atomcamp

Statistics

OVERVIEW & PURPOSE

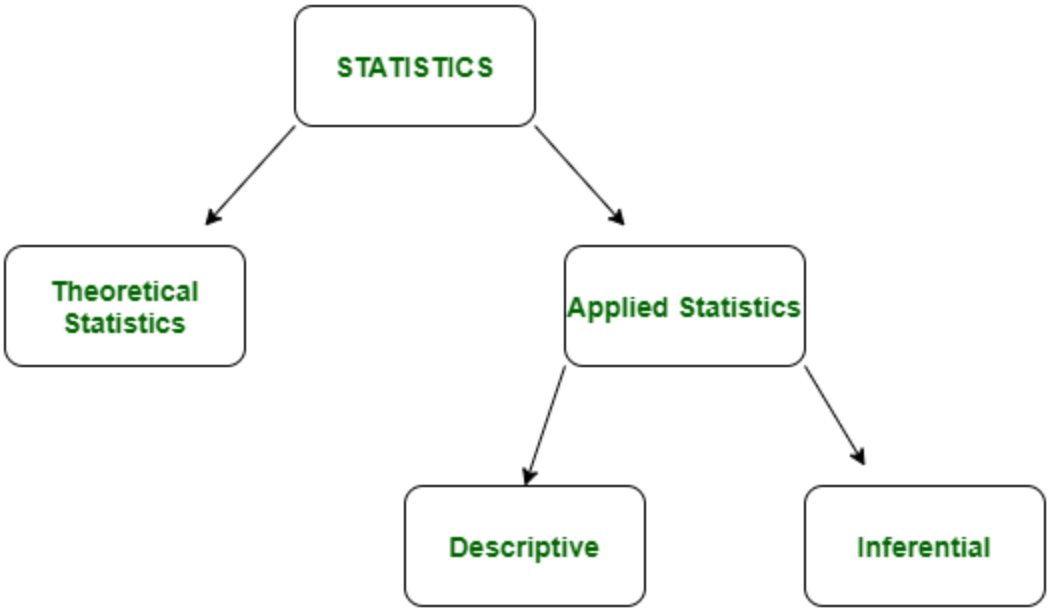
In this session, participants will have an Introduction to Statistics for data sciences

OBJECTIVE

- **Introduction to Statistics**
- **Data Types and Levels of Measurement**
- **Descriptive Statistics**
- **Probability Basics**
- **Sampling and Sampling Distributions**
- **Hypothesis Testing**
- **Correlation and Regression**
- **Introduction to Advanced Topics (Overview)**

Introduction to Statistics and Its Importance in Data Science

What is statistics?



Description: Statistics is the scientific discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.
Example: Using past sales data to forecast future sales trends.
Importance: Facilitates evidence-based decision-making by translating raw data into meaningful insights.

Role of statistics in data science and everyday life.
Description: In data science, statistics underpins models, predictions, and analyses. In everyday life, it helps in making informed decisions.
Example: Supermarkets using statistics to determine stock levels based on past sales.
Importance: Statistics bridges raw data with actionable conclusions, both in professional settings and everyday scenarios.

Descriptive vs. inferential statistics.

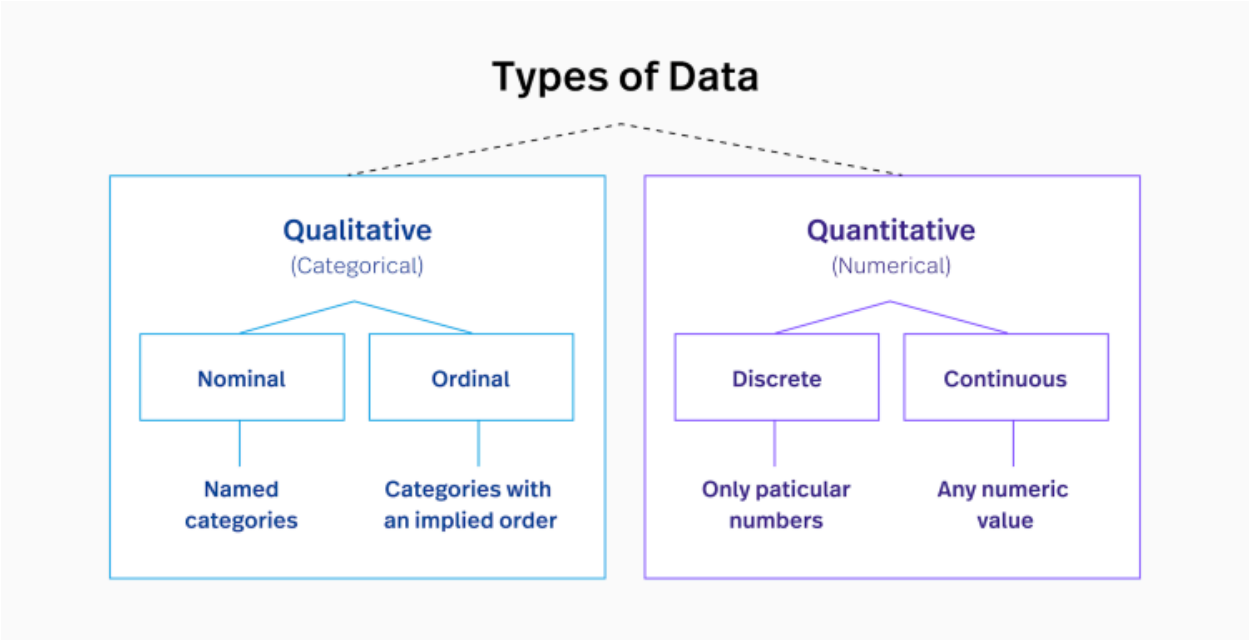
S. No	Descriptive Statistics	Inferential Statistics
1	Concerned with the describing the target population	Make inferences from the sample and generalize them to the population.
2	Organize, analyze and present the data in a meaningful manner	Compares, test and predicts future outcomes.
3	Final results are shown in form of charts, tables and Graphs	Final result is the probability scores.
4	Describes the data which is already known	Tries to make conclusions about the population that is beyond the data available.
5	Tools- Measures of central tendency (mean/median/ mode), Spread of data (range, standard deviation etc.)	Tools- hypothesis tests, Analysis of variance etc.

Description: Descriptive statistics provide a summary of the main aspects of data, while inferential statistics allow for conclusions or predictions based on data.
Example: Descriptive: Finding the average height of students in a class. Inferential: Predicting the average height of all students in a country based on a sample.
Importance: Descriptive stats provide a snapshot of data, while inferential stats provide the basis for prediction and decision-making.

