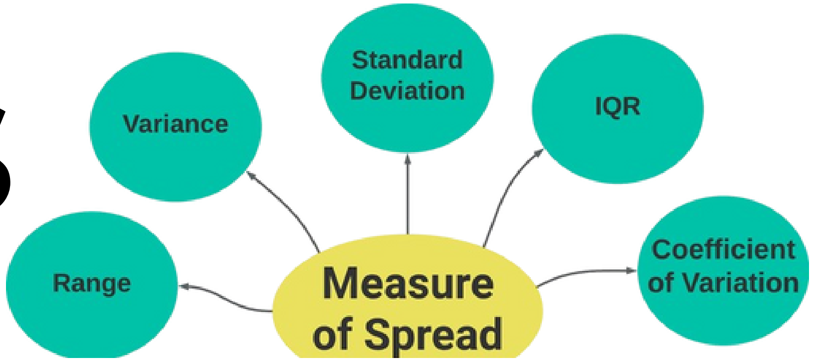
Data Analyst Perspective

STATISTICS

DESCRIPTIVE STATISTICS



Measure of Spread

Range

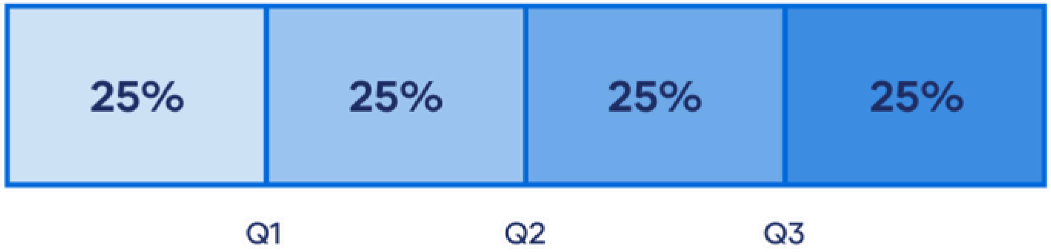
Variance

Standard Deviation

IQR

Coefficient of Variation

Variability Part-2



25%

25%

25%

25%

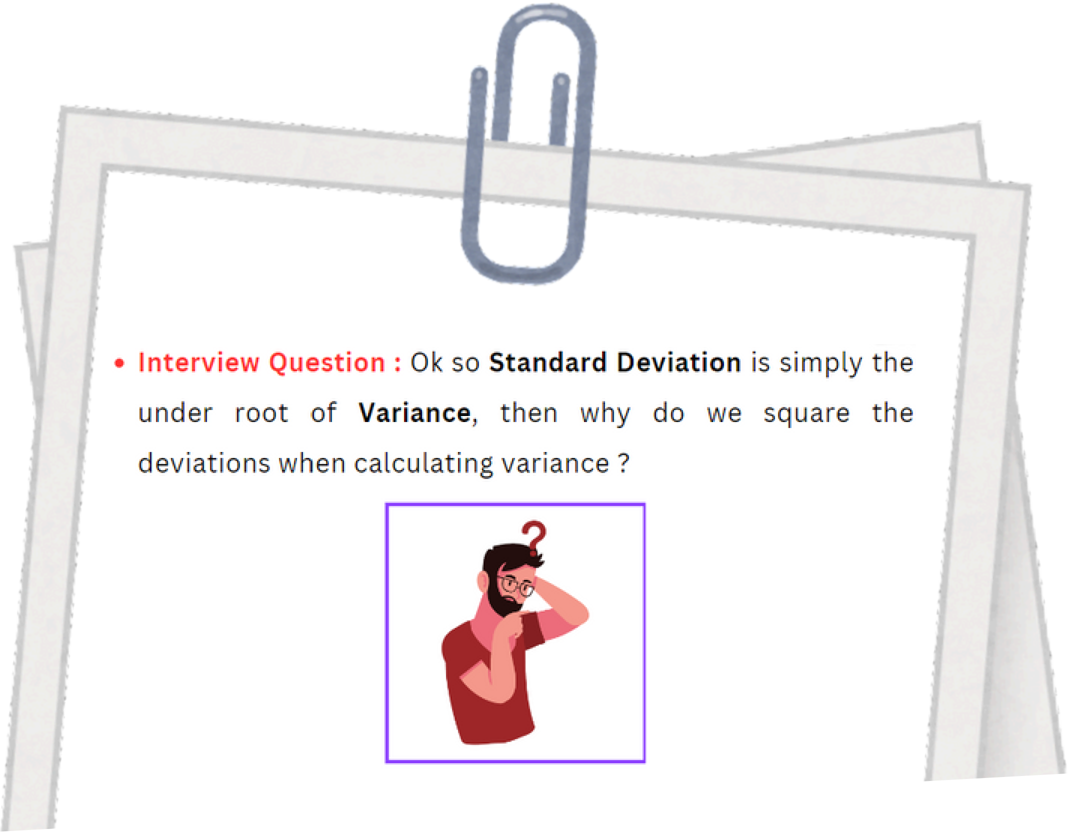
Q1

Q2

Q3

Agenda

- INTERVIEW QUESTION
- REMAINING KEY COMPONENTS OF VARIABILITY
- GET TO KNOW EACH COMPONENTS
- CHARACTERISTICS & APPLICATIONS OF EACH COMPONENTS
- PRACTICE TASK



- Interview Question : Ok so **Standard Deviation** is simply the under root of **Variance**, then why do we square the deviations when calculating variance ?

