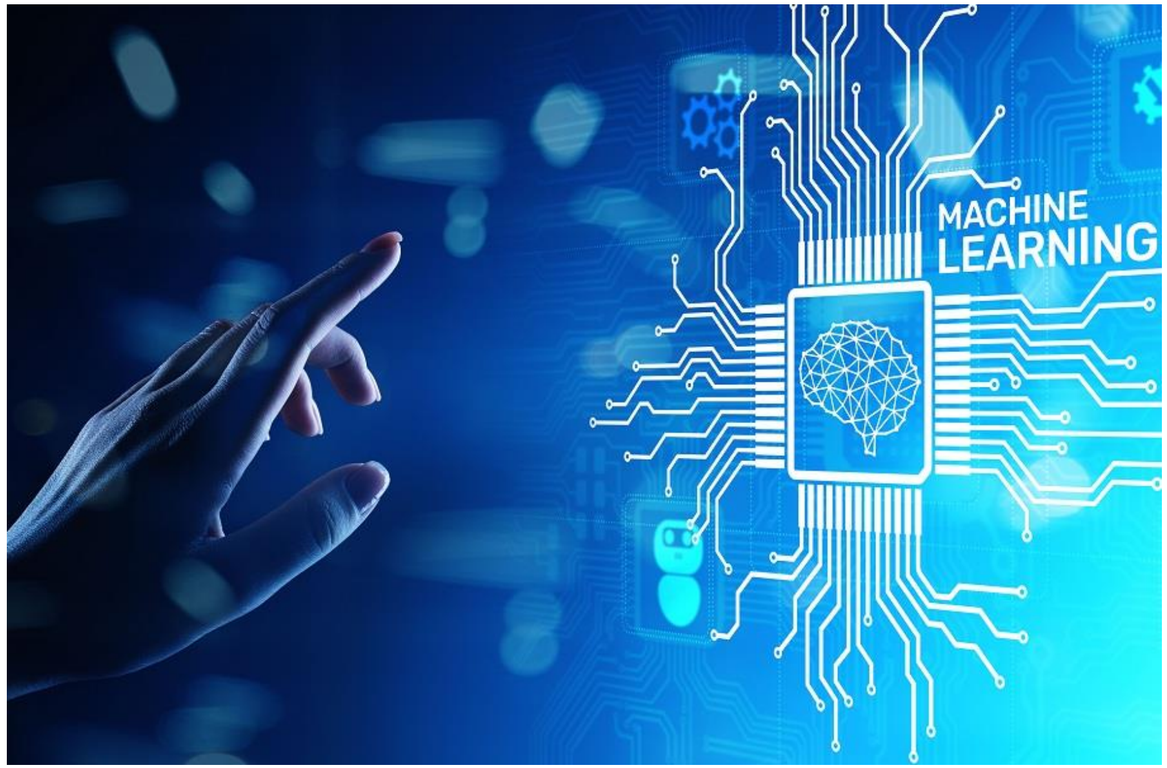
Introduction



Learning for Today (09.12.2025)

- Box Plot
  - Data Distribution
  - Outlier Detection
  - Variability before Modeling
  - Need for Pre-Processing
- Linear Regression

Types of Data: Terminologies

- Structured

A	B	C	D	E
0	1	1	F	0
1	6	5	E	1

- Unstructured

You are looking good
All the best for your second semester
What are you eating?

Structured Data

- Various data formats that can be stored
- Number
- String
- Boolean
- Date
- Images
- And many more…..
- Accordingly, algorithms are evolving……

Types of Data

- Quantitative data are measures of values or counts and are expressed as numbers.
- Quantitative data are data about numeric variables (e.g. how many; how much; or how often).
- Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.
- Qualitative data are data about categorical variables (e.g. what type).

