Basic Statistics and Descriptive Statistics

Course: Foundations of Statistics for Data Analytics and Machine Learning Using Excel

@ **DV Data Analytics, Bangalore**

Instructor:
Dr. Ujwal Deep Kadiyam

Introduction To Statistics

Why is Statistics Needed?

A pharmaceutical company develops a new drug to lower blood pressure.

- How do they determine if the drug is effective?
- They conduct a study with two groups: one receiving the drug, the other a placebo.
- Statistical analysis helps determine if the observed difference in blood pressure is due to the drug or random chance.
- Without statistics, we cannot confidently say whether the drug works.

Why is Statistics Needed?

Suppose a company wants to know if a new marketing strategy increases sales.

- Without statistics, they might rely on intuition.
- Sales can change due to many factors, not just marketing.
- Statistics helps determine if the observed increase is significant.

What is Statistics?

- Statistics is the science of collecting, analyzing, interpreting, and presenting data.
- It helps in making informed decisions based on data rather than guesswork.
- This is done by uncovering underlying patterns in the data.
- Used in various fields like Engineering, economics, medicine, and business.

Statistics

It is the science of Data.

Statistical Analysis Tools

- **R** (Free) - Widely used for statistical computing and graphics.
- **Python (with libraries like NumPy, SciPy, Pandas, and Statsmodels)** (Free) - Popular for data analysis and machine learning.
- **SPSS** (Commercial) - Used extensively in social sciences and business analytics.
- **SAS** (Commercial) - Powerful tool for data management and analytics.
- **Excel** (Commercial, but widely available) - Basic statistical functions for general users. Used for Data collection.
- **Stata** (Commercial) - Frequently used in economics, epidemiology, and social sciences.
- **MATLAB** (Commercial) - Used in engineering and scientific computing.
- **Minitab** (Commercial) - Popular for quality improvement and industrial statistics.

How to Master Statistics and Machine learning

Introduction

- Statistics and machine learning can seem hard at first, but you can definitely learn them!
- This presentation will give you practical tips and techniques to make the learning process easier.
- Stay persistent, stay curious, and remember that everyone struggles at first—you're not alone!

Start with the Basics

- Begin by mastering the fundamental concepts: probability, distributions, and regression.
- Focus on understanding why things work, not just memorizing formulas.
- Once you're comfortable with the basics, you can move on to more complex topics like machine learning algorithms.

Make Connections with Everyday Life

- Relate concepts to real-life situations: think of how Netflix recommends shows based on your preferences, or how spam filters work.
- Understanding how machine learning and statistics apply to everyday tools will make them easier to grasp.
- Create your own examples, like predicting the weather or analyzing a social media trend—this will help solidify your learning.

Break Down Complex Topics

- Don't get overwhelmed by big topics—break them into smaller, manageable parts.
- For example, start by understanding simple linear regression before diving into more complex models.
- Focus on mastering one small piece at a time, then move forward as you build your confidence.

Use Visualizations

- Visualize the data and the models you're working with to make abstract concepts clearer.
- Plot graphs, decision boundaries, and errors to see how things are working.
- Use tools like Jupyter Notebooks to experiment and see results immediately—this makes it easier to learn.

Practice and Experiment

- Practice regularly, even with small datasets and easy tasks.
- Apply what you're learning to real-world problems—build simple projects, such as predicting exam scores based on study hours.
- Don't be afraid to make mistakes. You'll improve by iterating on your models and experimenting with new ideas.

Learn from Mistakes

- Don't fear failure—mistakes are part of the learning process.
- When things don't work as expected, take the time to understand why. Debugging and problem-solving will help you grow.
- If a model doesn't work, try tweaking it. Every mistake is an opportunity to learn and improve.

Use Clear Resources

- Use beginner-friendly books, online courses, and tutorials to guide you.
- For coding, stick to tools like Python libraries (e.g., scikit-learn) to implement concepts.
- Don't try to learn everything at once—take it step by step and make sure you understand one concept before moving on.

Learn with Others

- Team up with classmates or online peers to work on projects together.
- Join discussion forums like Stack Overflow or Kaggle to ask questions and share knowledge.
- Learning with others will motivate you and help you understand concepts better.

Understand the Power of Machine Learning

- Keep in mind that machine learning can solve real-world problems in fields like healthcare, finance, and entertainment.
- Seeing how these tools apply to actual situations will keep you motivated and help you understand the potential of what you're learning.
- You're learning skills that are in high demand and can make a significant impact.

Keep Pushing Forward

- Celebrate small wins—whether it's completing a project or understanding a tough concept.
- Stay patient and persistent. Mastery takes time, and every step you take is progress.
- Remember, you'll improve with each challenge you face. Stick with it!

Focus on Real-World Problems

- Apply your knowledge to practical problems, such as analyzing data from a sports game or building a simple recommendation system.
- Focusing on real-world applications makes learning more interesting and relevant.
- Solving real problems with machine learning and statistics will help you see the value of what you're learning.

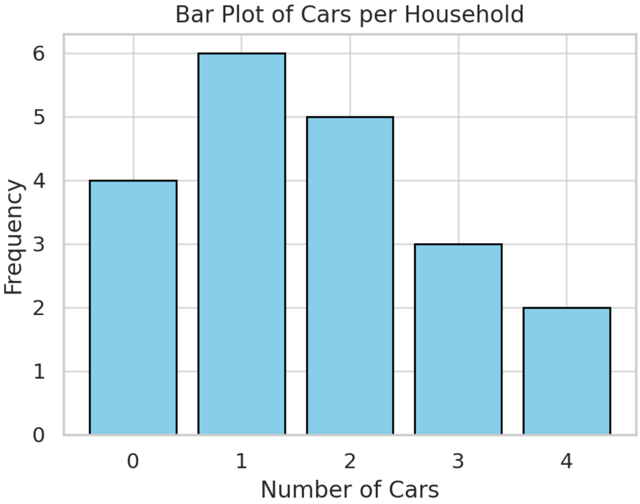
Conclusion

- Statistics and machine learning may seem difficult, but with persistence and the right approach, you can master them.
- Focus on understanding the basics, practicing regularly, and learning from your mistakes.
- The skills you develop will open many doors—keep learning and stay curious!

Statistical Visualizations of Data

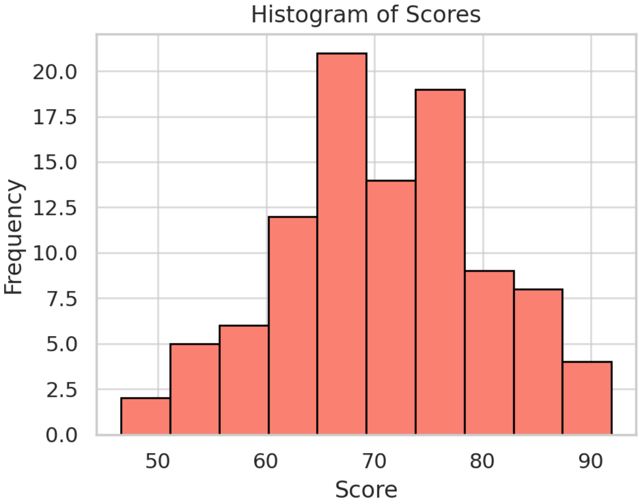
Statistical Visualizations: 1. Bar Plot

- Used to compare categories using rectangular bars.
- Example: Number of cars per household.



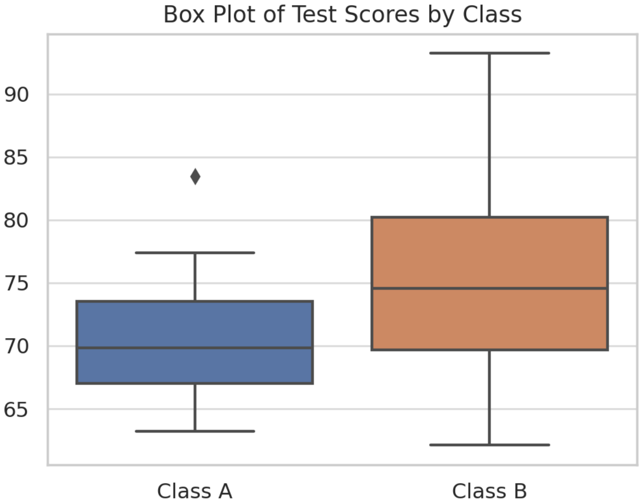
2. Histogram

- Shows the distribution of numerical data using bins.
- Example: Student test scores.



