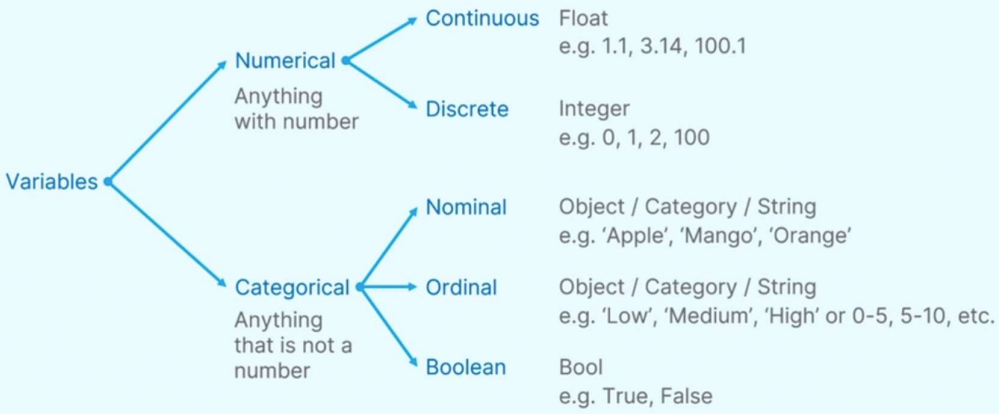
Module 2 | Topic 1 | Video 2

Variable Types Used in Statistical Data Analysis

Learning Objective

- Variable types in Statistics
- How variables are expressed in python

Variables for Statistical Data Analysis



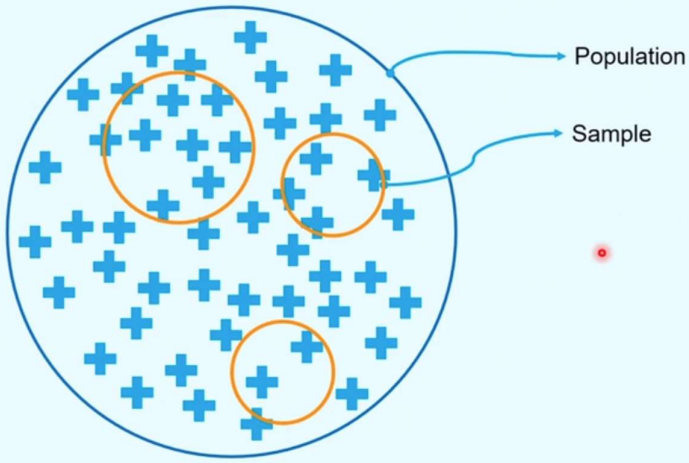
Module 2 | Topic 1 | Video 3

Population and Sampling

Learning Objective

- Population Vs Samples
- Learn why sampling
- Common sampling methods

Population Vs Samples

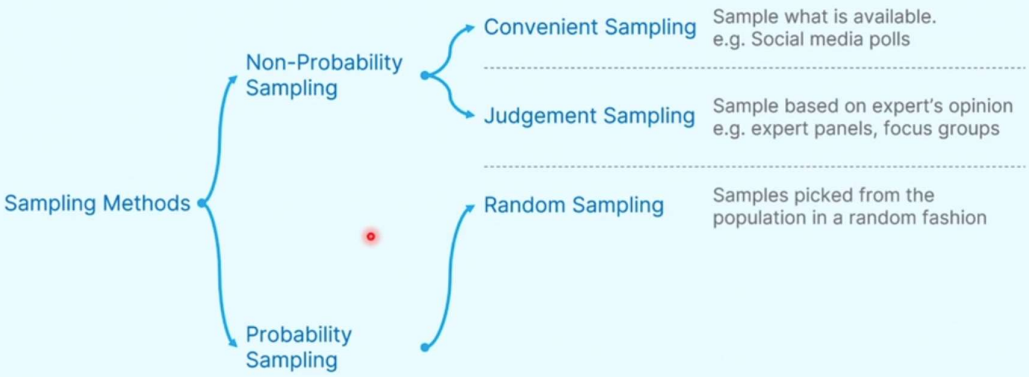


Why Sampling?

