Statistics -1

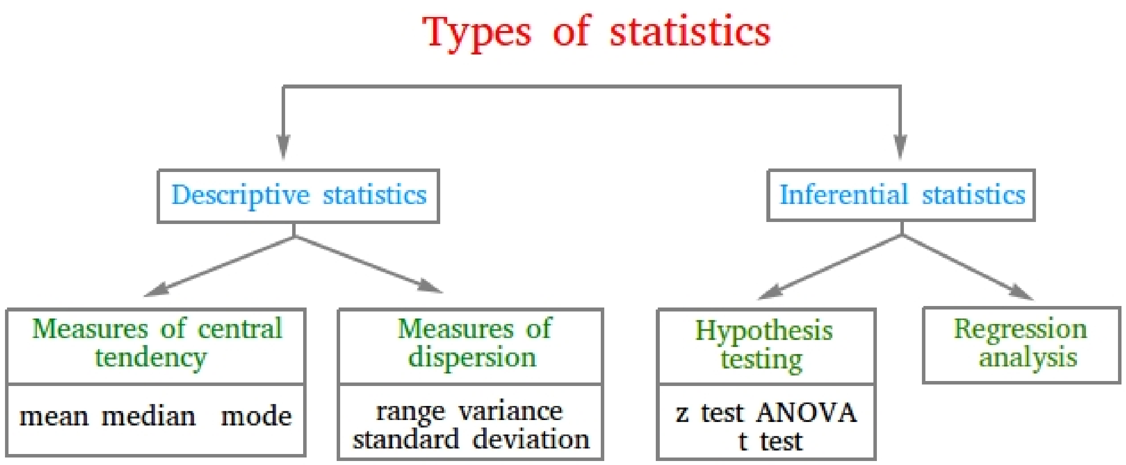
What is **Statistics** ?

- Statistics is a branch of mathematics that involves collecting, analysing, interpreting, and presenting data.
- It provides tools and methods to understand and make sense of large amounts of data and to draw conclusions and make decisions based on the data.
- In practice, statistics is used in a wide range of fields, such as business, economics, social sciences, medicine, and engineering. It is used to conduct research studies, analyse market trends, evaluate the effectiveness of treatments and interventions, and make forecasts and predictions.

Example :

1. Business - Data Analysis (Identifying customer behavior) and Demand Forecasting
2. Medical - Identify efficiency of new medicines (Clinical trials),
3. Identifying risk factor for diseases (Epidemiology)
4. Government & Politics - Conducting surveys, Polling
5. Environmental Science - Climate research

Types of Statistics :



1.Descriptive Statistics :

Descriptive statistics is all about working with data to make it easier to understand. It focuses on **summarizing and organizing information** so we can see what the data shows, without trying to guess or predict anything beyond that data.

Here's what it does:

1. **Collecting Data**: Gathering information, like test scores, survey results, or sales numbers.
2. **Organizing Data**: Arranging the data in a clear way, like in tables or charts.
3. **Analyzing Data**: Calculating things like averages, percentages, or ranges to understand the data better.
4. **Presenting Data**: Using graphs, charts, or summaries to show the data's main points.
   It's like taking a big pile of information and turning it into a simple story about what's happening in the data—without making guesses about the bigger picture

Population vs Sample

1. **Population**: The **entire group** of people or things we want to study.

    - Example: If you're studying students in a school, the population is **all the students in that school**.
    - Another example: If you're studying cars in a city, the population is **all the cars in that city**.
2. **Sample**: A **small group taken from the population** to study because it's hard to study the entire population.
    - Example: Instead of studying all the students, you might randomly pick **100 students** as your sample.
    - Another example: You could study **cars on one road** instead of the entire city.

Why Use a Sample?

- It's faster, cheaper, and easier than studying the whole population.
- Samples help us estimate what is true for the population.

Parameter vs Statistic

1. **Parameter**: A fact about the entire population (often unknown).
    - Example: The **average height** of all students in a school (population).
2. **Statistic**: A fact about the sample that we use to estimate the parameter.
    - Example: The **average height** of 100 students (sample) is a statistic used to estimate the parameter.

Key Things to Remember for a Good Sample

1. **Sample Size**: The sample should be big enough to give accurate results.
2. **Random Selection**: Everyone or everything in the population should have an equal chance of being chosen.
3. **Representative**: The sample should reflect the diversity of the population (e.g., if 50% of students are girls, the sample should include girls and boys in the same proportion).

Real-Life Examples

1. **Cricket Fans**:
    - **Population**: All cricket fans in the world.
    - **Sample**: Fans sitting in the stadium for one match.
2. **Students**:
    - **Population**: All students in a college.
    - **Sample**: Students who attend lectures on a specific day

1.Inferential Statistics :

What is Inferential Statistics?

Inferential statistics is about making guesses or predictions about a big group (population) by studying a smaller group (sample). Instead of looking at everyone, we take a sample, analyze it, and then try to understand what's true for the whole population.

It's like tasting a small bite of a dish to guess the overall flavor.

