Probability || project

ENG\ Abdelrahman Ibrahem.

ENG\Youssef sherif.

ENG\ Amr Sherif

Dot plot

Dot plot

Data itself cannot describe the insights stored in it we always need visualization for in-depth understanding to make useful impacts during analysis, one of the most common visualization plots is dot plot!

we will cover:

1-Usage of dot plot

2-Structure

3-Use cases

4-Advantages

5-Drawbacks

Dot plot

1-Uses of Dot plot

Visualizing in 2D manner which gives the same insights like functions in mathematics making points point to corresponding value

It helps identifying clusters, gaps and outliers in the data

You can measure aggregate functions to understand the trend in the data sample you have like(frequencies, counts, or other statistics across categories or groups)

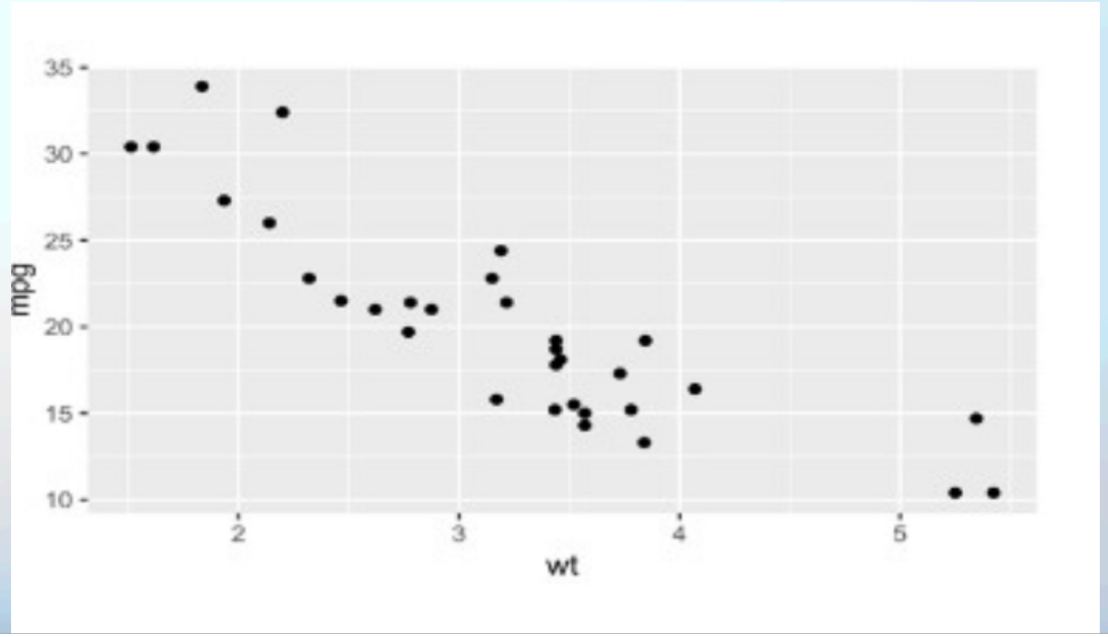
Dot plots are often used in teaching because of their simplicity and ability to clearly show individual data values.

Dot plot

1-Uses of Dot plot

Horizontal Axis (X-axis): Represents the categories or the data values.

Vertical Position (Y-axis): Generally not labeled, as the number of dots in a column represents frequency or counts.



Dot plot

1-Dot Plot Use Cases

Exam Scores of Students:

Each dot could represent a student’s score on an exam.

Age Distribution in a Small Population:

Each dot could represent a person’s age.

Sales Comparison:

Comparing sales counts for multiple products over a specific period.

Dot plot

1-Advantages of Dot Plots

1-Easy to use and simple to gain insights from it, due to its simplicity

Effectively represents small datasets.

2-Useful for identifying patterns and clusters.

3-Maintains granularity, showing every data point.

Dot plot

1-Limitations of Dot Plots

Not suitable for large datasets (clustering can make it unreadable).

Can be less effective for very detailed or complex comparisons due to limited dimensional parameters

Dot plot

How to implement ?

1- Understand Your Data:

Determine whether your data is categorical (e.g., product types) or numerical (e.g., test scores or age).

Ensure your dataset is small or summarized enough for clear visualization.

2- Choose an Axis:

Select a horizontal axis (X-axis) to represent your data values or categories.

3- Sort the Data (optional part):

If the data is numerical, sort it in ascending or descending order.

If categorical, decide on the order of categories (e.g., alphabetically or by frequency).

4- Map the Data:

For each unique value (numerical or categorical), count its frequency.

Dot plot

How to implement

5- Place Dots:

For each data point or frequency, place a dot above the corresponding value on the X-axis.

If two or more dots belong to the same value, stack them vertically.

6- Label the Plot:

Add appropriate labels to the X-axis (values or categories).

If necessary, add a title to clarify the purpose of the dot plot.

Add annotations to explain key observations or trends.

7- Review and Adjust:

Ensure the dots are evenly spaced and clearly visible.

If dealing with very high frequencies, consider using alternative visualizations like histograms.

Dot plot

conclusion

A dot plot is a great and simple tool for visualizing distributions and frequencies in small datasets. It provides clarity by showing individual data points, making it excellent for exploratory data analysis and teaching purposes. However, it may not be suitable for large datasets or highly detailed comparisons.

Box plot

Box plot

Box plots are used to show distributions of numeric data values, especially when you want to compare them between multiple groups.

Use-Cases of Box Plot

Box plots provide a visual summary of the data with which we can quickly identify the average value of the data, how dispersed the data is, whether the data is skewed or not (skewness).

The Median gives you the average value of the data.

