:

a. Meaningful for continuous variables
b. Affected by outliers or extreme values, which can heavily skew the result.

2- Median:

The median is the middle value of a dataset when it is ordered from least to greatest. If there is an even number of values, the median is the average of the two middle values.

Example: For the dataset {3,1,5,7,9}, when ordered, becomes {1,3,5,7,9} and the median is 5.

Properties:

a. Less influenced by extreme values, making it a better measure of central tendency for skewed distributions.
b. Meaningful for ordinal, ratio, and interval data.

3- Mode:

The mode is the value that occurs most frequently in a data set. A data set may have no model (if no value is repeated), one mode (If one value is repeated more than others), or multiple models (if more than one value is repeated with the same frequency)

Example: In the set {4,2,8,6,2,9,2}, the mode is 2 because it appears more frequently than any other value.

DataCrats

Properties:

a. Meaningful for categorical values. For numerical values, if unique values is small then it can also be used.

Summary:

1- Choose the measure based on the distribution of the data.
2- For normally distributed data, the mean is often appropriate.
3- For skewed, or data with outliers, consider the mean.
4- In case of categorical variables, mode is used.

Measures of Dispersion

Dispersion is the degree of variation in the data. Measures of dispersion, also known as variability or spread quantify the extent to which individual data points in a dataset differ from the central tendency (mean, median or mode). They provide important insights into the spread, scatter or distribution of the data. Two datasets of the same variable may exhibit similar positions of center but may be remarkably different with respect to variability. The main measures of dispersion include the range, IQR, variance, quartiles, percentiles and standard deviation.

1. Range:

   The range is the simplest measure of dispersion and is calculated as the difference between the maximum and minimum values in a dataset.

   Range = Max – Min

   Properties:

   a. Sensitive to extreme values
   b. Doesn’t consider the distribution of values
   c. Used when a quick assessment of the spread is needed and suitable for small datasets

2. Quartiles

   Quartiles divide a dataset into four equal parts, with three quartiles, Q1, Q2 (median) and Q3. Q1 is the value below which 25% of the data falls, Q2 is the median and the 50% of the data falls below it, and Q3 is the value below which 75% of the data falls.

DataCrats

   Q1 is at position $\frac{n+1}{4}$

   Q2 is at position $\frac{n+1}{2}$

   Q3 is at position $\frac{3(n+1)}{2}$

   Properties:

   a. Useful for identifying the central tendency and spread of specific sections of the data.

3. Percentiles:

   Percentiles divide a dataset into 100 equal parts, with specific percentiles representing the percentage of data below a given value. The 25th, 50th, and 75th percentiles are equivalent to the Q1, Q2 and Q3 quartiles, respectively.

$$P^{th}\ Percentile: Value = \frac{P}{100} X(Number\ of\ observations + 1)$$

Properties:

DataCrats

a. Useful for comparing the position of a particular data point relative to the entire dataset.

4. IQR (Interquartile Range)

   IQR is the range of the middle 50% of the data, representing the spread of the central portion of the distribution.

   IQR = Q3-Q1

   Properties:

   a. Less sensitive to extreme values than the range.
   b. Useful for identifying the spread of the central part of the data.

5. Variance

   Variance measures the average squared deviation of each data point from the mean.

   Variance = $\frac{\sum_{i=1}^{N}(Xi - \mu)^2}{N}$

DataCrats

   Where Xi is the individual data points, N is the number of data points and µ is the mean of the data.

   Properties:

   a. When a detailed understanding of the variability is needed.
   b. Sensitive to extreme values.

6. Standard deviation

   Standard deviation is the square root of the variance and is expressed in the same units as the original data.

   Standard deviation = $\sqrt{Variance}$

   It's good to use std as it is in the same units as x and variance is in square units.

Note:

- Average paints a partial picture of the data

- Average statistics is incomplete without std/var.

DataCrats

Example: Smartphone price analysis

Retailer A: $800

Retailer B: $850

Retailer C: $820

Retailer D: $855

Retailer E: $870

Retailer F: $855

Retailer G: $825

Retailer H: $865

Retailer I: $840

Retailer J: $810

List = [800,850,820,855,870,855,825,865,840,810]

DataCrats

Ordered list = [800, 810, 820, 825, 840, 850, 855, 855, 865, 870]

Mean = 839

Median = 845

Mode = 855

Range = 70

Q1 = 25th percentile = 821.25

Q3 = 75th percentile = 855

Standard Deviation = 24.01

Variance = 576.67

Frequency

Frequency is the number of times a value of the data occurs. It is commonly used for categorical data where you want to know how often each category occurs. Table listing all classes and their frequencies is called a frequency distribution table.

Example: 1,3,3,2,4,1,2,2,1,2,3,5,4,1,2,1,3,1,4,1

Data Value	Frequency
1	7
2	5
3	4
4	3
5	1

A relative frequency is the ratio of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. In other words, percentage or proportion of the data value present in the dataset.

Relative frequency = $\frac{Frequency\ in\ the\ class}{Total\ number\ of\ observation} * 100$

DataCrats

Cumulative frequency is a way to show the running total of frequencies as you move through categories. It provides information about the number of data points that are less than or equal to a certain value or category.

Example: Customer Complaints at a Company

Imagine you work in the customer service department of an e-commerce company, and you have collected data on the types of complaints received over a month. You have categorized the complaints into four main types: Shipping Delays, Product Quality, Billing Issues, and Returns. Here is a summary of the data:

Shipping Delays: 15 complaints

Product Quality: 10 complaints

Billing Issues: 8 complaints

Returns: 12 complaints

Data Value	Frequency	Relative Frequency	Cumulative Frequency

Shipping Delays	15	0.3125	15
Product Quality	10	0.2083	25
Billing Issues	8	0.1667	33
Returns	12	0.3125	45

Graphical Representations

Graphical representations play a crucial role in descriptive statistics, providing visual insights into the distribution, frequency, central tendency and dispersion of data.

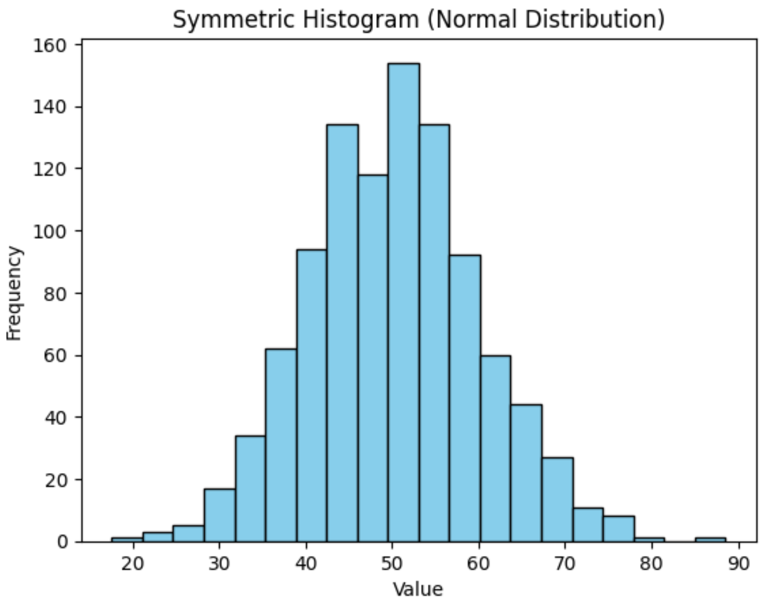
1. Histograms

- isplay the distribution of continuous data.
- Divides data into intervals (bins) and represents the frequency or density of observations in each bin.
- The number of bins can impact the appearance of the histogram.
- Vertical axis represents the frequency in each bin.
- Helps visualize the shape, central tendency and spread of the data
- Useful for identifying patterns, potential outliers, and skewness.

DataCrafts

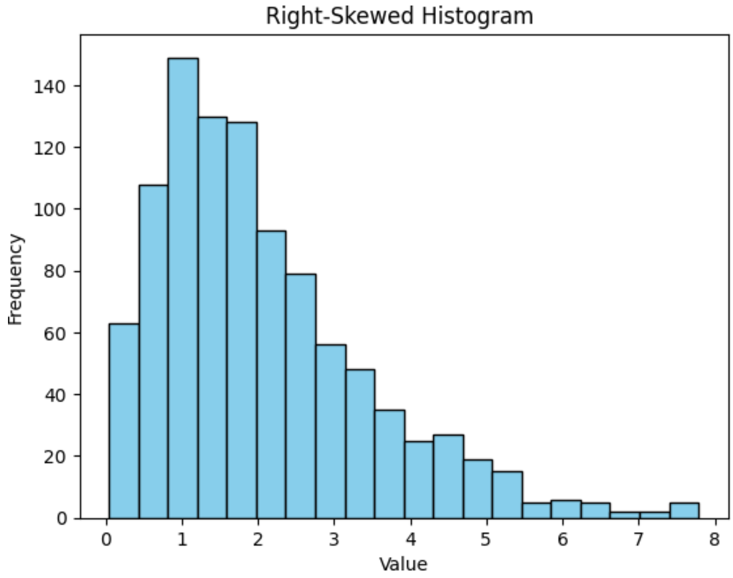
Types of skewed histograms:

- Symmetric (Normal Distribution):
    - Bell Shaped curve
    - Mean, median and mode are at the centre.
    - The data is evenly distributed on both sides of the mean.



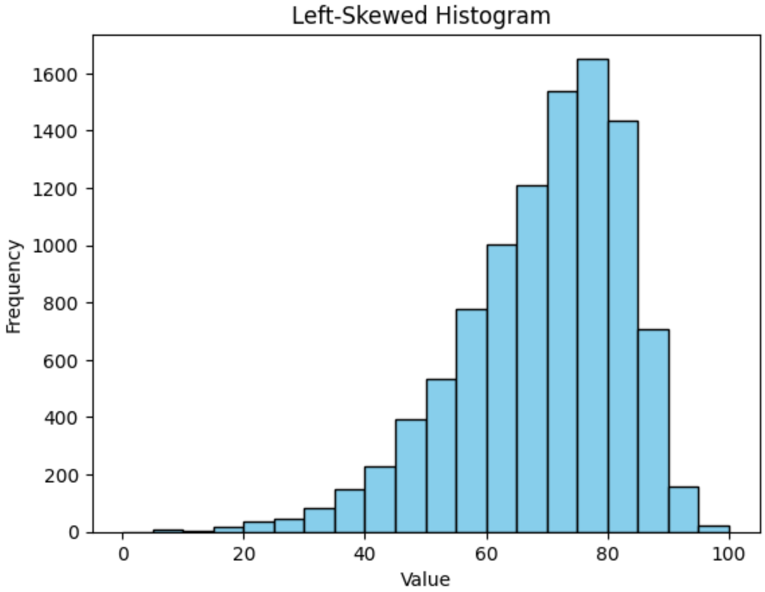
- Positively skewed (Right skewed) Distribution:
  - Tail on the right side
  - Mean is greater than the median
  - Data concentrates on the left side, with a tail extending to the right.

DataCrats



- Negatively skewed (Left skewed) Distribution:
  - Tail on the left side.
  - Mean is less than the median.

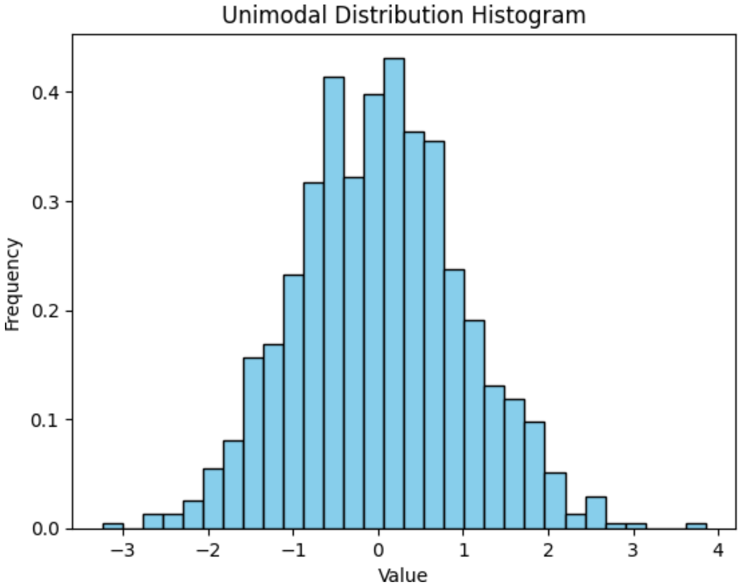
- Data concentrates on the right side, with a tail extending to the left.