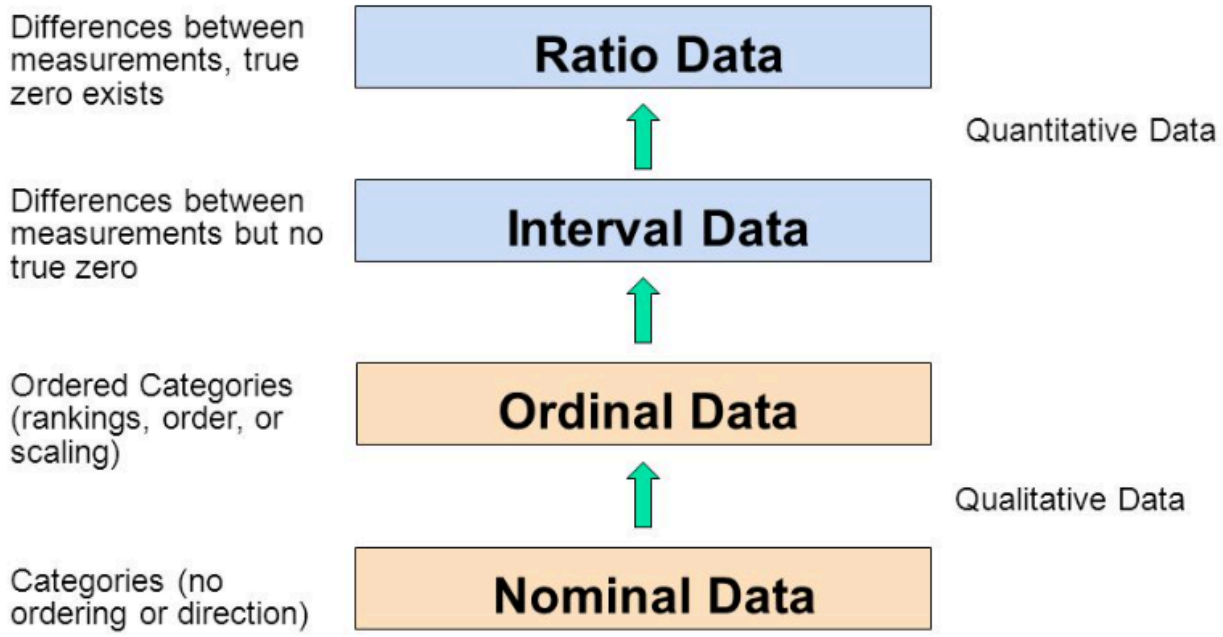
Data Types and Levels of Measurement

Types of data: qualitative (categorical) vs. quantitative (numerical).
Description: Qualitative data is non-numerical and often textual, while quantitative data is numerical.
Example: Qualitative: Colors of shirts. Quantitative: Prices of shirts.
Importance: The type of data dictates the kind of analysis and visualization techniques used.

atomcamp



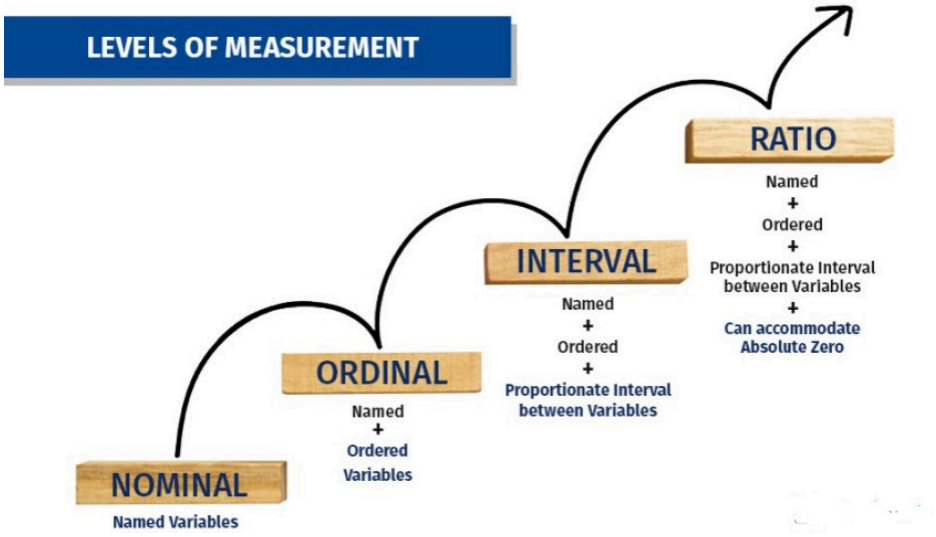
Levels of measurement: nominal, ordinal, interval, and ratio.



Description: Nominal data is categorical without order; ordinal has order; interval has order and equal differences but no true zero; ratio data has a true zero.

Example: Nominal: Brands of cars. Ordinal: Economic classes (lower, middle, upper). Interval: Temperature in Celsius. Ratio: Bank balance.

Importance: The level of measurement determines which statistical tests and measures are appropriate.



Categorical- ordinal/ nominal

Numerical- discrete/ continuous

Age Group (): Categories like 'Child', 'Teen', 'Adult', 'Senior'.

Temperature in Celsius (): Decimal values representing temperature.

Customer Satisfaction Rating (): Ranks like 'Very Unsatisfied' to 'Very Satisfied'.

Number of Children in Family (): Integer values like 0, 1, 2, 3, etc.

Types of Cuisine (): Categories such as 'Italian', 'Chinese', 'Mexican'.

Height in Centimeters (: Decimal values representing height.

Income Level (): Categories like 'Low', 'Medium', 'High'.

Frequency of Exercise per Week (: Integer values like 0, 1, 2, etc.

Color Preference (): Choices such as 'Red', 'Blue', 'Green'.

Time Spent on Homework Daily (): Decimal values in hours.

Educational Level (l): Ranks like 'High School', 'Bachelor's', 'Master's'.

Number of Pets Owned (: Integer values like 0, 1, 2, etc.

Favorite Movie Genre (): Categories like 'Action', 'Romance', 'Sci-Fi'.

Body Mass Index (BMI) (: Decimal values representing BMI.

Satisfaction with Public Transport (): Ratings like 'Poor', 'Fair', 'Good', 'Excellent'.

Number of Books Read in a Month (): Integer values like 0, 1, 2, etc.

Type of Housing (): Categories like 'Apartment', 'House', 'Condo'.

Price category of Type of Housing (): Categories like 'Apartment-low', 'House'-high, 'Condo-medium'.-ordinal category

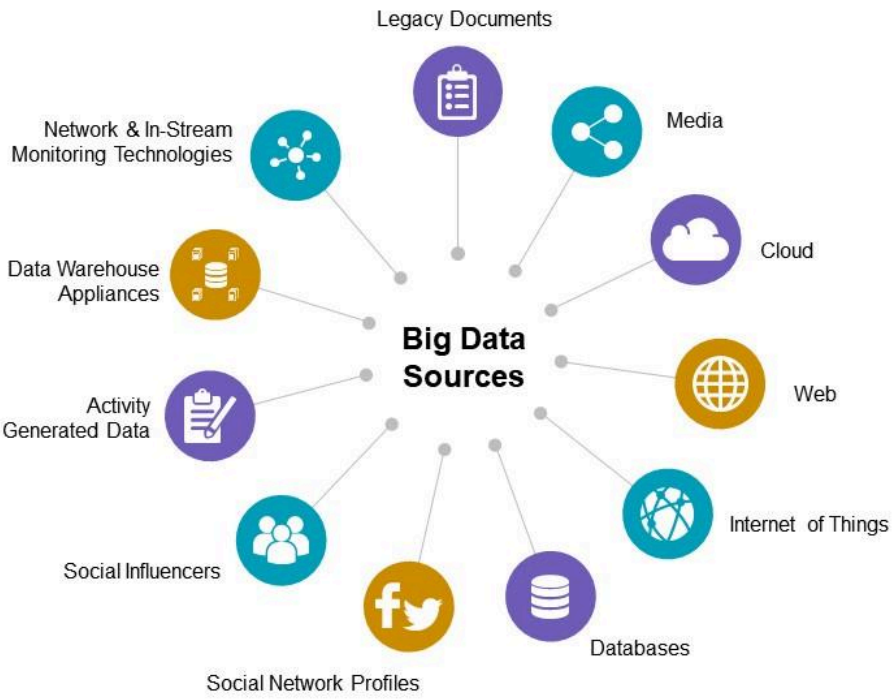
Price of Type of Housing (): Categories like 'Apartment', 'House', 'Condo'.-numerical continuous