Box Plots shows Skewness of the data

Box plot

A box plot gives a five-number summary of a set of data which is-

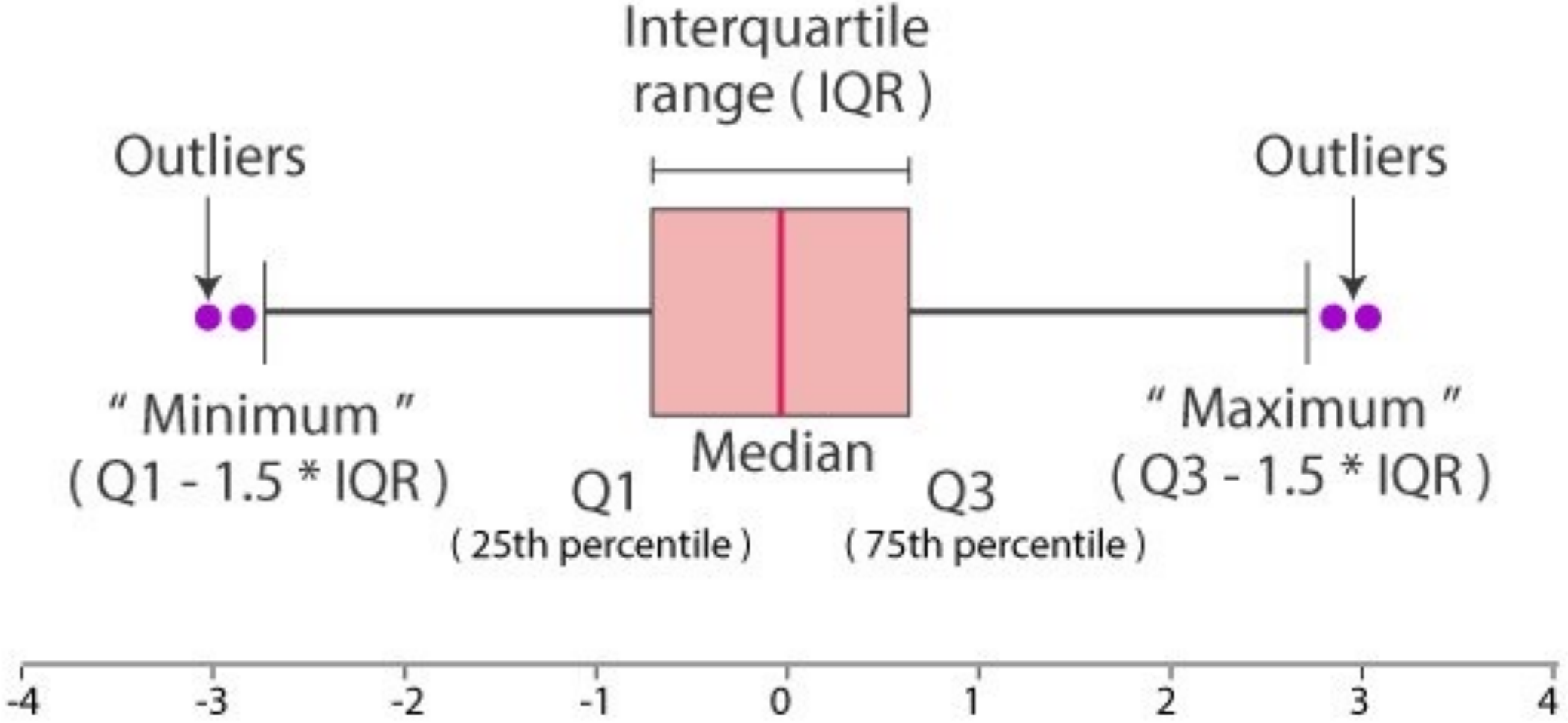
1-Minimum – It is the minimum value in the dataset excluding the outliers.

2-First Quartile (Q1) – 25% of the data lies below the First (lower) Quartile.

3-Median (Q2) – It is the mid-point of the dataset. Half of the values lie below it and half above.

4-Third Quartile (Q3) – 75% of the data lies below the Third (Upper) Quartile.

5-Maximum – It is the maximum value in the dataset excluding the outliers.



Different parts of boxplot

Box plot

Compare the Medians and the Dispersion or Spread of data

Compare the Medians — If the median line of a box plot lies outside the box of the other box plot with which it is being compared, then we can say that there is likely to be a difference between the two groups. Here the Median line of the plot B lies outside the box of Plot A.

Compare the Dispersion or Spread of data — The Inter Quartile range (length of the box) gives us an idea about how dispersed the data is. Here Plot A has a longer length than Plot B which means that the dispersion of data is more in plot A as compared to plot B. The length of whiskers also gives an idea of the overall spread of data. The extreme values (minimum &maximum) give the range of data distribution. Larger the range more scattered the data. Here Plot A has a larger range than Plot B.