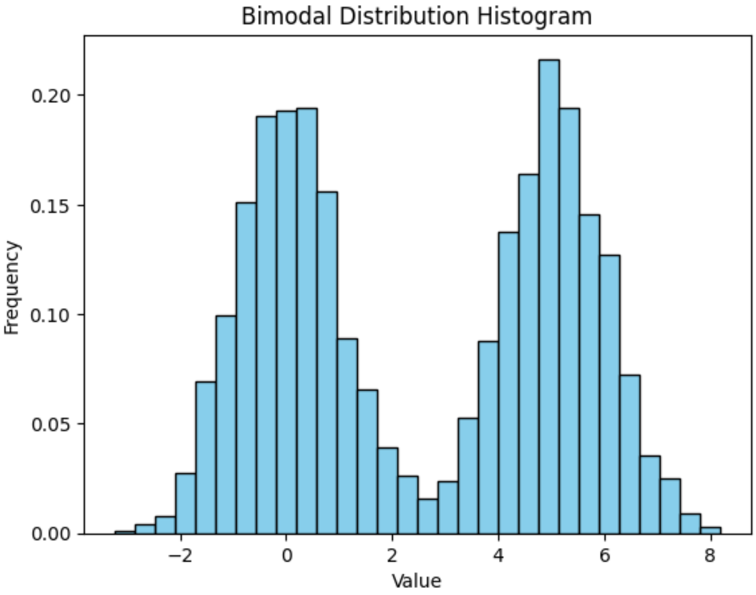
Types based on number of modes:

DataCrats

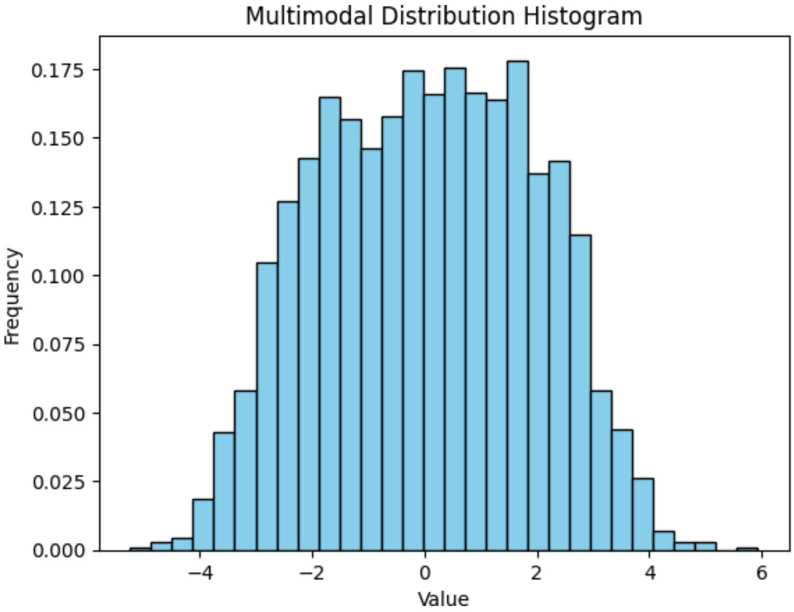
- Unimodal Distribution
  - One clear peak



- Bimodal Distribution
  - Two distinct peaks



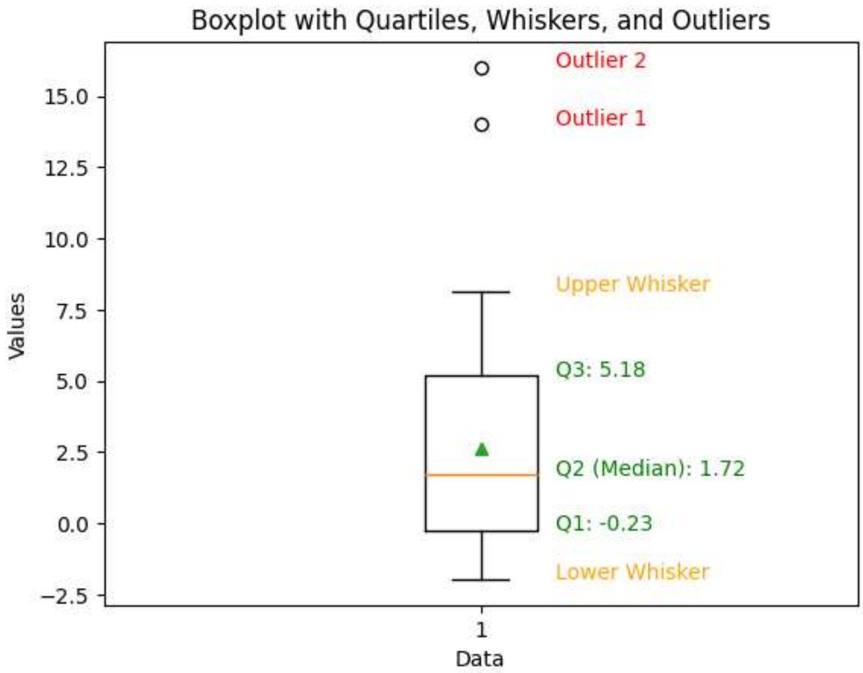
- Multimodal Distribution
  - More than two peaks



2. Box Plots (Whisker Plots)

- Show the spread of the data and identify outliers
- Box represents the IQR with the median line inside
- Whiskers extend to the minimum and maximum values within a certain range
- Outliers are plotted individually, aiding in the identification of extreme values.

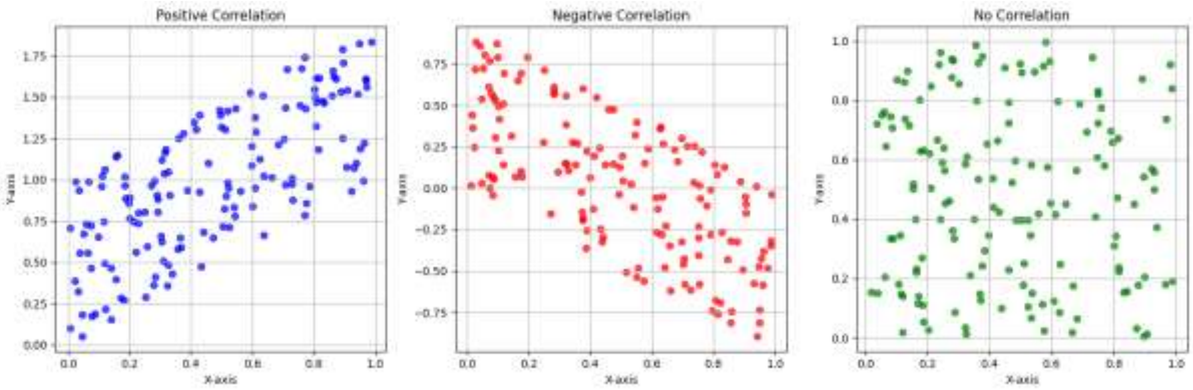
- Useful for comparing distributions and identifying skewness
- The range is typically set to 1.5 times the IQR. Values beyond this range are considered potential outliers.

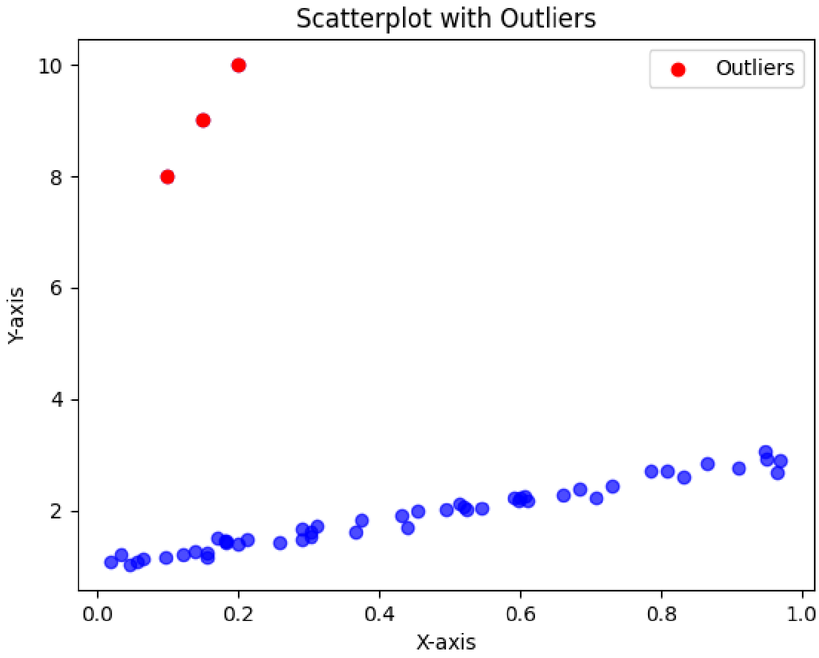


DataCrafts

3. Scatter Plots

- Visualize the relationship between the two continuous variables.
- Each point represents a pair of values, showcasing the joint distributions.
- Helps identify patterns, trends and correlation between variables
- Useful for detecting outliers and understanding the strength and direction of relationships
- Relationships can be Positive, Negative or No correlation.
- Can be enhanced with regression lines to highlight trends.





Outliers and understanding their impact

Outliers are data points that deviate significantly from the rest of the dataset. These values are notably different from the majority of the observations, can be unusually high or low and often potential to skew the overall interpretation of the data. Identifying and understanding outliers is a crucial aspect of data analysis, as they can have a substantial impact on statistical analysis.

1. Identification of Outliers:

Outliers can be identified through various statistical methods and visualizations. Common techniques include:

- Boxplots: Outliers are often visible as points beyond the "whiskers" of a boxplot.

- Z-Scores: Calculating the z-score for each data point helps identify values that deviate significantly from the mean.
- IQR (Interquartile Range): Outliers can be detected using the IQR by considering values outside a certain range.

2. Impact on Descriptive Statistics:

Outliers can heavily influence summary statistics:

- Mean and Standard Deviation: Outliers, especially those in the tails of a distribution, can distort the mean and inflate the standard deviation.
- Median and Quartiles: Robust statistics like the median and quartiles are less sensitive to outliers and provide a more reliable measure of central tendency and dispersion.

3. Effects on Inferential Statistics:

Outliers can affect the validity of statistical inferences:

- arametric Tests: Outliers may violate assumptions of normality in parametric tests, leading to inaccurate results.
- Regression Analysis: Outliers can disproportionately impact regression coefficients, affecting the model's predictive performance.

DataCrafts

4. Data Distribution and Modelling:

Outliers can impact the distribution of data:

- Skewness and Kurtosis: Outliers can introduce skewness and kurtosis, altering the shape of the distribution.
- Model Assumptions: Outliers may violate assumptions of linear models, leading to biased predictions and inaccurate model evaluations.

Correlation and Covariance

Covariance is a statistical measure that describes how much two variables change together. It indicates whether an increase in one variable is associated with an increase or decrease in another variable.

If the variables tend to increase or decrease together, the covariance is positive. If one variable tends to increase as the other decreases, the covariance is negative.

$$Cov(x, y) = \frac{\sum_{i=1}^{N}(Xi-\mu x)(Yi-\mu y)}{N}$$

It indicates the direction of the relationship; it does not quantify the strength of the relationship.

Correlation is a statistical measure that describes the extent to which two variables change together. In other words, it quantifies the degree to which a change in one variable is associated with a change in another variable. Correlation does not imply causation, but it helps us understand the relationship between variables in a dataset. It is a statistical technique used to measure the strength and direction of a linear relationship between two variables. The result is a correlation coefficient, a value between -1 and 1.

- Positive Correlation (1): When one variable increases, the other variable tends to increase. Correlation coefficient closer to 1 indicates a strong positive correlation.
- Negative Correlation (-1): When one variable increases, the other variable tends to decrease. Correlation coefficient closer to -1 indicates a strong negative correlation.
- No Correlation (0): There is no pattern or relation between the variables.

DataCrats

How to calculate correlation?

The most common method to calculate correlation is Pearson's correlation coefficient. It is calculated by dividing the covariance of the two variables by the product of their standard deviations.

$$\rho(x, y) = \frac{Cov(x,y)}{\sqrt{var(x)}\,\sqrt{var(y)}} = \frac{Cov(x,y)}{std(x)std(y)}$$

Considerations and Cautions:

- Correlation does not imply causation: Just because two variables are correlated does not mean that one causes the other.
- Outliers: Extreme data points can influence correlation, so its essential to check for outliers.
- Non-linear Relationships: Correlation specifically measures linear relationships. Non-linear relationships may not be accurately represented.

Note: Causation refers to the relationship between cause and effect. In a causal relationship, a change in the independent variable is directly responsible for a change in the dependent variable. Unlike correlation, which simply describes a relationship between two variables, causation implies a direct connection in which one variable influences the other.

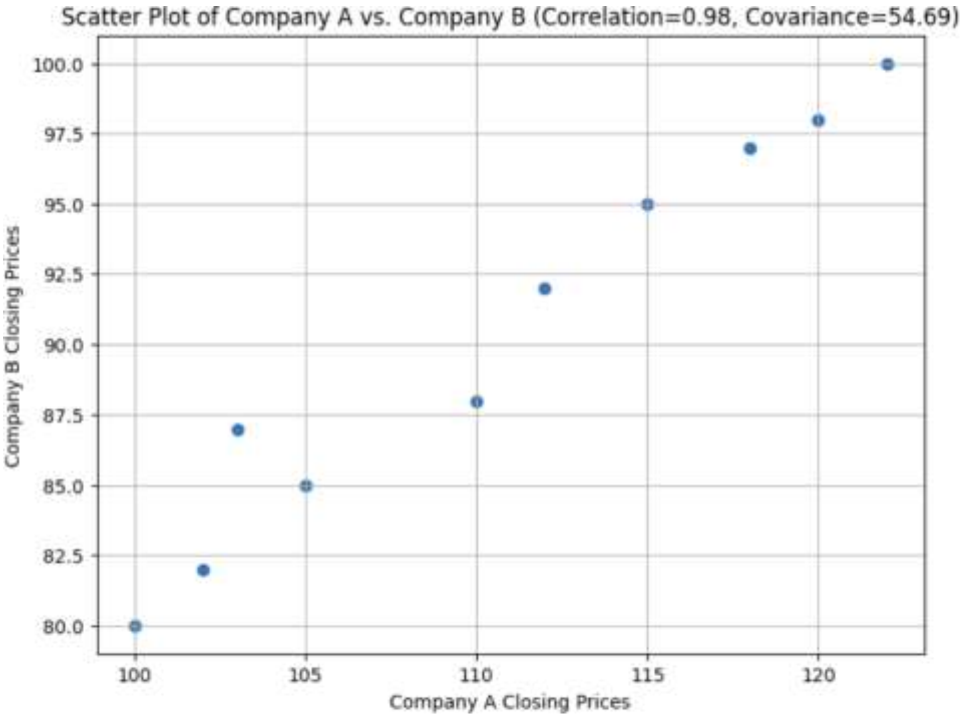
DataCrats

Example: Correlation and Covariance in Stock Market Analysis

Suppose you are a financial analyst studying the relationship between the stock prices of two technology companies, Company A and Company B, over the past year. You have collected daily closing prices for both companies and want to analyzewhether there is a correlation or covariance between their stock performance.