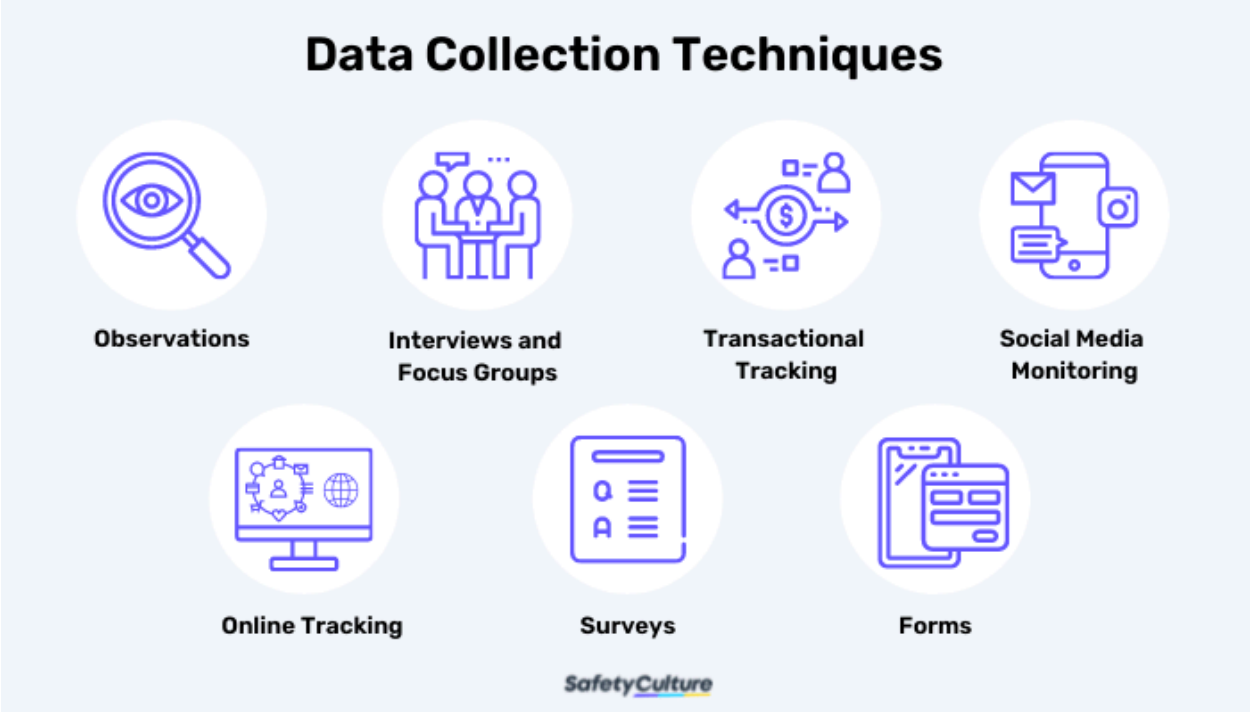
Blood Pressure Reading (): Decimal values representing blood pressure.

Level of Agreement with a Statement : Scales like 'Strongly Disagree', 'Neutral', 'Strongly Agree'.

Number of Countries Visited :Integer values like 0, 1, 5, 10,

Introduction to data sources and data collection methods.

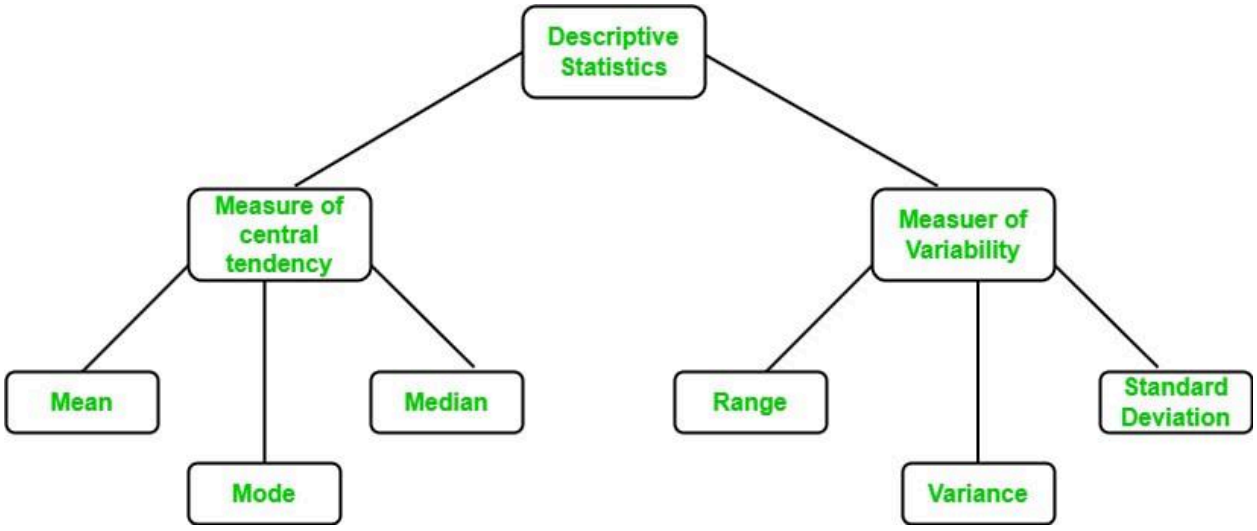




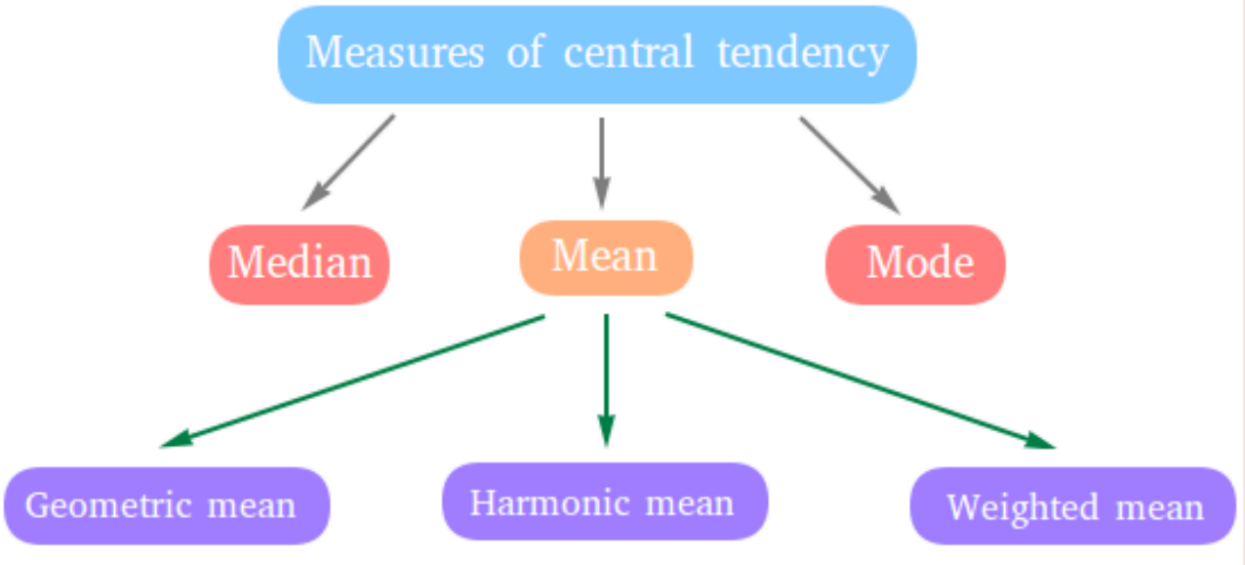
Description: The origins of data and the techniques used to gather it. Example: Surveys, experiments, online databases.

