Statistics and Its Applications

Statistics is a branch of mathematics that involves the collection, analysis, interpretation, presentation, and organization of data. It provides tools and methodologies for making informed **decisions based** on data.

Applications of Statistics:

1. **Machine Learning:** Statistics forms the backbone of machine learning algorithms, enabling data-driven predictions and insights.
2. **Healthcare:** Clinical trials and medical research rely on statistical methods to evaluate the effectiveness of treatments.
3. **Economics:** Econometrics uses statistics to analyze economic data and forecast trends.
4. **Marketing:** Companies analyze consumer data to optimize marketing strategies and improve customer satisfaction.
5. **Social Sciences:** Statistics is crucial for studying human behavior and societal trends.
6. **Quality Control:** Industries use statistical tools like Six Sigma to ensure product quality.

Real-Life Use Cases:

1. **Predicting Customer Churn in Businesses:**
    - Companies analyze past purchase data, support interactions, and usage trends to predict which customers might leave their services.
    - Example: Using logistic regression to identify high-risk customers based on features such as *"frequency of complaints"* and *"time since last purchase"*.
2. **Weather Forecasting:**
    - Meteorologists use historical weather data and statistical models to predict future weather conditions.
    - Example: Analyzing patterns of temperature, pressure, and humidity to forecast rainfall.
3. **Sports Analytics:**
    - Teams leverage player performance statistics to make strategic decisions and improve game outcomes.
    - Example: Evaluating a baseball player's batting average to decide the batting order.
4. **Public Health Studies:**
    - Governments track disease spread using statistical models to implement timely interventions.
    - Example: Estimating the reproduction number (R0) of a virus to predict the scale of an outbreak.
5. **Stock Market Analysis:**

- Investors use time-series data and statistical indicators to forecast stock prices and assess market trends.
- Example: Applying moving averages to detect potential buy or sell signals.

Types of Statistics

Statistics is broadly divided into two categories:

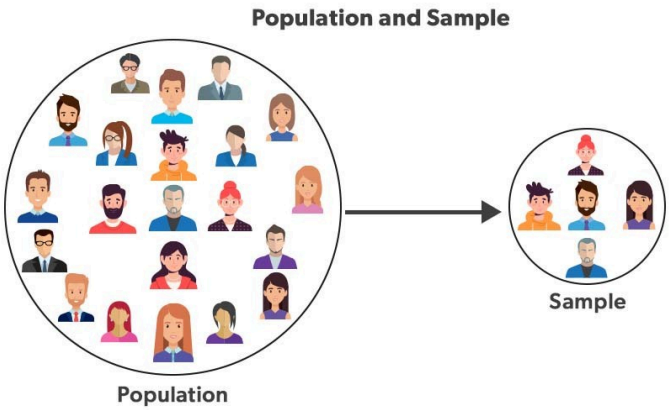
1. **Descriptive Statistics:**
   - Focuses on **summarizing** and organizing data.
   - Key tools:
     - **Measures of Central Tendency:** Mean, median, mode.
     - **Measures of Dispersion:** Range, variance, standard deviation.
     - **Visualization Tools:** Histograms, bar charts, pie charts.
2. **Example:**
   - A dataset of students' scores: [75, 85, 90, 95, 100].
   - Mean = (75 + 85 + 90 + 95 + 100) / 5 = 89.
   - Range = 100 - 75 = 25.
3. **Inferential Statistics:**
   - Uses a **sample** of data to make **inferences** about a larger population.
   - Sample -> inferences (Population).
   - Key concepts:
     - Hypothesis testing.
     - Confidence intervals.
     - Regression analysis.
4. **Example:**
   - Testing the effectiveness of a new drug using a sample of 100 patients to infer results for the entire population.

Population vs. Sample Data

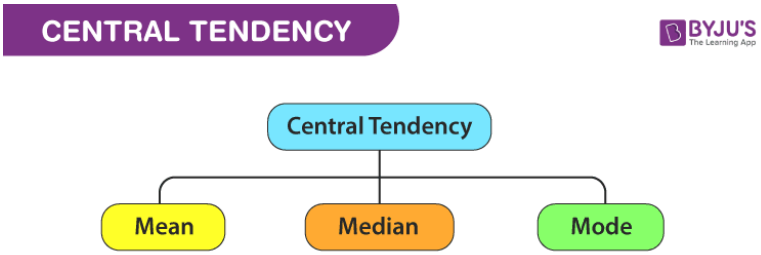
1. **Population:**
   - Refers to the entire group of individuals or observations of interest.
   - Denoted by capital letters (e.g. N).
2. **Example:** The height of all adults in a country.
3. **Sample:**
   - A subset of the population used for analysis.
   - Denoted by lowercase letters (e.g. n).
4. **Example:** The height of 1,000 adults selected randomly from the country.