Company A prices = [100, 102, 105, 103, 110, 112, 115, 118, 120, 122]

Company B Prices = [80, 82, 85, 87, 88, 92, 95, 97, 98, 100]



Probability is a measure of the likelihood of a particular event occurring. It is defined as a chance of something happening or likelihood of an event is to be happened. It is expressed as a number between 0 and 1, where 0 indicates impossibility, 1 indicates certainty, and values in between represent varying degrees of likelihood.

Probability basics

- Sample space: The set of all possible outcomes of a random experiment.

  Example: When rolling a six-sided die, the sample space is {1,2,3,4,5,6}.

- Random experiment: An experiment or a process for which the outcome cannot be predicted with certainty.

  Example: Tossing a coin, Rolling a die

- Events: Subsets of the sample space, representing specific outcomes or combination of outcomes.

  Example: In the die example, the event A could be getting an even number {2,4,6}

- Probability function: It assigns a probability to each event in the sample space, denoted by P(A), where A is an event.

  Example: P (getting a 3) = 1/6 for a fair six-sided die.

- Complement of an event: It consists of all outcomes not in A.P(A') = 1 – P(A)Example: If A is getting an even number on a die, then A' is getting an odd number.

Types of events

DataCrats

1. Disjoint Events (Mutually Exclusive):

   These are the events that cannot occur at the same time or do not have common outcomes.

   Probability of events A and B are disjoint, if their intersection is zero.

   P (A and B) = 0

   Example:

   - Single card drawn from a deck cannot be a king and a queen.
   - Event of getting head and tail.
   - Getting 2 and 5 when rolling a six-sided die
   - Drawing a red card and a club

   Suppose we select a random card from a deck, let event A be the event that the card is a Spade or a Club and let event B be the event that the card is a Heart or a Diamond.

A = {Spade, Club}

B = {Heart, Diamond}

There is no overlap between the two sample spaces, thus these are disjoint events.

The probability that either event occurs can be calculated by adding up their individual probabilities.

$P\ (A \text{ or } B) = P\ (A) + P\ (B)$

2. Joint Event (Non-Mutually Exclusive):

   These are the events that occurs at the same time or have common outcomes.

   Example: Consider two dice rolls. A joint event could be “both dice showing even numbers,” which occurs when both dice show, for instance, 2 and 4.

   The probability of both events A and B occurring

   $P\ (A \text{ and } B) = P\ (A) * P\ (B)$

DataCrats

   The probability of events A or B occurring:

   $P\ (A \text{ or } B) = P\ (A) + P\ (B) – P\ (A \text{ and } B)$

3. Dependent Events:

   If the occurrence of one event does affect the probability of the other occurring then the events are dependent.

   $P\ (A \text{ and } B) = P(A) * P(B|A)$

   Example: If you draw a card and do not replace it, the events of drawing a red card and then drawing a black card are dependent.

4. Independent Events:

   Two events are independent if the occurrence of one does not change the probability of the other occurring.

   $P\ (A \text{ and } B) = P(A) * P(B)$

Example: Rolling a die twice results in independent events as outcome of first roll has no impact on the outcome of the second roll.

Conditional Probability

Probability of an event or outcome based on the occurrence of a previous event or outcome. In other words, the probability of an event occurring given that another event has already occurred.

P(A|B) = Probability of event A occurring given that event B occurs.

P(A|B) = $\frac{P(A\ and\ B)}{P(B)}$

Example: Medical Diagnosis

Consider a medical scenario where we want to determine the probability that a patient has a specific disease (event A) given that they have tested positive for a certain medical test (event B).

P(A): The overall probability that a random person has the disease.

P(B|A): The probability that a person with the disease tests positive for the medical test.

DataCrats

P(B): The probability that a random person tests positive for the medical test.

The conditional probability P(A|B) represents the likelihood that a person has the disease if they tested positive. In this case, P (A and B) is the probability that a person both has the disease and tests positive.

Bayes Theorem

Bayes theorem, named after the Reverend Thomas Bayes, is a mathematical formula that enables us to update the probability of an event based on new evidence or information.

$$P(B) = \frac{P(A)*P(A)}{P(B)}$$

P(A|B): This is the probability of event A occurring given that event B has occurred. It is the updated probability based on new evidence.

P(A): The prior probability of event A, which is our initial belief about the likelihood of A.

P(B|A): The probability of observing event B given that event A has occurred. It represents how well the evidence (event B) supports our initial belief (event A).

P(B): The marginal probability of event B, which is the overall probability of B occurring.

1. Medical Diagnosis:

Bayes' theorem is extensively used in medicine to calculate the probability of a patient having a disease given the results of diagnostic tests. It helps clinicians make informed decisions and assess the likelihood of a condition in a patient.

2. Spam Email Classification:

In email filtering systems, Bayes' theorem helps classify emails as spam or not spam based on the presence of certain keywords and other features. It's a fundamental component of spam filters.

3. Recommendation Systems:

In recommendation algorithms, Bayes' theorem can be applied to suggest products, movies, or music to users. It takes into account user preferences, past behavior, and item characteristics.

DataCrats

4. A/B Testing and Conversion Optimization:

In A/B testing, Bayes' theorem helps analyse the results of experiments where you test two or more variations of a web page, app, or product to determine which one is more effective in achieving a specific goal.

5. Risk Assessment and Insurance:

Insurance companies use Bayes' theorem to assess the risk of insuring individuals or properties. It helps determine premium rates and estimate claim probabilities based on historical data and new information.

6. Fraud Detection:

In the financial and e-commerce industries, Bayes' theorem is used to detect fraudulent transactions. It considers patterns of legitimate and fraudulent activities to identify suspicious behaviour.

Example: Suppose there are 1000 people from which 900 are healthy and 100 are sick and they diagnosed using a test that is 99% affective.

	Diagnosed Healthy	Diagnosed Sick
Healthy	891	9
Sick	1	99

P(S) is the prior probability of a person being sick, which is 0.1 (100 out of 1000).

P(P|S) is the probability of testing positive if the person is sick, is 0.99.

P(¬S) is the prior probability of a person being healthy, which is 0.9 (900 out of 1000).

P(P|¬S) is the probability of testing positive if the person is healthy which is 0.01.

P(P) is the probability of person tested positive which is 0.108 after test (108 out of 1000)

We can now use Bayes' theorem to calculate P(S|P), the probability that a person is sick given they tested positive:

DataCrats

$$P(P) = \frac{P(S)*P(S)}{P(S)*P(S)+P(\neg S)*P(\neg S)}$$

$$P(P) = \frac{0.99*0.1}{0.99*0.1+0.01*0.9} = 0.9167$$

Or,

$$P(P) = \frac{P(S)*P(S)}{P(P)} = \frac{0.99*0.1}{0.108} = 0.9167$$

So, using Bayes' theorem, the probability that a person is sick (S) given that they tested positive (P) is approximately 0.9167 or 91.67%. This means that even with a 99% effective test, if a person tests positive, there is still a relatively high probability that they are sick, given that the disease is relatively common in the population.

Probability Distributions

Random Variables: A random variable is a variable whose values are outcomes of a random experiment or all the possible events in some partition of the sample space. It associates a numerical value with each outcome.

Discrete Random variables: Finite number of values (countable range of values).

DataCrats

Example: X = {0,1}

Continuous Random Variables: Infinite number of values (uncountable range of values, takes all values on an interval)

Example: X = (0,1)

Probability distribution for a random variable is an assignment of probability to each of the possible values.

Probability Density Function (PDF) and Probability Mass Function (PMF):

PMF (Discrete Variables): It gives the probability of a discrete random variable.

PDF (Continuous Variables): It gives the probability of a continuous random variable.

When tossing four fair coins, we have the following possible outcomes for the number of heads (H):

0 heads (TTTT)

1 head (HTTT, THTT, TTHT, TTTH)

DataCrats

2 heads (HHTT, HTHH, TTHH, THTH, HTTH, THHT)

3 heads (HHHT, HTHH, TTHH, HHTH)

4 heads (HHHH)

So, the PMF for the number of heads when tossing four fair coins is as follows:

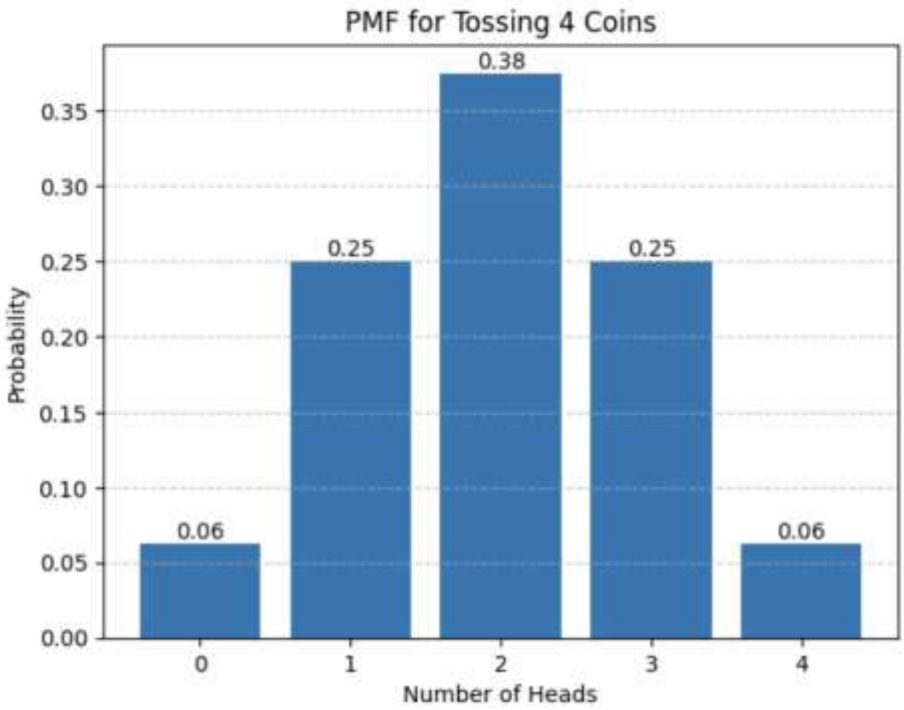
P (X = 0) = 1/16

P (X = 1) = 4/16

P (X = 2) = 6/16

P (X = 3) = 4/16

P (X = 4) = 1/16



magine a call centre that receives a large volume of customer inquiries and service requests. Customers are placed on hold until they can be connected to a customer service representative. The time a customer spends on hold is referred to as "wait time."

DataCrats

The mean wait time (μ) is estimated to be 5 minutes, indicating the average time customers spend on hold.
The standard deviation (σ) is estimated to be 2 minutes, reflecting the variability or dispersion in wait times.

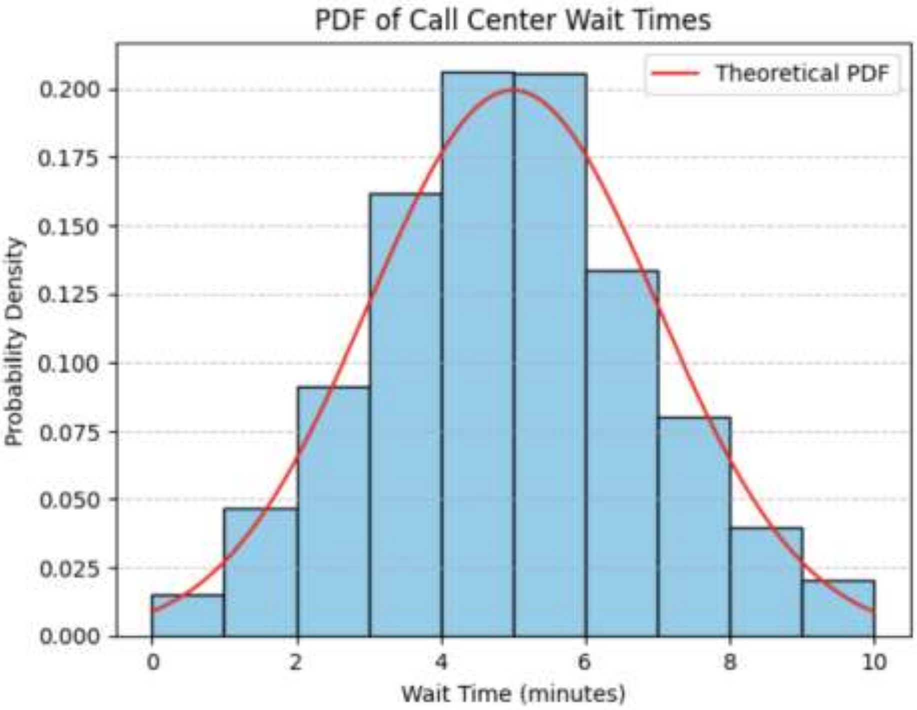
$fx(x)$ = area under curve = $\int_{a}^{b} f(x)dx$

Cumulative Distribution Function (CDF) shows how much probability the variable accumulated until a certain value.