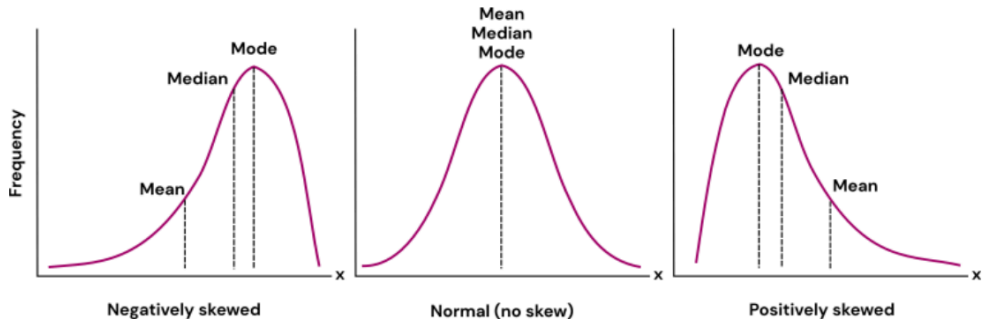
Importance: Quality and relevance of conclusions heavily rely on data quality and collection methods.

Descriptive Statistics



Measures of central tendency: mean, median, mode.





Description: Statistical measures that identify the center or typical value of a dataset.
Example: In a dataset of ages (23, 25, 25, 26, 29), the mean is 25.6, the median is 25, and the mode is 25.
Importance: Gives a summary view of data, aiding in quick interpretation.

Questions:

atomcamp

1. **Mean Calculation**: Given the data set $3, 7, 7, 15, 18, 20, 23, 23, 25, 28$, calculate the mean.
2. **Mode Identification**: Find the mode of the following data set: $12, 15, 12, 18, 19, 12, 17, 18, 15$.
3. **Median Determination**: Determine the median of this set of numbers: $5, 3, 8, 9, 4, 6, 7, 10$.
4. **Mixed Calculation**: For the data set $8, 22, 17, 6, 13, 19, 24, 17, 17$, calculate the mean, median, and mode.
5. **Even Numbered Set Median**: Calculate the median for the following even-numbered data set: $14, 6, 10, 8, 12, 16$.
6. **Odd Numbered Set Mean**: Find the mean of this odd-numbered data set: $11, 14, 9, 5, 7$.
7. **No Mode Scenario**: Identify the mode in the data set $21, 22, 23, 24, 25$, and explain why it might be a special case.
8. **Large Data Set Mean**: Calculate the mean of this large data set: $2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47$.
9. **Skewed Data Median and Mode**: For the skewed data set $33, 34, 34, 35, 36, 40, 45, 50, 55, 60$, calculate the median and mode. Discuss how skewness might affect these measures.
10. **Real-World Application**: A teacher recorded the number of books read by students in a month: $5, 3, 7, 5, 2, 5, 9, 4, 6, 5$. Calculate the mean, median, and mode. Discuss which measure best represents the central tendency of this data.

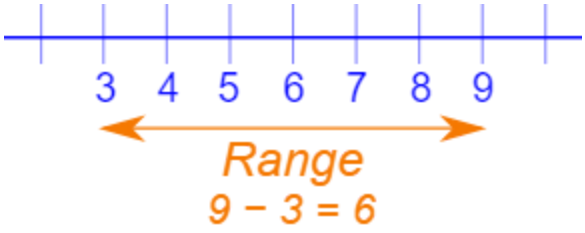
Measures of variability: range, variance, standard deviation.

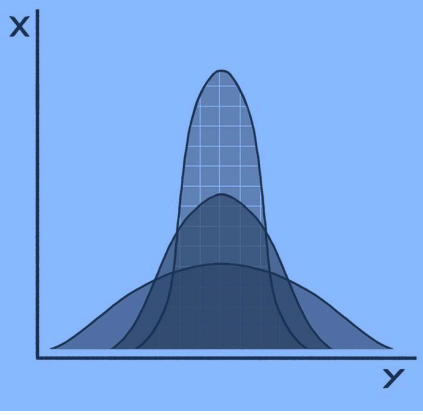
Description: Metrics that describe the spread or dispersion of data points in a dataset.

Example: For the ages above, range = 29-23 = 6, variance = 5.3, and standard deviation ≈ 2.3.

Importance: Understand the spread and consistency in the data which can aid in understanding its reliability.

atomcamp





Variance

['ver-ē-ən(t)s]

A measurement of how far each number in a data set is from the mean (average), and thus from every other number in the set.

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

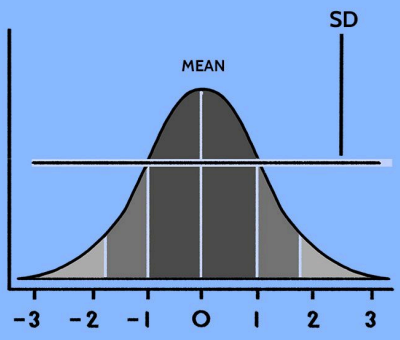
S^2 = Variance

n = The Number of data Point

X_i = Each of the values of the data

$\bar{X}$ = The Mean of X_i

atomcamp

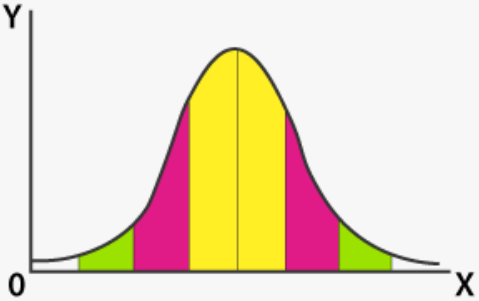


Standard Deviation

['stan-dərd dē-vē-'ā-shən]

A statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.