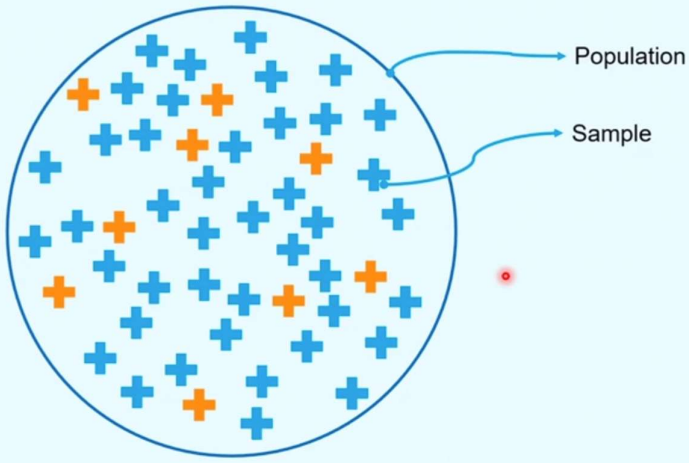
- Population keeps growing e.g. population census
- Analyzing population is expensive and almost always impossible

- Solution: Sampling
- Assumption: Properties of the sample are statistically identical to that of the population

Sampling Methods



Population Vs Samples



Sampling Methods

- Sampling Methods
  - Non-Probability Sampling
    - Convenient Sampling — Sample what is available. e.g. Social media polls
    - Judgement Sampling — Sample based on expert's opinion e.g. expert panels, focus groups
  - Probability Sampling
    - Random Sampling — Samples picked from the population in a random fashion
    - Systematic Sampling — Rule based sampling e.g. every third item, first 10 entries per day etc.
    - Stratified Sampling — Population broken into strata a.k.a. groups. Samples picked from each strata per quota.
    - Cluster Sampling — Similar to stratified sampling but population broken in the natural way e.g. by demographics etc.

What did we learn?

- Population Vs Samples
  - Population: All data. Keeps Growing. Impossible to collect.
  - Sample: Subset of the population
- Learn why sampling
- Sampling methods
  - Probability:
    - Random sampling, Systematic sampling, Stratified sampling, Cluster sampling
  - Non-Probability:
    - Convenient sampling, Judgement sampling