Fx(x) = P(X<=x)

F(x) is the probability that X is less than or equal to a given value x

P(X≤x) represents the cumulative probability up to x



Example: CDF for tossing 4 coins:

DataCrats

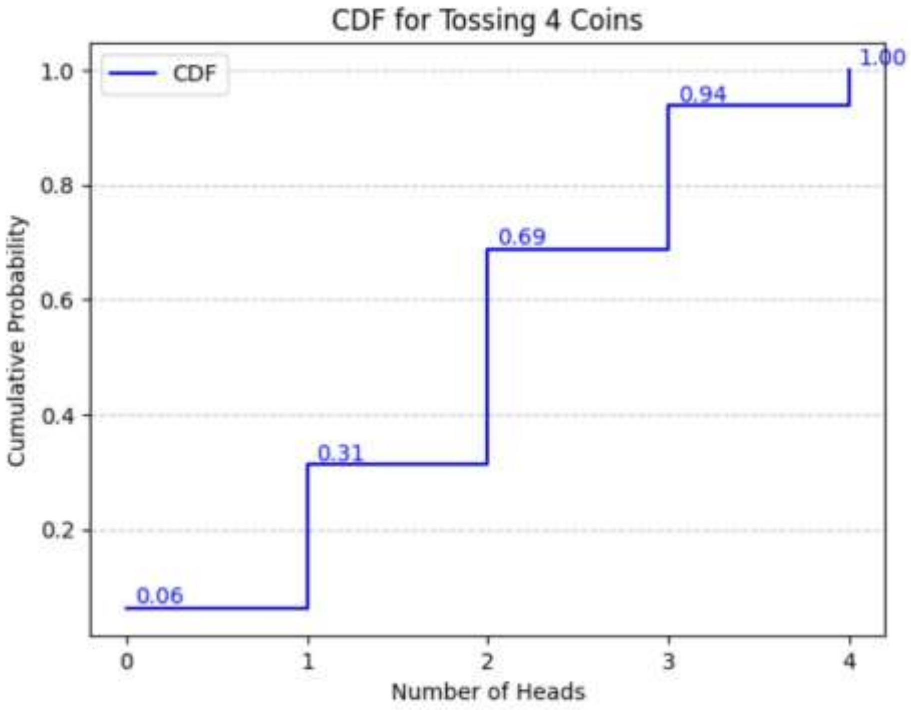
F(0) = P(X<=0) = Probability of 0 heads = 1/16

F(1) = P(X<=1) = Probability of 0 or 1 heads = 1/16 + 4/16 = 5/16

F(2) = P(X<=2) = Probability of 0,1 or 2 heads = 5/16 + 6/16 = 11/16

F(3) = P(X<=3) = Probability of 0, 1, 2 or 3 heads = 11/16 + 4/16 = 15/16

F(4) = P(X<=4) = Probability of 0, 1, 2, 3, or 4 heads = 15/16 + 1/16 = 1



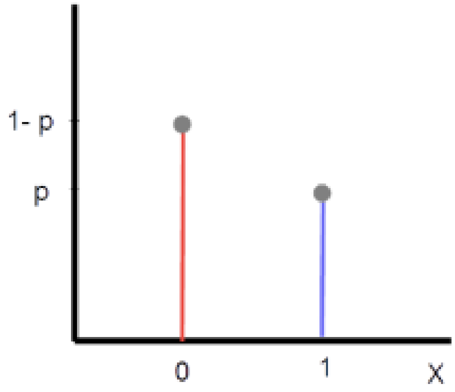
Discrete Distributions

DataCrats

Bernoulli Distribution:

A Bernoulli distribution models a binary outcome where an event can have one of two possible outcomes, typically labelled as success (1) or failure (0). Such an event is called a Bernoulli trial, and a Bernoulli distribution assesses only one trial. (n =1)

PMF = f(x,p) = $\{p \qquad if\ x = 1\ q = 1 - p\ if\ x = 0$



We can also express this formula as,

F(x,p) = $p^{x}(1-p)^{1-x}$, $x \in \{0, 1\}$

Mean = p

Variance = p*q = p(1-p)

Binomial Distribution:

Binomial distribution models the number of successful outcomes in a fixed number of independent Bernoulli trials.

PMF = f(x:n,p) = $\frac{n!}{(n-x)!x!} p^{x} (1-p)^{n-x}$

Example: Customer Conversion Rate

Imagine you're an e-commerce business owner, and you want to analyze the conversion rate of visitors to your website. In this scenario, conversion means making a purchase. You're interested in understanding the likelihood of a specific number of purchases (successes) out of a fixed number of website visits (trials) during a given period.

n (number of website visits) = 100

DataCrats

p (probability of making a purchase on each visit) = 0.75 (5% conversion rate)

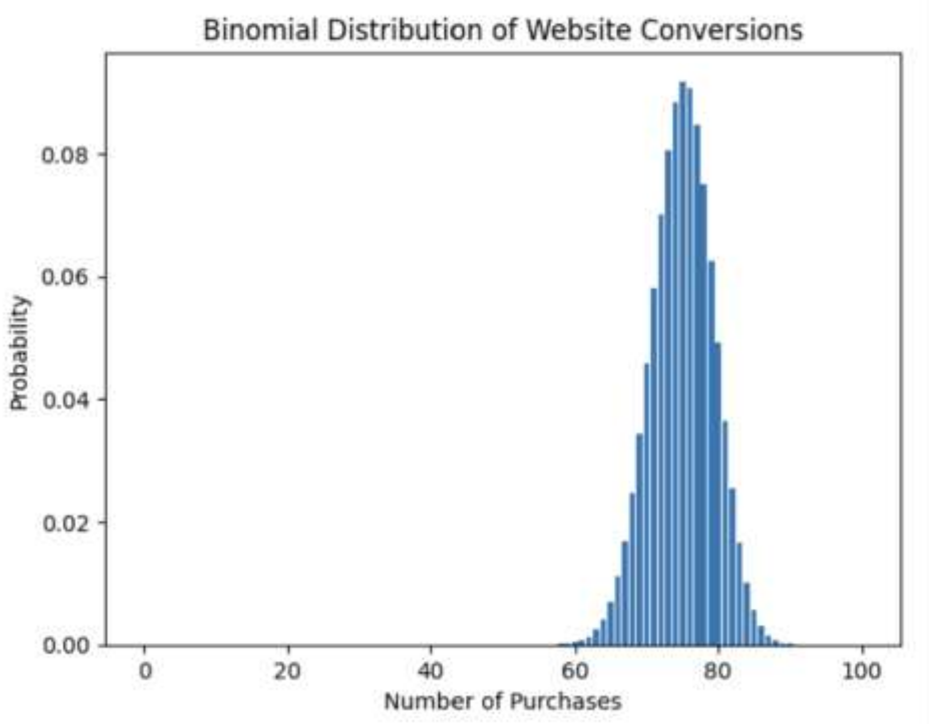
x (number of purchases) = 75

Question: What is the probability of exactly 75 visitors making a purchase out of 100 website visits given a 75% conversion rate?

Using the binomial distribution, you can calculate this probability:

P(X=50) = $\frac{100!}{(100-75)!75!} 0.75^{75} (1-0.75)^{100-75}$ = 0.0917

The probability of exactly 75 customers making a purchase out of 100 website visits with a 75% conversion rate is approximately 0.0917, or 9.17%



Multinomial distribution generalizes the binomial distribution to more than two categories or outcomes.

PMF = $\frac{n!}{x1!x2!x3!.......xk!} p1^{x1}\ p2^{x2}\ p3^{x3}\ ....\ pk^{xk}$

DataCrats

n is the total number of trials.

x is a vector of counts, with xi representing the number if observations in category i.

k is the number of categories

pi is the probability of category i.

xi! Represents the factorial of xi.

Example: Suppose two students play chess against each other. The probability that student A wins a given game is 0.5, the probability that student B wins a given game is 0.3, and the probability that they tie in a given game is 0.2. If they play 10 games, what is the probability that player A wins 4 times, player B wins 5 times, and they tie 1 time?

n is the total number of trials (in this case, 10 games).

x1 is the number of times player A wins (4 times) x2 is the number of times player B wins (5 times).

DataCrats

x3 is the number of ties (1 time).

p1 is the probability that player A wins a game (0.5).

p2 is the probability that player B wins a game (0.3).

p3 is the probability of a tie (0.2)

PMF = 0.038272

Uniform Distribution:

Uniform distribution is characterized by a constant probability density over a specified interval. All values within the interval are equally likely to occur.

PDF:

$f(x) = 1/(b-a)$ for $a<=x<=b$

CDF: $F(x) = x-a/b-a$, where $a<=x<=b$

Mean = a+b/2

Variance = (b-a)**2/12

DataCrats

Uniform distributions are used for generating random numbers. They ensure that each number in a given range has an equal chance of being selected.

In many lottery systems and games of chance, uniform distributions are used to ensure fairness. This guarantees that each possible outcome, such as drawing a specific card or rolling a die, has an equal probability of occurring.

Uniform distributions can be used in market research and economics to model the equally likely purchase or selection of various products or services within a given range. This information is valuable for stock management, marketing, and pricing strategies.

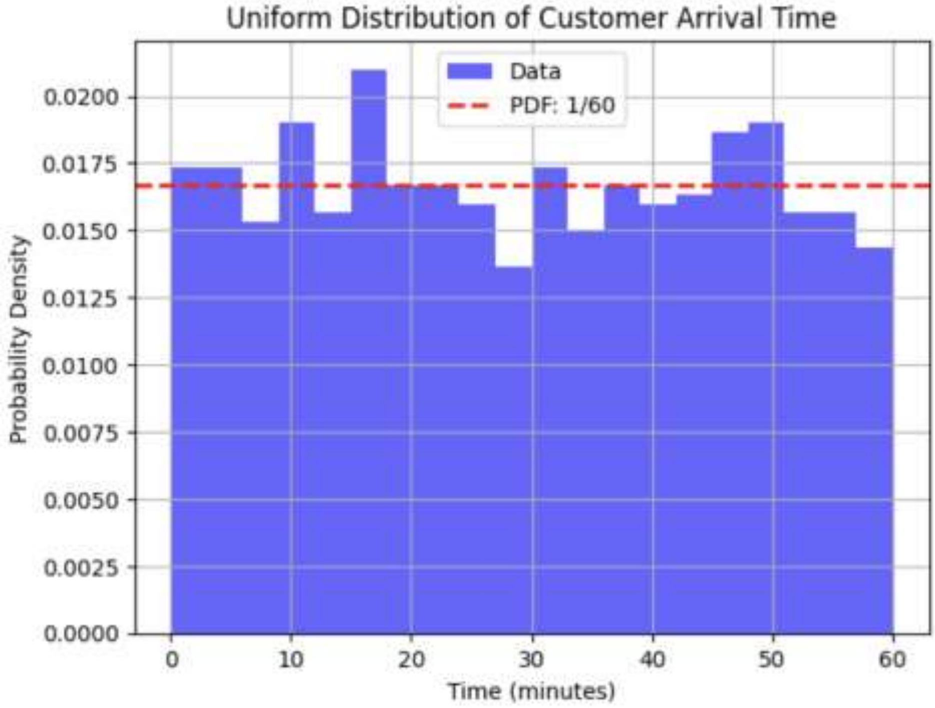
In hypothesis testing, a uniform distribution might be used as a null hypothesis to test whether data is uniformly distributed. If observed data significantly deviates from a uniform distribution, it can indicate a specific pattern or effect.

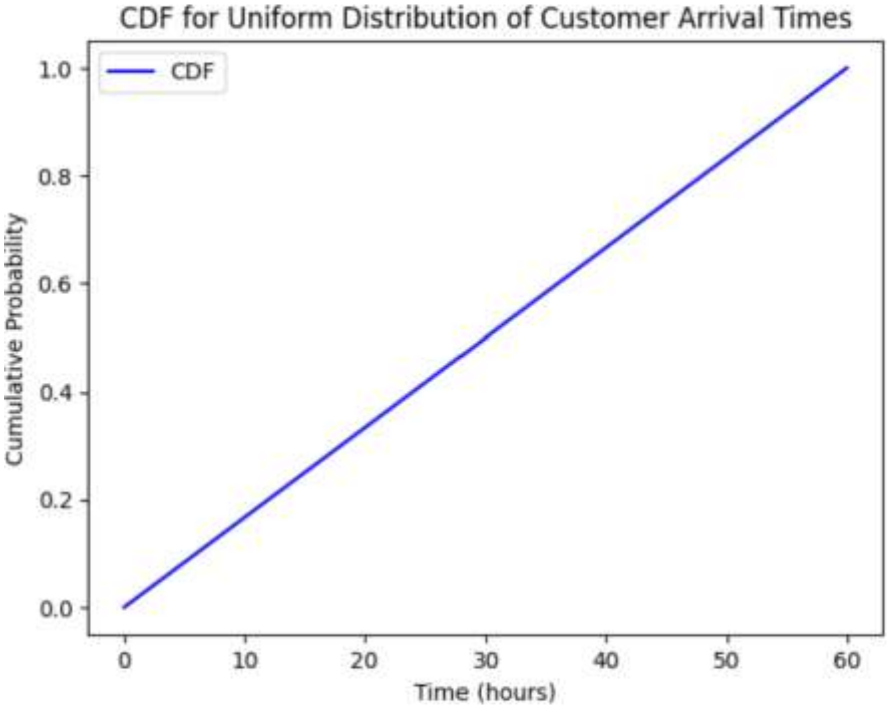
Real-World Scenario:

Imagine you're a manager at a coffee shop, and you want to predict the arrival time of customers during a particular hour of operation. You believe that customers arrive at a relatively constant rate throughout the hour. To model this scenario, you can use a uniform distribution.

a = 0, b = 60 as shop operates from 9:00 AM to 10:00 AM (60 minutes)

f(x) = 1/60 for all x within this interval





Normal distribution:

Also known as the Gaussian Distribution, it is a symmetric probability distribution that is characterized by a bell-shaped curve. Mean, Median and mode all coincide at the centre of the distribution.