Box plot

Comparing Outliers and Compare Skewness

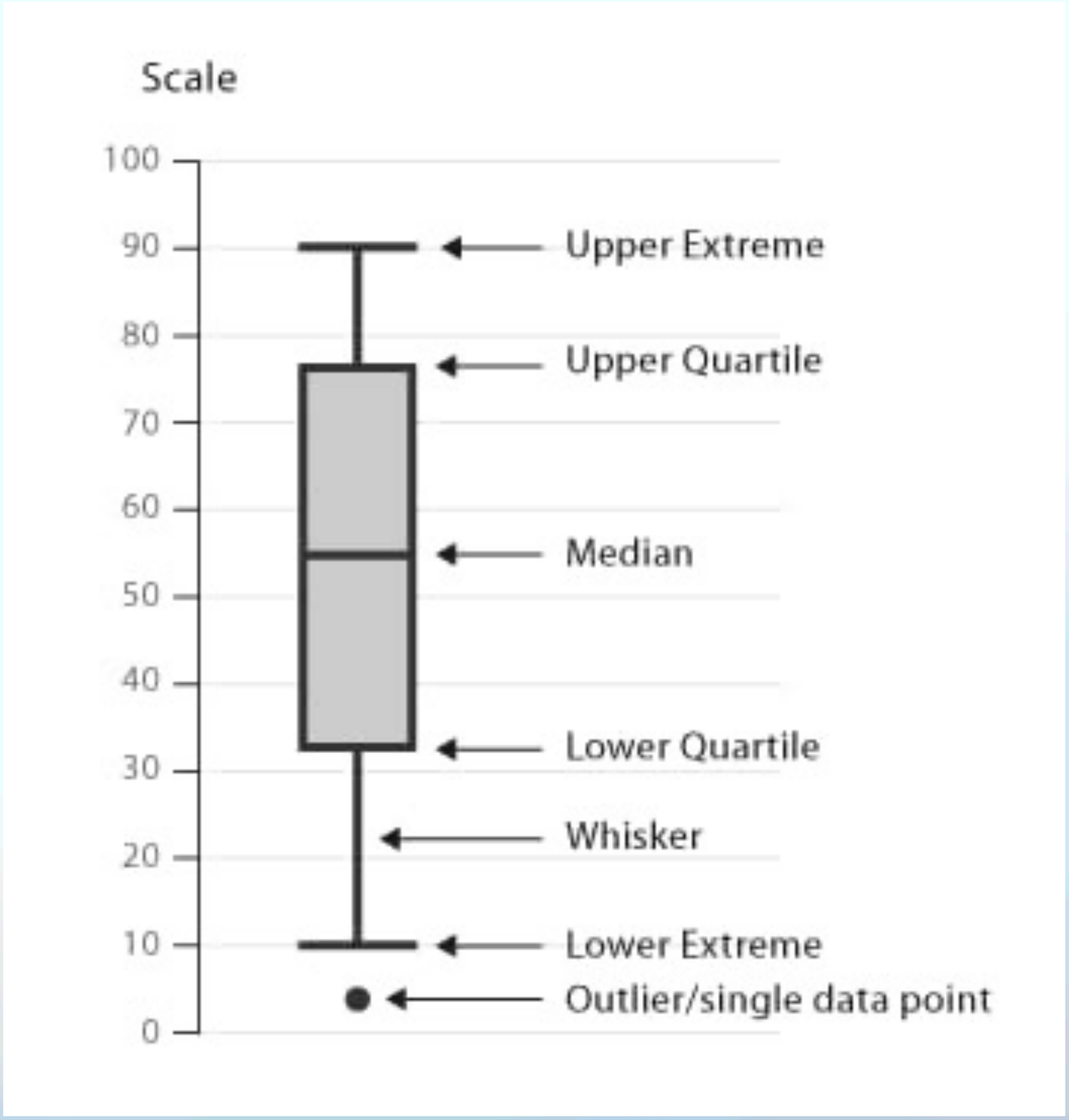
Comparing Outliers — The outliers give the idea of unusual data values which are distant from the rest of the data. More number of Outliers means the prediction will be more uncertain. We can be more confident while predicting the values for a plot which has less or no outliers.

Compare Skewness — Skewness gives us the direction and the magnitude of the lack of symmetry. We have discussed above how to identify skewness. Here Plot A is Positive or Right Skewed and Plot B is Negative or Left Skewed.

Box plot

Code to make Box plot

```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
# Sample data
data = {
  'Group A': np.random.normal(50, 10, 100),
  'Group B': np.random.normal(60, 15, 100),
  'Group C': np.random.normal(55, 20, 100)
}
# Prepare data for boxplot
groups = list(data.keys())
values = list(data.values())
```



Box plot

Code to make Box plot

```
# Creating a box plot

plt.figure(figsize=(8, 6))

sns.boxplot(data=values)

# Adding labels

plt.xticks(ticks=range(len(groups)), labels=groups)

plt.title("Box Plot Example")

plt.ylabel("Values")

plt.xlabel("Groups")

# Show the plot

plt.show()
```

Steam and Leaf

Steam and leaf

Stem and leaf plots are a simple and effective way to organize and display data. It displays numerical data by splitting it into a stem and a leaf. Generally, the leading digit or digits represent the stem, and the last digit represents the leaf. By splitting each number into a "stem" and a "leaf," this method provides a clear and concise view of numerical data.

In this article, we will discuss the concept of stem and leaf plots in detail, including their key features, methods to create the plot, and various solved and unsolved examples.

Steam and leaf

What are Stem and Leaf Plots?

A stem and leaf plot is a graphical representation used to organize and display quantitative data in a semi-tabular form. It helps in visualizing the distribution of the data set and retains the original data values, making it easy to identify the shape, central tendency, and variability of the data.

A stem and leaf plot splits each data point into a "stem" and a "leaf." The "stem" represents the leading digits, while the "leaf" represents the trailing digit. This separation makes it easy to organize data and see patterns.

Steam and leaf

For example: For the data set: 23, 25, 27, 32, 34, 35, 41, 42

This data can be represented as following table:

Stem	Leaves
2	3, 5, 7
3	2, 4, 5
4	1, 2

Steam and leaf

Key Features of Stem and Leaf Plots

Some of the key features of steam and leaf plots are:

1-Stem and leaf plots display numerical data by splitting each data point into a "stem" (usually the first digit or digits) and a "leaf" (usually the last digit).

Steam and leaf

2-The stems are listed vertically, and the leaves are written next to their corresponding stem in numerical order.

A key is provided to explain what the stem and leaf represent for that particular plot.

3-Stem and leaf plots are useful for displaying the shape, spread, and central tendency of a data distribution. They can highlight outliers and help identify the mod

Steam and leaf

Stems with no leaves are still included to preserve the horizontal axis scaling and highlight gaps in the data.