Examples

Data unit	Numeric variable	= Quantitative data	Categorical variable	= Qualitative data
A person	"How many children do you have?"	4 children	"In which country were your children born?"	India
	"How much do you earn?"	Rs.50 lk/Annum	"What is your occupation?"	Photography
	"How many hours do you work?"	38 hours per week	"Do you work full-time or part-time?"	Full-time
A house	"How many square metres is the house?"	2000 square metres	"In which city or town is the house located?"	Chennai
A business	"How many workers are currently employed?"	300 employees	"What is the industry of the business?"	Retail
A farm	"How many milk cows are located on the farm?	36 cows	"What is the main activity of the farm?"	Dairy

Scale of Measurement

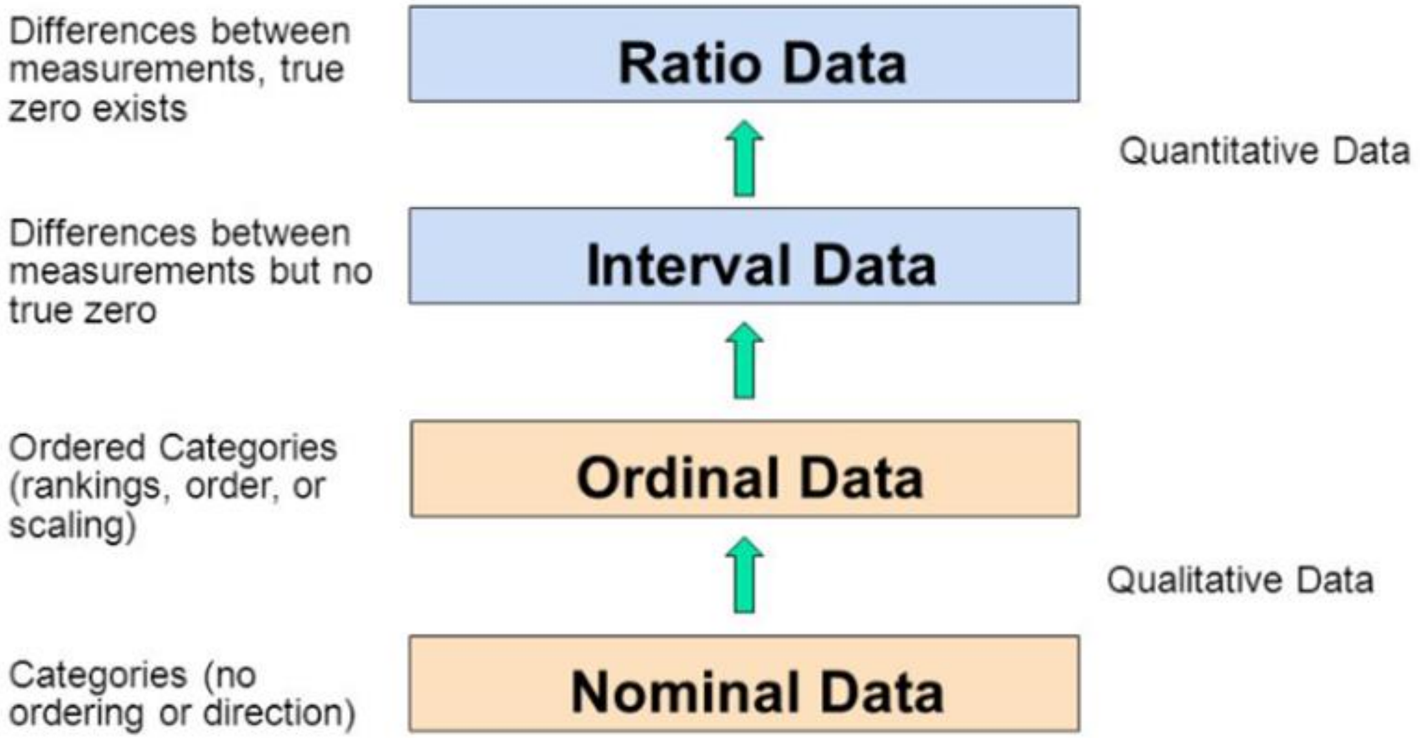
- **Nominal Scale**
  - Items are assigned to groups.
  - No ordering.
  - No number is generated.
  - Eg., Male/Female, Black/Brown/White