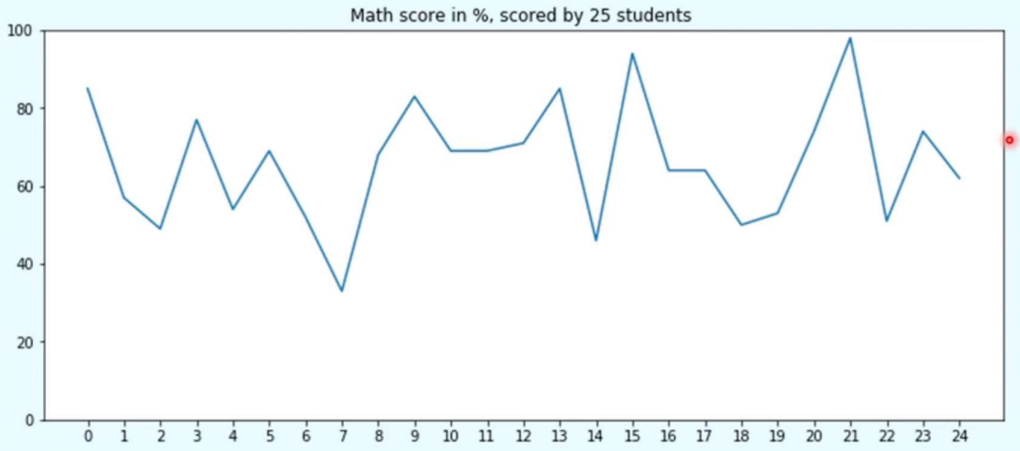
Module 2 | Topic 2 | Video 1

Central Tendency Measures

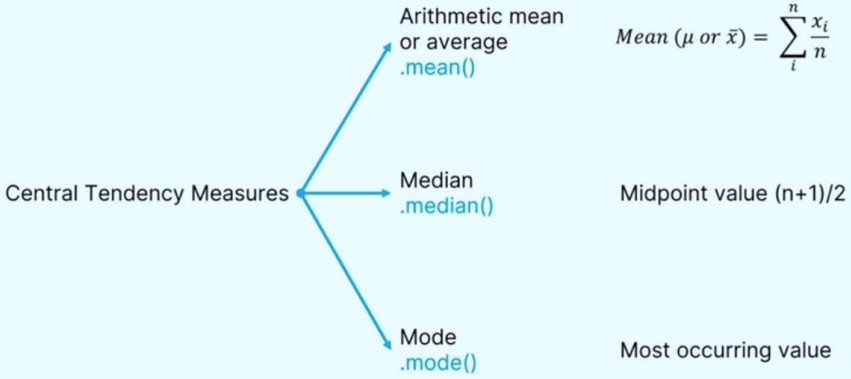
Learning Objective

- Need to summarize data
- Central Tendency Measures to summarize data
- Mean, Median and Mode
- When to use what

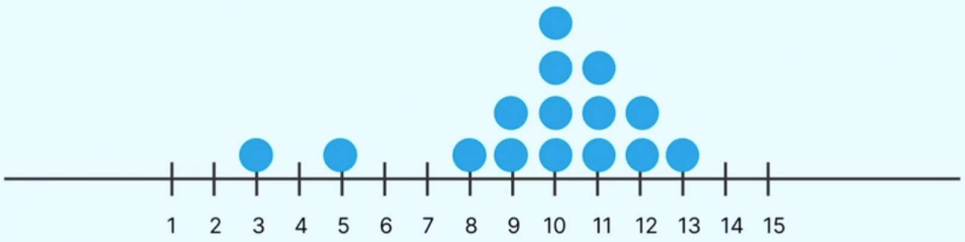
How To Summarize?



Central Tendency Measures



Mean, Median and Mode



Median = midpoint of { 3, 5, 8, 9, 9, 10, 10, 10, 10, 11, 11, 11, 12, 13 } = 10

When to use what?

- What's the cost per unit
- What's the daily revenue
- What's the temperature

- Who is the most popular movie star
- Which is the most popular movie genre
- Which is the most visited restaurant in the town

Mean Vs Median

Mean

- Less reliable when data is skewed or when dispersion is high
- Works with most statistical methods and therefore convenient

Median

- More reliable when data is skewed or when dispersion is high
- Few statistical methods use median

What did we learn?

- Need to summarize data
- Central Tendency Measures to summarize data
  - Mean: arithmetic center point
  - Median: geometric center point
  - Mode: most occurring value
- When to use what
  - Mean and median to summarize numerical data
  - Mode for categorical data