Main Topics in Inferential Statistics (Simplified)

1. **Hypothesis Testing:**
    - **Testing ideas or claims about a population using sample data.**
    - **Example: Checking if the average height of people in a city is more than 5 feet 5 inches.**
2. **Confidence Intervals:**
    - **Estimating a range where the true population value might lie.**
    - **Example: Saying, "The average height of people in a city is between 5.4 and 5.6 feet with 95% confidence."**
3. **Analysis of Variance (ANOVA):**

   - **Comparing averages of different groups to see if they are significantly different.**
   - **Example: Comparing the average test scores of students in three different schools.**

4. **Regression Analysis:**
   - **Finding relationships between variables to make predictions.**
   - **Example: Predicting sales based on advertising spend.**
5. **Chi-Square Tests:**
   - **Checking if two categories are connected.**
   - **Example: Testing if gender and favorite type of music are related.**
6. **Sampling Techniques:**
   - **Choosing a sample in a way that represents the population fairly.**
   - **Example: Randomly selecting 50 students from a school to study their eating habits.**
7. **Bayesian Statistics:**
   - **Updating what we believe about something as we get new information.**
   - **Example: If someone tests positive for a disease, updating the probability they actually have it based on the test accuracy.**

Why Machine Learning (ML) is Related to Statistics?

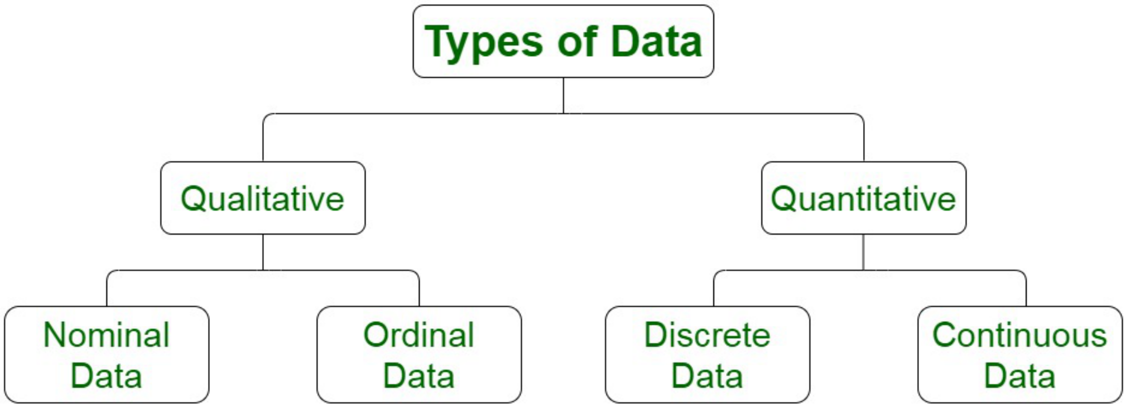
Machine learning and statistics are closely connected because they both deal with data and making predictions. Here's why:

- ML uses statistical techniques (like regression and probability) to find patterns in data.
- Both focus on analyzing data to make decisions or predictions.

- For example, in ML, predicting house prices is similar to using regression analysis in statistics.

Types of Data :

Data can be broadly classified into two main types: **Qualitative (Categorical)** and **Quantitative (Numerical)**. Let's explore these in detail:



1. Qualitative Data (Categorical Data)

This type of data represents categories or labels and is not numerical. It tells us "what kind" of something.

Types of Qualitative Data:

- **Nominal Data: Data that cannot be ranked or ordered.**
    - **Example: Colors of cars (red, blue, green), types of fruits (apple, banana).**
- **Ordinal Data: Data that has a specific order or rank, but the difference between values isn't meaningful.**
    - **Example: Customer satisfaction levels (satisfied, neutral, dissatisfied), ranks in a competition (1st, 2nd, 3rd).**

2. Quantitative Data (Numerical Data)

This type of data represents numbers and can be measured. It tells us "how much" or "how many" of something.

Types of Quantitative Data:

- **Discrete Data: Data that can only take certain specific values (countable).**
  - **Example: Number of students in a class, number of cars in a parking lot.**
- **Continuous Data: Data that can take any value within a range (measurable).**
  - **Example: Height of students, temperature of a city, weight of a bag.**

Types of Data (Simplified Explanation with Examples)

Data can be broadly classified into two main types: **Qualitative (Categorical)** and **Quantitative (Numerical)**. Let’s explore these in detail:

---

1. Qualitative Data (Categorical Data)

This type of data represents categories or labels and is not numerical. It tells us **"what kind"** of something.

Types of Qualitative Data:

- **Nominal Data**: Data that cannot be ranked or ordered.

    - **Example**: Colors of cars (red, blue, green), types of fruits (apple, banana).
- **Ordinal Data**: Data that has a specific order or rank, but the difference between values isn’t meaningful.
    - **Example**: Customer satisfaction levels (satisfied, neutral, dissatisfied), ranks in a competition (1st, 2nd, 3rd).