- **Ordinal Scale**
  - Ordering.
  - Higher number represents higher value (Usually).
  - numbers are used only for ordering.
  - Eg., Very High/High/Fair/Bad/Very bad

Scale of Measurement

- **Interval Scale**
  - Used when interval that separates groups is known.
  - can add and subtract
  - But can’t multiply and divide
  - Eg., No of educated/Uneducated on year basis
- **Ratio Scale**
  - Ordering, interval and ratio.
  - Multiplication and division possible.
  - Value zero represents the absence of value measured.
  - Eg., Temperature of a city for a period

Scale of Measurement



- http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+quantitative+and+qualitative+data

Measures of Central Tendency

- Measures of central tendency reflect the central tendencies of a distribution
  - Mode reflects the attribute with the greatest frequency
  - Median reflects the attribute that cuts the distribution in half
  - Mean reflects the average; sum of attributes divided by # of cases

Measures of Dispersion

- Measures of dispersion reflect the spread or distribution of the distribution
  - Range is the difference between largest & smallest scores; high – low
  - Variance is the average of the squared differences between each observation and the mean
  - Standard deviation is the square root of variance

Descriptive data summarization

- Identify Properties and characteristics of data
- Measure of Central Tendency
  - Mean, Median, Mode, Midrange
- Dispersion of data
  - Range, Quartiles , Inter Quartile Range and Variance

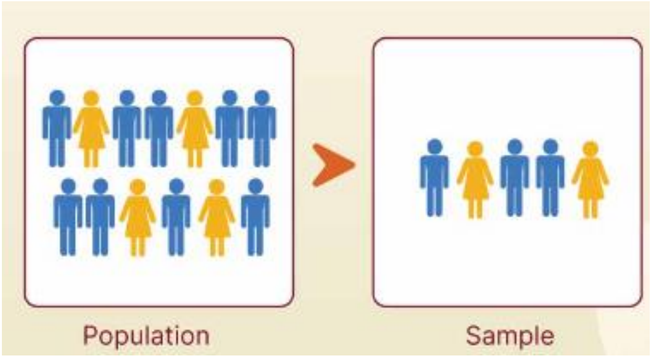
Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):
  - Weighted arithmetic mean
  - Trimmed mean: chopping extreme values
- Median: A holistic measure
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*)
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

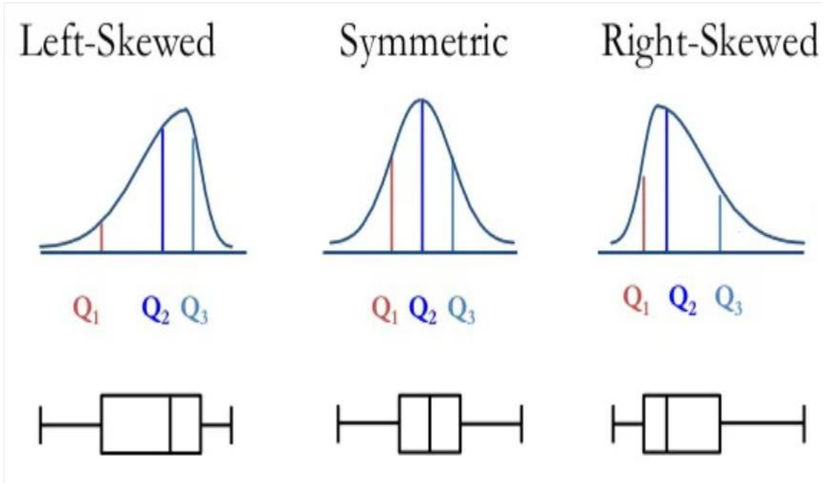
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