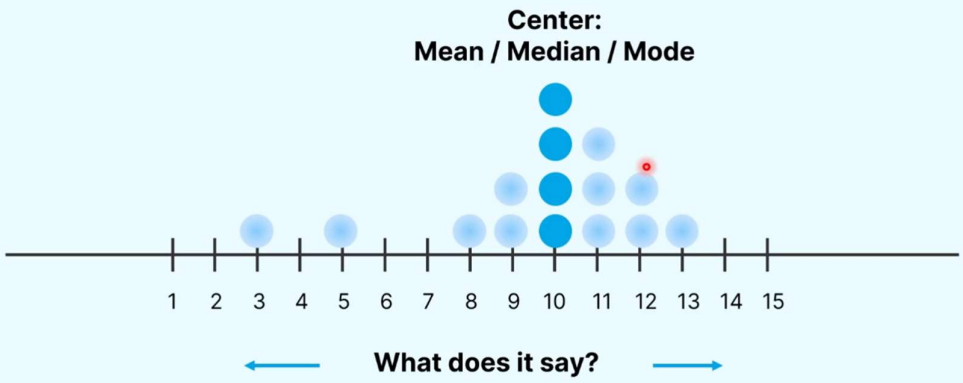
Module 2 | Topic 2 | Video 2

Dispersion Measures

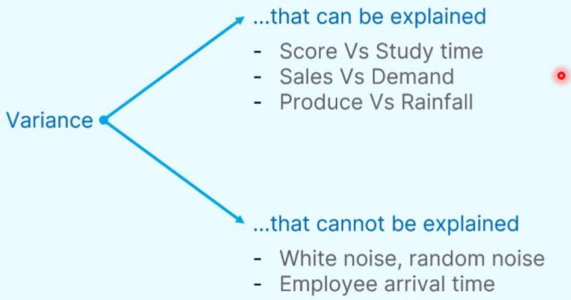
Learning Objective

- Why central tendency measures are not enough
- What is dispersion?
- How to measure dispersion?
- Types of dispersion measures

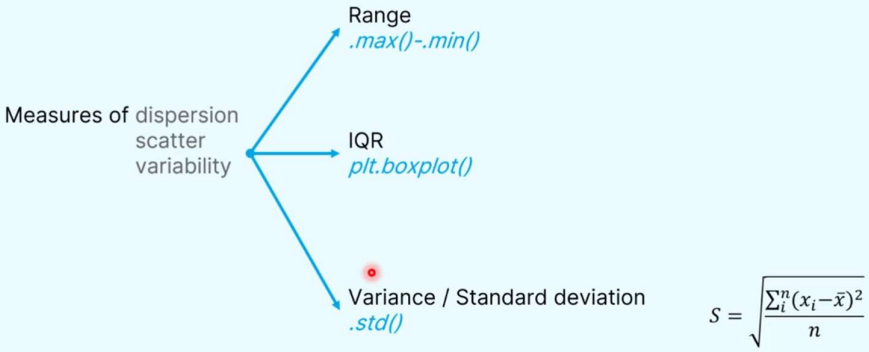
Dispersion



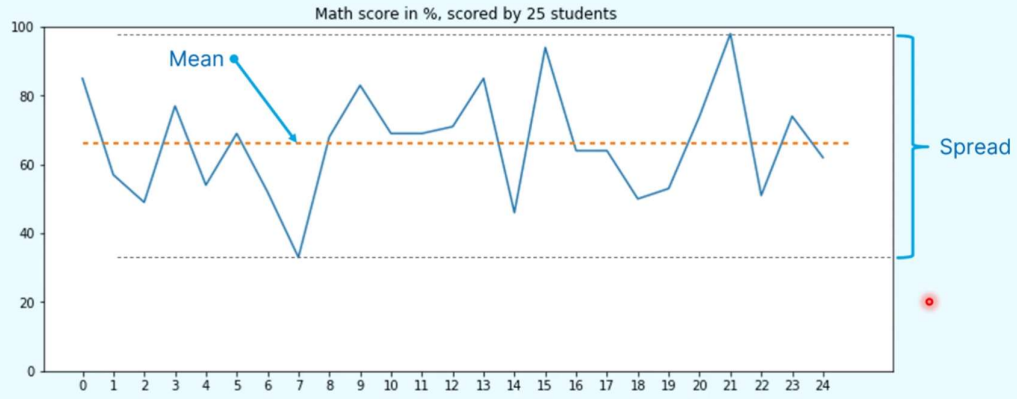
Sources of variance. a.k.a. Spread, Deviation