Why Use Samples?

- Collecting data from the entire population can be impractical or impossible.
- Sampling saves time and resources while providing reliable insights.



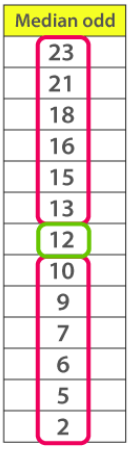
Measure of Central Tendency

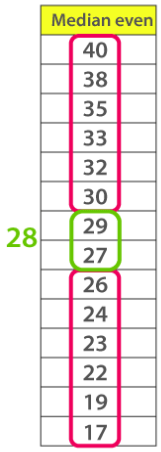


1. **Mean (Arithmetic Average):**
   - Formula:

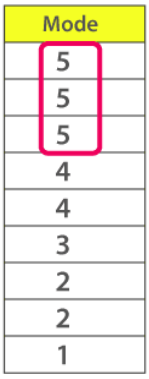
Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

   -
   - Example: Scores [10, 20, 30], Mean = 20.
2. **Median:**
   - The middle value in an ordered dataset.
   - Example: [10, 20, 30] → Median = 20.





3. **Mode:**
    - The most frequently occurring value(s) in the dataset.
    - Example: [10, 10, 20, 30] → Mode = 10.



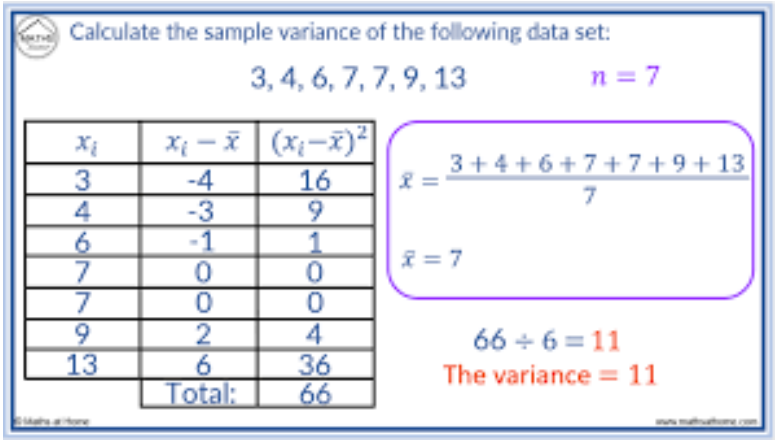
Measure of Dispersion

Dispersion measures the spread or variability in a dataset.

1. **Range:**
    - Formula: It is simply the difference between the maximum value and the minimum value given in a data set.
    - Example: 1,3,5,6,7 => Range = 7 -1= 6
    - Example: [10, 20, 30] → Range = 30 - 10 = 20.
2. **Variance:**
    - Formula (Population): Deduct the mean from each data in the set, square each of them and add each square and finally divide them by the total no of values in the data set to get the variance.

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element $\bar{x}$ = sample mean n = sample size

    - **σ²** (sigma squared) for the population variance.
    - **s²** for the sample variance.
    - Measures the **spread** or **dispersion** of data points around the mean.
    - A high variance indicates data points are spread out, while a low variance shows they are clustered closely around the mean.



x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
3	-4	16
4	-3	9
6	-1	1
7	0	0
7	0	0
9	2	4
13	6	36
	Total:	66

3. **Standard Deviation (or):**
    - Formula:

Standard Deviation Formula

cuemath
THE MATH EXPERT

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

Sample Variance: Why Divide by n-1?

Dividing by n−1 when calculating **sample variance** (instead of n) is done to account for **bias** in the estimation of the population variance from a sample. This adjustment, called **Bessel's correction**, ensures that the sample variance is an **unbiased estimator** of the population variance.

Key Concepts

1. **Population Variance Formula**:

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

   Here, μ is the population mean, and N is the population size.