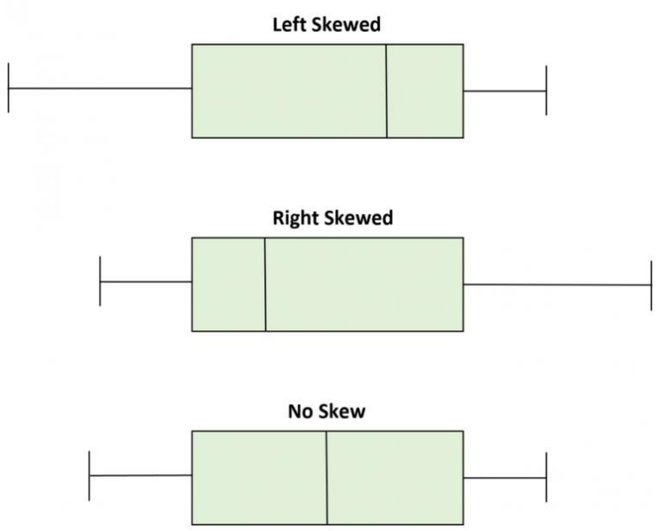
Measuring the Dispersion of Data

- Quartiles, outliers, and box plots
  - Quartiles: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five number(point) summary: min, Q_1, M, Q_3, max
  - Boxplot: ends of the box are the quartiles, the median is marked, whiskers, and plot outlier individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ)*
  - Variance: (algebraic, scalable computation)
  - Standard deviation *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*



Types of Skew Analysis





Quartile Calculation

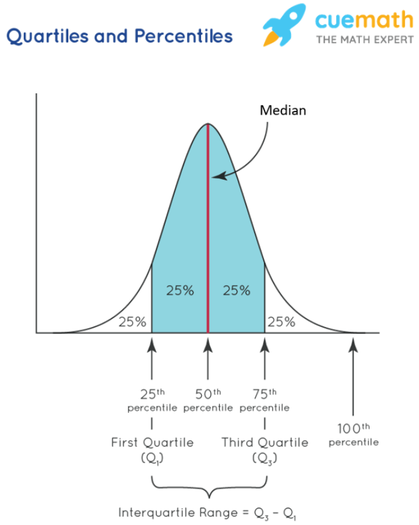
First Quartile(Q1) = $((n + 1)/4)^{th}$ Term.

Term = $((n + 1)/4)^{th}$ Term.

For better accuracy = Term+ (next term- term)

Second Quartile(Q2) = $((n + 1)/2)^{th}$ Term.

Third Quartile(Q3) = $(3(n + 1)/4)^{th}$ Term.



Lab Examples

1. Calculate the median, lower quartile, upper quartile, and interquartile range of the following data set of values: 20, 19, 21, 22, 23, 24, 25, 27, 26. Draw the box-plot for the above data and mention the five-point summary of data distribution.

```
import matplotlib.pyplot as plt
data = [20, 19, 21, 22, 23, 24, 25, 27, 26]
plt.boxplot(data)
plt.title("Box Plot of the Given Data")
plt.ylabel("Values")
plt.show()
```

Lab Examples

2. Calories in Vanilla-Flavored Ice Cream Bars : 342 ,377 ,319, 353 , 295 ,234 ,294 ,286 ,377 ,182 ,310 ,439 ,111 ,201, 182 ,197 ,209 ,147 ,190, 151 ,131 ,151. Analyze the dispersion of data (Mean, Median, Mode, Five Point summary, Outliers. List down the inferences

```
import pandas as pd
import matplotlib.pyplot as plt
Data=pd.read_csv('/content/Vanilla.csv')
# Creating plot
plt.boxplot(Data)
#plt.boxplot(Data, vert=False)
# show plot
plt.show()
```

Lab Examples

3. Outlier

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
Data =pd.read_csv('/content/Lab1_Tips.csv')
Data.boxplot(column =['total_bill'])
plt.show()
```

Lab Examples

4. Outlier and Grid