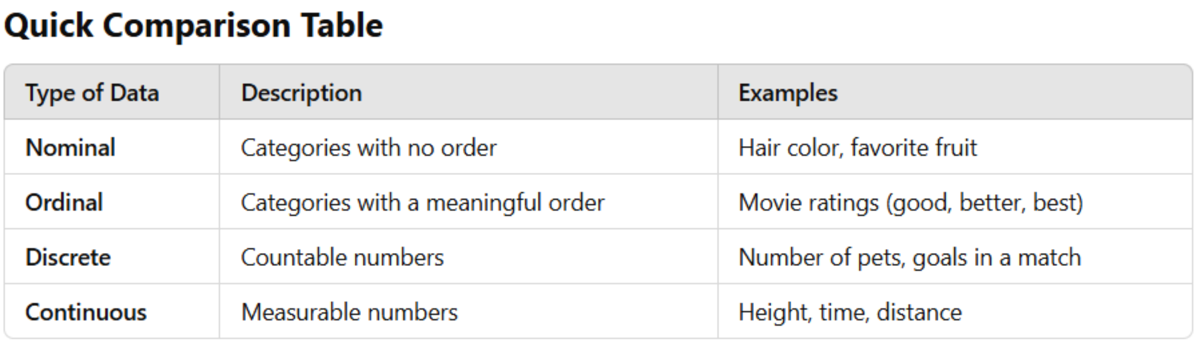
---

3. Quantitative Data (Numerical Data) :

This type of data represents numbers and can be measured. It tells us **"how much"** or **"how many"** of something.

Types of Quantitative Data:

- **Discrete Data**: Data that can only take certain specific values (countable).
    - **Example**: Number of students in a class, number of cars in a parking lot.
- **Continuous Data**: Data that can take any value within a range (measurable).
    - **Example**: Height of students, temperature of a city, weight of a bag.



Quick Comparison Table

Type of Data	Description	Examples
Nominal	Categories with no order	Hair color, favorite fruit
Ordinal	Categories with a meaningful order	Movie ratings (good, better, best)
Discrete	Countable numbers	Number of pets, goals in a match
Continuous	Measurable numbers	Height, time, distance

Measures of Central Tendency

The **measure of central tendency** is a way to summarize a set of data by identifying the center or middle value. It gives us an idea of what a "typical" data point looks like. The three main measures of central tendency are **Mean**, **Median**, and **Mode**.

---

1. Mean (Average)

The **mean** is the sum of all the data points divided by the number of data points.

Formula:

Mean=Sum of all valuesNumber of values\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}Mean=Number of valuesSum of all values

Example:

Scores: 10, 20, 30, 40, 50

Mean=10+20+30+40+505=30\text{Mean} = \frac{10 + 20 + 30 + 40 + 50}{5} = 30Mean=510+20+30+40+50=30

Key Points:

- **Best for** numerical data without extreme outliers.
- **Sensitive** to outliers (very large or very small numbers can distort the mean).

---

2. Median

The **median** is the middle value when the data is arranged in ascending order.

- If the number of values is **odd**, the median is the middle number.
- If the number of values is **even**, the median is the average of the two middle numbers.

Example 1 (Odd):

Scores: 10, 20, 30, 40, 50
Median = 30 (middle value)

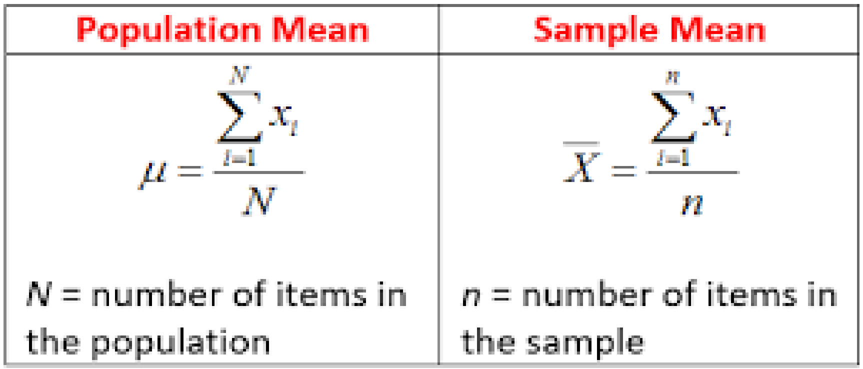
Example 2 (Even):

Scores: 10, 20, 30, 40 taking two middle point
Median = (20+30)/2= 25

Key Points:

- **Best for** skewed data or when there are outliers.
- Not affected by extreme values

Formula :



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

---

3. Mode

The **mode** is the value that occurs most frequently in the data.

- There can be **no mode**, **one mode**, or **multiple modes**.

Example 1 (Single Mode):

Scores: 10, 20, 20, 30, 40
Mode = 20 (appears twice)

Example 2 (Multiple Modes):

Scores: 10, 10, 20, 20, 30
Modes = 10 and 20 (both appear twice)

Key Points:

- **Best for** categorical data (e.g., favorite colors or brands).

- May not represent the "center" well if data is spread out.

When to Use Each Measure?

Measure	Use When
Mean	Data has no extreme outliers, and you want an overall average.
Median	Data has outliers or is skewed, and you want a value unaffected by extreme values.
Mode	Data is categorical, or you want to know the most common value.