2. **Sample Variance Formula**:

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

   Here, $\bar{x}$ is the sample mean, and n is the sample size.

3. When using a sample, the mean $\bar{x}$ is used instead of the true population mean μ. Because $\bar{x}$ is calculated from the sample, it tends to underestimate variability. Dividing by $n - 1$ corrects this bias.

Example: Comparing Dividing by n and $n-1$

Suppose we have a small population: $[3, 5, 7]$. The true population variance is:

$$\sigma^2 = \frac{1}{N}\sum(x_i - \mu)^2 = \frac{1}{3}[(3-5)^2 + (5-5)^2 + (7-5)^2] = \frac{1}{3}[4+0+4] = 2$$

Now, consider a random sample of size 2: $[3, 5]$.

Step 1: Sample Mean

$$\bar{x} = \frac{3+5}{2} = 4$$

Step 2: Variance Calculation

- If we divide by $n = 2$:

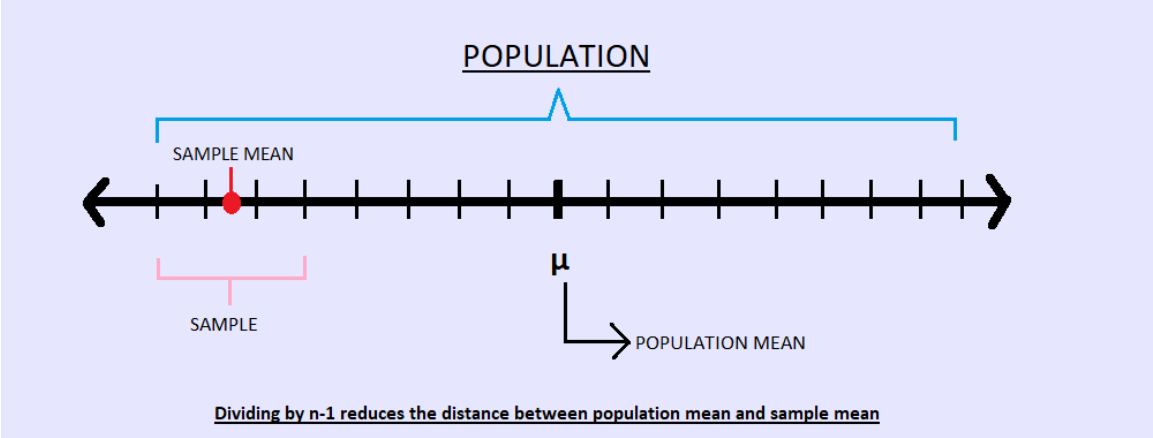
$$s^2 = \frac{1}{2}\sum(x_i - \bar{x})^2 = \frac{1}{2}[(3-4)^2 + (5-4)^2] = \frac{1}{2}[1+1] = 1$$

- If we divide by $n - 1 = 1$:

$$s^2 = \frac{1}{1}\sum(x_i - \bar{x})^2 = \frac{1}{1}[1+1] = 2$$

Step 3: Compare with Population Variance

The sample variance using $n-1$ matches the population variance (2), making it an **unbiased estimator**. Dividing by n underestimates the variance.



Standard deviation (SD)

The **standard deviation (SD)** is a statistical measure that quantifies the amount of variation or dispersion in a dataset. It tells you how much individual data points deviate, on average, from the mean (average) of the dataset.

Key Points

1. **Definition**:
   Standard deviation is the square root of the variance.
2. **σ** (lowercase sigma) for the population standard deviation.
3. **s** for the sample standard deviation.
4. $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \quad \text{(Population SD)}$$
   $$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{(Sample SD)}$$
5. **Interpreting Standard Deviation**:
   a. **Low SD**: Data points are close to the mean (less variability).
   b. **High SD**: Data points are spread out from the mean (more variability).
6. **Units**:
   a. Standard deviation is expressed in the same units as the original data, unlike variance, which is in squared units.
7. **Why Important**:
   a. It helps measure the consistency of data.
   b. It is widely used in finance, science, quality control, and more to understand data variability.

Variables

A **variable** is a characteristic or property that can take on different values. It can be measured, observed, or classified.