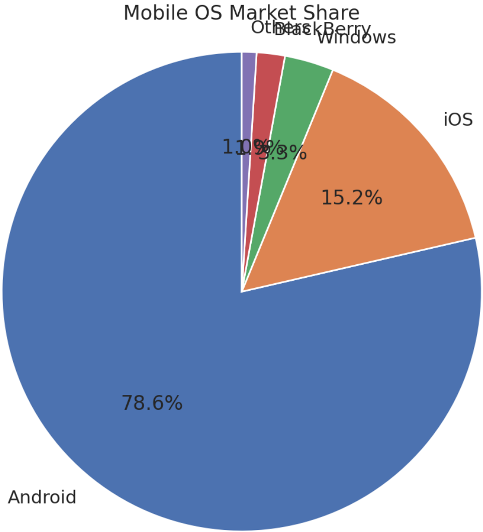
3. Box Plot

- Visualizes the distribution, median, quartiles, and outliers.
- Example: Compare test scores across classes.



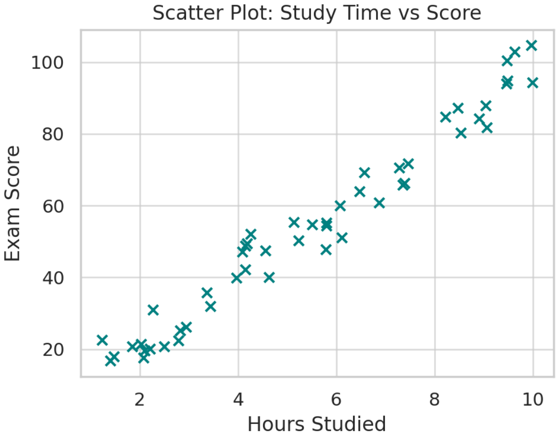
4. Pie Chart

- Displays proportions of a whole.
- Example: Market share of mobile OS.



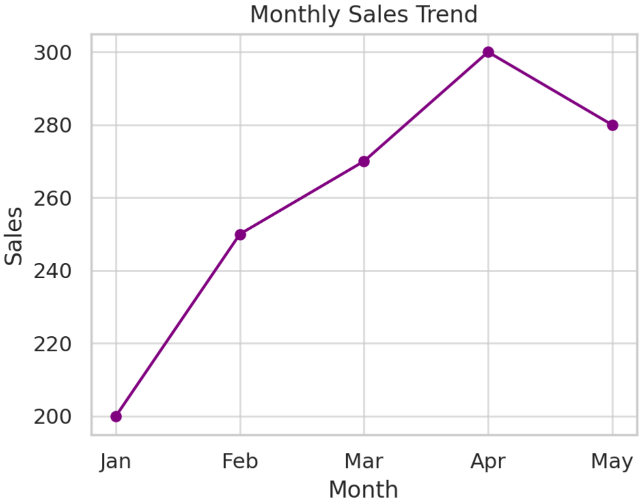
5. Scatter Plot

- Shows relationship between two continuous variables.
- Example: Hours studied vs. exam score.



6. Line Plot

- Tracks changes over time or sequence.
- Example: Monthly sales data.



Summary of Plots

- **Bar Plot:** Category comparison
- **Histogram:** Distribution of numerical data
- **Box Plot:** Summary statistics and outliers
- **Pie Chart:** Proportional breakdown
- **Scatter Plot:** Relationship between variables
- **Line Plot:** Trend over time

Types of Statistics

Types of Statistics

- **Descriptive Statistics**: Summarizes data using measures like mean, median, and standard deviation.
  - Descriptive statistics summarize and organize data so that it can be easily understood.
  - Example: A teacher calculates the average test score of a class to summarize student performance.
- **Inferential Statistics**: Draws conclusions and makes predictions based on data samples.
  - Inferential statistics allow us to make predictions or inferences about a population based on a sample.
  - Involves hypothesis testing, confidence intervals, and regression analysis.
  - Example: A survey of 500 people is used to predict the voting preferences of an entire country.

Population And Sample

Population and Sample

Population

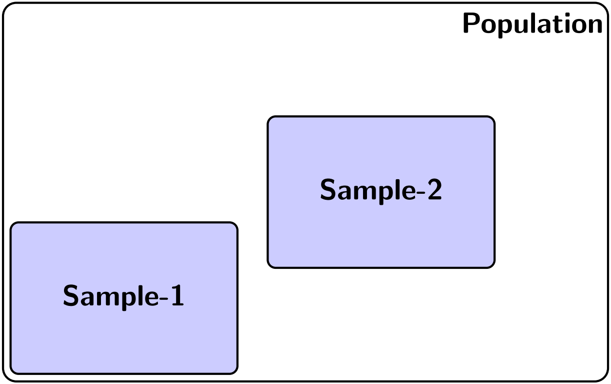
It is the collection or set of all objects or measurements that are of interest to the collector.

- **Population**: The entire group that is being studied.
- **Sample**: A subset of the population used to make inferences.
- Example:
  - A university wants to study the average GPA of all its students (population).
  - Instead of surveying all students, they select 200 students at random (sample).
  - The statistics from the sample help estimate characteristics of the entire population.

Sample

It is a subset of data selected from a population. The **size** of a sample is the number of elements in it.

Diagram: Population and Sample



- Population: All Individuals of interest (e.g., all students in a country).
- Sample: A subset of the population (e.g., 500 surveyed students).

Characteristics of a Good Sample

- **Representative**: The sample should reflect the diversity of the population.
- **Randomly Selected**: Ensures each member of the population has an equal chance of being included.
- **Sufficiently Large**: A larger sample size increases accuracy and reduces sampling error.
- **Unbiased**: Avoids systematic errors that could misrepresent the population.
- **Example**: If studying student GPAs at a university, selecting students from different departments, grade levels, and backgrounds ensures the sample is representative.

Steps For Achieving the Objective of Inferential Statistics (Mendenhall)

- **Defining the Population**: Clearly specify the group being studied.
- **Selecting a Representative Sample**: Ensure the sample accurately reflects the population.
- **Applying Appropriate Statistical Techniques to analyze sample information**: Use an appropriate method of analysis to obtain the information that the sample contains.
- **Making Inferences**: Draw conclusions about the population based on the analysis of the sample.
- **Assessing the Reliability of Inferences**: Evaluate the accuracy and precision of conclusions using probability theory.

Application of Statistics: An Example Illustration

Application of Statistics

- Statistics is a powerful tool for making sense of data.
- Helps in decision-making under uncertainty.
- It allows businesses, scientists, and policymakers to make informed decisions.
- Essential in business for market analysis and forecasting.
- Understanding statistics helps in critical thinking and problem-solving.
- **Statistical modeling** involves three key components:
  - **Design**
  - **Description**
  - **Inference**
- Let's explore each with a real-world example.

Scenario: Online Learning Tool Effectiveness

- **Question:** Do students using an online learning platform perform better in final exams?
- **Population:** High school students taking Mathematics.
- **Groups:**
  - Group A: Used the platform
  - Group B: Did not use the platform

Step 1: Design

- **Goal:** Determine the effect of the online tool on student performance.
- Plan:
    - Collect final exam scores
    - Record platform usage (hours/week)
    - Include GPA and attendance as control variables
- **Importance:** Ensures data collected is relevant and reliable.

Step 2: Description

- Use descriptive statistics:
    - Mean scores, standard deviation
    - Boxplots, histograms
- **Example:**
    - Group A: Mean = 82.4, SD = 6.3
    - Group B: Mean = 75.1, SD = 7.0
- **Importance:** Summarizes patterns in the data before inference.

Step 3: Inference

- Use statistical tests (e.g., t-test):
  - Is the difference in scores statistically significant?
  - Confidence intervals for mean difference
  - Regression to predict performance
- **Result:**
  - Significant difference found ($p < 0.01$)
  - Platform users performed better
- **Importance:** Supports conclusions and predictions beyond sample data.