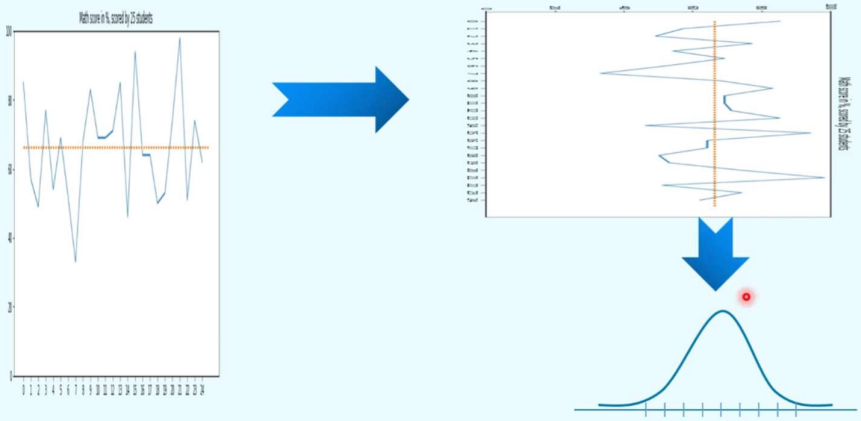
Measures of Dispersion



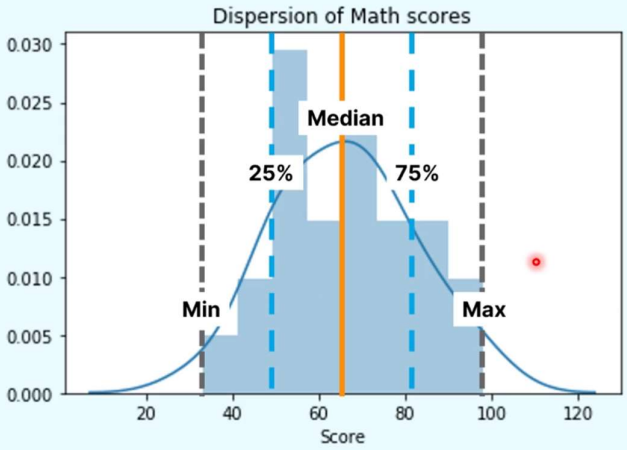
How to Summarize?



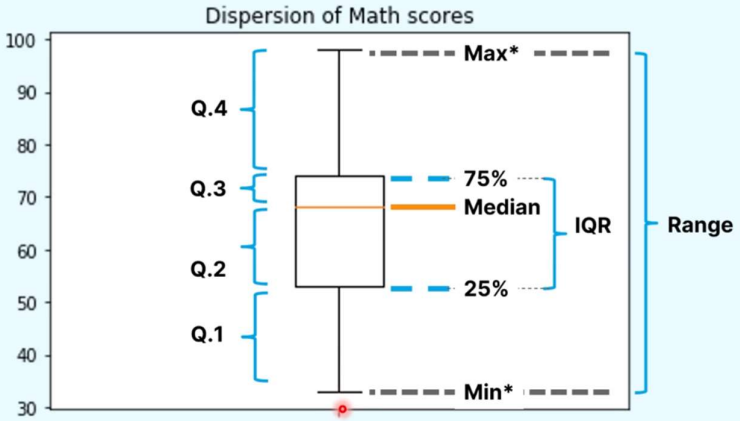
How to Summarize?



Histogram



Box Plot and IQR



*End of whiskers are not always max and min

What did we learn?

- Why central tendency measures are not enough to summarize data and not completely reliable when dispersion is high
- Dispersion measures the spread or variability of the data
  - Range
  - Interquartile range