Types of Variables

a. Quantitative Variables (Numerical)

- Represent numerical values that can be measured or counted.
- **Subtypes**:
  - **Continuous**: Can take any value within a range (e.g., height, weight, temperature).
  - **Discrete**: Can take only specific values, often integers (e.g., number of cars, students in a class).

b. Qualitative Variables (Categorical)

- Represent categories or groups.
- Binary (two categories, e.g., Yes/No) or multi-level (e.g., Red, Blue, Green).
- **Subtypes**:
    - **Nominal**: Categories with no intrinsic order (e.g., colors, gender, types of fruits).
    - **Ordinal**: Categories with a meaningful order, but the intervals are not uniform (e.g., education levels, satisfaction ratings like poor, average, excellent).

Key Differences

Aspect	Quantitative	Qualitative
Definition	Represents numeric values	Represents categories or labels
Subtypes	Continuous, Discrete	Nominal, Ordinal
Examples	Age, salary, temperature	Gender, country, education level
Operations	Arithmetic operations (e.g., sum)	Counting, categorization

Understanding variable types helps in:

- **Choosing the right analysis techniques** (e.g., mean for quantitative, mode for categorical).
- **Data visualization** (e.g., scatter plots for quantitative, bar charts for categorical).
- **Statistical testing** (e.g., t-test for quantitative, chi-square for categorical).

Random Variable Basics

A **random variable** is a variable that takes on different values based on the outcome of a random process. It is fundamental in probability and statistics to describe numerical outcomes of random phenomena.

Types of Random Variables

a. Discrete Random Variable

- Takes on a **countable** number of distinct values (e.g., integers).
- Examples:
    - The number of heads in 3 coin flips.
    - The number of students in a class.

b. Continuous Random Variable

- Takes on an **uncountable** number of values within a range (e.g., real numbers).
- Examples:
    - The time it takes to complete a task.
    - The height of a person.

Histograms Basics

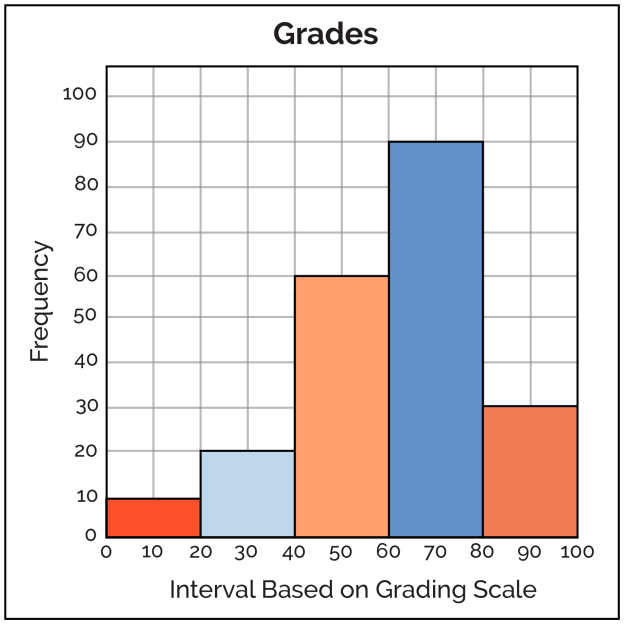
A **histogram** is a graphical representation of the distribution of a dataset. It displays the frequency (or count) of data points within specified intervals, called **bins**.

Key Characteristics

1. **Purpose**:
    - To visualize the distribution, spread, and shape of numerical data.
    - To identify patterns, such as skewness, peaks, or outliers.
2. **Structure**:
    - **Bins (or Intervals)**: Represent the range of values. Each bin has a width.
    - **Frequency**: Height of the bar represents the count (or proportion) of data points within each bin.

How to Create a Histogram

1. **Collect Data**: Gather your numerical dataset.
2. **Choose Bin Ranges**: Divide the data range into intervals (bins).
    - Example: If data ranges from 0 to 100, you might create bins like 0–10, 10–20, ..., 90–100.
3. **Count Frequencies**: Count how many data points fall into each bin.



Kernel Density Estimation (KDE)

Kernel Density Estimation (KDE) is a non-parametric way to estimate the probability density function (PDF) of a random variable. Unlike histograms, which group data into bins, KDE provides a smooth and continuous estimate of the distribution.