Standard Deviation Formula

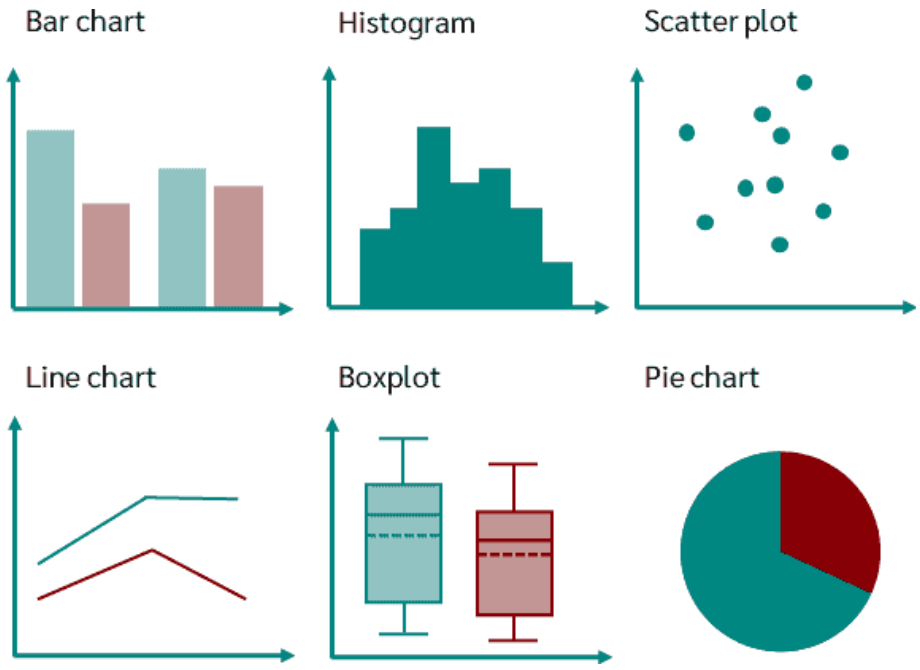


$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N}}$$

atomcamp

1. **Range Calculation**: Find the range for the following data set: $8, 12, 5, 19, 22, 7, 14$.
2. **Basic Variance Computation**: Compute the variance for the set of numbers: $4, 8, 6, 10, 2$.
3. **Standard Deviation for a Small Set**: Calculate the standard deviation for the data set: $15, 9, 12, 17, 13$.
4. **Combined Range and Variance Question**: For the data set $3, 7, 7, 15, 18, 20$, calculate both the range and the variance.
5. **Large Data Set Standard Deviation**: Find the standard deviation for these numbers: $10, 20, 30, 40, 50, 60, 70, 80$.
6. **Variance in a Real-World Context**: A teacher recorded the scores of students on a test as follows: $72, 88, 95, 70, 80, 85, 90$. Calculate the variance of these scores.
7. **Comparing Ranges**: Compare the range of the following two data sets: Set A - $5, 15, 25, 35, 45$ and Set B - $10, 20, 30, 40, 50$. Discuss any observations.
8. **Standard Deviation and Data Spread**: For the data set $12, 15, 12, 18, 19, 12, 17, 18, 15$, calculate the standard deviation and discuss what it tells you about the spread of the data.
9. **Variance in Even and Odd Sets**: Calculate the variance for each of these sets and compare: Set 1 (odd number of data points) - $4, 8, 12$, Set 2 (even number of data points) - $3, 7, 11, 15$.
10. **Applying Standard Deviation in a Practical Scenario**: A researcher is studying the heights (in cm) of a particular plant specie[illegible]he recorded heights are: $40, 42, 38, 45, 50, 37, 39$. Calculate the standard deviation and explain how it might

Visualization: histograms, box plots, and scatter plots.



Description: Graphical representations of data. Histograms show frequency distributions, box plots depict data spread and outliers, scatter plots show relationships between two variables.
Example: Using a scatter plot to visualize the relationship between age and salary.
Importance: Visual aids for understanding and interpreting complex data sets.

Probability Basics

Introduction to probability: what it is and why it matters.
Description: Probability measures the likelihood of an event occurring, ranging from 0 (impossible) to 1 (certain).
Example: The probability of flipping heads on a fair coin is 0.5.
Importance: Provides the foundational concepts behind statistical inference and predictions.

What is the probability of rolling a 4 on a fair six-sided die?
If you flip a fair coin three times, what is the probability of getting exactly two heads?
A deck of cards contains 52 cards. What is the probability of drawing a red card (either a heart or a diamond)?
If you have a bag with 5 red balls and 3 blue balls, what is the probability of drawing a red ball without replacement?
If you spin a spinner divided into 8 equal parts, what is the probability of landing on an even number?
If you have a jar with 10 marbles, 3 of which are green and 7 are yellow, what is the probability of randomly selecting a green marble?
In a class of 30 students, 15 are male and 15 are female. What is the probability of randomly selecting a female student?
A box contains 8 red balls, 6 green balls, and 4 blue balls. What is the probability of drawing a blue ball if you pick one ball randomly?

atomcamp

If you roll two fair six-sided dice, what is the probability that the sum of the two dice is 7?
You have a bag with 10 chocolates, 3 of which are dark chocolate and 7 are milk chocolate. What is the probability of drawing two dark chocolates in a row without replacement?

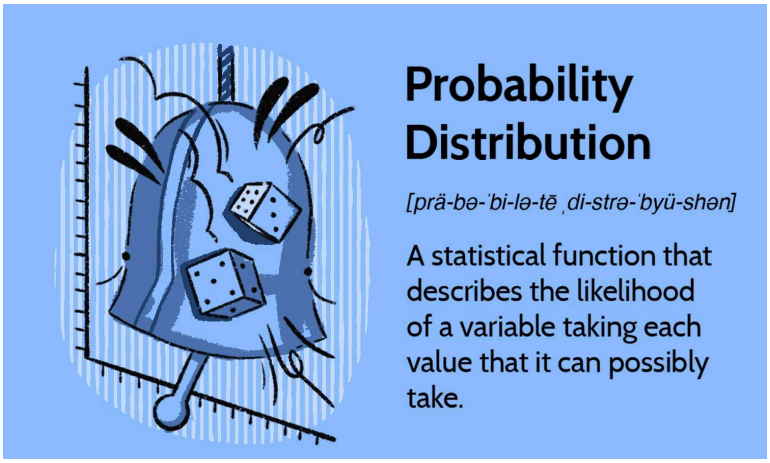
Basic rules of probability.

Description: Principles governing the computation of probabilities in various situations, including the addition rule and multiplication rule.
Example: In a deck of cards, the probability of drawing an Ace and then a King without replacement is (4/52) * (4/51).
Importance: Enables the calculation of likelihoods in complex scenarios.

1. **Coin Toss**: What is the probability of getting heads when flipping a fair coin?
2. **Dice Roll**: If you roll a fair six-sided die, what is the probability of rolling a number greater than 4?
3. **Card Selection**: From a standard deck of 52 playing cards, what is the probability of drawing an Ace?
4. **Multiple Dice Rolls**: What is the probability of rolling two six-sided dice and getting a total of 7?
5. **Marbles in a Bag**: In a bag containing 5 red marbles, 3 blue marbles, and 2 green marbles, what is the probability of randomly picking a blue marble?
6. **Birthdays**: Assuming each day of the year is equally likely, what is the probability that a randomly selected person has their birthday on a Tuesday?
7. **Coin Toss Sequence**: What is the probability of flipping a fair coin three times and getting two heads and one tail, in any order?
8. **Drawing Cards**: If two cards are drawn from a standard deck of 52 playing cards without replacement, what is the probability that both are Kings?
9. **Even Number on a Die**: When rolling a fair six-sided die, what is the probability of rolling an even number?
10. **Colored Balls in a Bag**: A bag contains 4 red balls, 5 yellow balls, and 6 green balls. If one ball is drawn from the bag, what is the probability that it is neither red nor yellow?