Steam and leaf

Steps to implement steam and leaf plot :

Step 1: Arrange your data set in ascending order.

Step 2: Identify the "stems" by separating the leading digits of the numbers in your data set.

For example, if your data includes numbers like 23, 25, and 27, the stem would be 2.

Step 3: Identify the "leaves" by taking the trailing digits of the numbers.

For instance, in the number 23, the leaf would be 3.

Step 4: Write down the stems in a vertical column.

Step 5: Next to each stem, write down the corresponding leaves in ascending order.

Steam and leaf

Conclusion :

In conclusion, stem and leaf plots are a simple yet powerful tool for visualizing data. They help in quickly identifying the shape, spread, and central tendency of a dataset. By organizing data in a structured format, these plots make it easy to see patterns and outliers. Whether for educational purposes or data analysis, stem and leaf plots offer a clear and concise way to understand numerical data.

How to define skewness type from graph

How to define skewness type from graph

Skewness refers to the asymmetry of the distribution of data. To determine the type of skewness from a graph (typically a histogram or a boxplot), observe the shape of the distribution:

1- Symmetrical Distribution (No Skewness):

The graph is evenly distributed around the central peak.

The left and right sides are mirror images.

Example: A normal distribution.

How to define skewness type from graph

2-Positive Skewness (Right-Skewed):

The tail on the right side (higher values) is longer than the tail on the left side.

Most of the data points are concentrated on the left.

Example: Income distribution in many populations.

How to define skewness type from graph

3-Negative Skewness (Left-Skewed):

The tail on the left side (lower values) is longer than the tail on the right side.

Most of the data points are concentrated on the right.

Example: Scores on an easy test where most students perform well.

How to define skewness type from graph

Steps to Define Skewness from a Graph:

1-Visual Inspection:

Look at the tails of the distribution and identify the longer tail.

Identify where the bulk of the data is concentrated.

How to define skewness type from graph

2-Compare the Mean and Median:

If the mean is greater than the median, the distribution is positively skewed.

If the mean is less than the median, the distribution is negatively skewed.

Use Descriptive Statistics (if available):

Skewness can be quantified with a skewness coefficient.

A value of 0 indicates symmetry, positive values indicate positive skewness, and negative values indicate negative skewness.

Example box plot

How to define skewness type from graph

Here are the boxplots illustrating different types of skewness:

1-Symmetric Distribution (No Skewness):

The median is centered in the box, and the whiskers are of roughly equal length.

2-Positive Skew (Right-Skewed):

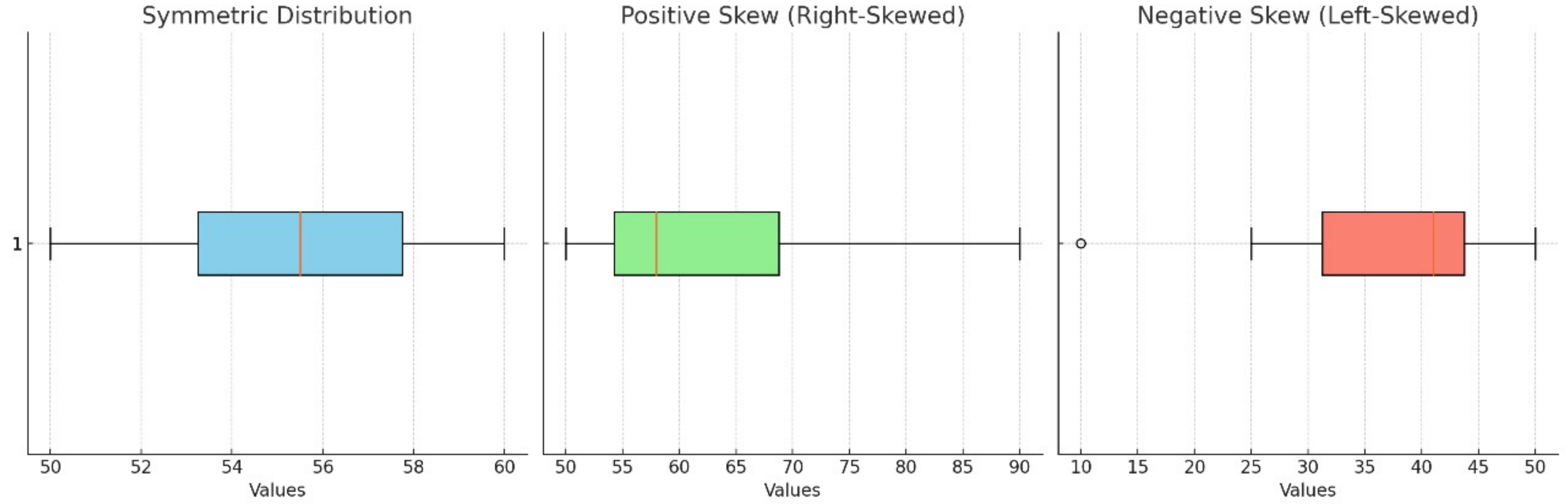
The median is closer to the lower quartile, and the upper whisker (right side) is longer.

3-Negative Skew (Left-Skewed):

The median is closer to the upper quartile, and the lower whisker (left side) is longer.

These visualizations demonstrate how skewness can be identified from a boxplot.

How to define skewness type from graph



ANOVA: Analysis Of Variances

ANOVA: Analysis Of Variances

When it used?

You might use ANOVA when you want to test a particular hypothesis between groups, determining – in using one-way ANOVA – the relationship between an independent variable and one quantitative dependent variable. An example could be examining how the level of employee training impacts customer satisfaction ratings.

ANOVA: Analysis Of Variances

When it used?