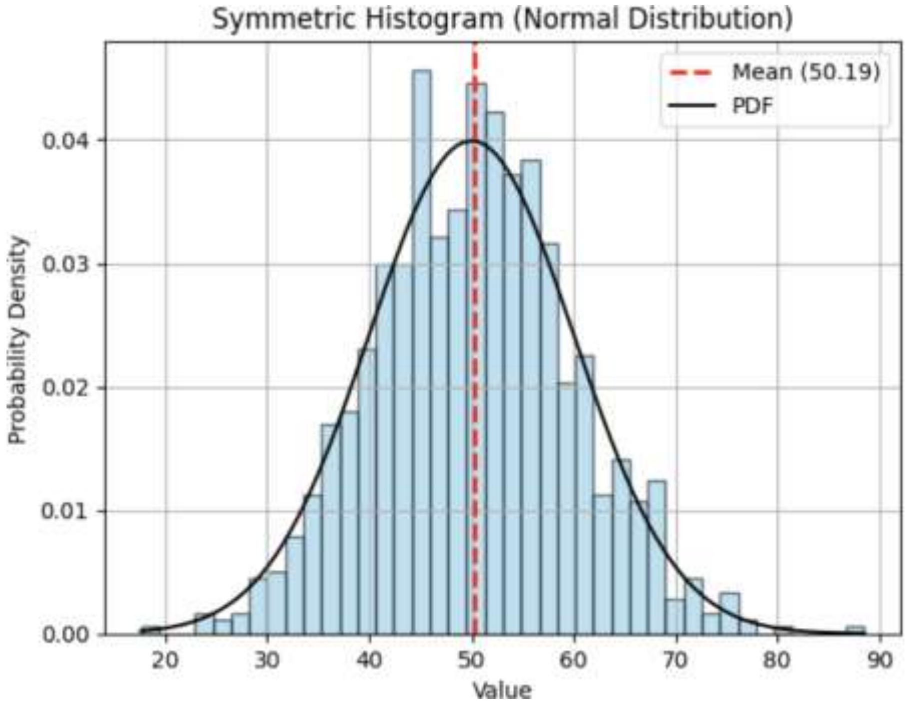
PDF : $\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^{2}}$

E[X] = u

Var(x) = σ**2

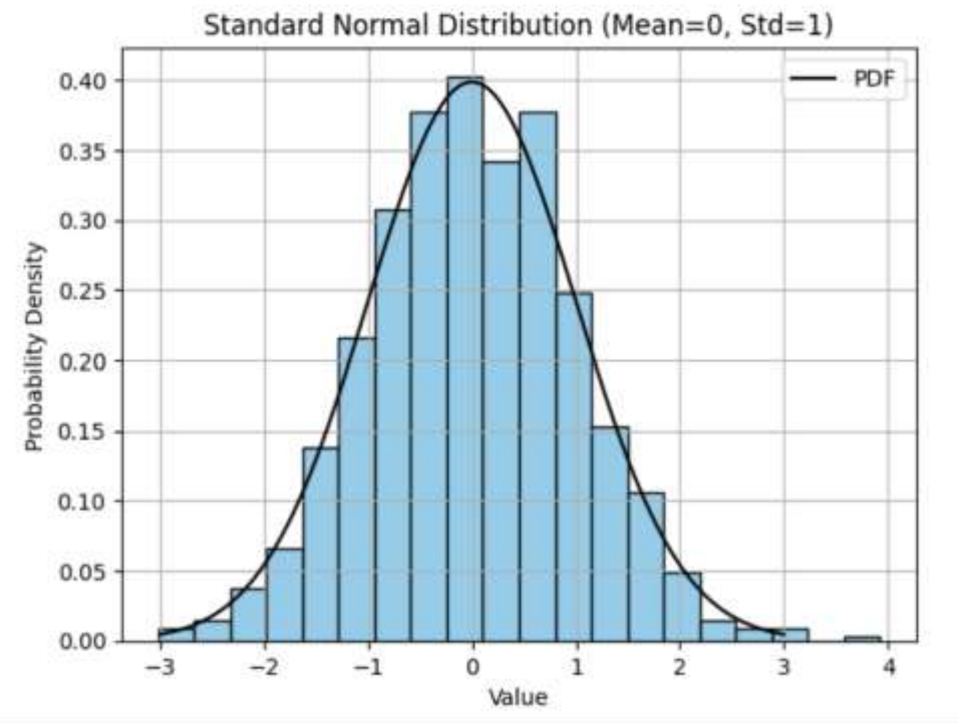
mean = 50

std_dev = 10



Standard normal distribution:

Standard normal distribution also called **the z distribution** is a special normal distribution where the mean is 0 and the standard deviation is 1. Any normal distribution can be standardized by converting its values into z scores. Z scores tell you how many standard deviations from the mean each value lies.



Standardization

When you standardize a normal distribution, the mean becomes 0 and the standard deviation becomes 1.

DataCrats

A positive z score means that your x value is greater than the mean.

A negative z score means that your x value is less than the mean.

A z score of zero means that your x value is equal to the mean.

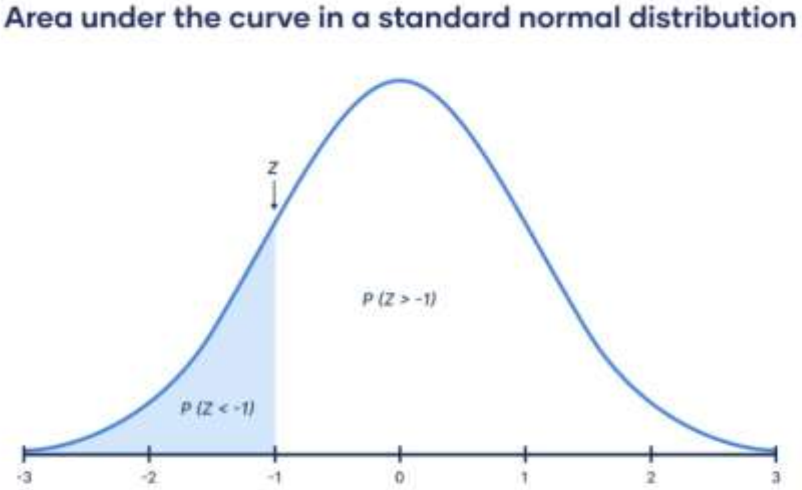
It is possible to change each normal random variable X into a z score through the following standard normal distribution formula

Z = (X- μ)/σ

X represents a normal random variable

μ is the mean of X,

σ is the standard deviation of X



Normalization rescales a dataset so that each value falls between 0 and 1.

$\mathbf{x_{new} = (x_i - x_{min}) / (x_{max} - x_{min})}$

where:

- $\mathbf{x_i}$: The i^{th} value in the dataset
- $\mathbf{x_{min}}$: The minimum value in the dataset
- $\mathbf{x_{max}}$: The maximum value in the dataset

Typically, we **normalize** data when performing some type of analysis in which we have multiple variables that are measured on different scales and we want each of the variables to have the same range. This prevents one variable from being overly influential, especially if it's measured in different units (i.e., if one variable is measured in inches and another is measured in yards).

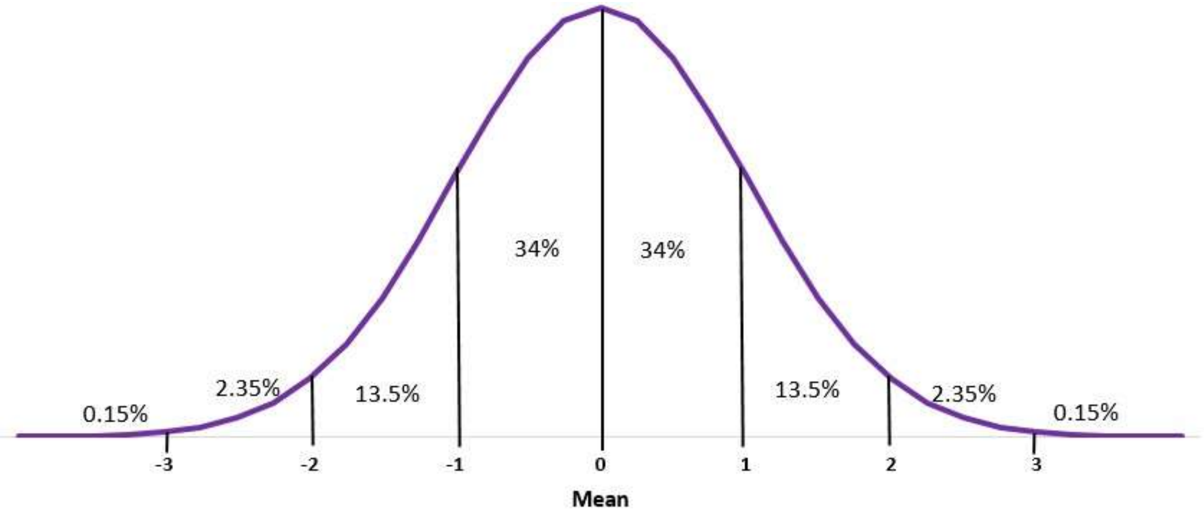
On the other hand, we typically **standardize** data when we'd like to know how many standard deviations each value in a dataset lies from the mean.

Normalization	Standardization
Minimum and maximum value of variables are used for scaling.	Mean and standard deviation of variables are used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between 0 and 1	It is not bounded to a certain range.
It is affected by outliers	It is less effected by outliers
It is useful when we don't know about the distribution.	It is useful when the feature distribution is normal or gaussian.

Maps the data to a specific range, which may lead to loss of information about the original distribution.	Preserves the shape of the original data distribution but shifts and scales it.

A standard normal distribution has the following properties:

- About 68% of data falls within one standard deviation of the mean
- About 95% of data falls within two standard deviations of the mean
- About 99.7% of data falls within three standard deviations of the mean



This is known as the **Empirical Rule** and is used to understand the distribution of values in a dataset.

For example, suppose the height of plants in a certain garden are normally distributed with a mean of 47.4 inches and a standard deviation of 2.4 inches.

According to the Empirical Rule, what percentage of plants are less than 54.6 inches tall?

The Empirical Rule states that for a given dataset with a normal distribution, 99.7%of data values fall within three standard deviations of the mean. This means that 49.85% of values fall between the mean and three standard deviations above the mean.

In this example, 54.6 is located three standard deviations above the mean. Since we know that 50% of data values fall below the mean in a normal distribution, a total of 50% + 49.85% = 99.85% of values fall below 54.6.

Thus, **99.85%** of plants are less than 54.6 inches tall.

The empirical rule, often referred to as the 68-95-99.7 rule, is a fundamental statistical concept used to understand the distribution of data in a normal or approximately normal distribution. It provides valuable insights into the spread and concentration of data points around the mean. This rule is highly useful in data analysis, particularly in the context of descriptive statistics and quality control. Let's delve into the details of the empirical rule:

Why and where is it Used?

Understanding Data Distribution: The empirical rule is primarily used to gain a quick and intuitive understanding of how data is distributed around the mean in a normal or nearly normal distribution.

Quality Control: It's commonly applied in quality control and manufacturing to assess variations and deviations in products. For instance, in a manufacturing process, you can use the rule to estimate the percentage of products that fall within certain quality specifications.

Descriptive Statistics: Data analysts and statisticians use the empirical rule to provide a simple summary of data distribution in presentations and reports.

How is it Useful in Data Analysis?

The empirical rule is valuable in data analysis for several reasons:

Rapid Assessment: It allows for a quick visual or mental assessment of how data is dispersed around the mean.

Approximate Probabilities: Without performing detailed statistical calculations, you can make rough estimates of the likelihood of data falling within specific ranges.

Identifying Outliers: It can help identify potential outliers or data points that are significantly distant from the mean.

Example: Applying the Empirical Rule

Let's consider an example to illustrate the use of the empirical rule:

Scenario: Imagine you are analyzing the exam scores of a large group of students, and you suspect that the scores follow a normal distribution. The mean score is 75, and the standard deviation is 10.

According to the empirical rule:

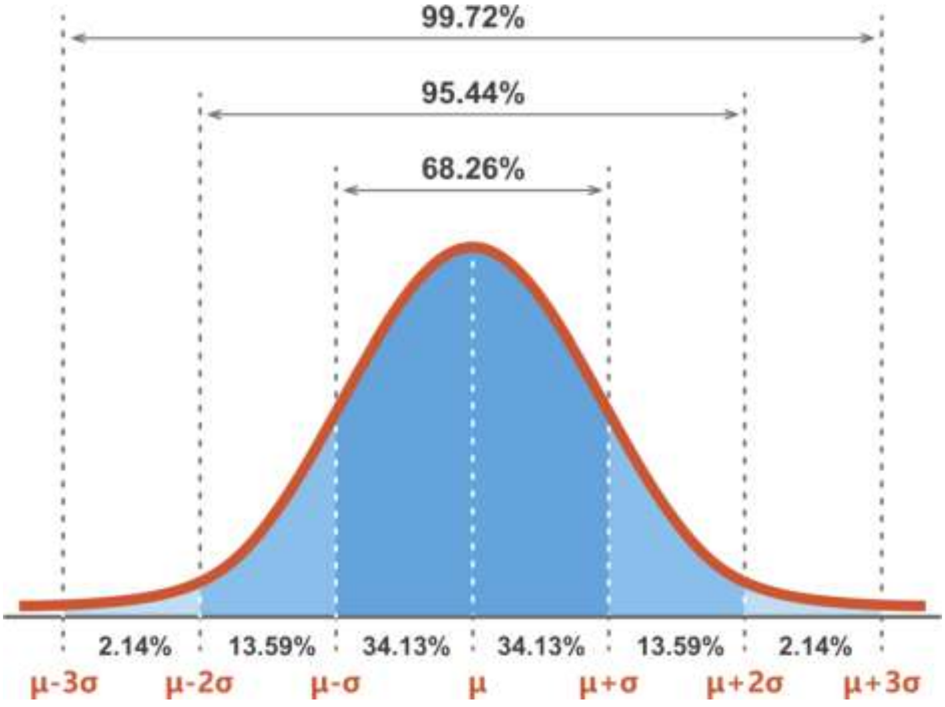
Approximately 68% of students scored between 65 and 85 (within one standard deviation of the mean).