Percentage

- **Definition**: A way of expressing a number as a fraction of 100.
  Formula:

$$\text{Percentage} = \left(\frac{\text{Part}}{\text{Whole}}\right) \times 100$$

  Example: If you scored 75 out of 100 in a test, your percentage is:

$$\frac{75}{100} \times 100 = 75\%$$

-

Percentile

- **Definition**: A value below which a given percentage of data falls. It divides a dataset into 100 equal parts.

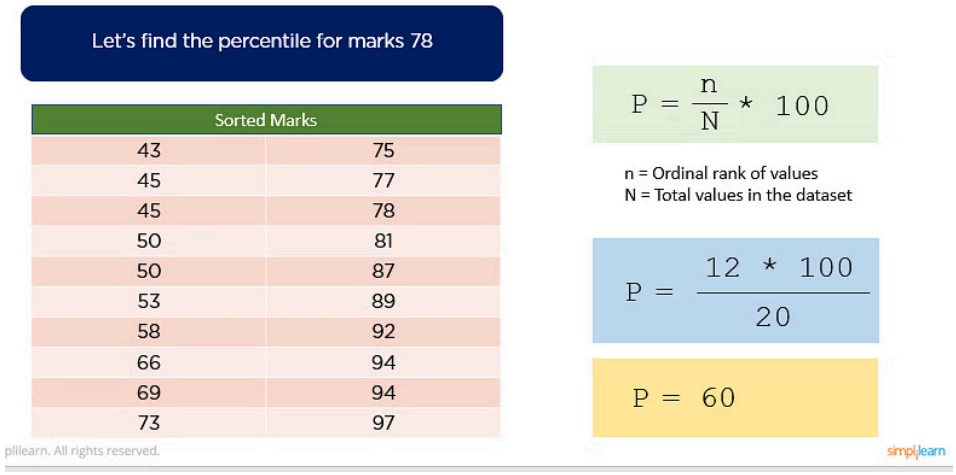
Percentile Formula

cuemath
THE MATH EXPERT

$$\text{Percentile} = \frac{\text{Number of Values Below "x"}}{\text{Total Number of Values}} \times 100$$

$$Value\ \# = \frac{Percentile}{100}(n+1)$$

$$Value\ \# = \frac{25}{100}(20+1) = 5.25$$

- **Example**:
  - In a class of 100 students, if your score is at the 90th percentile, it means you scored higher than 90% of the students.
  - 

- **Applications**:
  - Grading systems.
  - Analyzing distributions (e.g., income levels, test scores).

Quantile

- **Definition**: Points that divide a dataset into equal parts. It generalizes percentiles to divide data into any number of groups.
- **Types**:
  - **Quartiles**: Divide data into 4 equal parts.
    - Q1 (25th percentile): First quartile.
    - Q2 (50th percentile): Second quartile (Median).
    - Q3 (75th percentile): Third quartile.

    - **Deciles**: Divide data into 10 equal parts.
    - **Percentiles**: Divide data into 100 equal parts.
- **Example**:
  For a dataset: [10,20,30,40,50]
    - Q1 = 20 (25% of data is below this value).
    - Q2 = 30 (50% of data is below this value, i.e., the median).
    - Q3 = 40 (75% of data is below this value).
    - 

Aspect	Percentage	Percentile	Quantile
Definition	Fraction out of 100.	Value below which a percentage of data lies.	Points dividing data into equal groups.
Focus	Relative comparison.	Position in a dataset.	Partitioning data.
Usage	Test scores, statistics.	Ranking, statistical summaries.	Advanced data analysis.
Example	75% score on a test.	90th percentile = scored better than 90%.	Quartiles, deciles, percentiles.

Applications

1. **Percentage**: Comparing parts to the whole (e.g., grades, growth rates).
2. **Percentile**: Measuring rank or position in a group.
3. **Quantile**: Summarizing and analyzing distributions.

Five-Number Summary Basics

The **five-number summary** is a simple and effective way to summarize a dataset. It consists of five key statistics that provide insights into the distribution and spread of data.

The Five Numbers

1. **Minimum**: The smallest value in the dataset.
2. **First Quartile (Q1)**: The median of the lower half of the data (25th percentile).
3. **Median (Q2)**: The middle value in the dataset (50th percentile).
4. **Third Quartile (Q3)**: The median of the upper half of the data (75th percentile).
5. **Maximum**: The largest value in the dataset.