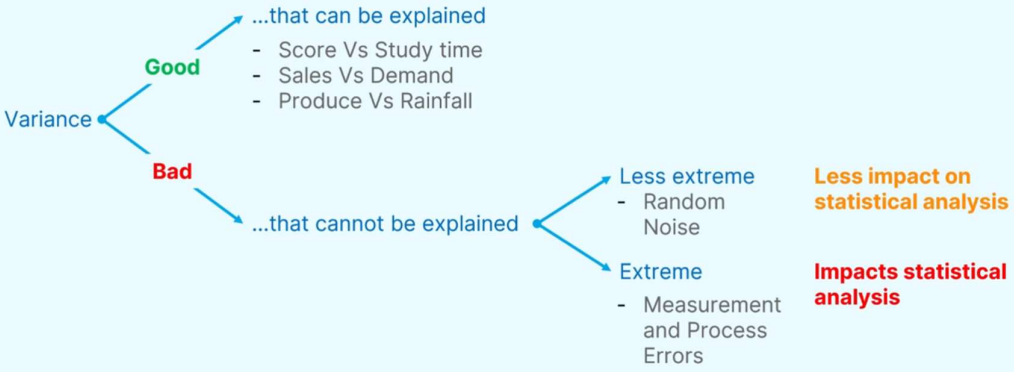
Module 2 | Topic 3 | Video 5

Outliers

Learning Objective

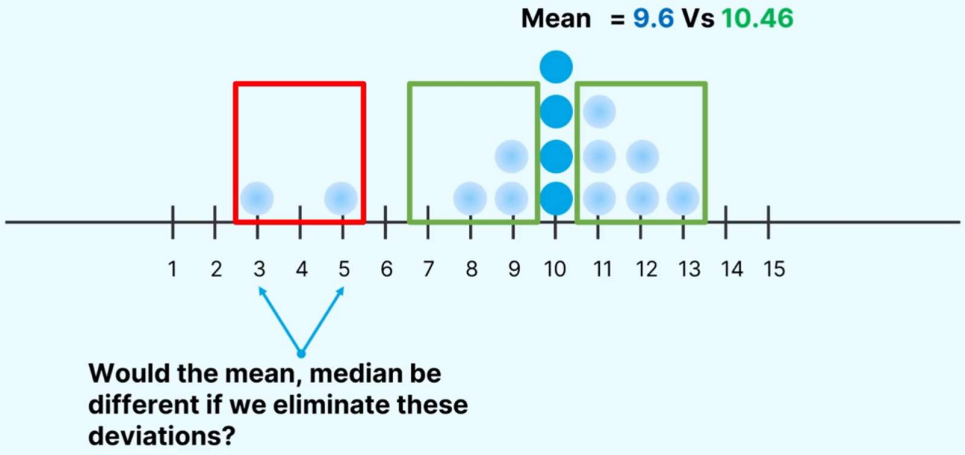
- What outliers are?
- How to detect outliers?
- How to deal with them?

Outliers and Variance

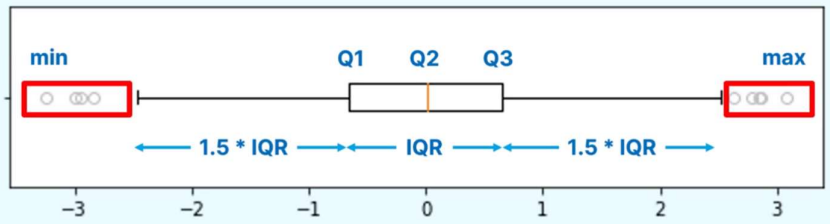


Outliers are extreme deviations. They are significantly different from the remaining observations. May indicate error in measurement or in the process.

Spread: Good or Bad?



Inter Quartile Range and Outliers



IQR = Q3 – Q1

Lower Limit = Q1 – 1.5*IQR

Upper Limit = Q3 + 1.5*IQR

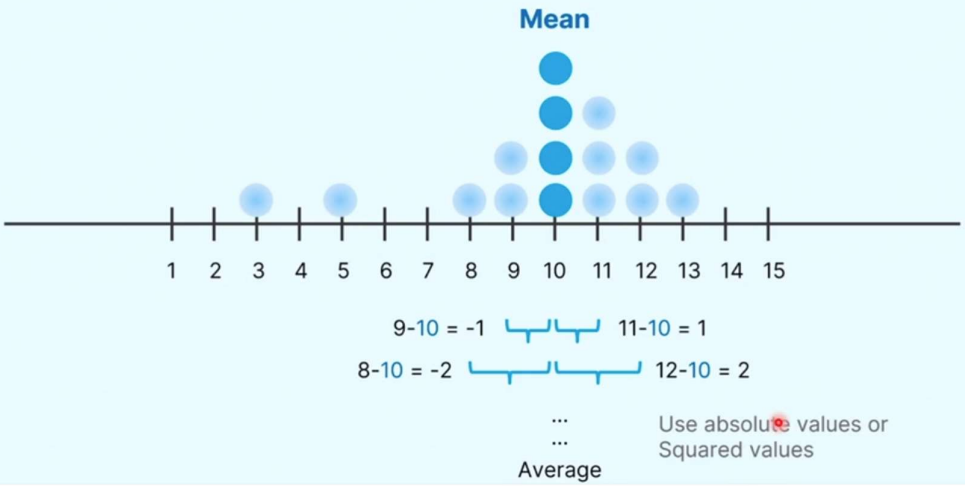
What did we learn?