```
import pandas as pd
import matplotlib.pyplot as plt
Data =pd.read_csv('/content/Lab1_Tips.csv')
Data.boxplot(by ='day', column =['total_bill'], grid=False)
plt.ylabel('Total Bill')
plt.xlabel('Days')
plt.show()
```

Lab Examples

5. With Seaborn

```
import seaborn
seaborn.set(style='whitegrid')
tip = seaborn.load_dataset('tips')
seaborn.boxplot(x='day', y='tip', data=tip)
```

Lab Examples

6. Variability Example

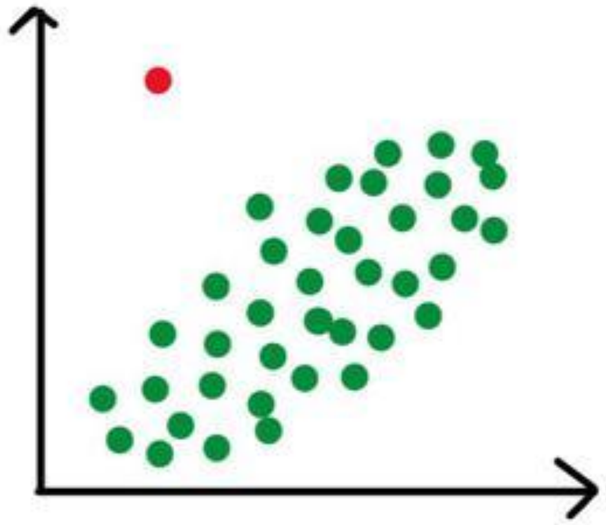
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
# ------------------------
# 1. Load Kaggle dataset
# ------------------------
df = pd.read_csv("/content/Lab1_StudentsPerformance.csv")
print("First 5 rows:")
print(df.head())
# ------------------------
# 2. Select numeric features
# ------------------------
numeric_cols = ["math score", "reading score", "writing score"]
df_numeric = df[numeric_cols]
print("\nSummary Statistics:")
print(df_numeric.describe())
# ------------------------
# 3. Box plots to show variability
# ------------------------

plt.figure(figsize=(12, 5))

for i, col in enumerate(numeric_cols):
    plt.subplot(1, 3, i+1)
    sns.boxplot(y=df[col])
    plt.title(f"Variability in {col}")

plt.tight_layout()
plt.show()
# ------------------------
# 4. Interpretation (printed)
# ------------------------

print("\nInterpretation Guide:")
print("- The height of each box shows the spread of the middle 50% of scores.")
print("- Whiskers show the overall range of performance.")
print("- Dots (if any) indicate outliers—students with unusually high/low scores.")
print("- Seeing unequal spreads helps understand variability before modeling.")
```