One-Way ANOVA :

This test is used to see if there is a variation in the mean values of three or more groups. Such a test is used where the data set has only one independent variable. If the test statistic exceeds the critical value, the null hypothesis is rejected, and the averages of at least two different groups are significant statistically

ANOVA: Analysis Of Variances

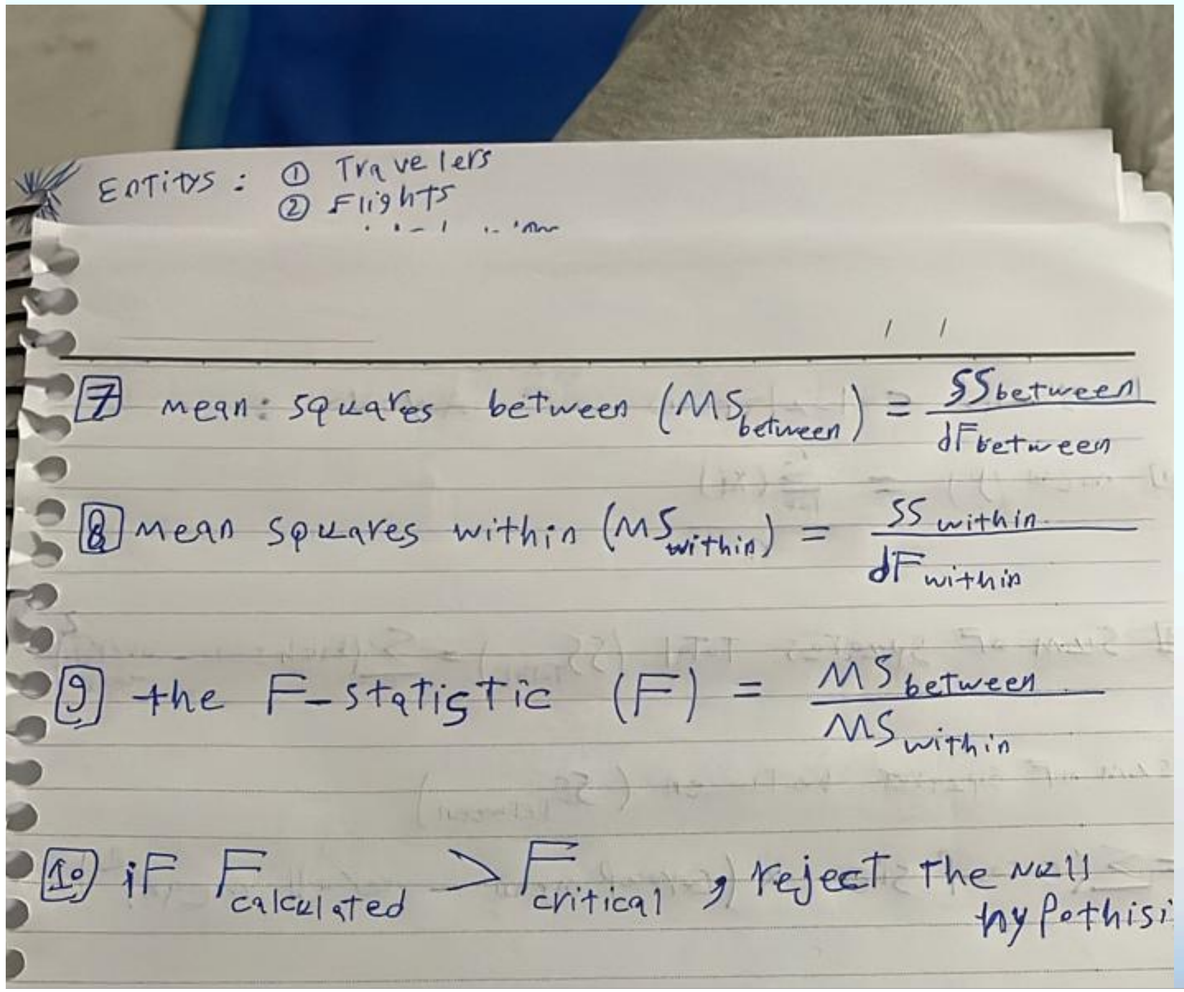
When it used?