Summary

- **Mean** gives the average.
- **Median** gives the middle value.
- **Mode** shows the most frequent value.

4.Weighted Mean

The **weighted mean** is a type of average that takes into account the importance or frequency of each value. Instead of treating all values equally, it gives more weight to some values based on their significance.

Formula: Weighted Mean= ∑(Value×Weight) / ∑(Weights)

Example:

Imagine you are calculating your grade for a course where assignments, exams, and projects have different weights:

- Assignment score = 80 (weight = 20%)
- Exam score = 90 (weight = 50%)
- Project score = 85 (weight = 30%)

Weighted Mean=

((80×0.2)+(90×0.5)+(85×0.3)) / 0.2+0.5+0.3 = 16+45+25.51= 86.5

Use Case:

The weighted mean is useful when certain values are more important than others, such as grades, stock portfolios, or survey results where responses are weighted by the number of participants.

5.Trimmed Mean

A **trimmed mean** is calculated by removing a percentage of the smallest and largest values in a dataset and then finding the mean of the remaining values.

How It Works:

1. Decide on the **trimming percentage** (e.g., 10%).
2. Remove that percentage of the smallest and largest values from the dataset.

3. Calculate the mean of the remaining values.

Example:

Dataset: [5, 10, 15, 20, 25, 30, 35]

- Trim 10% from each end (round to 1 value from each end).
- New dataset: [10, 15, 20, 25, 30]
- Mean = (10+15+20+25+30)/5=20

Use Case : Trimmed means are helpful when the dataset has **extreme outliers** that could distort the regular mean. For example, calculating average income or test scores while excluding unusually high or low values

Measure of Dispersion? (Simple Explanation)

A measure of dispersion tells us how spread out or scattered the data is around the central value (like the mean or median). It shows whether the data points are close to each other or far apart.

In simple words:

- **If the data is tightly packed, the dispersion is small.**
- **If the data is spread out, the dispersion is large.**

Why is it Important?

Knowing the spread helps us understand how consistent the data is. For example:

- **Two students might both have an average score of 75, but one student's scores might vary a lot (e.g., 50, 100) while the other's scores are consistent (e.g., 70, 80).**

Real-World Examples of Dispersion

1. **Exam Scores:**
   - **Class A: Scores are 70, 72, 74, 76, 78 (small dispersion – scores are close).**
   - **Class B: Scores are 50, 60, 70, 80, 90 (large dispersion – scores are spread out).**
   - **Even if both classes have the same average score, Class A has more consistent performance.**
2. **Weather:**
   - **City A: Temperatures in a week: 20°C, 21°C, 22°C, 21°C, 20°C (low dispersion – consistent weather).**
   - **City B: Temperatures in a week: 10°C, 30°C, 25°C, 15°C, 35°C (high dispersion – unpredictable weather).**

Common Measures of Dispersion

1. **Range:**
   - **Difference between the highest and lowest values in a datasets.but it affected from outliers.**
   - **Example: For scores 10, 20, 30, 40, 50 → Range = 50 - 10 = 40.**
2. **Variance:**

Variance is a measure of **how spread out the data is around the mean** (average). It shows how much the values in a dataset differ from the mean.

- If the variance is **small**, most data points are close to the mean.
- If the variance is **large**, the data points are spread out far from the mean.

How is Variance Calculated?

1. Find the mean (average) of the dataset.
2. Subtract the mean from each data point to find the difference (how far each value is from the mean).
3. Square these differences (to make them positive).
4. Find the average of these squared differences.

Formula:

Variance=∑(Data Point−Mean) / Number of Data Points

Real-World Example

Example 1: Exam Scores

Suppose five students scored: 60, 70, 80, 90, 100.

1. **Step 1: Find the mean**

Mean=(60+70+80+90+100)/5 = 80

2. **Step 2: Subtract the mean from each score and square the result**

(60 - 80)^2 = 400,\ (70 - 80)^2 = 100,\ (80 - 80)^2 = 0,\ (90 - 80)^2 = 100,\ (100 - 80)^2 = 400

3. **Step 3: Find the average of these squared differences**

Variance=(400+100+0+100+400)/5=200

So, the variance is **200**.

Why is Variance Useful?

- **Low variance**: Scores like 78, 79, 80, 81, 82 → Scores are close to the mean, indicating consistency.
- **High variance**: Scores like 50, 60, 70, 90, 100 → Scores are spread out, indicating variability.

3. Standard Deviation :

- Standard Deviation: The standard deviation is the square root of the variance.
- It is a widely used measure of dispersion that is useful in describing the shape of a distribution.
- If the standard deviation is **small**, most data points are close to the mean (data is consistent).
- If the standard deviation is **large**, the data points are spread out far from the mean (data is more varied).