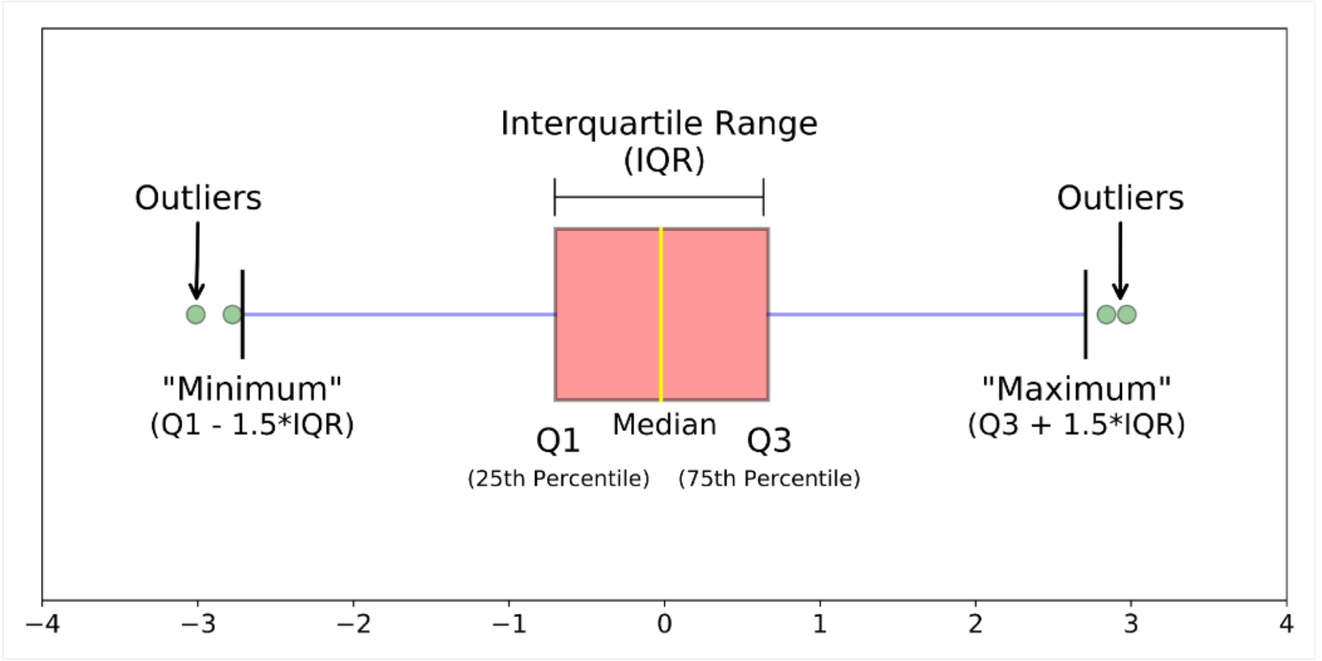
Lab Exercise

1. *Apply variability to tips.csv*

What is Outlier & How to calculate it?

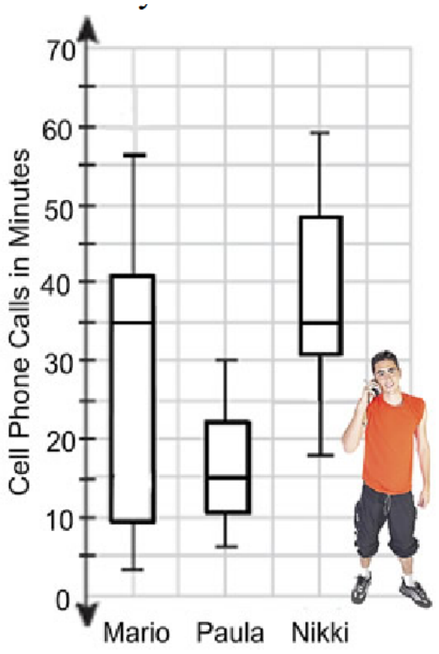
An outlier is an object that deviates significantly from the rest of the objects. The analysis of outlier data is referred to as outlier analysis or outlier mining.

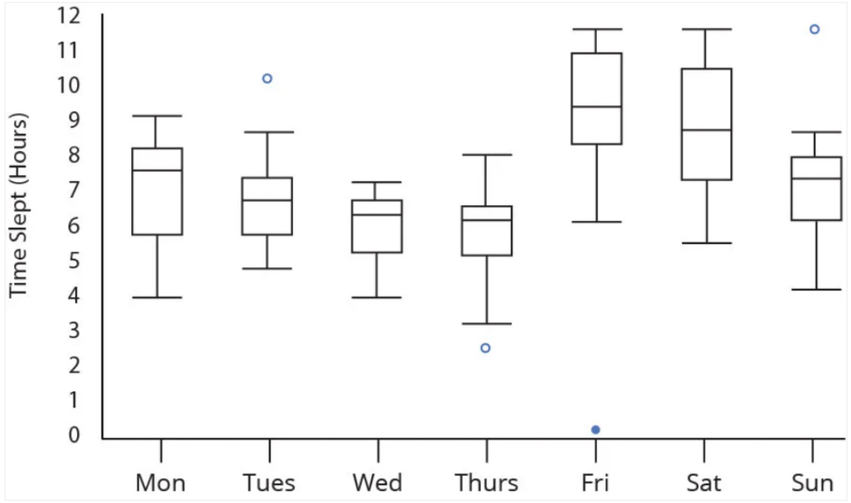




Source: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

How to analyze and Interpret this ?