These questions cover a range of basic pro ↓ ility concepts, including single events,

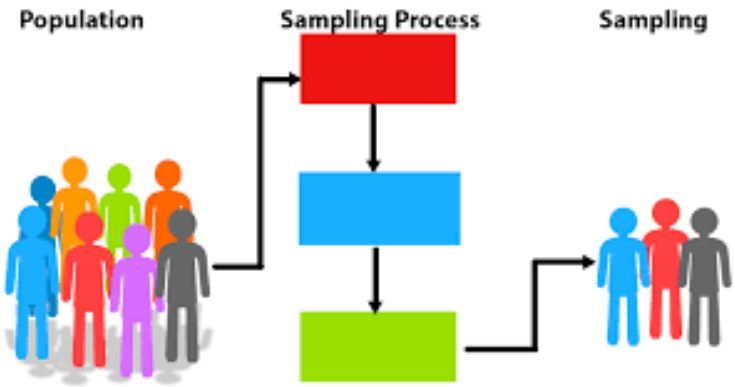
Probability distributions: uniform, normal, binomial, and others.



Description: Functions that list all possible values of a random variable and their corresponding probabilities.
Example: The number of heads when flipping a coin three times follows a binomial distribution.
Importance: Forms the basis for many statistical tests and models.

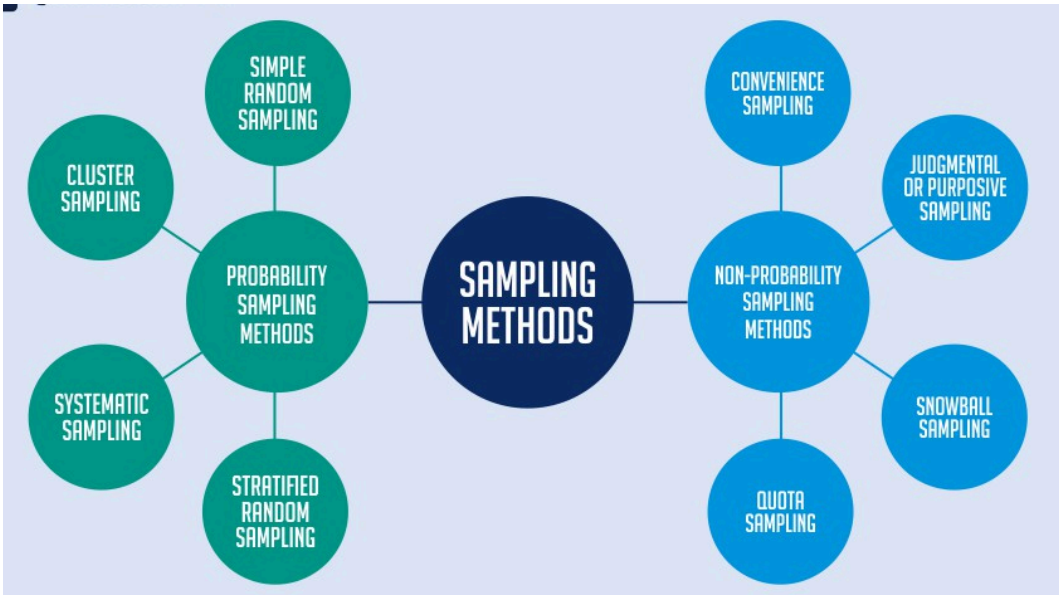
Sampling and Sampling Distributions

Concepts of population vs. sample.



Description: A population contains all members of a specified group, while a sample is a subset of the population.
Example: If we want to know the average height of adults in a city, the city's adult residents are the population, while a group of 1,000 randomly chosen city residents would be a sample.
Importance: Inference about a whole population is typically made from samples due to the impracticality of surveying an entire population.

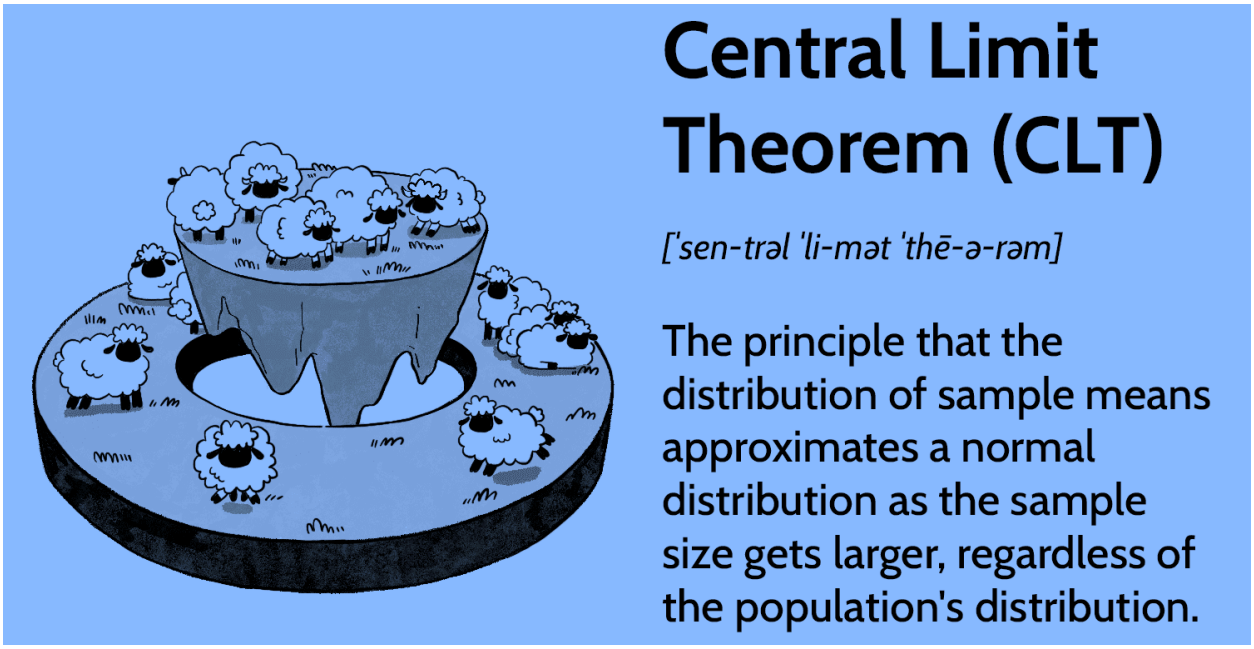
Sampling methods:



Description: Techniques for selecting a subset from a larger set.
Example: Stratified sampling might be used to ensure equal representation from various age groups in a study.

Importance: Ensures that samples are representative, which aids in drawing valid conclusions.