Steps to Calculate the Five-Number Summary

1. **Arrange the Data**: Sort the data in ascending order.
2. **Find the Minimum**: The smallest number.
3. **Find the Maximum**: The largest number.
4. **Find the Median (Q2)**:
    - If the number of data points is odd, the median is the middle value.
    - If the number of data points is even, the median is the average of the two middle values.
5. **Find the First Quartile (Q1)**: The median of the lower half (excluding the overall median if the dataset has an odd number of values).
6. **Find the Third Quartile (Q3)**: The median of the upper half (excluding the overall median if the dataset has an odd number of values).

Applications

1. **Descriptive Statistics**: Summarizing the distribution of data.
2. **Box Plots**: Visualizing the five-number summary.
3. **Data Analysis**: Identifying outliers, spread, and central tendency.

Example

Consider the following dataset:
4,7,10,15,16,20,23,24,25,304, 7, 10, 15, 16, 20, 23, 24, 25, 304,7,10,15,16,20,23,24,25,30

Step 1: Sort the data (already sorted):
4,7,10,15,16,20,23,24,25,304, 7, 10, 15, 16, 20, 23, 24, 25, 304,7,10,15,16,20,23,24,25,30

Step 2: Find the minimum and maximum:

- **Minimum** = 4
- **Maximum** = 30

Step 3: Find the median (Q2):

- There are 10 data points, so the median is the average of the 5th and 6th values: Median=16+20\2=18

Step 4: Find the first quartile (Q1) and third quartile (Q3):

- The lower half of the data (excluding the median): 4,7,10,15,164, 7, 10, 15, 164,7,10,15,16
    - The median of the lower half (Q1) is 10.
- The upper half of the data (excluding the median): 20,23,24,25,3020, 23, 24, 25, 3020,23,24,25,30
    - The median of the upper half (Q3) is 24.

Five-Number Summary:

- Minimum: 4
- Q1: 10
- Median (Q2): 18
- Q3: 24
- Maximum: 30

Interquartile Range (IQR)

The **Interquartile Range (IQR)** is a measure of statistical dispersion, or how spread out the values in a dataset are. It represents the range within which the middle 50% of the data lies.

How to Calculate the IQR

The IQR is calculated as the difference between the **third quartile (Q3)** and the **first quartile (Q1)**:

$$\mathbf{IQR = Q3 - Q1}$$

Where:

- **Q1 (First Quartile)**: The median of the lower half of the data (25th percentile).
- **Q3 (Third Quartile)**: The median of the upper half of the data (75th percentile).

Interpretation

- The **IQR** gives you a sense of how spread out the central 50% of your data is. A larger IQR indicates more variability, while a smaller IQR indicates that the data points are closer to the median.

In statistics, **outliers** are data points that significantly differ from the rest of the dataset. One way to detect outliers is by using **fences**—the higher and lower bounds beyond which values are considered outliers. These fences are determined using the **Interquartile Range (IQR)**.

1. Lower Fence and Higher Fence

- **Lower Fence**: The lower bound for identifying outliers, calculated as:
  Lower Fence=Q1−1.5×IQR
  Where Q1 is the first quartile and IQR is the interquartile range.
- **Higher Fence**: The upper bound for identifying outliers, calculated as:
  Higher Fence=Q3+1.5×IQR
- Where Q3 is the third quartile and IQR is the interquartile range.

2. Identifying Outliers

- **Outlier**: A data point is considered an outlier if it is either:
  - **Below the Lower Fence** (smaller than Q1−1.5×IQRQ1 - 1.5 \times \text{IQR}Q1−1.5×IQR).
  - **Above the Higher Fence** (larger than Q3+1.5×IQRQ3 + 1.5 \times \text{IQR}Q3+1.5×IQR).

Values falling outside these fences are considered unusually high or low relative to the rest of the data.

Applications of Fences and Outlier Detection

1. **Identifying Outliers**: Outliers are flagged using the lower and higher fences.
2. **Visualizing Outliers**: In a **box plot**, data points outside the fences are marked as individual dots, representing outliers.
3. **Robust Statistics**: Removing or adjusting outliers can improve the robustness of certain statistical analyses.

Covariance

Covariance is a statistical measure that indicates how two random variables change together. It helps determine whether an increase in one variable corresponds to an increase (or decrease) in another.

The covariance between two variables X and Y measures the **direction** of their linear relationship.