Graphic Display of Descriptive Summaries

- Histograms
- Bar charts
- Quantile plot
- Quantile-Quantile (Q-Q Plot)
- Scatter plot
- LOESS Curve (LOWESS -locally weighted scatterplot smoothing)

Standard Deviation and Variance

Standard Deviation Formula	
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*

For samples:

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

$$\text{standard deviation} = s = \sqrt{s^2}$$

Calculating Formula

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

For populations:

$$\text{variance} = \sigma^2 = \frac{\Sigma (x-\bar{x})^2}{n}$$

$$\text{standard deviation} = \sigma = \sqrt{\sigma^2}$$

Calculating Formula

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

If the variance is high, that means you have larger variability in your dataset. In the other way, we can say more values are spread out around your mean value.

Standard deviation represents the average distance of an observation from the mean

The larger the standard deviation, the larger the variability of the data.

Types of data sets

Univariate, bivariate and multivariate are the various types of data that are based on the number of variables. Variables mean the number of objects that are under consideration as a sample in an experiment.

- UNIVARIATE DATA:

Univariate data is used for the simplest form of analysis. It is the type of data in which analysis are made only based on one variable. For example, there are sixty students in class VII. If the variable marks obtained in math were the subject, then in that case analysis will be based on the number of subjects that fall into defined categories of marks.

Types of data sets

- BIVARIATE DATA:

Bivariate data is used for little complex analysis than as compared with univariate data. Bivariate data is the data in which analysis are based on two variables per observation simultaneously.

- MULTIVARIATE DATA:

Multivariate data is the data in which analysis are based on more than two variables per observation. Usually multivariate data is used for explanatory purposes.

Three types of analysis

- Univariate analysis
  - the examination of the distribution of cases on only one variable at a time

Heights (in cm)	164	167.3	170	174.2	178	180	186

- Bivariate analysis
  - the examination of two variables simultaneously

TEMPERATURE(IN CELSIUS)	ICE CREAM SALES
20	2000
25	2500
35	5000
43	7800

- Multivariate analysis
  - the examination of more than two variables simultaneously

Comparison between Datasets

Univariate	Bivariate	Multivariate
It only summarize single variable at a time.	It only summarize two variables	It only summarize more than 2 variables.
It does not deal with causes and relationships.	It does deal with causes and relationships and analysis is done.	It does not deal with causes and relationships and analysis is done.
It does not contain any dependent variable.	It does contain only one dependent variable.	It is similar to bivariate but it contains more than 2 variables.
The main purpose is to describe.	The main purpose is to explain.	The main purpose is to study the relationship among them.
The example of a univariate can be height.	The example of bivariate can be temperature and ice sales in summer vacation.	Example, Suppose an advertiser wants to compare the popularity of four advertisements on a website. Then their click rates could be measured for both men and women and relationships between variable can be examined

Independent & Dependent Variables

- An independent variable is a variable that stands alone and isn't changed by the other variables you are trying to measure. For example, someone's age might be an independent variable.
- A dependent variable is something that depends on the independent variable.

For instance, rent depends upon the square feet of an apartment building. In this example, rent is a dependent variable that depends upon square feet which is an independent variable.

The dependent variable is also denoted as “Y” variable. The independent variable is denoted as the “X” variable.

Identify the correct Statements

You are buying boxes of cookies at a bakery. Each box of cookies costs Rs.40. Which of the following statements are true?

Choose all answers that apply:

- The dependent variable is the number of boxes of cookies you buy.
- The independent variable is the number of boxes of cookies you buy.
- The dependent variable is the amount of money you spend on the cookies.
- The independent variable is the amount of money you spend on the cookies.