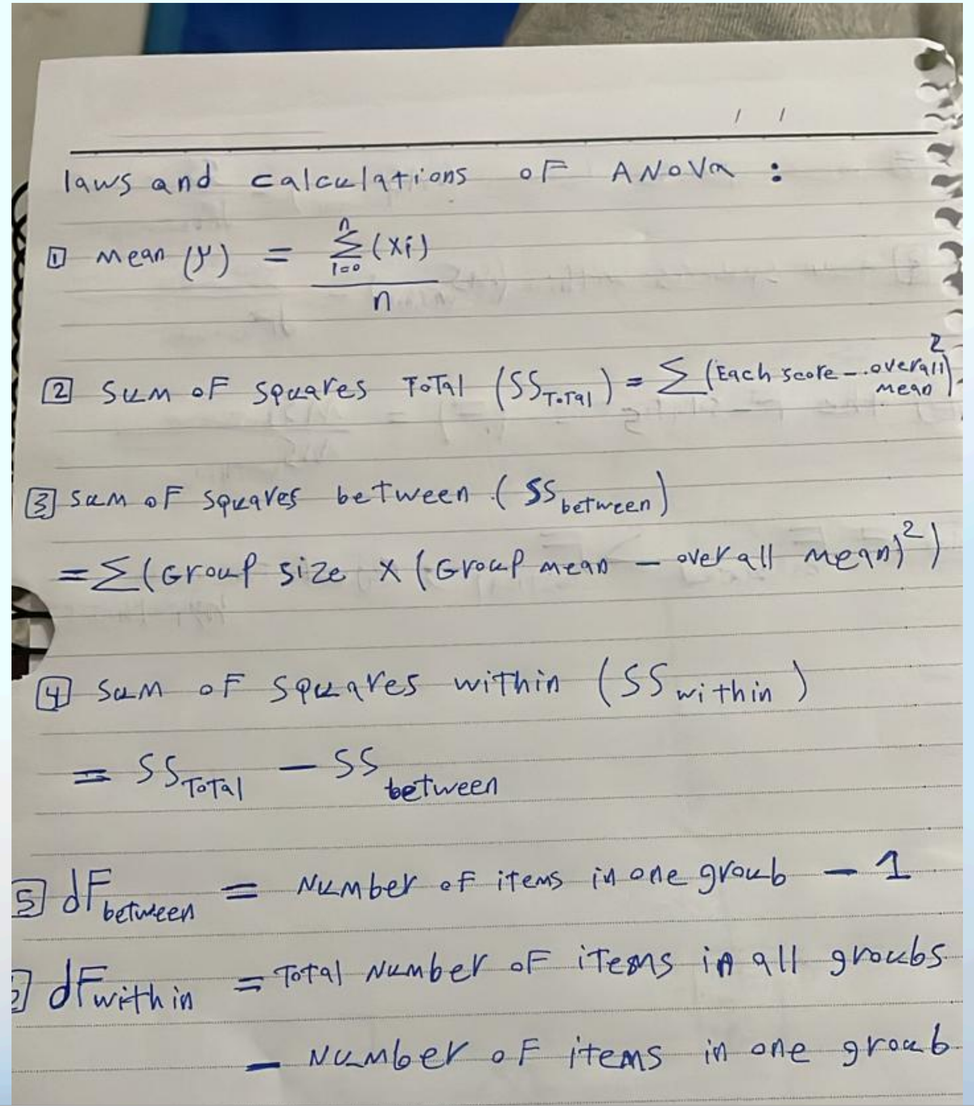
Two-Way ANOVA

Two independent variables are used in the two-way ANOVA. As a result, it can be viewed as an extension of a one-way ANOVA in which only one variable influences the dependent variable. A two-way ANOVA test is used to determine the main effect of each independent variable and whether there is an interaction effect. Each factor is examined independently to determine the main effect, as in a one-way ANOVA. Furthermore, all components are analyzed at the same time to test the interaction impact.

ANOVA: Analysis Of Variances

Laws and calculations





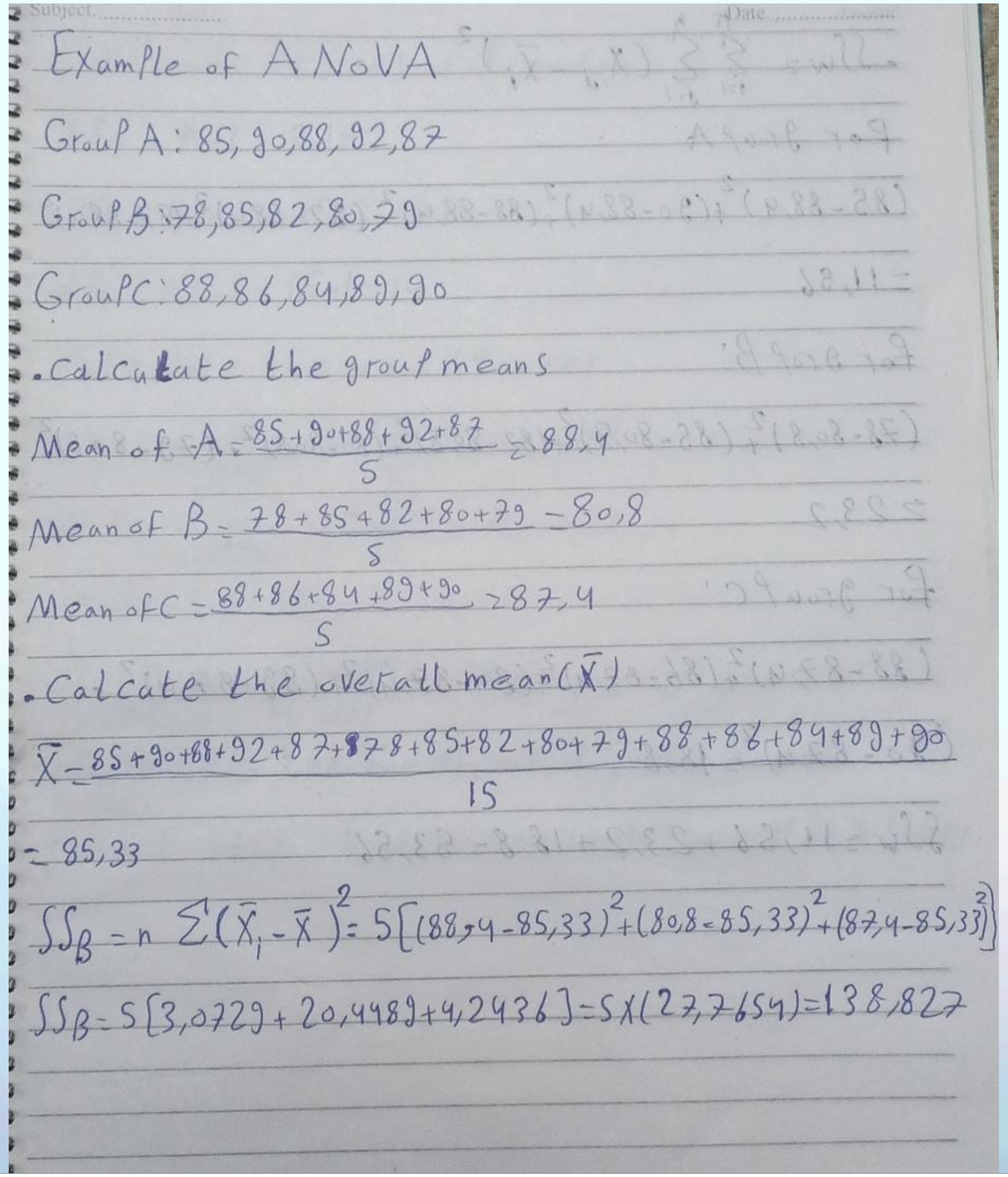
ANOVA: Analysis Of Variances

Example

Student Group	Scores
Method A	85, 90, 88, 92, 87
Method B	78, 85, 82, 80, 79
Method C	88, 86, 84, 89, 90