Summary of the Three Steps

Step	What It Does	Why It's Useful
Design	Plan data collection	Ensures validity and relevance
Description	Summarize the data	Reveals patterns and trends
Inference	Draw conclusions	Enables predictions and decisions

Table: Three Key Steps in Statistical Modeling

- Each step—Design, Description, Inference—is essential in a sound statistical investigation.
- Together, they help answer real-world questions with evidence-based confidence.
- Always begin with a clear question and build from there!

Types of Data

Types of Data

- Data is the foundation of statistical analysis.
- It is crucial to understand different types of data to apply appropriate statistical methods.
- Broadly categorized into:
    - **Qualitative (Categorical) Data**
    - **Quantitative (Numerical) Data**

Types of Data (Ramachandran and Tsokos)

Data can be classified into different types based on characteristics:

- **Qualitative (Categorical) Data**: Data that represents categories or groups.
  - Example: Gender (Male, Female), Eye color (Blue, Green, Brown)
  - Describes attributes or characteristics. Cannot be measured numerically.
  - Types:
    - **Nominal**: No natural order (e.g., Gender, Nationality)
    - **Ordinal**: Has a meaningful order (e.g., Satisfaction rating: Poor, Fair, Good)
- **Quantitative (Numerical) Data**: Data that represents numeric values.
  - Represents measurable quantities.
  - Can be subjected to arithmetic operations.
  - Types:
    - **Discrete**: Countable values (e.g., Number of students)
    - **Continuous**: Measurable on a scale (e.g., Height, Weight)

Dataset with Different Data Types

Person	Gender (Nominal)	Satisfaction (Ordinal)	Children (Discrete)	Height (Continuous)
1	Male	Very Satisfied	2	175.3 cm
2	Female	Satisfied	0	160.5 cm
3	Other	Neutral	1	168.2 cm
4	Female	Unsatisfied	3	158.9 cm
5	Male	Very Unsatisfied	2	182.7 cm

Table: Example dataset containing all 4 types of data

Scales of Measurement

- **Nominal Scale**: Categorizes data without a specific order.
  - Example: Types of fruits (Apple, Banana, Orange).
- **Ordinal Scale**: Categorizes data with a meaningful order but no fixed intervals.
  - Example: Customer satisfaction levels (Satisfied, Neutral, Dissatisfied).
- **Interval Scale**: Numeric data with equal intervals but no true zero.
  - Example: Temperature in Celsius or Fahrenheit.
- **Ratio Scale**: Numeric data with equal intervals and a true zero.
  - Example: Height, weight, and distance.

Example: Understanding Scales

- **Nominal**: Types of pets (Dog, Cat, Bird).
- **Ordinal**: Education levels (High School, Bachelor's, Master's, PhD).
- **Interval**: Years (2000, 2010, 2020) where differences matter but there is no absolute zero.
- **Ratio**: Age in years (Has a true zero and equal intervals).

Example Dataset with Nominal, Ordinal, Interval, and Ratio Scales

Person	Blood Type (Nominal)	Pain Level (Ordinal)	Temperature (Interval)	Income ($) (Ratio)
1	A+	Severe	38°C	45000
2	B-	Moderate	36.5°C	52000
3	AB+	Mild	37°C	60000
4	O-	None	36°C	0
5	A-	Moderate	39°C	72000

Table: Example dataset showing all four measurement levels

Summary Table

Type	Sub-type	Description	Examples
Qualitative	Nominal	No order or ranking	Gender, Color, Nationality
	Ordinal	Ordered categories	Education level, Satisfaction
Quantitative	Discrete	Countable numbers	Number of children, Cars owned
	Continuous	Measurable, infinite values	Height, Temperature, Weight

Table: Summary of Types of Data

- Knowing data types helps in selecting the right statistical tools.
- Always begin analysis by identifying the type of data you're working with.

Excel For Data analysis

Data Analysis tool pack in excel

- The Data Analysis ToolPack is an add-in for Microsoft Excel.
- It provides statistical and data analysis tools , such as regression, t-tests, histograms, and more.
- It is widely used in finance, research, and engineering for quick analysis.

Key Features

- **Descriptive Statistics** – Summary statistics (mean, variance, standard deviation, etc.).
- **Regression Analysis** – Linear regression models for predictive analysis.
- **T-Tests & ANOVA** – Statistical hypothesis testing tools.
- **Histograms & Sampling** – Data visualization and sampling methods.
- **Moving Averages & Exponential Smoothing** – Time series analysis.

Installation and Verification

- **Installation For Windows:**

1. Open Excel and go to File − > Options .
2. Select Add-ins from the left panel.
3. In the Manage box, choose Excel Add-ins and click Go .
4. Check Analysis ToolPack and click OK .

- **Installation For Mac:**

1. Open Excel , go to Tools − > Add-ins .
2. Check Analysis ToolPack and click OK .

- **Verification:**

  - Once installed, go to Data tab in Excel.
  - You should see a new section called Data Analysis .
  - Click Data Analysis to access available tools.

- **Lab: Excel-1a- summary statistics**
- **Lab: Excel-1b- Formulae**

Formula

Sum

1. Drag cells for Sum function.
2. Enter formula in formula bar.
3. use formula function.

Range

1. Range of cells.
2. Name range of cells. Edit names.

Relative and Absolute References

Relative Reference

- Default behavior in Excel.
- Adjusts automatically when copied to another cell.
- Example: If `B2` contains `=A1`, copying it to `C2` changes it to `=B1`.

Absolute Reference

- Fixed cell reference using $ symbol.
- Does not change when copied to another cell.
- Example: `=$A$1` always refers to cell `A1`, even when copied.
- Mixed references: `=A$1` (row fixed) or `=$A1` (column fixed).

Converting Relative to Absolute Reference

- Select the cell with the formula.
- Click inside the formula bar.
- Press `F4` to cycle through reference types:
    - `A1` (relative)
    - `$A$1` (absolute)
    - `A$1` (row absolute)
    - `$A1` (column absolute)

Basic Descriptive Statistics

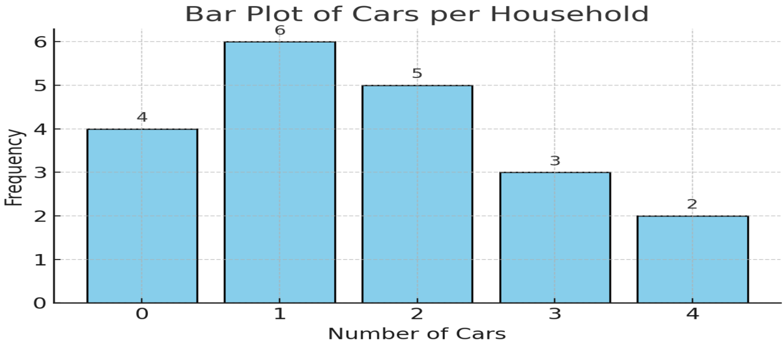
Basic Descriptive statistics

- In 20 homes, people were asked how many cars were registered to their households. The results were recorded as follows: 1, 2, 1, 0, 3, 4, 0, 1, 1, 1, 2, 2, 3, 2, 3, 2, 1, 4, 0, 0

Number of Cars	**Frequency**
0	4
1	6
2	5
3	3
4	2

Table: Frequency of Registered Cars in 20 Households

Basic Descriptive statistics



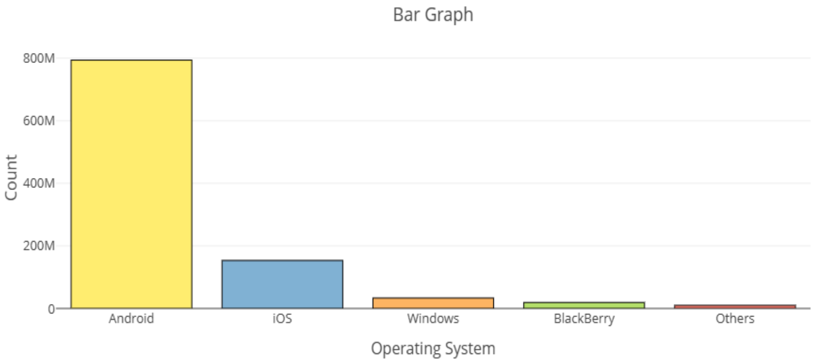
- How many houses have 1 car, 2 cars.. ?
- How many houses have at least 3 cars?
- How many houses have at most 2 cars ?