- Remember this guy from my **previous post**, let’s help him
- Let’s find out **why we do use square method** ,while calculating deviations or variability ???

- Let’s understand first why we are squaring to calculate deviation
- Example :

  Data : 10,12,14,16,18

  Step 1 - Mean Calculation Result = 14

  Step 2 - Calculate Deviation from the mean for each datapoint

  $10-14=-4$, $12-14=-2$, $14-14=0$, $16-14=2$, $18-14=4$

  Step 3 - Sum the Deviation **(without squaring) :**

  $-4+(-2)+0+2+4=0$

- **Observation** :The sum of the deviations is 0, even though the data points are clearly spread out.
- **Reason** : Negative and Positive values cancelled out each other.

→ **Solution......**

Approach 1 :Mean Absolute Deviation (MAD)

Calculate the absolute value (modulus) of the deviations

Step 2 - Calculate **Absolute Deviation** from the mean for each datapoint

|10− 14| =4 , |12 - 14| = 2, |14− 14| = 0, |16−14| = 2, |18−14   = |4|

Step 3 - Sum the **Absolute Deviations**

4+2+0+2+4=12

Step 4 - Mean Absolute Deviation

12/5 = 2.4 ✓

So why don’t we use this ???

WHY

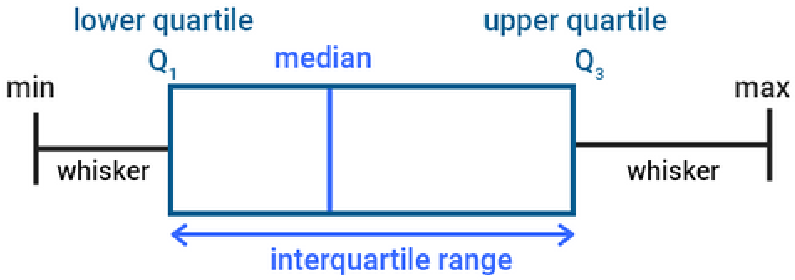
MAD is not used

- **Less Sensitivity to Outliers**: MAD is less sensitive to outliers, which can be a drawback if outliers are important to your analysis.
- **(V.IMP) Non-differentiability:** The absolute value function is not differentiable at zero, which can complicate mathematical operations, particularly in optimization problems or calculus-based methods. This makes MAD less convenient for more advanced statistical modeling and analysis compared to variance and standard deviation.
- **Lack of Relation to Normal Distribution:** Many statistical techniques, particularly those based on the normal distribution (like confidence intervals and hypothesis testing), rely on variance and standard deviation. MAD does not have the same direct relationship with these concepts, making it less suitable in these contexts.

Let’s Continue our Journey and explore remaining components of Variability :

- Interquartile Range (IQR)
- Coefficient of Variation (CV)

5 NUMBER SUMMARY



Minimum: The smallest data point in the dataset.
Q1 (First Quartile): The median of the lower half of the dataset (25th percentile).
Median (Q2 or Second Quartile): The middle value of the dataset (50th percentile).
Q3 (Third Quartile): The median of the upper half of the dataset (75th percentile).
Maximum: The largest data point in the dataset.

INTERQUARTILE RANGE (IQR)

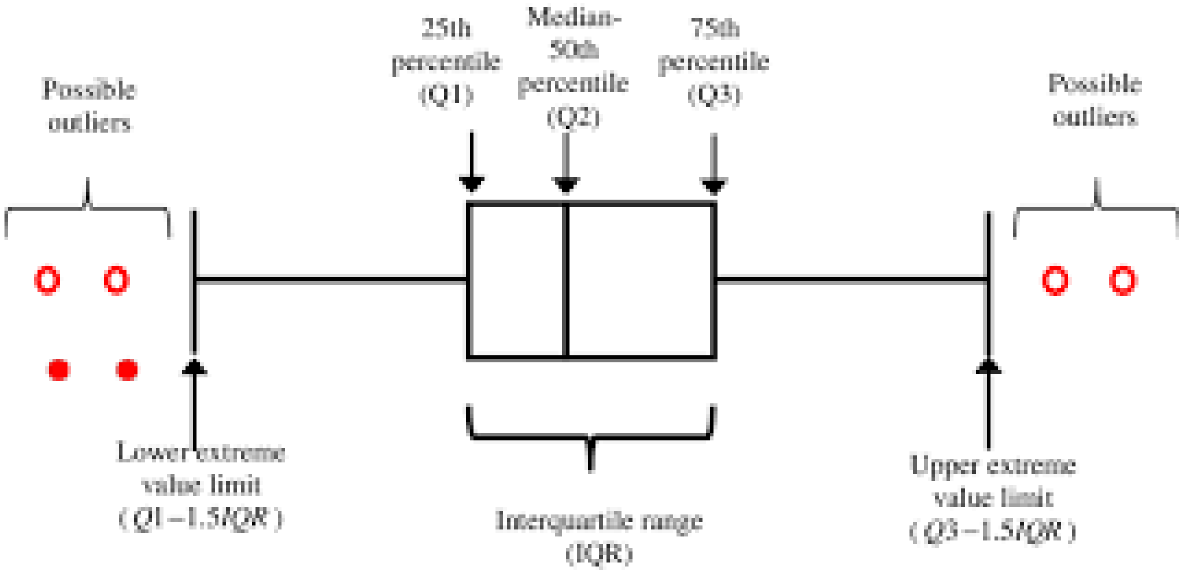
- Interquartile Range (IQR) is **derived from the five-number summary**, which represents the range within which the central **50% of a dataset** lies.
- It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1):