About 95% of students scored between 55 and 95 (within two standard deviations of the mean).

Nearly 99.7% of students scored between 45 and 105 (within three standard deviations of the mean).

Usage in Data Analysis:

Based on this knowledge, you can quickly assess the distribution of scores and communicate to stakeholders that the majority of students scored within a certain range. This information can be vital for setting performance benchmarks, identifying exceptional students, or assessing the quality of teaching.



Inferential Statistics

The Relationship Between Descriptive and Inferential Statistics:

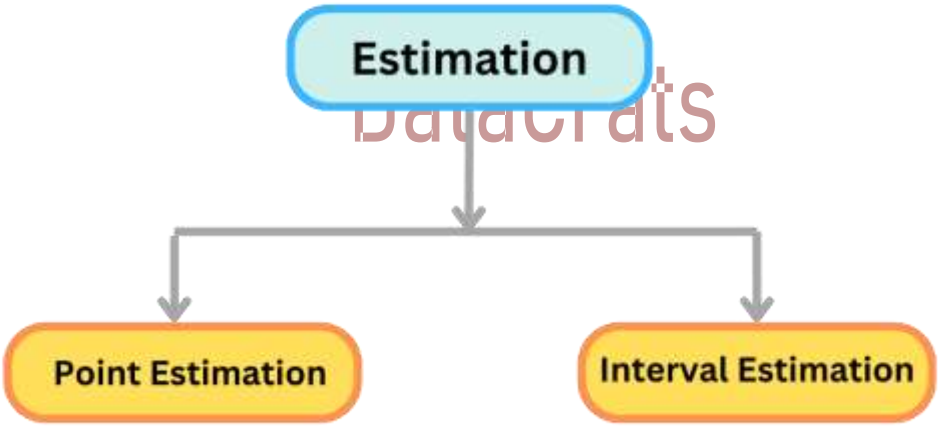
- **Sequential Process:** Typically, data analysis starts with descriptive statistics. Analysts explore and summarize the dataset to gain insights into its characteristics and patterns. These insights inform the subsequent inferential analysis.
- **Data Exploration:** Descriptive statistics help identify potential relationships, trends, or outliers in the data. This can guide the formulation of research questions and hypotheses for inferential analysis.
- **Sampling:** When conducting inferential statistics, the choice of an appropriate sample is often informed by the insights gained from descriptive statistics. Descriptive statistics help determine the characteristics of the population and inform how representative the sample should be.

- **Assumption Checking:** Inferential statistics often rely on assumptions about the distribution of data. Descriptive statistics are used to check whether these assumptions hold, ensuring the validity of inferential methods.
- **Interpretation:** After conducting inferential analyses, the results need to be interpreted and communicated. Descriptive statistics can be used to provide context and help make the inferential findings more understandable and actionable.

In summary, descriptive statistics provide a foundation for understanding the nature of data, while inferential statistics allow us to make broader conclusions and inferences about populations based on samples. Together, these two branches of statistics are crucial for comprehensive data analysis, ensuring that we not only understand our data but can also draw meaningful insights and make informed decisions based on it.

Point and Interval Estimation

Have you asked yourself how statisticians determine parameters such as mean age of an entire country's population? It is obvious that they can't get data from every single member of the population to calculate this statistic.



Datacrafts

Estimation is the prediction of characteristics of population with the help of sample or is the process of using sample data to make informed guesses or approximations about unknown population parameters. It plays a pivotal role in statistical inference, where we seek to draw conclusions about a population based on the information available in a sample.

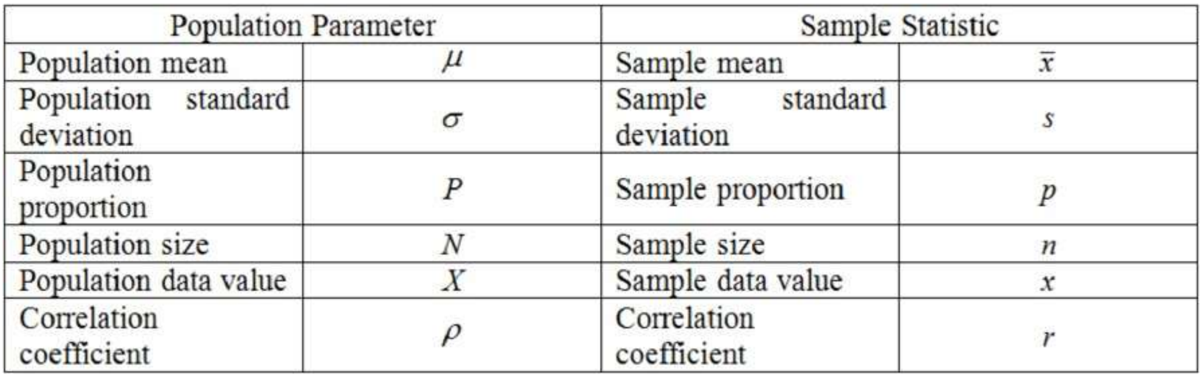
The process of finding an appropriate value of some parameter of a population from random samples of the population.

1. Population Parameters:

In many statistical analyses, the goal is to learn something about a specific characteristic of a population, known as a population parameter. These parameters represent the true, but often unknown, properties of the entire population. Examples of population parameters include the mean, variance, proportion, or regression coefficients.

2. Sample Statistics:

To estimate these population parameters, we typically collect data from a subset of the population, known as a sample. From this sample, we calculate sample statistics, which are numerical values that serve as estimates of the corresponding population parameters. Common sample statistics include the sample mean, sample variance, sample proportion, and so on.



Population Parameter		Sample Statistic	
Population mean	μ	Sample mean	$\bar{x}$
Population standard deviation	σ	Sample standard deviation	s
Population proportion	P	Sample proportion	p
Population size	N	Sample size	n
Population data value	X	Sample data value	x
Correlation coefficient	ρ	Correlation coefficient	r

Point Estimation

This is the simplest form of estimation, where you provide a single value as your best guess for the unknown parameter. For example, if you calculate the average income in a sample of households and use that as an estimate for the average income of all households in a city, you're performing point estimation. The sample data of a population is used to find a point estimate or a statistic that can act as the best estimate of an unknown parameter that is given for a population.

Properties of point estimation:

It is desirable for a point estimate to be the following:

Consistent -the larger is the sample size, the more accurate is the estimate.

Unbiased - The expectation of the observed values of various samples equals the corresponding population parameter. Let’s take, for example, we can say that sample mean is an unbiased estimator for the population mean.

When the two properties above are met for an estimator, you have the most efficient or best unbiased estimator.

- Formulae that Can be Used to Measure Point Estimators:

  Maximum Likelihood Estimation or MLE

  Jeffrey Estimation

  Wilson Estimation

  Laplace Estimation

- What are the Values Needed to Calculate Point Estimators?

  The number of successes is shown by S.

  The number of trials is shown by T.

  The Z–score is shown by z.

Once You Know All the Values Listed Above, You Can Start Calculating the Point Estimate According to the Following Given Equations:

Maximum Likelihood Estimation: MLE = S / T

DataCrats

Laplace Estimation: Laplace equals (S + 1) / (T + 2)

Jeffrey Estimation: Jeffrey equals (S + 0.5) / (T + 1)

Wilson Estimation: Wilson equals $(S + z^2/2) / (T + z^2)$

Drawback of Point Estimates

The drawback of point estimate is that no information is available regarding its reliability, i.e., how close it is to its true population parameter. In fact, the probability that a single sample statistic actually equals the population parameter is extremely small. For this reason, point estimates are rarely used alone to estimate population parameters. It is better to offer a range of values within which the population parameters are expected to fall so that reliability (probability) of the estimate can be measured. This is the purpose of **interval estimation.**

Interval Estimation

A point estimate does not provide information about ‘how close is the estimate’ to the population parameter unless accompanied by a statement of possible sampling error involved based on the sampling distribution of the statistic. It is therefore important to know the precision of an estimate before depending on it to make a decision. Thus, decision makers prefer to use an interval estimate (i.e., the range of values defined around a sample statistic) that is likely to contain the population parameter value.

The evaluation of a parameter of a population by computing an interval, (or range of values) within which the parameter is most likely to be located. More accurate than point estimate.

Confidence Interval:

However, it is also important to state ‘how confident’ one should be that the interval estimate contains the parameter value. Hence an interval estimate of the population parameter is a confidence interval with a statement of confidence that the interval contains the parameter value. In other words, a confidence interval estimation is an interval of values computed from sample data that is likely to contain the true population parameter value. Intervals are chosen such that parameter falls within a 95% or 99% probability called the confidence coefficient. Hence the intervals are called Confidence Interval. The end points of such an interval called upper and lower confidence limits. A confidence level is the probability that the interval estimate will include the population parameter (such as the mean)

CI = point estimate +- margin of error

DataCrats

Where Margin of error = $zc \times Standard\ error\ of\ a\ particular\ statistic$

zc= critical value of standard normal variable that represents confidence level

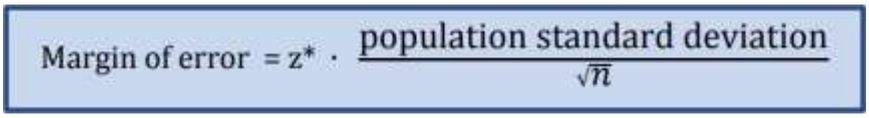
(probability of being correct) such as 0.90, 0.95, and so on.

Margin of Error:

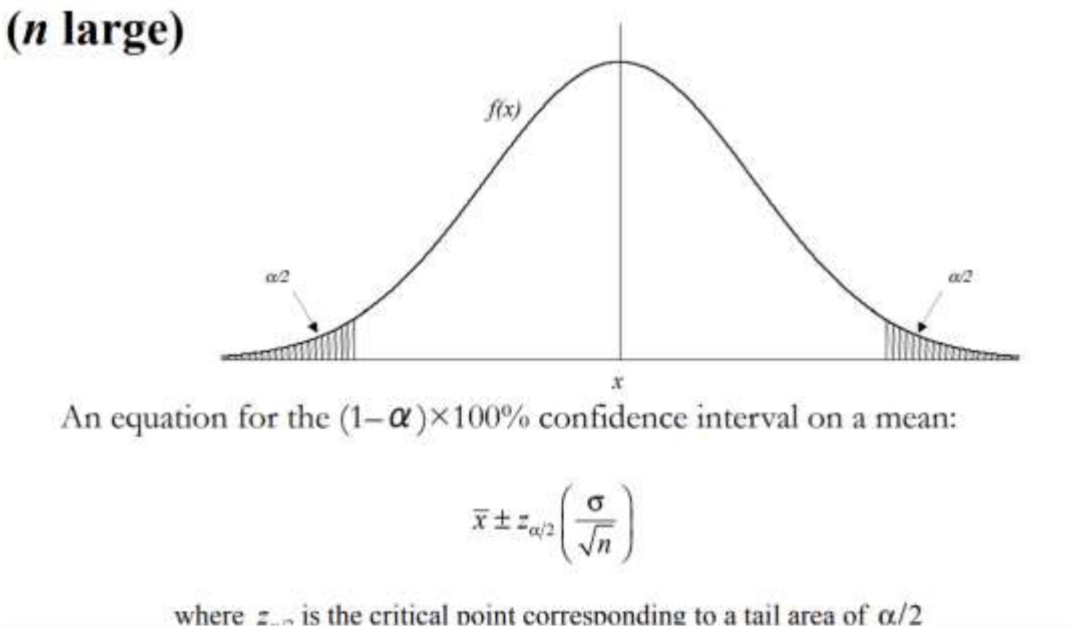
- In order to find a confidence interval, the margin of error must be known.

- The margin of error depends on the degree of confidence that is required for the estimation.