Example: Frequency Table: Operating Systems

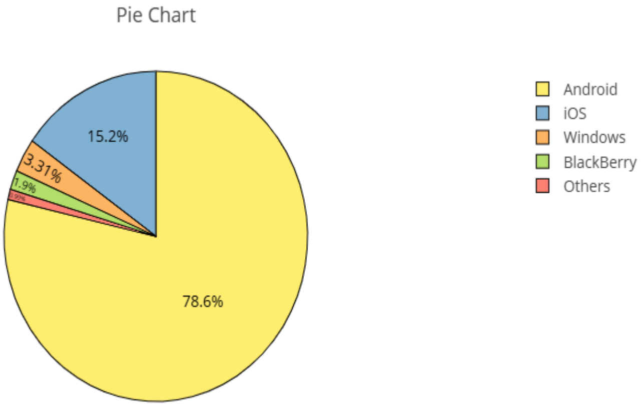
Operating System	**Frequency**	**Proportion**	**Percent (%)**
Android	793,600,000	0.7861	78.61
iOS	153,400,000	0.1519	15.19
Windows	33,400,000	0.0331	3.31
BlackBerry	19,200,000	0.0190	1.90
Others	10,000,000	0.0099	0.99
Total	**1,009,600,000**	**1.0000**	**100.00**

Table: Distribution of Mobile Operating Systems

Example


Bar Graph
Count
800M
600M
400M
200M
0
Android
iOS
Windows
BlackBerry
Others
Operating System

Example


Pie Chart
15.2%
3.31%
1.9%
0.99%
78.6%
Android
iOS
Windows
BlackBerry
Others

Key Descriptive Statistics in Data Science

- **Measures of Central Tendency**: Mean, Median, Mode.
- **Measures of Dispersion**: Variance, Standard Deviation, Range, Interquartile Range.
- **Shape of Distribution**: Skewness, Kurtosis.
- **Relationships Between Variables**: Correlation, Covariance.

Basic Descriptive Statistics: Mean and Median

Central Tendencies: Mean and Median

Mean

- The arithmetic mean
- Sum of values/ Count of values
- Gives a quick idea on average of a variable

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Mean in Python:

- Import “Census Income Data/Income_data.csv”.
- gain_mean = Income[“capital-gain”].mean()

Mean and extreme data points

- 10.18.Guess the mean: $\{1.5, 1.7, 1.9, 0.8, 0.8, 1.2, 1.9, 1.4, 99, 0.7, 1.1\}$.
- 90% of data is less than two. Mean is _____________. It doesn't make sense.
- There is an unusual value in the above data vector i.e 99.
- It is known as outlier.
- Average Income in India: Ambani, Adani etc are outliers.
- Outliers are not similiar to most of the data. They are not part of the data.
- Eg. Age:-240.

Lab: Excel-1c- Averages

Median

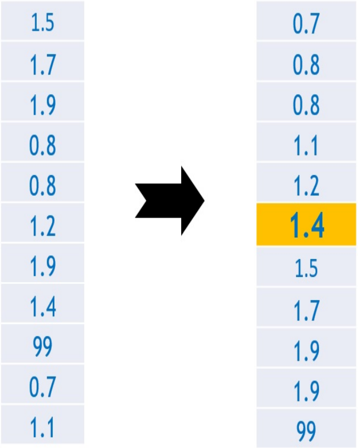
- Mean is not a good measure in presence of outliers
- Mean is not the true middle value in presence of outliers. Mean is very much effected by the outliers.
- We use median, the true middle value in such cases
- Sort the data either in ascending or descending order
- $x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_n$
- If the size n is:
  - odd : the median is the value at position p where

$$p = \frac{n+1}{2}, \quad \tilde{x} = x_p.$$

  - even : the median is the average of the values at positions p and p + 1 where

$$p = \frac{n}{2}, \quad \tilde{x} = \frac{x_p + x_{p+1}}{2}.$$

Median and Outliers



1.5
1.7
1.9
0.8
0.8
1.2
1.9
1.4
99
0.7
1.1

0.7
0.8
0.8
1.1
1.2
1.4
1.5
1.7
1.9
1.9
99

- Mean of the data is 2
- Median of the data is 1.4
- Even if we have the outlier as 990, we will have the same median
- Median is a positional measure, it doesn't really depend on outliers

Median Calculation: Even vs Odd Number of Elements

- **Example 1: Odd Number of Elements**

$$\{5, 8, 12, 14, 18, 21, 24\}$$

  - Number of elements: 7 (odd)
  - Median: Middle value = 14

- **Example 2: Even Number of Elements**

$$\{5, 8, 12, 14, 18, 21\}$$

  - Number of elements: 6 (even)
  - Median: Average of the 3rd and 4th values $= \frac{12+14}{2} = 13$

Median and Outliers

Median and Median

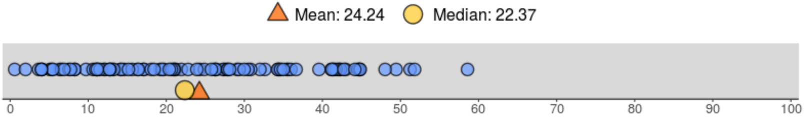
- Mean is calculated using actual data values.
- Median is a positional measure.

- **Lab: Excel-1d- Median**

Median and Outliers

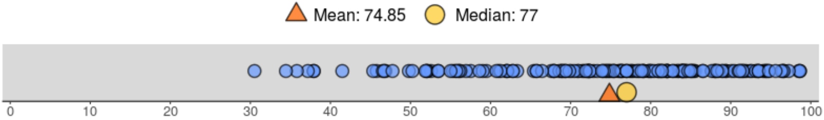
- When there are no outliers, mean and median will be nearly equal.
- When mean is not equal to median it gives us an idea about presence of outliers in the data

Mean vs Median: Visual Representation



Descriptive Statistics:

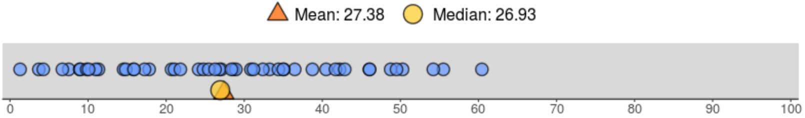
Sample Size	Mean	Standard Deviation	Minimum	1st Quartile	Median	3rd Quartile	Maximum	IQR
100	24.24	14.07	0.58	12.59	22.37	35.11	58.58	22.52



Descriptive Statistics:

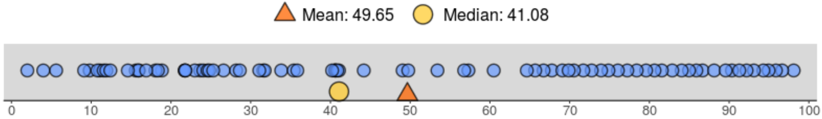
Sample Size	Mean	Standard Deviation	Minimum	1st Quartile	Median	3rd Quartile	Maximum	IQR
190	74.85	15.08	30.51	67.65	77.00	85.11	98.65	17.46

Mean vs Median: Visual Representation



Descriptive Statistics:

Sample Size	Mean	Standard Deviation	Minimum	1st Quartile	Median	3rd Quartile	Maximum	IQR
50	27.38	15.26	1.29	15.07	26.93	38.17	60.44	23.10



Descriptive Statistics:

Sample Size	Mean	Standard Deviation	Minimum	1st Quartile	Median	3rd Quartile	Maximum	IQR
83	49.65	30.22	1.99	22.40	41.08	77.82	100.06	55.42

Basic Descriptive Statistics: Mode

What is Mode?

- **Definition:** The mode is the value that appears most frequently in a dataset.
- A dataset may have:
  - No mode (if all values appear with equal frequency).
  - One mode (Unimodal).
  - Multiple modes (Bimodal or Multimodal).
- **Example:**
  - Dataset: $\{3, 7, 3, 2, 9, 3, 7, 10, 7\}$
  - Modes: 3 and 7 (Bimodal distribution)

- **Lab: Excel-1e- Mode**

Applications and Limitations of Mode

Applications:

- **Market Research:** Identifying the most popular product sold.
- **Education:** Determining the most common grade in an exam.
- **Medicine:** Finding the most frequently diagnosed disease.
- **Traffic Analysis:** Identifying the most common accident locations.