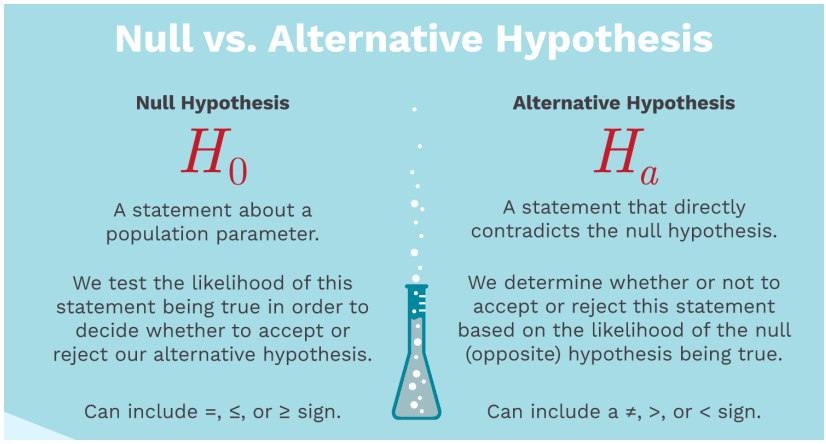
The Central Limit Theorem (CLT).



Description: States that, given a sufficiently large sample size, the sampling distribution of the mean will be approximately normally distributed, regardless of the population's distribution.
Example: Averaging dice rolls many times will result in a normal distribution, even though a die's roll is uniformly distributed.
Importance: CLT provides a foundation for many statistical methods, especially hypothesis testing.

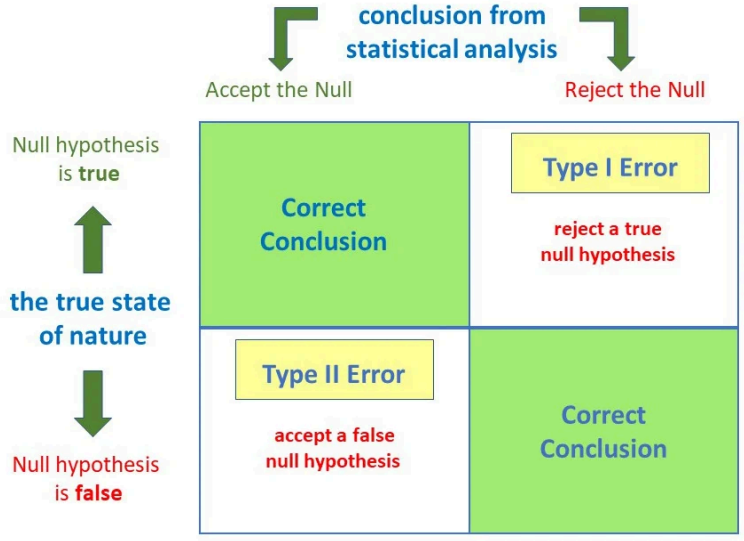
Hypothesis Testing

Null and alternative hypotheses.



Description: Initial claims about a population parameter (null) and what the test aims to prove (alternative).
Example: Testing a new drug, the null might be "The drug has no effect" vs. the alternative "The drug has an effect."
Importance: Sets the stage for statistical testing to validate or refute assumptions.

Types of errors (Type I and Type II).

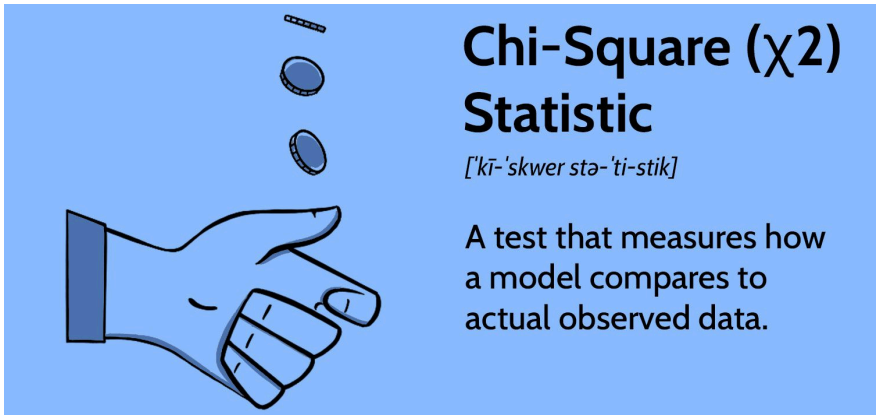


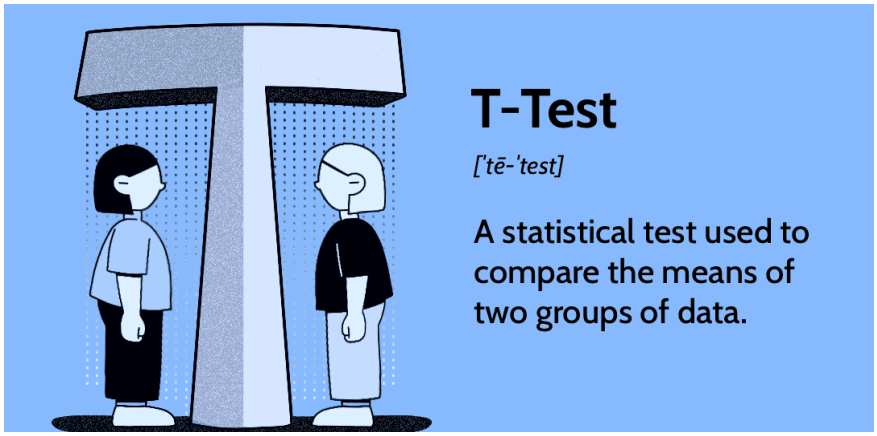
Description: Type I is a false positive, concluding something is true when it's not. Type II is a false negative, failing to detect a true effect.

Example: Claiming a drug works when it doesn't (Type I) vs. overlooking a drug's effect (Type II).
Importance: Understanding errors is crucial for the interpretation of test results.

Significance level, p-values, and confidence intervals.
Description: Significance level (alpha) is the threshold for considering results significant. P-value is the probability of observing results at least as extreme as the ones in the test, assuming the null hypothesis is true. Confidence intervals provide a range for parameter estimates.
Example: If a drug trial yields a p-value of 0.03, it's significant at an alpha of 0.05, suggesting the drug has an effect.
Importance: Aids in determining the validity and precision of results.

Basic tests like the t-test and chi-square test.