-The level of confidence is represented by z* (called z star)



$$\text{Margin of error } = z^* \cdot \frac{\text{population standard deviation}}{\sqrt{n}}$$



An equation for the (1−α)×100% confidence interval on a mean:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

where $z_{\alpha/2}$ is the critical point corresponding to a tail area of α/2

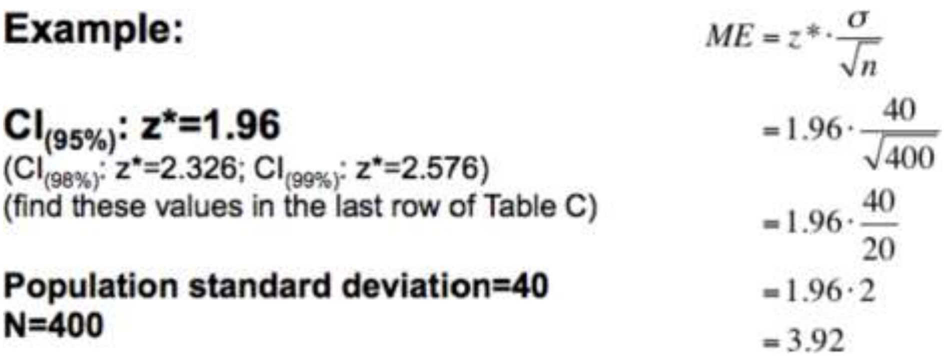
1-alpha = confidence interval

Alpha = 0.05

Alpha = significance value

DataCrats

Z value is the no. of std from the sample mean. Za/2 = confidence coef or critical value



Example:

$CI_{(95\%)}$: z*=1.96
($CI_{(98\%)}$: z*=2.326; $CI_{(99\%)}$: z*=2.576)
(find these values in the last row of Table C)

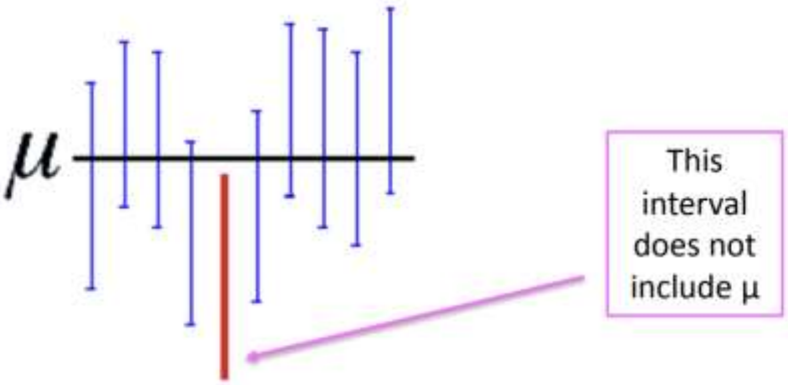
Population standard deviation=40
N=400

$$\begin{aligned} ME &= z^* \cdot \frac{\sigma}{\sqrt{n}} \\ &= 1.96 \cdot \frac{40}{\sqrt{400}} \\ &= 1.96 \cdot \frac{40}{20} \\ &= 1.96 \cdot 2 \\ &= 3.92 \end{aligned}$$

- In the above example, a confidence level of 95% was selected.
- The value of z* for a specific confidence level is found using a table in the back of a statistics textbook. The value of z* for a confidence level of 95% is 1.96.
- After putting the value of z*, the population standard deviation, and the sample size into the equation, a margin of error of 3.92 is found.

- Sample means will follow the normal probability distribution for large sample sizes (n ≥ 30)

Select 10 samples and construct 90 % confidence intervals around each of the sample means Theoretically, 9 of the 10 intervals will contain the true population mean, which remains unknown



Careful not to misinterpret the definition of a confidence interval

✹ NOT Correct – “there is a 90 % probability that the true population mean is within the interval”

DataCrats

✹ CORRECT – “there is a 90 % probability that any given confidence interval from a random sample will contain the true population mean

What if n is small? (Student’s t Distribution) (n<30)

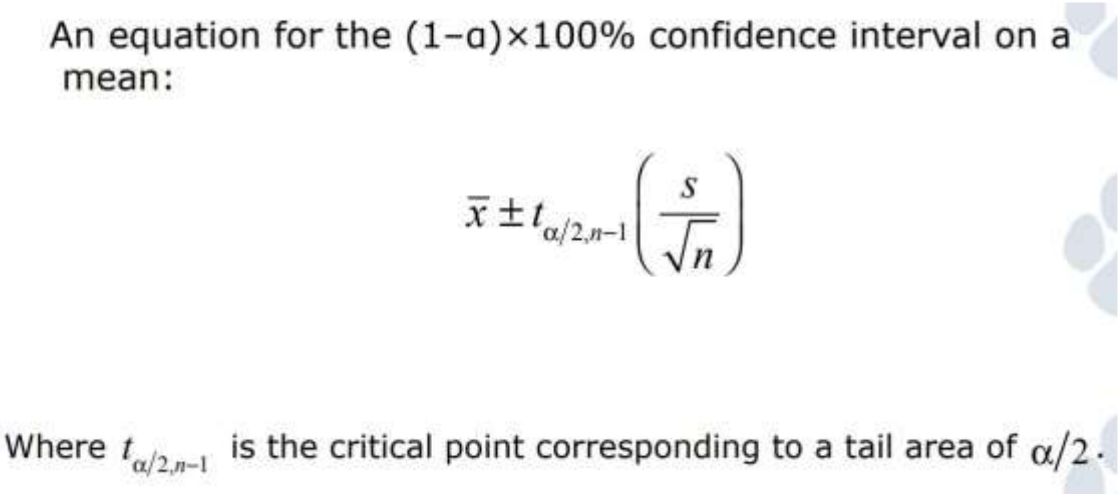
As the sample size becomes smaller, the sample standard deviation becomes an increasingly poor approximation of the population standard deviation. The end result is that a 95% confidence interval computed using s instead of σ may actually only contain the population mean 90% of the time, or 85% of the time, or even less. William Gosset developed a new probability distribution, which he called the t distribution, to describe the probabilities associated with the statistic.

The *t*-distribution describes the standardized distances of sample means to the population mean when the population standard deviation is not known, and the observations come from a normally distributed population.

- As the sample size increases, the *t*-distribution becomes more similar to a normal distribution.

- The shape of the *t*-distribution depends on the degrees of freedom (n-1). The curves with more degrees of freedom are taller and have thinner tails. All three *t*-distributions have “heavier tails” than the z-distribution.

- You can see how the curves with more degrees of freedom are more like a z-distribution.



An equation for the (1−α)×100% confidence interval on a mean:

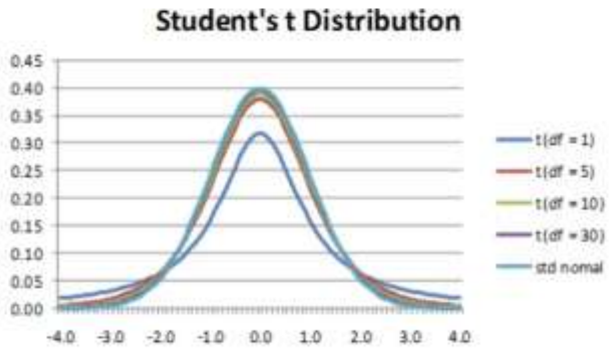
$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

Where $t_{\alpha/2,n-1}$ is the critical point corresponding to a tail area of $\alpha/2$.

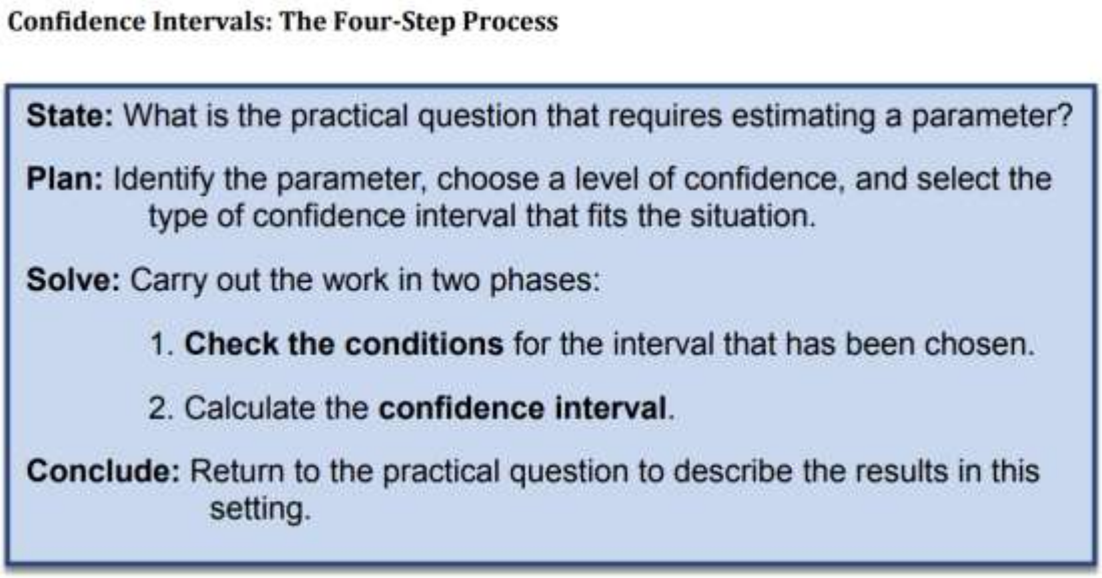
- So far we have discussed confidence intervals for the mean where n ≥ 30 and when population standard deviation is known, we are assuming the population is normally distributed and so we can follow the z distribution procedure for large sample sizes.
- When population standard deviation is unknown (more often the case!), we follow the t-distribution procedure and substitute s (sample standard deviation) in place of population standard deviation.

DataCrats

As degrees of freedom increase, the shape of t-distribution becomes similar to normal distribution.



1. The density curve looks like a standard normal curve, but the tails of the t-distribution are "heavier" than the tails of the normal distribution. That is, we are more likely to get extreme t-values than extreme z-values.
2. As the degrees of freedom r increases, the t-distribution appears to approach the standard normal z-distribution.



Confidence Intervals: The Four-Step Process

State: What is the practical question that requires estimating a parameter?

Plan: Identify the parameter, choose a level of confidence, and select the type of confidence interval that fits the situation.

Solve: Carry out the work in two phases:

1. **Check the conditions** for the interval that has been chosen.
2. Calculate the **confidence interval**.

Conclude: Return to the practical question to describe the results in this setting.

Datacrats

Hypothesis Testing

Let's say you have another medicine that you think is superior to the one you're taking. How can you demonstrate it? Hypothesis testing is useful in this situation. It issimilar to being a detective, looking for proof that can support up your claim.

A **hypothesis** is a claim about a population parameter.

Hypothesis testing is a statistical method used to evaluate a claim or hypothesis about a population parameter. It involves comparing sample data to a hypothesis about a population parameter to determine whether there is enough evidence to reject or fail to reject the hypothesis.

A hypothesis test is a formal procedure to check if a hypothesis is true or not. Without hypothesis & hypothesis tests, you risk drawing the wrong conclusions and making bad decisions

Types of Hypotheses:

Hypothesis testing is based on making two different claims about a population parameter.

DataCrats

Null Hypothesis (H0): The default assumption that there is no significant difference or effect.

Alternative Hypothesis (H1 or Ha): The assertion that there is a significant difference or effect. The alternative hypothesis is typically what we are trying to prove.

The two claims need to be **mutually exclusive**, meaning only one of them can be true.

For **example,** we want to check the following claim:

"The average height of people in Denmark is more than 170 cm." In this case, the parameter is the average height of people in Denmark. The null and alternative hypothesis would be:

Null hypothesis: The average height of people in Denmark is 170 cm.

Alternative hypothesis: The average height of people in Denmark is more than 170 cm.

- If the data supports the alternative hypothesis, we **reject** the null hypothesis and **accept** the alternative hypothesis.

DataCrafts

- If the data does **not** support the alternative hypothesis, we **keep** the null hypothesis.

Real-Life Examples:

Drug Efficacy: Null Hypothesis (H0): The new drug has no effect. Alternative Hypothesis (Ha): The new drug is effective.

A/B Testing: Null Hypothesis (H0): There is no difference between versions A and B. Alternative Hypothesis (Ha): There is a significant difference.

Manufacturing Quality: Null Hypothesis (H0): The manufacturing process meets standards. Alternative Hypothesis (Ha): The manufacturing process does not meet standards.

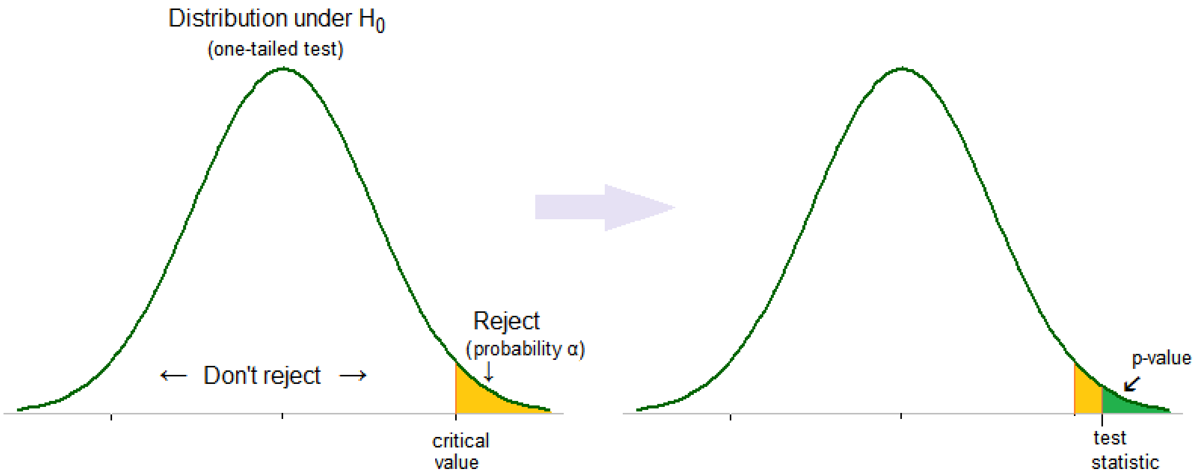
Level of significance and p value:

The **significance level**, denoted as **α (alpha),** is like setting a rule or standard for how much evidence you need before you believe a hypothesis. It's the maximum chance you're willing to take of being wrong when you reject a true null hypothesis. Its the predetermined threshold for deciding if the evidence is strong enough to rejectthe null hypothesis. Commonly used values for α are 0.05, 0.01, or 0.10.

A lower significance level means that the evidence in the data needs to be stronger to reject the null hypothesis.

The **p-value** is a measure that helps you decide if the evidence from your data is strong enough to reject the null hypothesis. A low p-value indicates stronger evidence against the null hypothesis. If the p-value is less than α, you might say theresult is "statistically significant.

Decision Rule: If the p-value is less than or equal to α, you reject the null hypothesis.