- Outliers are extreme values
- Inter quartile ranges and boxplots are useful in identifying the outliers
- Whiskers show the upper and lower limits
- Values outside the limits can be eliminated before analysis

Module 2 | Topic 3 | Video 6

Variance and Standard Deviation

How to quantify dispersion?



Variance and Standard Deviation

$$Var = \frac{\sum_i^n (x_i - \bar{x})^2}{n}$$

$$\sigma = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n}}$$

Standard deviation is the **unit distance** by which an observation (data point) is away **from the mean** of population it belongs to

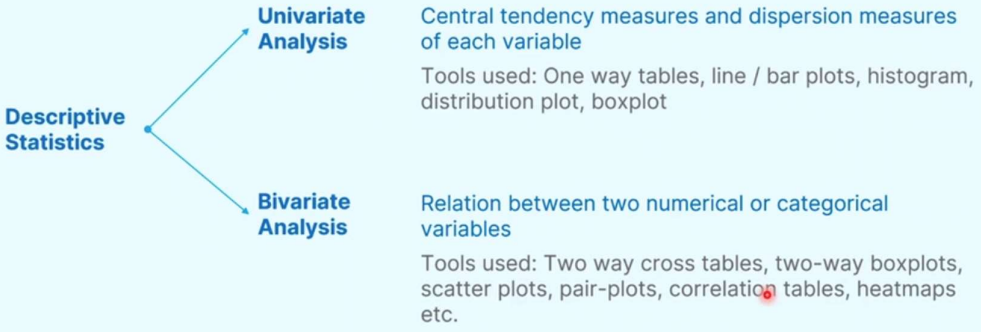
Module 2 | Topic 4 | Video 1

Introduction to Bivariate Analysis

Learning Objective

- What is bivariate analysis?
- Common tools for bivariate analysis: two-way boxplots and scatterplots

Tools for Descriptive Statistics



Bivariate Analysis

- Objective is to understand relation between two variables
- Often used to study the effect of one variable (independent) on the other variable (dependent)

		Impacted Variable *Dependent*	
		Categorical	**Numerical**
Impacting Variable *Independent*	**Categorical**	Two way frequency table and Heatmaps	
	Numerical		

Two way frequency table and heatmap

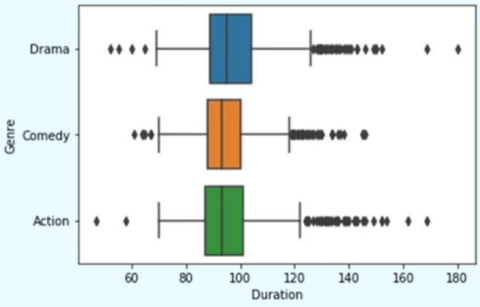
Ratings_cat	High	Low	Medium
Year			
2001	0.089820	0.520958	0.389222
2002	0.112903	0.564516	0.322581
2003	0.107914	0.597122	0.294964
2004	0.093333	0.580000	0.326667
2005	0.085714	0.619048	0.295238
2006	0.065068	0.630137	0.304795
2007	0.130282	0.542254	0.327465
2008	0.080769	0.634615	0.284615
2009	0.070336	0.663609	0.266055
2010	0.070922	0.602837	0.326241
2011	0.091463	0.664634	0.243902
2012	0.100000	0.643243	0.256757
2013	0.060209	0.675393	0.264398
2014	0.053435	0.681934	0.264631
2015	0.067010	0.667526	0.265464
2016	0.056000	0.704000	0.240000
2017	0.062176	0.668394	0.269430
2018	0.078652	0.651685	0.269663