Example:

- Survey on favorite ice cream flavors:
- Data: {"Vanilla", "Chocolate", "Strawberry", "Chocolate", "Chocolate", "Vanilla"}
- Mode: **Chocolate** (Most frequently chosen flavor)

Limitations:

- **Not Unique:** A dataset can be multimodal.
- **May Not Exist:** If all values appear equally, no mode exists.
- **Not Useful for Continuous Data:** Unlike mean/median, mode is less informative for non-repeating values.
- **Ignores Magnitude:** Mode does not consider the value sizes (e.g., in salaries, mean is often preferred).

Basic Descriptive Statistics: Range, Minimum and Maximum

Range, Minimum, and Maximum

Definitions:

- **Minimum:** The smallest value in a dataset.
- **Maximum:** The largest value in a dataset.
- **Range:** The difference between the maximum and minimum values.

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

Example:

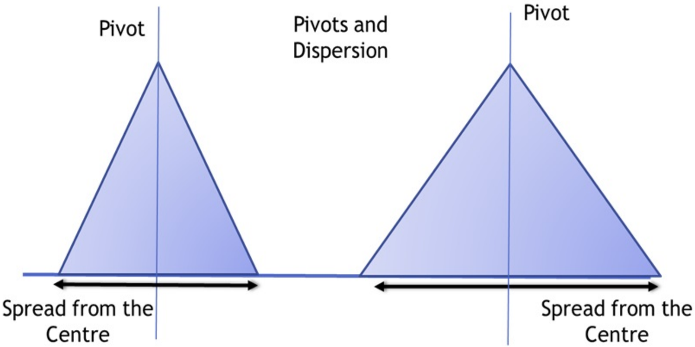
- Given dataset: $\{12, 5, 8, 19, 3, 7, 15\}$
- $\text{Minimum} = 3$
- $\text{Maximum} = 19$
- $\text{Range} = 19 - 3 = 16$

Importance:

- Measures the spread of data.
- Helps in identifying variability.
- Sensitive to extreme values (outliers).

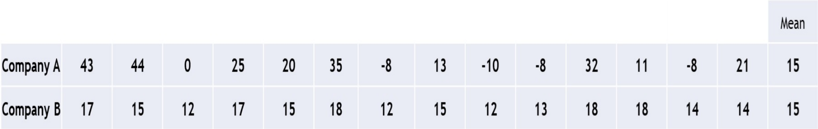
Basic Descriptive Statistics: Variance and Standard Deviation

Dispersion



- Mean acts the central/focal point of the data.
- That alone does not describe the data effectively.
- Variables can have same mean but may behave differently.

Dispersion



															Mean
Company A	43	44	0	25	20	35	-8	13	-10	-8	32	11	-8	21	15
Company B	17	15	12	17	15	18	12	15	12	13	18	18	14	14	15

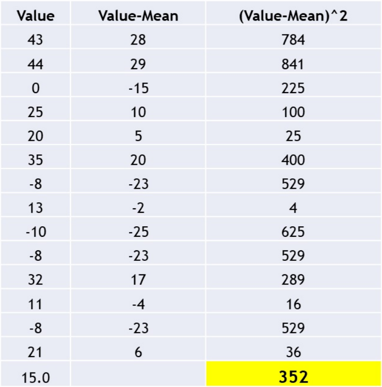
- Profit details of two companies A & B for last 14 Quarters.
- Average profit is 15 in both the cases.
- Which company has performed consistently ?
- Company A had losses.
- Measure of dispersion will describe this behaviour.

Variance and Standard deviation

- Variance **quantifies** the spread of each point from mean value.
- Steps to calculate variance:
    1. Calculate $z_i = x_i - \bar{x}$ for each i.
    2. variance $= \dfrac{(z_1)^2 + (z_2)^2 + \ldots + (z_n)^2}{n}$.
- Variance: $\sigma^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$.
- Variance is the average of squared distances of each point from the mean.
- It is a fairly good measure of dispersion.

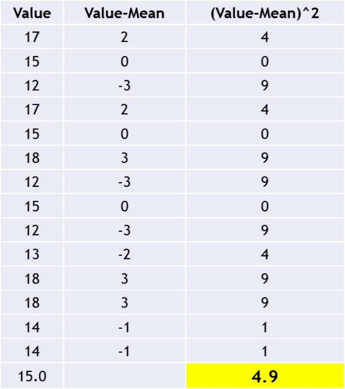
Variance of two companies

Variance of Company A



Value	Value-Mean	(Value-Mean)^2
43	28	784
44	29	841
0	-15	225
25	10	100
20	5	25
35	20	400
-8	-23	529
13	-2	4
-10	-25	625
-8	-23	529
32	17	289
11	-4	16
-8	-23	529
21	6	36
15.0		**352**

Variance of Company B



Value	Value-Mean	(Value-Mean)^2
17	2	4
15	0	0
12	-3	9
17	2	4
15	0	0
18	3	9
12	-3	9
15	0	0
12	-3	9
13	-2	4
18	3	9
18	3	9
14	-1	1
14	-1	1
15.0		**4.9**

- Why is $x_i - \bar{x}$ squared ?

Standard deviation

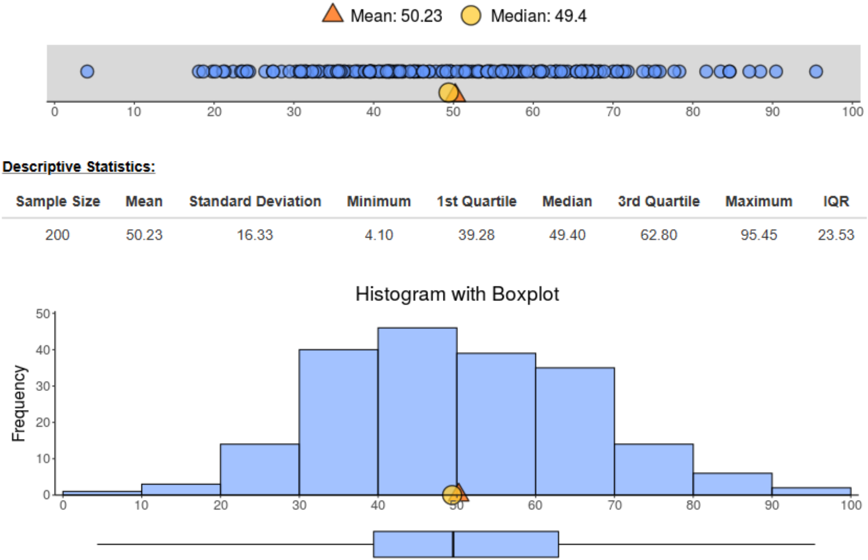
- Variance is **average** of the **squared** distance from the mean.
- Its units are squared.
- Take *square root of variance* to obtain dispersion in the **same units as the actual data**. This is called as standard deviation.

Standard Deviation:

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

.

Variance: Visual Representation

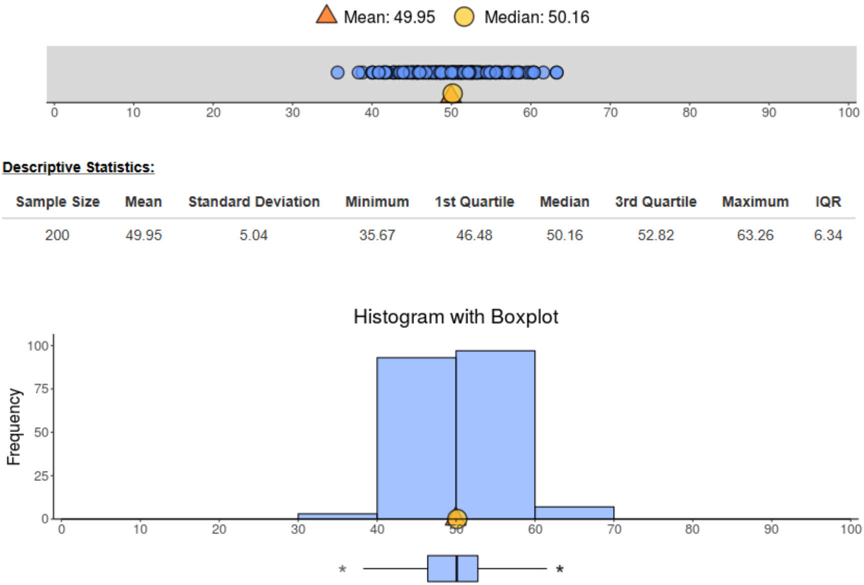


Descriptive Statistics:

Sample Size	Mean	Standard Deviation	Minimum	1st Quartile	Median	3rd Quartile	Maximum	IQR
200	50.23	16.33	4.10	39.28	49.40	62.80	95.45	23.53

Histogram with Boxplot

Variance: Visual Representation

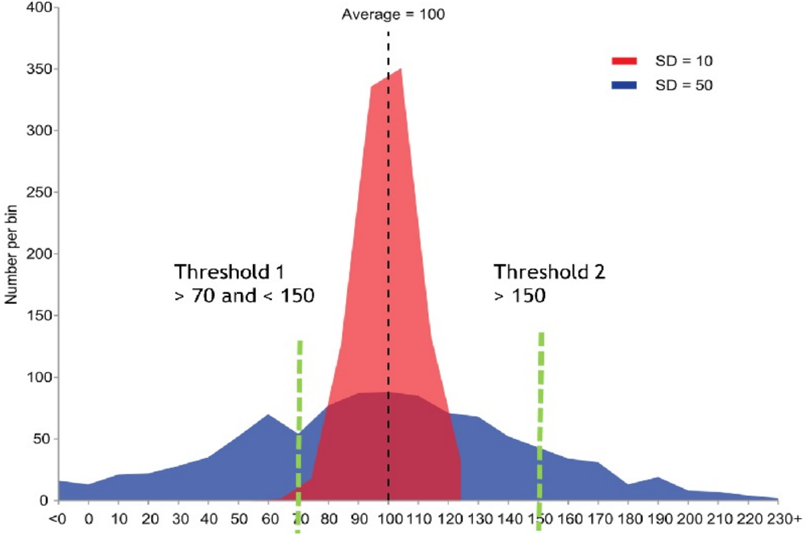


Descriptive Statistics:

Sample Size	Mean	Standard Deviation	Minimum	1st Quartile	Median	3rd Quartile	Maximum	IQR
200	49.95	5.04	35.67	46.48	50.16	52.82	63.26	6.34

Histogram with Boxplot

Is variance bad or good



- Is the distribution with high variance bad or good ?
- it depends on business context.

Variance

Which company to choose ?

- Short term investement: High risk:==>.
- Long term investment: Low risk:==>.

- Lab: Excel-1f- Range, Variance and Standard deviation
- Lab: Excel-1g- Count Functions

Additional Descriptive Statistics

- **Minimum and Maximum:**
  - Provide the smallest and largest values in the dataset.
- **Count and Sum:**
  - Provides the number of values and the sum of these in the dataset.
- **Quartiles:**
  - Divide the dataset into four equal parts, helping to understand the distribution.
- **Skewness:**
  - Indicates whether the data distribution is symmetric or skewed to the left/right.
- **Kurtosis:**
  - Measures the "tailedness" of the data distribution, indicating how extreme values behave.
- Skewness and Kurtosis describe the shape of a data distribution.

Basic Descriptive Statistics: Skewness

Skewness: Definition & Formula

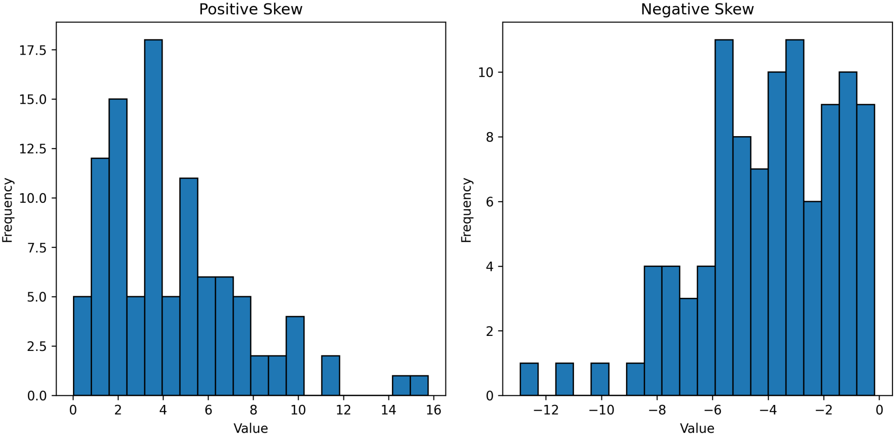
Definition: Skewness measures the asymmetry of a distribution.
Formula:

$$S = \frac{\sum (x_i - \bar{x})^3}{n\sigma^3}$$

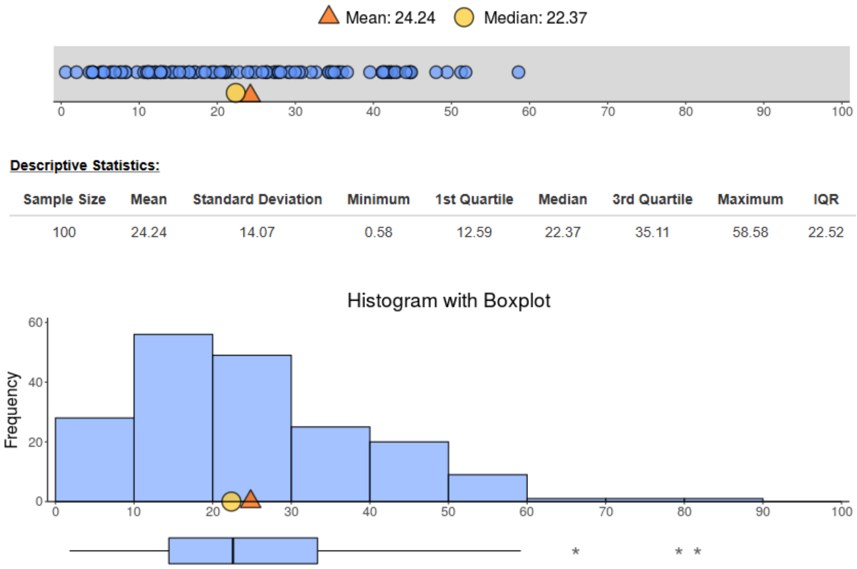
Interpretation:

- $S = 0$: Symmetric distribution.
- $S > 0$: Right-skewed (tail on the right).
- $S < 0$: Left-skewed (tail on the left).

Skewness: Visual Representation



Skewness: Visual Representation

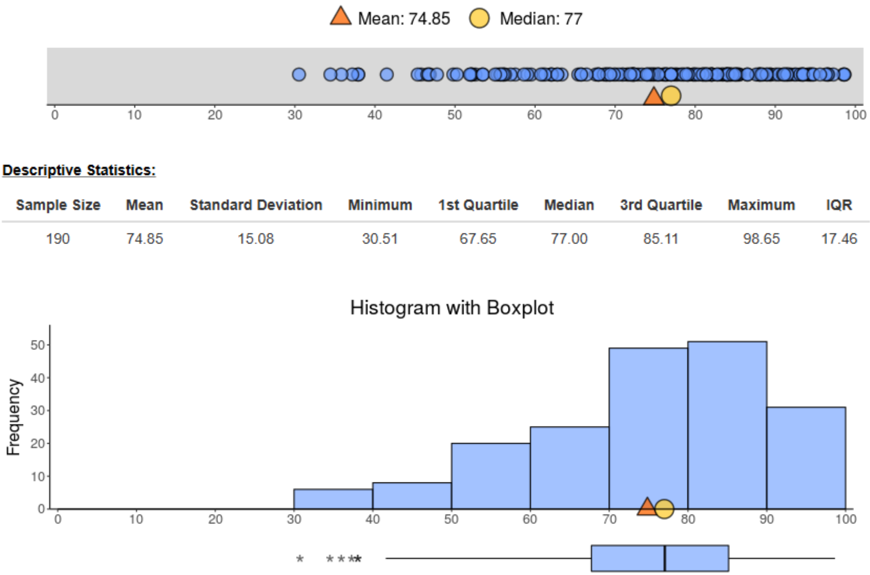


Descriptive Statistics:

Sample Size	Mean	Standard Deviation	Minimum	1st Quartile	Median	3rd Quartile	Maximum	IQR
100	24.24	14.07	0.58	12.59	22.37	35.11	58.58	22.52

Histogram with Boxplot

Skewness: Visual Representation



Descriptive Statistics:

Sample Size	Mean	Standard Deviation	Minimum	1st Quartile	Median	3rd Quartile	Maximum	IQR
190	74.85	15.08	30.51	67.65	77.00	85.11	98.65	17.46

Basic Descriptive Statistics: Kurtosis

Kurtosis: Definition & Formula

Definition: Kurtosis measures the "tailedness" of a distribution.