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$

Population Variance

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

Sample Variance

What is Standard Deviation? (Simple Explanation)

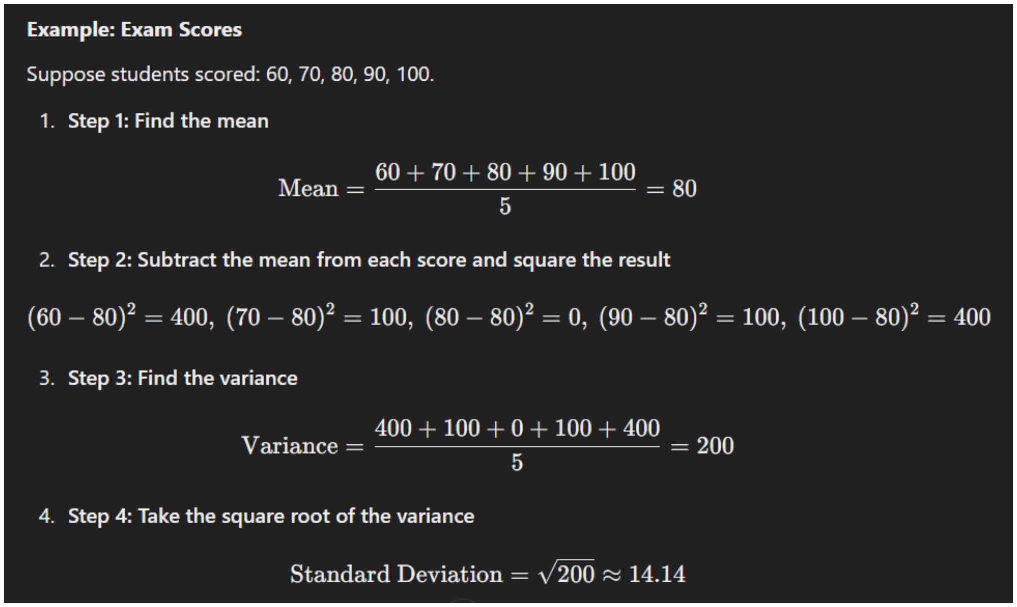
The **standard deviation** measures how spread out the data is around the mean. It is the square root of the **variance**, which makes it easier to understand because it's in the same unit as the data.

- If the standard deviation is **small**, most data points are close to the mean (data is consistent).
- If the standard deviation is **large**, the data points are spread out far from the mean (data is more varied).

How is Standard Deviation Calculated?

1. Find the **mean** (average) of the dataset.

2. Subtract the mean from each data point and square the result.
3. Find the **average of these squared differences** (this is the variance).
4. Take the **square root** of the variance to get the standard deviation.



Example: Exam Scores

Suppose students scored: 60, 70, 80, 90, 100.

1. **Step 1: Find the mean**

$$\text{Mean} = \frac{60 + 70 + 80 + 90 + 100}{5} = 80$$

2. **Step 2: Subtract the mean from each score and square the result**

$$(60-80)^2 = 400,\ (70-80)^2 = 100,\ (80-80)^2 = 0,\ (90-80)^2 = 100,\ (100-80)^2 = 400$$

3. **Step 3: Find the variance**

$$\text{Variance} = \frac{400 + 100 + 0 + 100 + 400}{5} = 200$$

4. **Step 4: Take the square root of the variance**

$$\text{Standard Deviation} = \sqrt{200} \approx 14.14$$

4. Coefficient of Variation

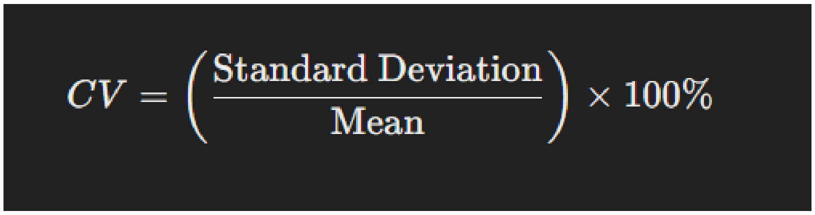
Coefficient of Variation (CV): The CV is the ratio of the standard deviation to the mean expressed as a percentage. It is used to compare the variability of datasets with different means and is commonly used in fields such as biology, chemistry, and engineering.

The coefficient of variation (CV) is a statistical measure that expresses the amount of variability in a dataset relative to the mean. It is a dimensionless quantity that is expressed as a percentage.

- A **higher CV** means the data is more variable compared to the mean.

- A **lower CV** means the data is more consistent compared to the mean.

Formula:



$$CV = \left(\frac{\text{Standard Deviation}}{\text{Mean}}\right) \times 100\%$$

Real-World Example: Comparing Salaries and Expenses

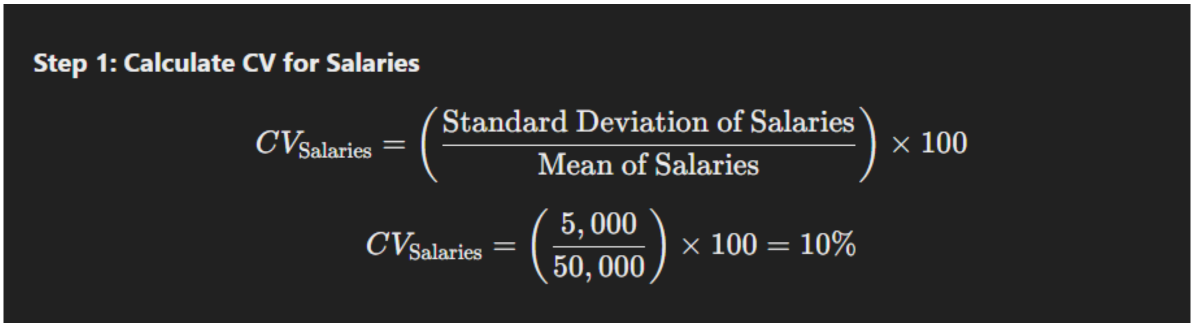
Scenario:

Suppose you want to compare the variability in **salaries** and **monthly expenses** for a group of employees:

Dataset:

- **Salaries** (in ₹): Mean = ₹50,000, Standard Deviation = ₹5,000
- **Monthly Expenses** (in ₹): Mean = ₹20,000, Standard Deviation = ₹4,000

Step 1: Calculate CV for Salaries

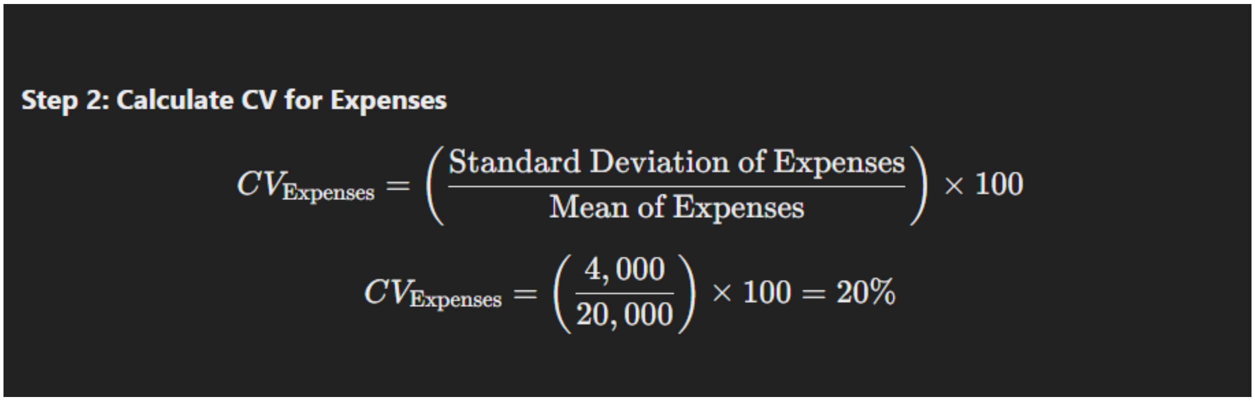


Step 1: Calculate CV for Salaries

$$CV_{\text{Salaries}} = \left(\frac{\text{Standard Deviation of Salaries}}{\text{Mean of Salaries}}\right) \times 100$$

$$CV_{\text{Salaries}} = \left(\frac{5,000}{50,000}\right) \times 100 = 10\%$$

Step 2: Calculate CV for Expenses



Step 2: Calculate CV for Expenses

$$CV_{\text{Expenses}} = \left(\frac{\text{Standard Deviation of Expenses}}{\text{Mean of Expenses}}\right) \times 100$$

$$CV_{\text{Expenses}} = \left(\frac{4,000}{20,000}\right) \times 100 = 20\%$$

Interpretation

- The **CV for Salaries is 10%**, meaning salaries are relatively consistent with only small variations compared to the average salary.
- The **CV for Expenses is 20%**, meaning expenses are more variable compared to the average expense.

When to Use CV?

1. **Comparing Different Units**:
   - CV is unitless, making it ideal for comparing variability in datasets with different units (like salaries in ₹ vs expenses in ₹).
2. **Fields like Biology, Chemistry, or Finance**:
   - Example: Comparing the consistency of test results in two labs with different average measurements.

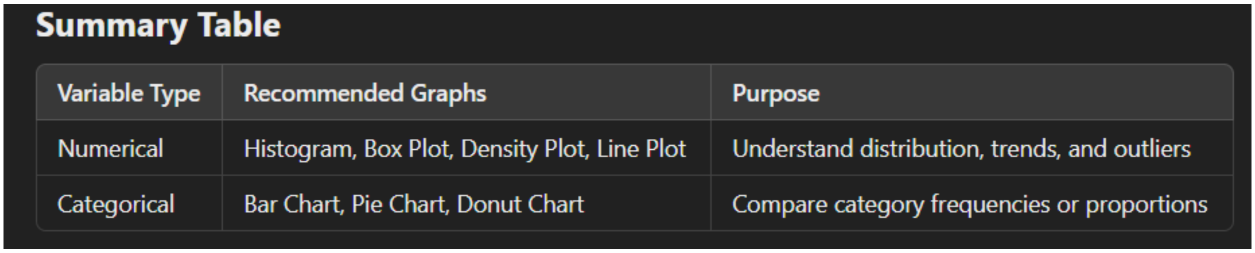
Graphs for Univariate Analysis

Univariate analysis focuses on analyzing one variable / single column at a time. The choice of graphs depends on whether the variable is numerical or categorical.