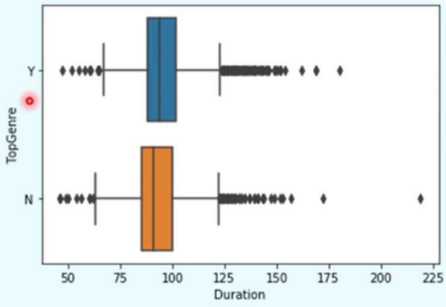
Bivariate Analysis

- Objective is to understand relation between two variables
- Often used to study the effect of one variable (independent) on the other variable (dependent)

		Impacted Variable *Dependent*	
		Categorical	**Numerical**
Impacting Variable *Independent*	**Categorical**	Two way frequency table and Heatmaps	Box plot
	Numerical		

Two-way box plots





Bivariate Analysis

- Objective is to understand relation between two variables
- Often used to study the effect of one variable (independent) on the other variable (dependent)

		Impacted Variable *Dependent*	
		Categorical	**Numerical**
Impacting Variable *Independent*	**Categorical**	Two way frequency table and Heatmaps	Box plot
	Numerical	Discriminant Analysis	Scatter plots, Pair plots, Variance – Covariance matrix, Correlation matrix, Heatmaps

What did we learn?

- In Bivariate analysis the Objective is to understand relation between two variables. Often used to study the effect of one variable (independent) on the other variable (dependent)
- Common tools for bivariate analysis of variables: two-way boxplots and scatterplots

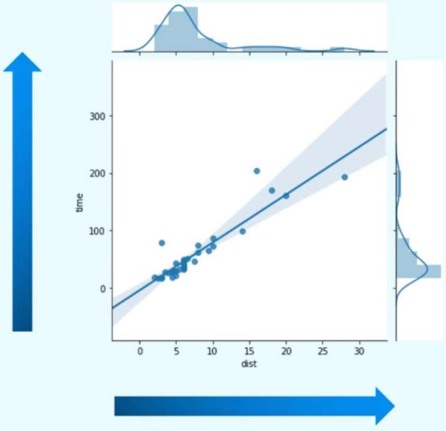
Module 2 | Topic 4 | Video 2

Covariance and Correlation

Learning Objective

- What covariance and correlation are?
- Mathematical intuition
- Calculating covariance and correlation in python

Covariance



How to mathematically model this relationship?

We know that variance is the mean squared distance from the variable's own mean

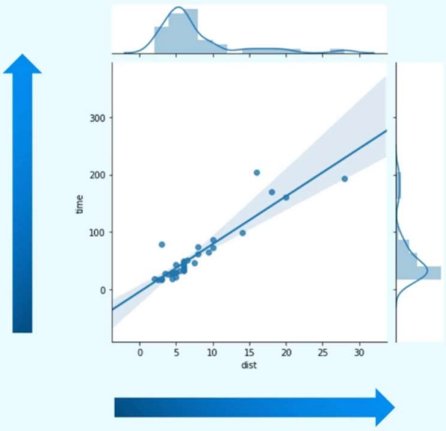
$$Var = \frac{\sum_i (x_i - \bar{x})^2}{N - 1}$$

Similarly we can calculate the covariance by multiplying the distance from mean for the variable pairs

$$CoVar = \frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N - 1}$$

Positive values indicate positive relation and Negative values indicate negative relation

Correlation



Correlation is a standardized measure

$$Correlation = \frac{\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sigma_x \sigma_y}$$

Values are between -1 and + 1

-1 indicate perfect negative correlation and
+1 indicate perfect positive correlation

What did we learn?

- Covariance is a measure to quantify the relationship between two numerical variables
- Correlation does the same but it is a standardized measure with values ranging between -1 and + 1
- -1 means negative correlation and +1 means positive correlation