- **Positive Covariance**: Both variables increase or decrease together.
- **Negative Covariance**: One variable increases while the other decreases.

Formula

For two variables X and Y with n observations:

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) (Y_i - \bar{Y})$$

Where:

- X_i, Y_i: Data points for X and Y.
- $\bar{X}, \bar{Y}$: Mean values of X and Y.
- n: Number of data points.

Interpretation

- $\text{Cov}(X, Y) > 0$: Variables have a **positive relationship** (as one increases, the other increases).
- $\text{Cov}(X, Y) < 0$: Variables have a **negative relationship** (as one increases, the other decreases).
- $\text{Cov}(X, Y) = 0$: No linear relationship between the variables.

$$X = [1, 2, 3, 4, 5], \quad Y = [2, 4, 6, 8, 10]$$

1. **Step 1**: Calculate the means:

$$\bar{X} = \frac{1+2+3+4+5}{5} = 3, \quad \bar{Y} = \frac{2+4+6+8+10}{5} = 6$$

2. **Step 2**: Compute the deviations for each pair:

$$(X_i - \bar{X}) = [-2, -1, 0, 1, 2], \quad (Y_i - \bar{Y}) = [-4, -2, 0, 2, 4]$$

3. **Step 3**: Multiply the deviations:

$$(X_i - \bar{X})(Y_i - \bar{Y}) = [8, 2, 0, 2, 8]$$

4. **Step 4**: Calculate the average:

$$\text{Cov}(X, Y) = \frac{1}{5} \sum_{i=1}^{5} (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{5}(8 + 2 + 0 + 2 + 8) = 4$$

Covariance is in the units of X×Y which makes its magnitude difficult to interpret directly.

This dependence on units is why covariance is often standardized into **correlation**, which is unitless.

Applications

1. **Portfolio Theory**:
   - Used in finance to measure how two assets move together.
2. **Data Analysis**:
   - Helps in understanding relationships between variables.
   - Example: Temperature and ice cream sales (positive covariance).

Limitations

1. **Scale Dependence**:
   - Covariance is affected by the units of the variables, making it difficult to compare across datasets.
2. **No Strength**:
   - Covariance only indicates the direction of the relationship, not its strength.
3. **Only Linear Relationships**:
   - Covariance does not capture non-linear relationships.

Correlation

Correlation measures the degree to which two variables are related. It quantifies the strength and direction of the relationship between variables in a dataset.

Key Features of Correlation

1. **Strength of Relationship**:
    - Indicates how closely the two variables are related.
    - Values range between **-1** and **+1**.
2. **Direction of Relationship/Types of Correlation**:
    - **Positive Correlation** (r>0r > 0r>0): Both variables increase together.
        i. Example: Hours studied and exam scores.
    - **Negative Correlation** (r<0r < 0r<0): One variable increases as the other decreases.
        i. Example: Speed of a car and travel time for a fixed distance.
    - **No Correlation** (r=0r = 0r=0): Variables do not have a linear relationship.
        i. Example: Shoe size and intelligence.
3. **Types of Variables**:
    - Correlation applies to **quantitative variables** (numeric data).

Correlation Coefficient

The **correlation coefficient (r)** is a standardized measure of correlation.
It is calculated as:

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Where:

- $\text{Cov}(X, Y)$: Covariance between X and Y.
- σ_X, σ_Y: Standard deviations of X and Y.

Types of Correlation Based on Methods

1. Pearson Correlation

- **Measures**: The linear relationship between two continuous variables.
- **Best for**:
    - Normally distributed data.
    - Linear relationships.

- **Range**: −1 to +1.
- **Sensitive to Outliers**: Yes.

2. Spearman Correlation

- **Measures**: The monotonic relationship (increasing or decreasing trend), using ranks instead of actual values.
- **Best for**:
    - Ordinal data or data not normally distributed.
    - Non-linear monotonic relationships.
- **Range**: −1 to +1.
- **Sensitive to Outliers**: Less than Pearson.