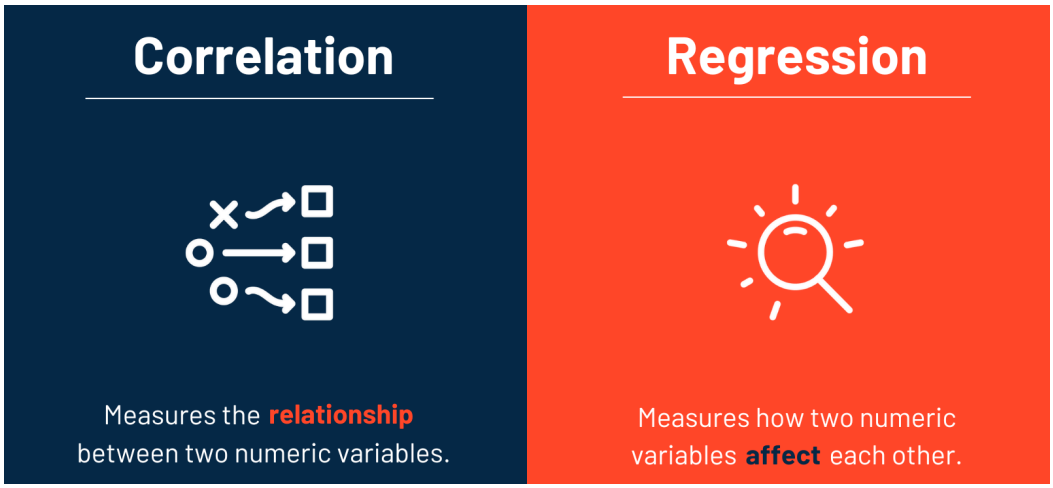


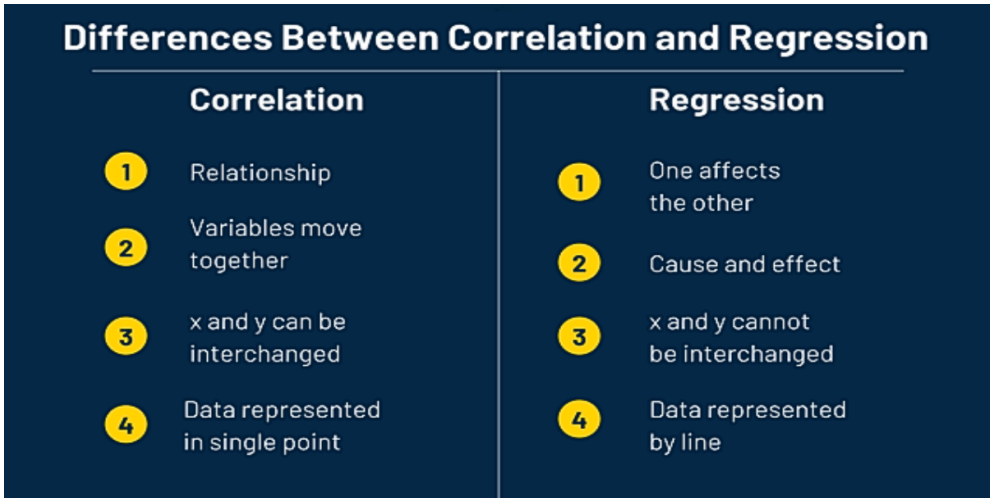


Description: The t-test compares means of two groups, while the chi-square tests relationships between categorical variables.
Example: Comparing the average test scores of two classrooms using a t-test.
Importance: Provide methodologies to test various data types and research questions.

Correlation and Regression

Understanding the relationship between two variables.

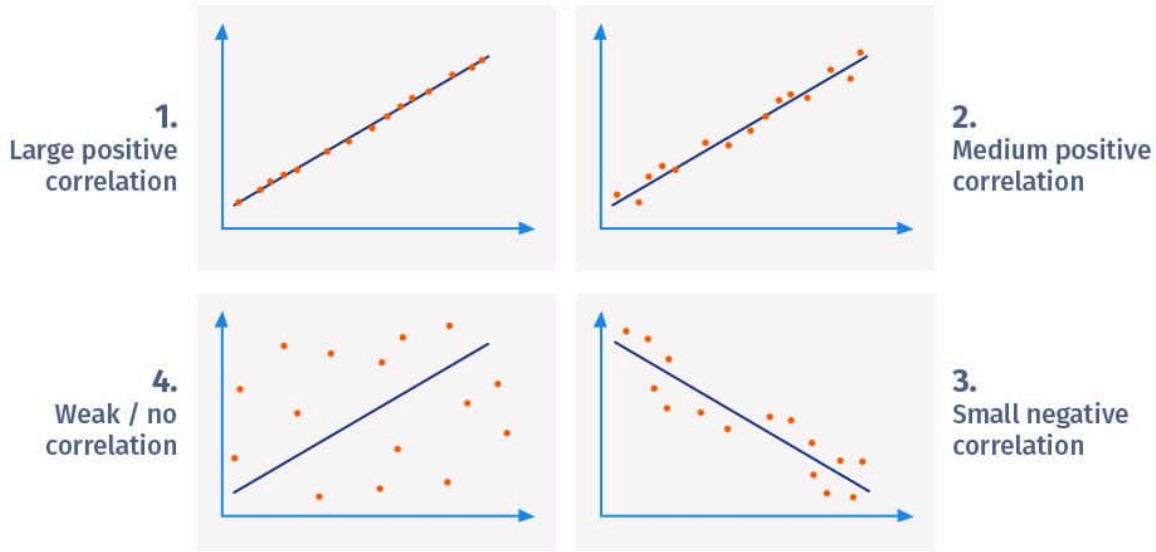




Description: Analyzing how two variables change together.
Example: Observing how sales change with advertising spend.
Importance: Determines associations and can hint at causation, guiding strategies.

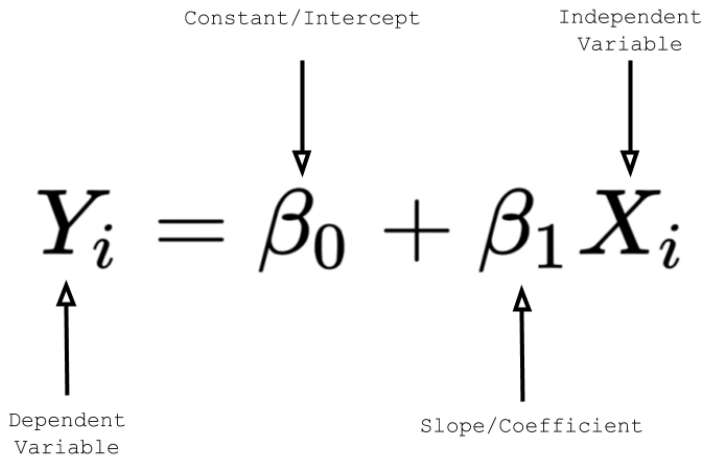
Pearson's correlation coefficient.

Description: A measure between -1 and 1 that represents the linear relationship between two variables.
Example: A coefficient of 0.9 might represent a strong positive relationship between hours studied and test scores.



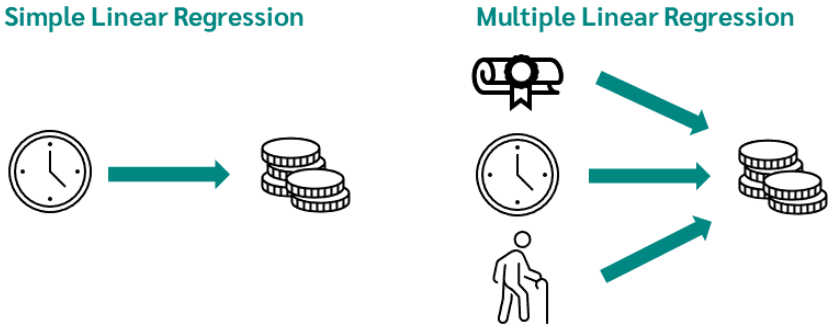
Importance: Quantifies the strength and direction of relationships.

Simple linear regression: interpreting slope and intercept, and understanding residuals.



Description: A method for modeling the relationship between a dependent and one independent variable. Slope indicates the rate of change, and intercept is the value when the independent variable is zero. Residuals are the differences between observed and predicted values.
Example: Predicting sales based on advertising spend.
Importance: Predictive modeling and understanding factors influencing outcomes.

Introduction to Advanced Topics (Overview)



Brief touch on multiple regression.

atomcamp

Description: Like simple regression, but models the relationship between one dependent and multiple independent variables.
Example: Predicting house prices based on square footage, location, and number of rooms.
Importance: Allows for more complex modeling by incorporating multiple factors.