ANOVA: Analysis Of Variances

Example



Example of ANOVA

Group A: 85, 90, 88, 92, 87

Group B: 78, 85, 82, 80, 79

Group C: 88, 86, 84, 89, 90

. Calculate the group means

$$\text{Mean of A} = \frac{85+90+88+92+87}{5} = 88,4$$

$$\text{Mean of B} = \frac{78+85+82+80+79}{5} = 80,8$$

$$\text{Mean of C} = \frac{88+86+84+89+90}{5} = 87,4$$

. Calcate the overall mean ($\bar{X}$)

$$\bar{X} = \frac{85+90+88+92+87+78+85+82+80+79+88+86+84+89+90}{15}$$

$$= 85,33$$

$$SS_B = n \sum (\bar{X}_i - \bar{X})^2 = 5[(88,4-85,33)^2 + (80,8-85,33)^2 + (87,4-85,33)^2]$$

$$SS_B = 5[3,0729 + 20,4489 + 4,2436] = 5 \times (27,7654) = 138,827$$

ANOVA: Analysis Of Variances

$$SSw = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2$$

For group A:

$(85-88.4)^2 + (90-88.4)^2 + (88-88.4)^2 + (92-88.4)^2 + (87-88.4)^2$

$= 11,56$

for group B:

$(78-80,8)^2 + (85-80,8)^2 + (82-80,8)^2 + (80-80,8)^2 + (79-80,8)^2$

$= 23,2$

for group C:

$(88-87,4)^2 + (86-87,4)^2 + (84-87,4)^2 + (89-87,4)^2 +$

$(90-87,4)^2 = 18,8$

$SSw = 11,56 + 23,2 + 18,8 = 53,56$

Calculate Degrees of freedom

$df_B = k - 1 = 3 - 1 = 2$

$df_w = N - k = 15 - 3 = 12$

$df_T = N - 1 = 15 - 1 = 14$

Calculate the mean Squares

$MS_B = \frac{SSB}{df_B} = \frac{138,827}{2} = 69,4135$

$MSw = \frac{SSw}{df_w} = \frac{53,56}{12} = 4,4633$

- Calculate the f-ration

$F = \frac{MSB}{MSw} = \frac{69,4135}{4,3633} = 15,55$

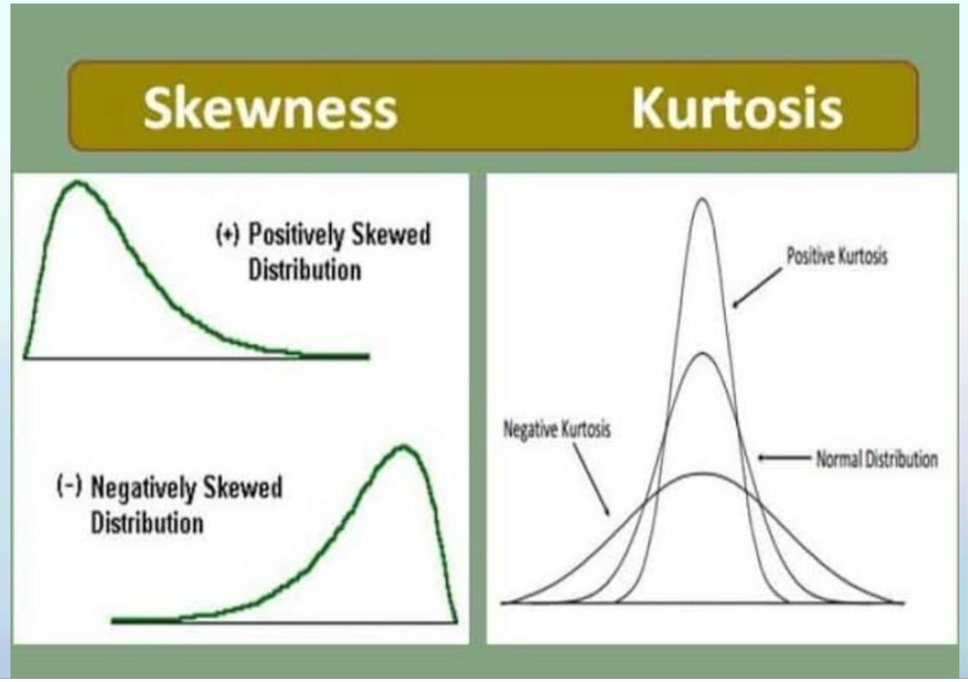
*$F_{critical} \approx 3,88$ for $df_1 = 2, df_2 = 12, \alpha = 0,05$

Since $F = 15,55 > 3,88$. We reject the null hypothesis

Skewness and kurtosis

Skewness and kurtosis

Skewness and kurtosis are statistical measures used to describe the distribution of a dataset. They provide insights beyond basic metrics like mean, median, or variance, focusing on the shape of the data's distribution.



Skewness and kurtosis

A) Skewness Measuring Symmetry

Skewness quantifies the asymmetry of a distribution around its mean.

Types of skewness :

1- Positive Skew (Right-Skewed):

The tail of the distribution extends more to the right.