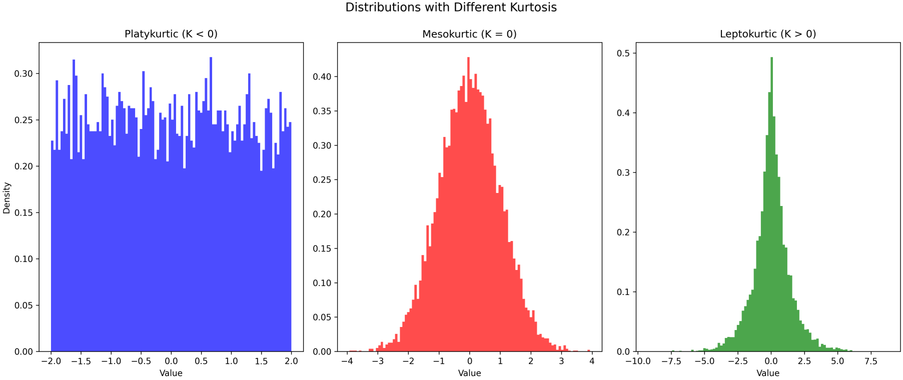
Formula:

$$K = \frac{\sum (x_i - \bar{x})^4}{n\sigma^4} - 3$$

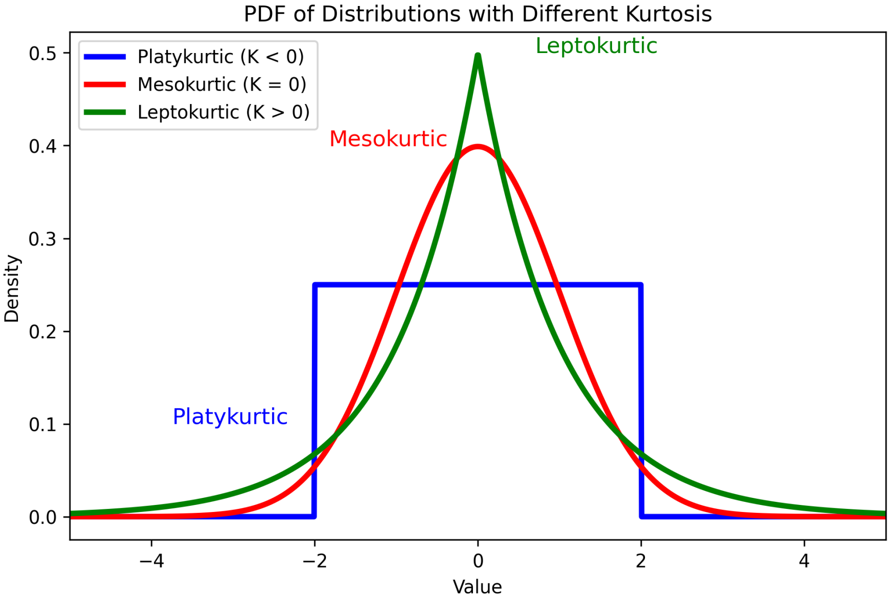
Interpretation:

- $K = 0$: Mesokurtic (Normal-like).
- $K > 0$: Leptokurtic (Heavy tails, extreme values).
- $K < 0$: Platykurtic (Flat, light tails).

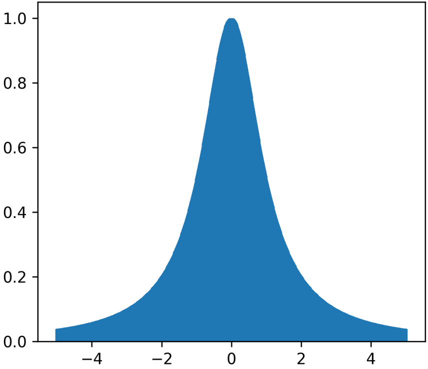
Kurtosis: Visual Representation

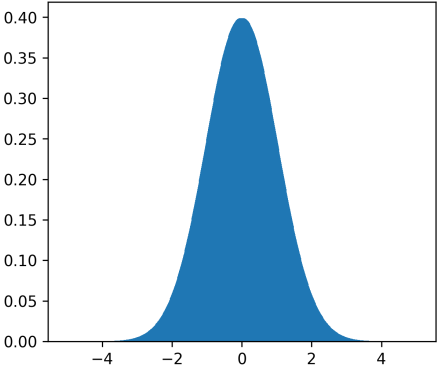


Kurtosis: Visual Representation



Kurtosis: Which one has heavy tail ?





Skewness and Kurtosis (Excel Functions)

Function	Query	Example Formula	Interpretation
SKEW()	What is the skewness of the sales across all products and regions?	`=SKEW(C2:C11)`	A skewness value between -0.5 and +0.5 is considered approximately symmetric. Positive values indicate right skew, and negative values indicate left skew.
KURT()	What is the kurtosis of the sales across all products and regions?	`=KURT(C2:C11)`	**Kurtosis = 3** indicates a normal (Mesokurtic) distribution. **Kurtosis** $>$ **3** (Leptokurtic) implies heavy tails, and **Kurtosis** $<$ **3** (Platykurtic) implies light tails.

D Functions in Excel

D Functions in Excel

- D functions work with structured data (databases/tables).
- Used for performing calculations on subsets of data meeting specific conditions.
- Syntax: `Dfunction(database, field, criteria)`
- Example: `=DSUM(A1:D10, "Sales", F1:F2)`

Advantages and Applications of D Functions in Excel

- Efficiently process large structured datasets.
- Allow conditional calculations without complex formulas.
- Reduce the need for manual filtering and sorting.
- Improve readability and maintainability of Excel models.
- Financial analysis - Summing sales data for specific regions.
- Inventory management - Counting available stock based on category.
- Employee management - Finding average salary of employees in a department.
- Statistical analysis - Calculating variance and standard deviation for filtered datasets.

D Functions Summary

Function	Description
DSUM	Adds values that match criteria.
DAVG	Calculates the average of matching values.
DCOUNT	Counts numeric values that match criteria.
DCOUNTA	Counts all non-blank values meeting criteria.
DVAR	Estimates sample variance for matching values.
DVARP	Calculates population variance for matching values.
DSTDEV	Estimates sample standard deviation for matches.
DSTDEVP	Calculates population standard deviation for matches.

Table: Overview of Excel D Functions

- **Lab: Excel-1h- D-Functions**

Basic Descriptive Statistics: Percentiles

Percentiles

An interesting Question:

- How to compare data point with other data points in the data set ?
- Variance compares a data point with mean.

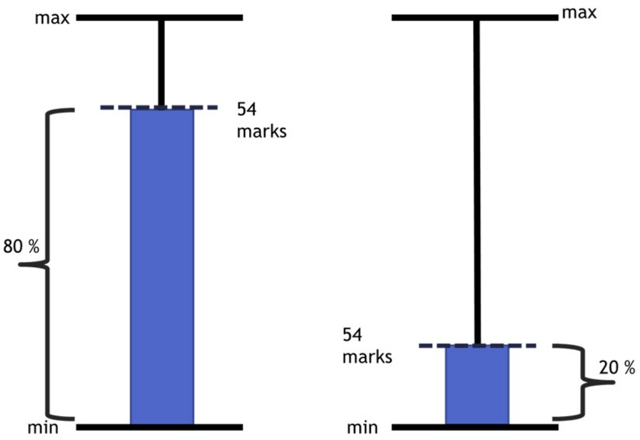
Percentiles:

- Percentiles are a way to describe the relative standing of a value within the dataset.
- 25^{th} percentile means 25% of the data is below that value.
- 90^{th} percentile means 90% of the data is below that value.