There are two type of Graphs :

1.Categorical

2.Numerical

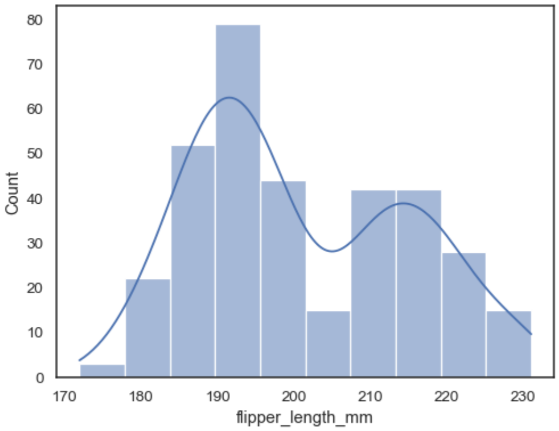


Summary Table

Variable Type	Recommended Graphs	Purpose
Numerical	Histogram, Box Plot, Density Plot, Line Plot	Understand distribution, trends, and outliers
Categorical	Bar Chart, Pie Chart, Donut Chart	Compare category frequencies or proportions

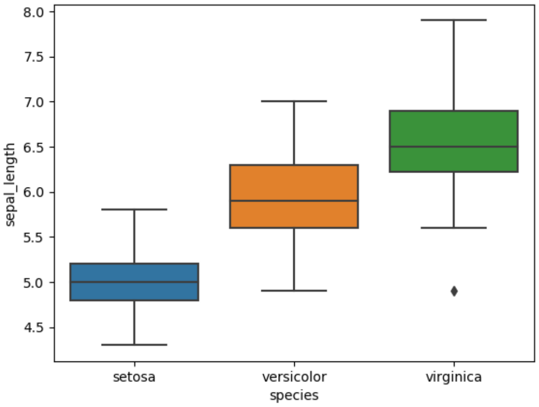
1. For Numerical Columns (Continuous or Discrete Data)

1. Histogram



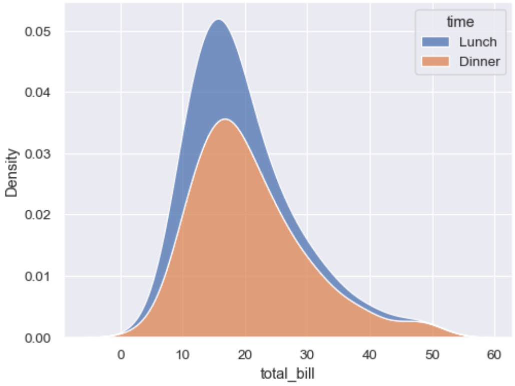
- **Use: Shows the frequency distribution of numerical data.**
- **Example: Examining the distribution of ages or salaries.**
- **Interpretation: Helps understand the shape of the data (normal, skewed, etc.).**

2. Box Plot



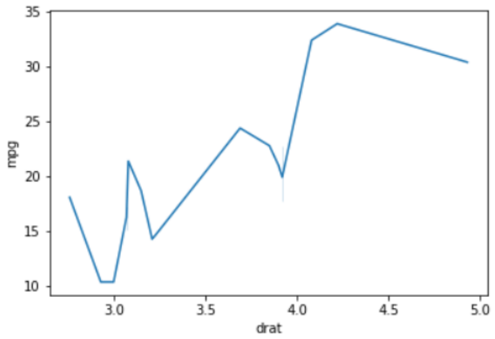
- ○ **Use: Displays the spread and identifies outliers.**
- ○ **Example: Analyzing income distribution and detecting outliers.**
- ○ **Interpretation: Shows the median, quartiles, and potential outliers.**

3. Density Plot(Kde_plot)



- ○ **Use: A smoothed version of the histogram to show the data distribution.**
- ○ **Example: Examining the probability distribution of heights.**
- ○ **Interpretation: Useful for understanding the data's underlying trends.**

4. Line Plot

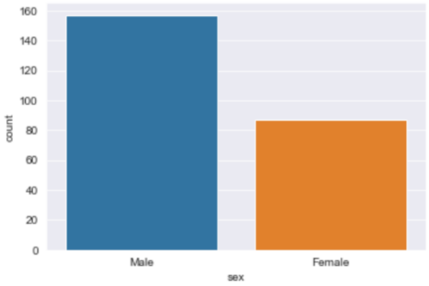


- Use: Shows trends or patterns for ordered numerical data.
- Example: Analyzing stock prices over time.
- Interpretation: Best for sequential or time-related data.

---

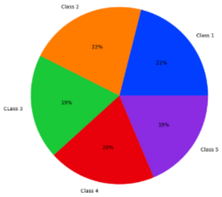
2. For Categorical Columns (Categories or Groups)

1. Bar Chart



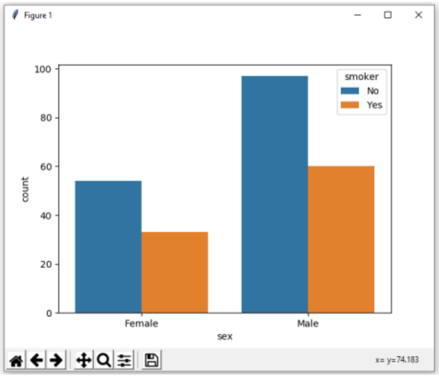
- Use: Displays the frequency of each category.
- Example: Analyzing gender distribution or product preferences.
- Interpretation: Highlights which category is most or least frequent.