Mean > Median > Mode.

Example: Income distribution (a few very high earners skew the data).

Skewness and kurtosis

2-Negative Skew (Left-Skewed):

The tail of the distribution extends more to the left.

Mean < Median < Mode.

Example: Scores in an easy exam (many high scores, few low scores)

3-Zero Skew (Symmetrical):

The distribution is perfectly symmetrical.

Mean = Median = Mode.

Example: Normally distributed data.

Skewness and kurtosis

Law

$$Skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - mean}{Stdev} \right)^3$$

xi: Data points

$x$bar: Mean of the data

s: Standard deviation

n: Number of data points

Skewness and kurtosis

Useful Cases:

Financial Data: Detecting whether returns are skewed towards profits or losses.

E-commerce: Understanding customer spending patterns (most people spend less, a few spend a lot).

Quality Control: Detecting asymmetry in product measurements.

Skewness and kurtosis

Kurtosis: Measuring Tail Heaviness

Definition:

B)Kurtosis measures the "tailedness" of a distribution, i.e., the presence of extreme values.

Types of Kurtosis:

1- Leptokurtic (High Kurtosis):

Heavy tails and sharp peak.

Indicates the presence of outliers.

Example: Stock market returns (few extreme returns).

Skewness and kurtosis

2- Platykurtic (Low Kurtosis):

Light tails and a flat peak.

Data is evenly distributed.

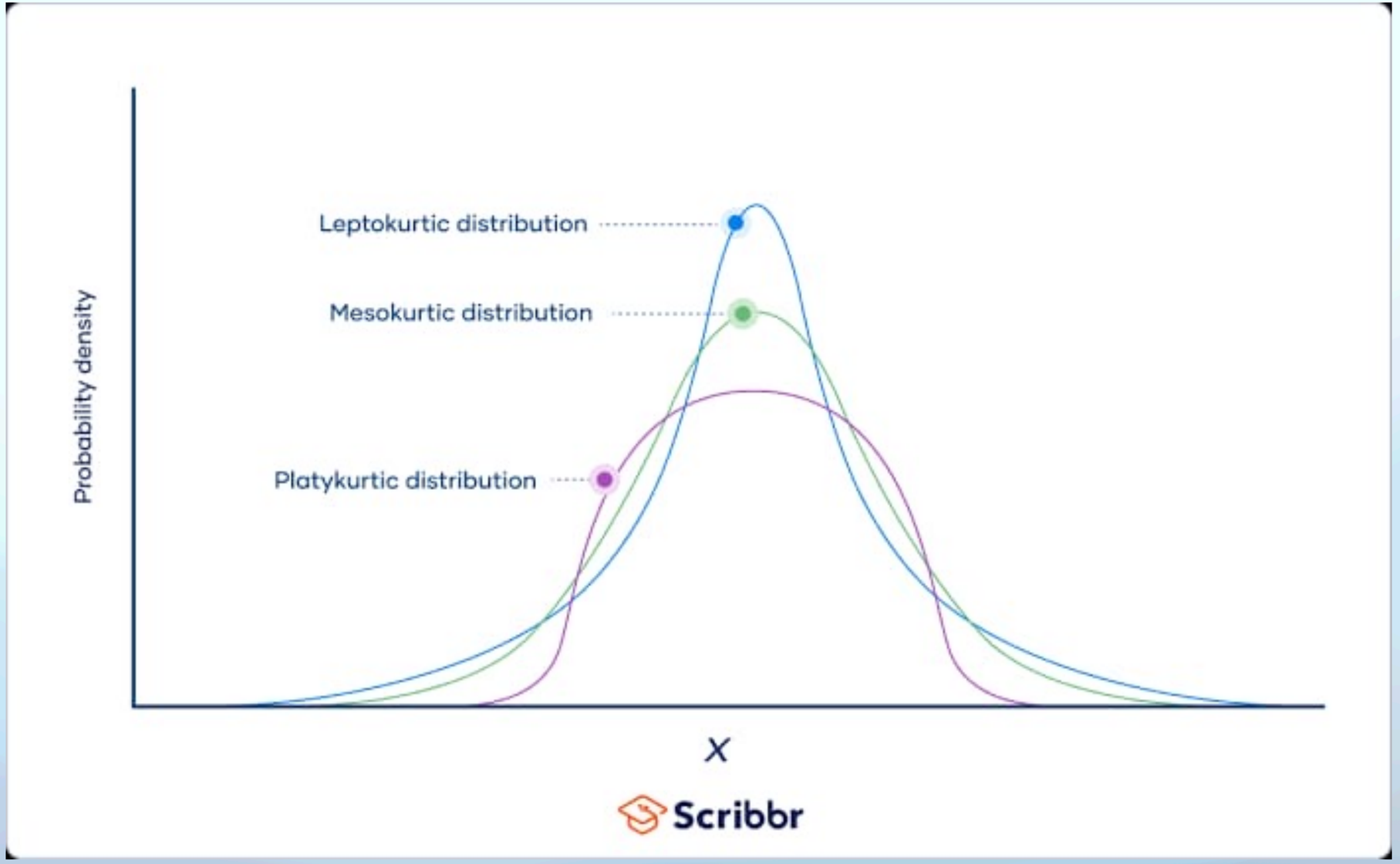
Example: Uniform distribution.

3-Mesokurtic (Normal Kurtosis):

Tails similar to a normal distribution.

Example: Normally distributed test scores.

Skewness and kurtosis



Skewness and kurtosis

Useful Cases:

Risk Management: Identifying datasets with extreme outliers (e.g., extreme market losses).

Insurance: Understanding rare, high-cost events.

Environmental Studies: Analyzing extreme weather conditions.

Skewness and kurtosis are simple yet powerful tool for visualization.

Thank you