3. Kendall's Tau

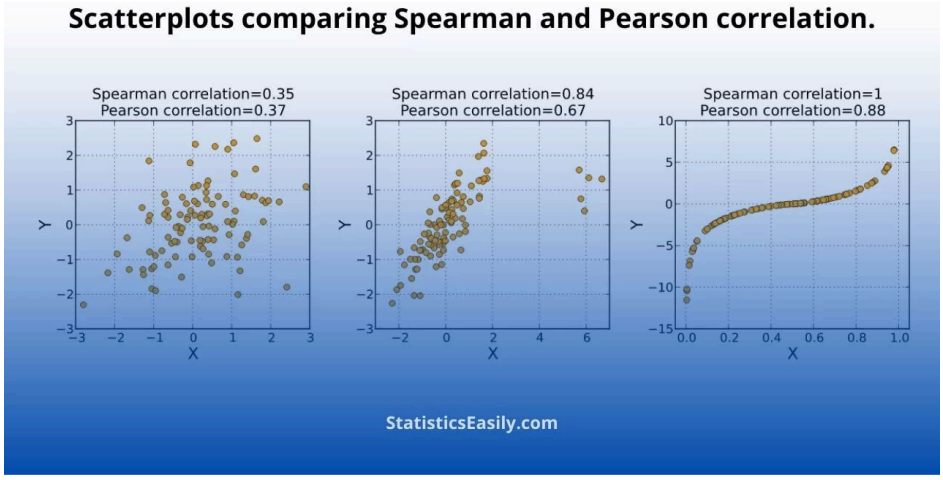
- **Measures**: The strength of a monotonic relationship, based on concordant and discordant pairs.
- **Best for**:
    - Small datasets.
    - Handling tied ranks better than Spearman.
- **Range**: −1 to +1.

4. Point-Biserial Correlation

- **Measures**: The relationship between a continuous variable and a binary variable.
- **Best for**:
    - Comparing test scores (continuous) against pass/fail (binary).
- **Range**: −1 to +1.

5. Phi Coefficient

- **Measures**: The relationship between two binary variables.
- **Best for**:
    - Categorical data encoded as 0 or 1.
- **Range**: −1 to +1.



Visualizing Correlation

Scatter Plot:

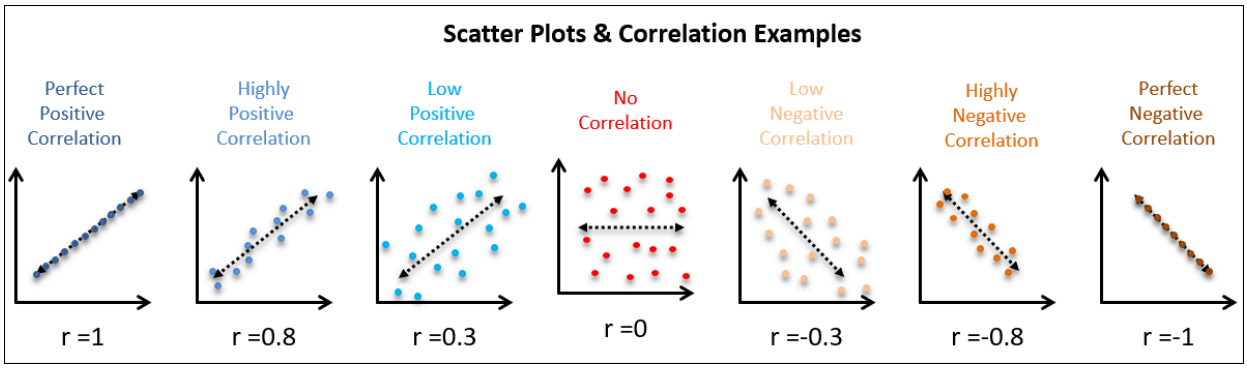
- Positive Correlation: Points cluster along an upward-sloping line.
- Negative Correlation: Points cluster along a downward-sloping line.
- No Correlation: Points are scattered randomly.

Heatmap:

- Displays pairwise correlation values in a matrix form, often used in multivariate analysis.

Applications

1. **Data Analysis**:
    - Understand relationships between variables.
2. **Feature Selection**:
    - In machine learning, highly correlated features may be redundant.
3. **Healthcare**:
    - Study relationships, e.g., between exercise and blood pressure.


Scatter Plots & Correlation Examples
Perfect Positive Correlation
Highly Positive Correlation
Low Positive Correlation
No Correlation
Low Negative Correlation
Highly Negative Correlation
Perfect Negative Correlation
r =1
r =0.8
r =0.3
r =0
r =-0.3
r =-0.8
r =-1