$$IQR=Q3-Q1$$

- IQR is a **robust measure of variability** because it only considers the middle 50% of the data, ignoring extreme values (outliers).

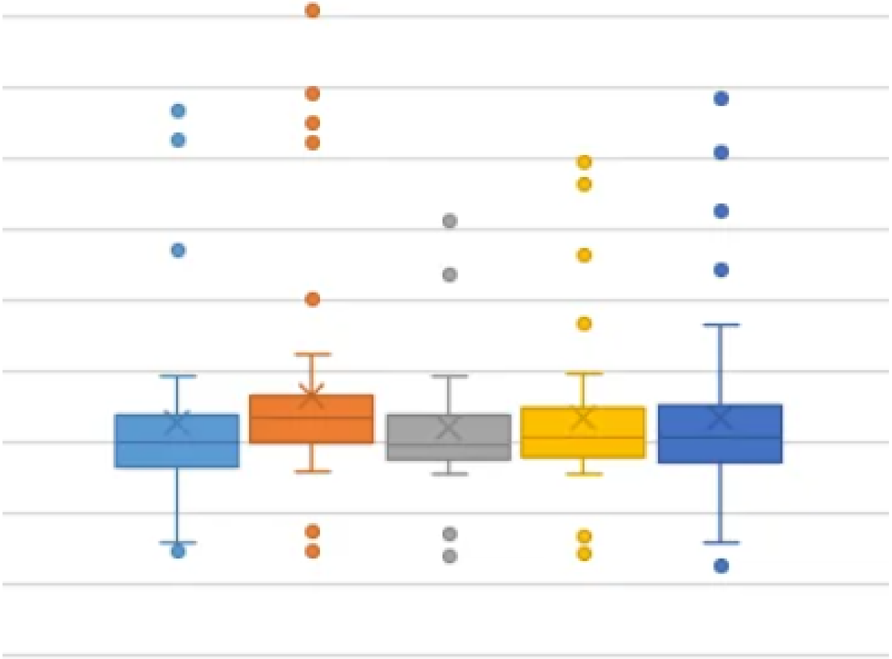
APPLICATION
OF IQR

- **Identifying Outliers:** Data points that lie **below Q1−1.5×IQR** or **above Q3+1.5×IQR** are often considered outliers.


25th percentile (Q1)
Median- 50th percentile (Q2)
75th percentile (Q3)
Possible outliers
Possible outliers
Lower extreme value limit ($Q1-1.5IQR$)
Interquartile range (IQR)
Upper extreme value limit ($Q3-1.5IQR$)

APPLICATION
OF IQR

- **Comparison of Data :**Useful for comparing the spread of different datasets visually using **BOXPLOT**



THANK YOU BRO FOR SHOWING ME OUTLIERS IN A GLANCE,SAVED LOT OF MY TIME
Data Analyst
BOX PLOT

COEFFICIENT OF VARIATION (CV)

$$\text{Coefficient of Variation Formula} = \frac{\text{Standard Deviation}}{\text{Mean}}$$

- Coefficient of Variation (CV) is a statistical measure of the **relative variability** of data points in a dataset.
- The CV allows for the **comparison of the degree of variation** between different datasets, even if they have **different units or widely different means.**

APPLICATION
OF CV

- **Comparing Risk:** In finance, the CV is often used to assess the risk-to-return ratio of an investment. A higher CV indicates more risk relative to the expected return.
- **Quality Control:** In manufacturing, the CV is used to assess the consistency of processes or product quality.
- **Unlike standard deviation,** which is an **absolute measure,** the CV provides a **relative measure** of variability in relation to the mean.

TASK FOR YOU

- Calculate 5 Number Summary for this data and also find out potential outliers :
  2,3,5,7,8,10,11,13,14,15,18,19,21,22,25,28,30,100,105,110

- Utilize these 2 buddies :

THANK YOU

Share your thoughts and feedback !!

👍 🔁