Example: You're testing a new drug, and the p-value is 0.03. This means there's a 3% chance that the observed results happened by random chance alone. If your α is0.05, which is common, you might decide the evidence is strong enough to reject theidea that the drug has no effect.

Types of errors:

When conducting hypothesis testing, there are two main types of errors that researchers need to be aware of: Type I error and Type II error. Let's break down each type with examples:

1. Type I Error (False Positive):

Type I error occurs when the null hypothesis is incorrectly rejected when it is actually true. In other words, it's a false alarm or a false positive.

Example: Suppose a pharmaceutical company is testing a new drug, and the null hypothesis is that the drug has no effect. If, based on the sample data, the researchers mistakenly conclude that the drug is effective (reject the null hypothesis), when in reality it is not, this is a Type I error.

2. Type II Error (False Negative):

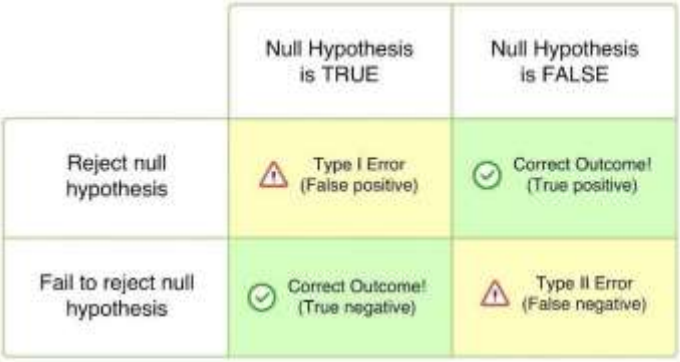
Type II error occurs when the null hypothesis is not rejected when it is actually false. It means failing to detect a real effect that exists.

Example: Using the same drug example, let's say the drug does have a significant effect, but the researchers fail to detect it based on the sample data. In this case, they accept the null hypothesis (fail to reject it), when they should have rejected it. This is a Type II error.

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

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

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



	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	⚠ Type I Error (False positive)	✓ Correct Outcome! (True positive)
Fail to reject null hypothesis	✓ Correct Outcome! (True negative)	⚠ Type II Error (False negative)

One Tailed and Two Tailed Test:

In hypothesis testing, the choice between a one-tailed and a two-tailed test depends on the specific research question and the directionality of the expected effect. Let's delve into each type with examples and decision rules:

One-Tailed Test:

In a one-tailed (or one-sided) test, the critical region for rejecting the null hypothesis is located in only one tail of the distribution (either the left or the right).

When to Use: When the research hypothesis specifies a direction (e.g., the effect is expected to be greater than or less than a certain value).

Example:

Research Hypothesis: The average time to complete a task with a new software is less than 10 minutes.

Null Hypothesis: The average time to complete a task with a new software is 10 minutes or more.

Two-Tailed Test:

In a two-tailed (or two-sided) test, the critical region for rejecting the null hypothesis is divided between both tails of the distribution.

When to Use: When the research hypothesis is non-directional or when you want to detect any significant difference, whether it's greater or less than a certain value.

Example:

Research Hypothesis: There is a difference in the average scores between two groups.

Null Hypothesis: There is no difference in the average scores between two groups.

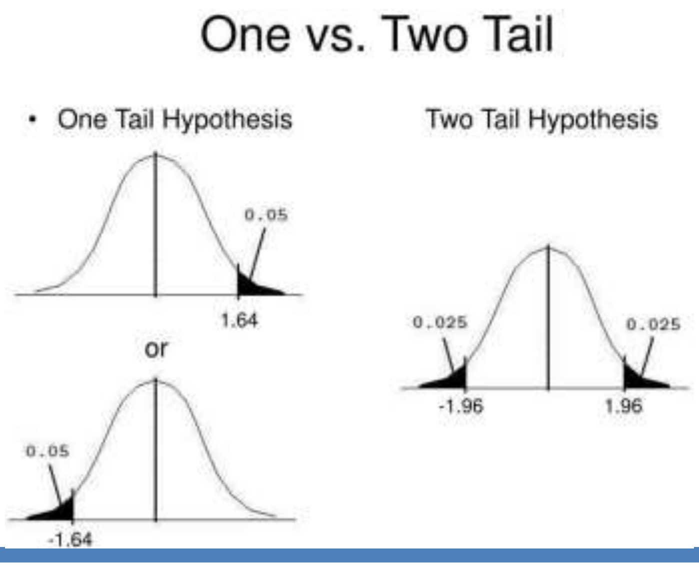
Decision Rules for One-Tailed Test:

DataCrats

- If the p-value is less than or equal to alpha, reject the null hypothesis.
- If the p-value is greater than alpha, do not reject the null hypothesis.

Decision Rules for Two-Tailed Test:

- If the p-value is less than or equal to alpha/2, reject the null hypothesis.
- If the p-value is greater than alpha/2, do not reject the null hypothesis.



Types of Tests in Statistics:

1. Z-Test

The Z-test is a statistical method used to determine if there is a significant difference between a sample mean and a known population mean, or between the means of two independent samples. It's particularly useful when working with large sample sizes (n>30) and when the population standard deviation is known. There are two main types of Z-tests: one-sample Z-test and two-sample Z-test.

- **One-Sample Z -Test**

When you have a single sample and want to determine if its mean is significantly different from a known population mean.

Formula:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

DataCrats

where,

$\overline{x}$ is the sample mean,

μ is the population mean

σ is the population standard deviation

n is the sample size

Example:

Imagine you're a social media manager, and you want to know if your latest campaign is a hit. You survey 100 followers and ask how many likes they give to your posts, on average. But how can you tell if this average (sample mean) reflects the true feelings of your entire audience (population)? This is where the one-sample z-test comes in!

Let's say your sample of 100 followers gives an average of 7 likes, and you hypothesize the average like for all followers is 5 likes ($\mu = 5$) with a known standard deviation of 1 like ($\sigma = 1$). You calculate a z-statistic of 2 (indicating the sample mean is 2 standard deviations higher than expected). If the p-value from a z-test calculator is 0.02, you would reject the null hypothesis. This suggests your campaign is likely a success, with your sample showing a significantly higher average like than the hypothesized average for all followers.

- **Two-Sample Z -Test**

The two-sample z-test is a statistical method used to compare the means of two independent samples to determine if there is a significant difference between them. It's particularly useful when working with numerical data from two different groups or populations.

Formula:

$$Z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n1}}}$$

Example: A marketing team tests a new website design on two groups of users (Group A and Group B). They measure the average time spent browsing the website (in minutes) for each group. They want to know if the new design (Group B) keeps users engaged for significantly longer than the old design (Group A).

DataCrafts

2. T-Test

The t-test is a statistical method used to determine if there is a significant difference between the means of two groups or to compare the mean of a sample to a known value. There are different types of t-tests, but the two most common ones are the independent samples t-test and the paired samples t-test. It is effective when dealing with small sample sizes(n<30) and when the population standard deviation is unknown.

- **Independent Samples T-Test**

It is used when comparing the means of two independent groups to determine if there is a statistically significant difference between them.

Formula:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n1} + \frac{s_2^2}{n2}}}$$

Example: A/B Testing for Website Conversion Rates

Imagine you work for an e-commerce company, and you're tasked with analyzing the effectiveness of two different website designs (A and B) in terms of their conversion rates. You want to know if there's a significant difference in the average conversion rates between the two designs.

- **Paired Samples T-Test**

It is used when comparing the means of two related groups such as before and after measurements on the same subjects.

Formula:

$$t = \frac{\overline{x}_d}{\frac{s_d}{\sqrt{n}}}$$

DataCrats

where, $\overline{x}_d$ is the mean of the difference between paired observations,

s_d is the standard deviation of the difference

Example: Scenario: Evaluating the Effectiveness of a Training Program

Imagine you work for a company that has implemented a training program to improve the performance of its employees. You want to assess whether the training program has led to a significant improvement in employee productivity.

It's important to note that the t-test assumes that the data is approximately normally distributed. If the sample size is large, the t-distribution approaches the normal distribution, and the t-test becomes robust against deviations from normality. If the sample size is small and the data is not normally distributed, other non-parametric tests may be more appropriate.

3. ANOVA (Analysis of Variance)

It is a statistical method used to compare means among three or more groups. ANOVA assesses whether the observed variances between group means are significantly greater than the expected variances if all the group means were the same. An ANOVA test can be either one-way or two-way depending upon the number of independent variables.

- ANOVA produces the F statistic as its outcome, which is crucial for analyzing differences among group means.

- The F statistic is a ratio that compares within-group variance tobetween-group variance.

- If the F-ratio is large enough to reject the null hypothesis, it suggests meaningful differences among groups.

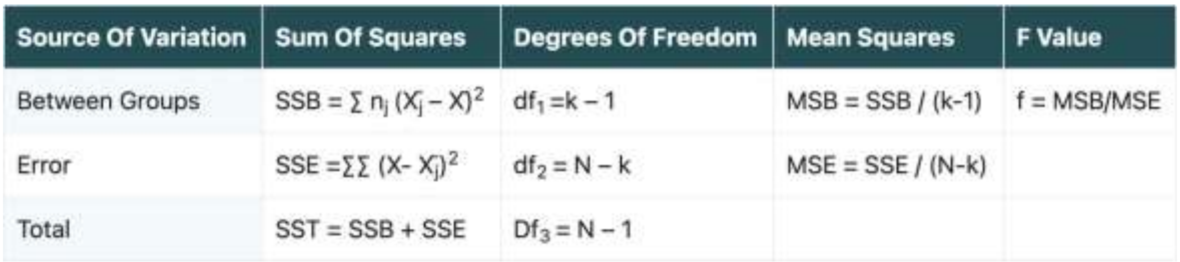
- Conversely, if the F-ratio is small and fails to reach significance, it indicates no meaningful differences among groups, and the null hypothesis is accepted.

In ANOVA terminology, an independent variable is called a factor variable and a dependent variable is called a response variable. One of the biggest challenges in machine learning is the selection of the most reliable and useful features that are used in order to train a model. ANOVA helps in selecting the best features to train a model. ANOVA minimizes the number of input variables to reduce the complexity of the model.

DataCrafts

Formula:

$$F = \frac{MSB}{MSW}$$



Source Of Variation	Sum Of Squares	Degrees Of Freedom	Mean Squares	F Value
Between Groups	$SSB = \sum n_j (X_j - X)^2$	$df_1 = k - 1$	MSB = SSB / (k-1)	f = MSB/MSE
Error	$SSE = \sum\sum (X - X_j)^2$	$df_2 = N - k$	MSE = SSE / (N-k)	
Total	SST = SSB + SSE	$Df_3 = N - 1$		

where, MSB is mean squares between groups

MSW or MSE is mean squares within groups of mean squares of errors

SSB = sum of squares between groups

SSE = sum of squares of errors

$\bar{X} j - \bar{X}$ = mean of the jth group,

X- $\bar{X}$ j = overall mean, and nj is the sample size of the jth group.X

= each data point in the jth group (individual observation)

N = total number of observations/total sample size,

and SST = Total sum of squares = SSB + SSE

If the value of F is near about 1, then there is insignificant variance between the means of the two groups of data set under observation.

- **One-way ANOVA**

It is used when we have only one factor (independent variable) with two or more levels (samples).

Example: Scenario: Comparison of Exam Scores Among Multiple Teaching Methods

DataCrats

Suppose you're a researcher interested in comparing the effectiveness of three different teaching methods (Traditional Lecture, Interactive Learning,and Online Modules) on student exam scores.

- **Two-way ANOVA**

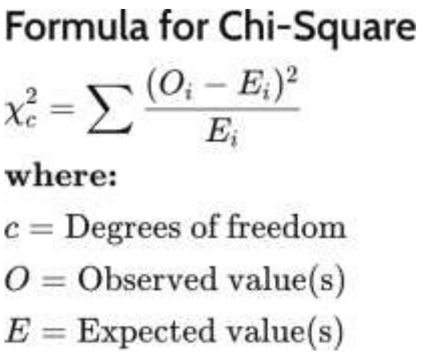
It is used when we have two or more independent variables. Each of these factors can have multiple levels.

Example: Scenario: Influence of Diet and Exercise on Weight Loss

Imagine you're studying the effects of diet type (Low Carb vs. Low Fat) and exercise intensity (High vs. Moderate) on weight loss in individuals. You want to investigate whether there are significant main effects of diet type and exercise intensity, as well as any interaction effect between the two factors.

4. Chi-Square Test

The Chi-Square test is a statistical method used to assess the association or independence between categorical variables. It is particularly useful when dealing with nominal or ordinal data, where data points fall into categories rather than numerical values. There are different variations of the Chi-Square test, such as the Chi-Square Test for Independence and the Chi-Square Goodness-of-Fit Test.



Formula for Chi-Square

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

c = Degrees of freedom

O = Observed value(s)

E = Expected value(s)

- **Chi-Square Test for Independence**

You can use a chi-square test of independence when you have two categorical variables. It allows you to test whether the two variables are related to each other. If two variables are independent (unrelated), the probability of belonging to a certain group of one variable isn’t affected by the other variable.

DataCrafts

Example: Scenario: Relationship Between Smoking Habits and Lung Cancer

Suppose you're investigating the relationship between smoking habits and the occurrence of lung cancer. You want to determine if there is a significant association between smoking status (Smoker vs. Non-Smoker) and the presence of lung cancer (Diagnosed with lung cancer vs. Not diagnosed with lung cancer).

- **Chi-Square Goodness of Fit Test**

You can use a chi-square goodness of fit test when you have one categorical variable. It allows you to test whether the frequency distribution of the categorical variable is significantly different from your expectations.

Example: Scenario: Distribution of Blood Types in a Population

Suppose you're interested in determining whether the distribution of blood types in a population follows the expected distribution based on national statistics.