Example

- A student attended an exam along with 1000 others.
- He got 54 marks? How good or bad did he perform in the exam?
- What will be his rank overall?

Percentiles



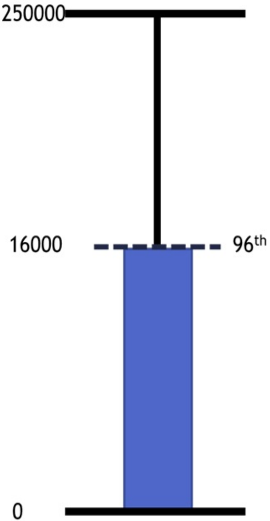
- If 80^{th} percentile is 54, it means that the student is better than 80% of other students.
- If 20^{th} percentile is 54, it means that the student is better than 20% of other students.

Percentiles and Outliers

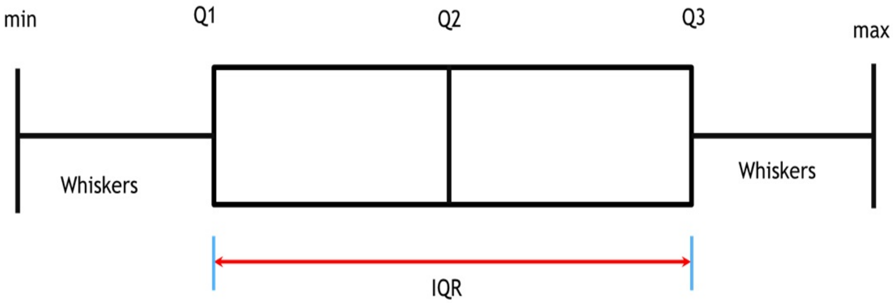
- Percentiles help us in getting an idea on outliers.
- For example the highest income value is 250,000 but 96th percentile is 1,000 only.
- That means 96% of the values are less than 16,000. So the values near 250,000 are clearly outliers

t^{th} Percentile

t^{th} percentile : $t\%$ of observations are below it and $(100-t)\%$ of observations are above it.



Quartiles and Box Plot



- **Min**: 0^{th} Percentile.
- **Q1**: The 25^{th} Percentile (First/Lower Quartile):
- **Q2**: The 50^{th} Percentile (Median):
- **Q3**: The 75^{th} Percentile (Third/Upper Quartile):
- **Max**: The 100^{th} Percentile:
- **Inter Quartile range**: $Q3 - Q1$.

Methods for calculating IQR

- Interquartile Range (IQR) is a measure of statistical dispersion, or spread, of a dataset.
- It is the difference between the 75th percentile (Q3) and the 25th percentile (Q1).
- There are two methods for calculating IQR: Inclusive and Exclusive.

Inclusive IQR Calculation

- Inclusive IQR includes the median when calculating the quartiles.
- It treats the dataset as a whole, including the middle value when dividing into quartiles.
- Steps:
  1. Arrange the data in ascending order.
  2. Find the median of the entire dataset (Q2).
  3. Split the data into two halves: one below Q2 and one above Q2.
  4. Calculate the lower quartile (Q1) and upper quartile (Q3) for these halves.
  5. IQR = Q3 - Q1.

Exclusive IQR Calculation

- Exclusive IQR excludes the median when calculating the quartiles.
- It divides the dataset into two halves excluding the median, then finds Q1 and Q3 from these halves.
- Steps:
  - Arrange the data in ascending order.
  - Find the median (Q2) and exclude it from the dataset.
  - Split the remaining data into two halves: one below Q2 and one above Q2.
  - Calculate the lower quartile (Q1) and upper quartile (Q3) for these halves.
  - IQR = Q3 - Q1.

Dataset

- Sorted dataset:
$$\{5, 8, 12, 14, 18, 21, 24, 30, 35\}$$
- Median (Overall): 18 (middle value)

Exclusive Approach: Q1, Q3, and IQR

- Q1 : Median of the lower half {5, 8, 12, 14}, excluding 18.
- Q1 $= \frac{8+12}{2} = 10$
- Q3 : Median of the upper half {21, 24, 30, 35}, excluding 18.
- Q3 $= \frac{24+30}{2} = 27$
- IQR = Q3 - Q1 = 27 - 10 = 17

Inclusive Approach: Q1, Q3, and IQR

- Q1 : Median of the lower half $\{5, 8, 12, 14, 18\}$, including 18.
- Q1 = 12
- Q3 : Median of the upper half $\{18, 21, 24, 30, 35\}$, including 18.
- Q3 = 24
- IQR = Q3 - Q1 = 24 - 12 = 12

Summary

- Exclusive Approach : $\text{IQR} = 17$
- Inclusive Approach : $\text{IQR} = 12$
- The difference arises from how the median is treated in each method.

Box Plot and Outliers

Box Plot and Outliers

- If there are no outliers, the box plot will be equally distributed.
- If there are any outliers in the data, the box plot will be compressed.

- **Lab: Excel-1i- IQR and Box Plots**

What is a Proportion?

Definition

A **proportion** is a number between 0 and 1 that represents the fraction of the total that has a particular attribute.

- Expressed as a fraction, decimal, or percentage.
- Helps describe parts of a whole in simple terms.

Formula for Proportion

$$\text{Proportion} = \frac{\text{Number of successes}}{\text{Total number of observations}}$$

Success simply refers to the outcome of interest — not necessarily something "good."

Proportion vs Percentage

- **Proportion** is usually a decimal (e.g., 0.25).
- **Percentage** is the proportion multiplied by 100 (e.g., $0.25 \times 100 = 25\%$).
- Both tell the same story but are used depending on the context.

Example 1: Favorite Subject Survey

Problem

In a survey of 200 students, 60 said Math is their favorite subject.

$$\text{Proportion} = \frac{60}{200} = 0.30 \quad \Rightarrow \quad 30\%$$

- 30% of students prefer Math.

Example 2: Quality Control

Problem

Out of 500 light bulbs produced, 15 were defective.

$$\text{Proportion} = \frac{15}{500} = 0.03 \quad \Rightarrow \quad 3\%$$

- Only 3% of bulbs were defective.

Types of Proportions

- **Sample Proportion** ($\hat{p}$): Based on a sample (subset).

  $$\hat{p} = \frac{x}{n}$$

  where x = number of successes, n = sample size.

Types of Proportions

- **Sample Proportion** $(\hat{p})$: Based on a sample (subset).

$$\hat{p} = \frac{x}{n}$$

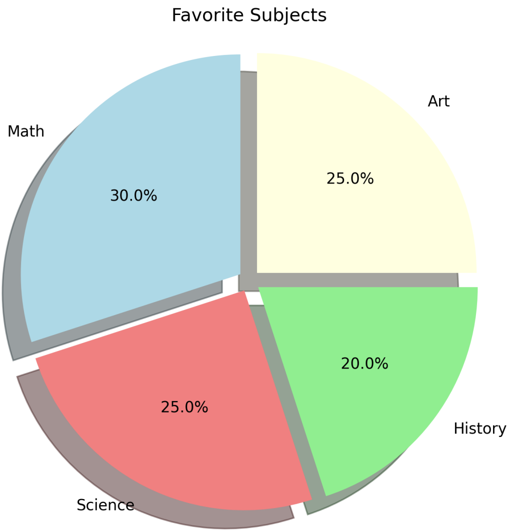
  where $x =$ number of successes, $n =$ sample size.

- **Population Proportion** (p): Based on the entire population.

Why are Proportions Important?

- Summarize data quickly.
- Compare groups easily (e.g., male vs female preferences).
- Essential in hypothesis testing (e.g., proportion tests).
- Used in business, healthcare, marketing, and quality control.

Visualizing Proportions



Pie charts and bar charts are popular ways to visualize proportions.

Mini Exercise

Question: In a class of 40 students, 12 students are left-handed. What is the proportion and percentage of left-handed students?

Think about:

- What is the formula?
- How to express it as a percentage?

Solution :

$$\text{Proportion} = \frac{12}{40} = 0.30$$

$$\text{Percentage} = 0.30 \times 100 = 30\%$$

Answer: 30% of the students are left-handed.

THANK YOU