2. Pie Chart



- **Use: Represents proportions of each category as a part of a whole.**
- **Example: Examining the percentage of market share by companies.**
- **Interpretation: Useful for understanding relative sizes, though less precise.**

3. **Frequency Table or Count Plot**



- **Use: A simple representation of category counts.**
- **Example: Showing counts of departments in an organization.**
- **Interpretation: Helps compare the sizes of categories.**

4. **Donut Chart (Variation of Pie Chart)**

- **Use: Similar to a pie chart but visually more appealing for percentages.**

Combination of Numerical and Categorical

If you'd like to compare numerical data for different categories:

- **Use a violin plot or box plot to show the distribution of numerical values for each category.**
- **Example: Compare the distribution of salaries by department.**

Frequency Distribution Table with Examples

Frequency Distribution Table

A frequency distribution table shows how often each category or value appears in a dataset.

- It summarizes raw data into manageable and understandable categories.

Example:
You survey 200 people about their favorite vacation type. The results are:

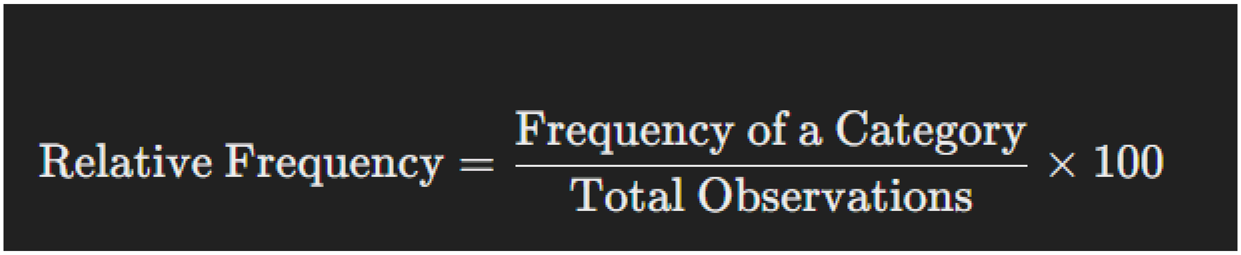
- Beach (50), City (40), Adventure (30), Nature (35), Cruise (25), Other (20).

Vacation Type	Frequency (Number of People)
Beach	50
City	40
Adventure	30
Nature	35
Cruise	25
Other	20
Total	200

Relative Frequency

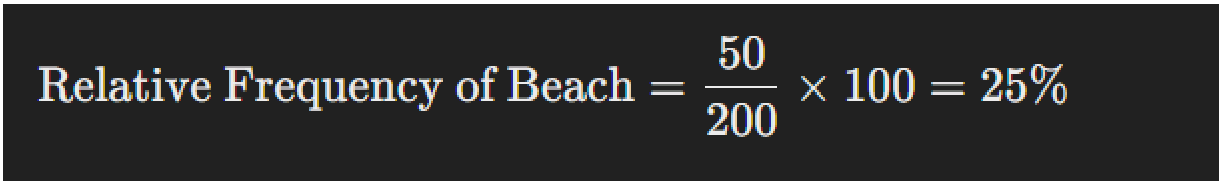
Relative frequency is the proportion of observations in a category, expressed as a fraction or percentage.

Formula:



$$\text{Relative Frequency} = \frac{\text{Frequency of a Category}}{\text{Total Observations}} \times 100$$

Example Calculation:
For the Beach category:



$$\text{Relative Frequency of Beach} = \frac{50}{200} \times 100 = 25\%$$

Relative Frequency Table:

Vacation Type	Frequency	Relative Frequency (%)
Beach	50	25%
City	40	20%
Adventure	30	15%
Nature	35	17.5%
Cruise	25	12.5%
Other	20	10%
Total	200	100%

Cumulative Frequency

Cumulative frequency is the running total of frequencies up to the current category.

How to Calculate:

- **Add the frequency of the current category to the total frequency of all previous categories.**

Example Calculation:

1. Beach: 50
2. Beach + City: 50+40=9050 + 40 = 9050+40=90
3. Beach + City + Adventure: 90+30=12090 + 30 = 12090+30=120
   ...and so on.

Cumulative Frequency Table:

Vacation Type	Frequency	Cumulative Frequency
Beach	50	50
City	40	90
Adventure	30	120
Nature	35	155
Cruise	25	180
Other	20	200

Summary of Terms with Examples

Term	Description	Example (Vacation Survey)
Frequency	Number of times a value occurs.	50 people prefer Beach vacations.
Relative Frequency	Proportion or percentage of a category.	25% prefer Beach vacations.
Cumulative Frequency	Running total of frequencies up to the current category.	90 people prefer Beach or City vacations combined.

This structured approach makes it easy to analyze and interpret survey data effectively!