Statistics

1. Introduction to Statistics

Statistics is the branch of mathematics dealing with the **collection, organization, analysis, interpretation, and presentation** of data.

- **Data:** A collection of facts, numbers, or measurements.
- **Discrete Data:** Data that can only take specific values (e.g., number of students in a class).
- **Continuous Data:** Data that can take any value within a range (e.g., height, weight, time).

2. Measures of Central Tendency

These are single values that attempt to describe a set of data by identifying the central position within that set.

A. The Mean ($\bar{x}$)

The average of all data points.

- **Formula (Ungrouped):** $\bar{x} = \frac{\sum x}{n}$
- **Formula (Grouped):** $\bar{x} = \frac{\sum fx}{\sum f}$
    - *Where f = frequency and x = class midpoint.*

B. The Median

The middle value when data is arranged in ascending or descending order.

- **Odd number of data (n):** $\text{Position} = \frac{n+1}{2}$
- **Even number of data (n):** The average of the two middle terms.

C. The Mode

The value that appears most frequently in a data set.

- **Bimodal:** A set with two modes.
- **Modal Class:** In grouped data, the class with the highest frequency.

3. Presentation of Data

To make data understandable, we use various visual tools:

1. **Frequency Distribution Table:** A table showing marks and how often they occur.
2. **Bar Chart:** Uses rectangular bars to represent categories (bars do not touch).
3. **Histogram:** Similar to a bar chart but for grouped data (bars **must** touch).
4. **Pie Chart:** A circle divided into sectors.
    - **Calculation:**

$$\text{Angle} = \frac{\text{Frequency}}{\text{Total Frequency}} \times 360^\circ$$

5. **Cumulative Frequency Curve (Ogive):** A graph showing the running total of frequencies. It is used to estimate the **Median**, **Quartiles**, and **Percentiles**.

4. Measures of Dispersion (Range & Variation)

These measures tell us how "spread out" the data is from the center.

A. Range

The difference between the highest and lowest values.

$$\text{Range} = \text{Highest Value} - \text{Lowest Value}$$

B. Mean Deviation

The average of the absolute differences between each data point and the mean.

$$\text{M.D} = \frac{\sum |x - \bar{x}|}{n}$$

C. Variance and Standard Deviation (σ)

Standard Deviation is the most reliable measure of dispersion.

- **Variance (σ^2):** $\frac{\sum(x - \bar{x})^2}{n}$
- **Standard Deviation (σ):** $\sqrt{\text{Variance}}$

5. Grouped Data Calculations

When dealing with large sets of data organized into classes (e.g., 10–19, 20–29):

Class Interval	Frequency (f)	Mid-point (x)	fx
1–5	2	3	6
6–10	4	8	32
Total	$\sum f = 6$		$\sum fx = 38$

Example Calculation:

Mean ($\bar{x}$) = $\frac{38}{6} = 6.33$

6. Quartiles and Percentiles

- **Lower Quartile (Q_1):** The 25^{th} percentile ($1/4$ of the data).
- **Middle Quartile (Q_2):** The Median ($1/2$ of the data).
- **Upper Quartile (Q_3):** The 75^{th} percentile ($3/4$ of the data).
- **Interquartile Range:** $Q_3 - Q_1$

Quick Study Tips

- **Always check if the data is ordered** before calculating the Median.
- **In Histograms,** the area of the bar represents the frequency.
- **Standard Deviation** is always the square root of the variance—don't forget the final step!

Practice Example 1: Calculating the Mean and Standard Deviation

Let's look at a typical exam-style question for a set of test scores: **2, 4, 6, 8, 10**.

Step 1: Find the Mean ($\bar{x}$)

$$\bar{x} = \frac{2 + 4 + 6 + 8 + 10}{5} = \frac{30}{5} = 6$$

Step 2: Calculate the Variance

We find the square of the difference between each score and the mean:

- $(2-6)^2 = (-4)^2 = 16$
- $(4-6)^2 = (-2)^2 = 4$
- $(6-6)^2 = (0)^2 = 0$
- $(8-6)^2 = (2)^2 = 4$
- $(10-6)^2 = (4)^2 = 16$

Sum of squares = $16 + 4 + 0 + 4 + 16 = 40$.

$$\text{Variance} = \frac{40}{5} = 8$$

Step 3: Find the Standard Deviation (σ)

$$\sigma = \sqrt{8} \approx 2.83$$

Practice Example 2: The Cumulative Frequency Curve (Ogive)

The Ogive is essential for estimating values without doing heavy algebra.

How to use it:

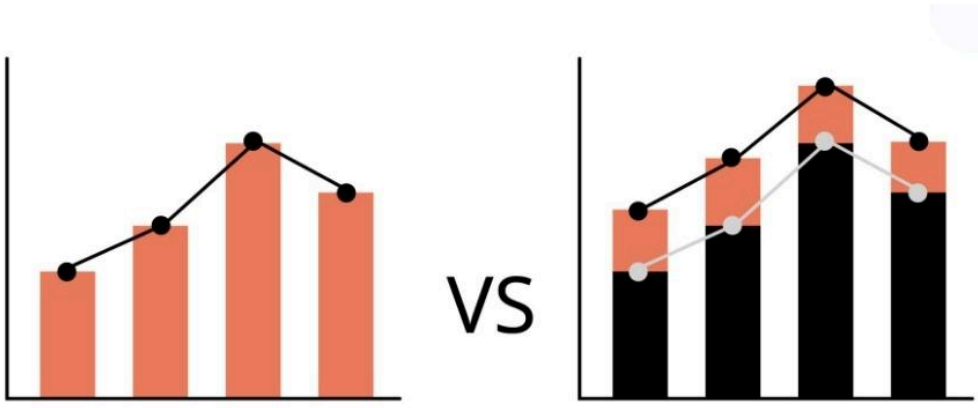
1. **To find the Median (Q_2):** Locate $\frac{N}{2}$ on the vertical (y) axis, move horizontally to the curve, and drop down to the horizontal (x) axis.

2. **To find the Interquartile Range (IQR):**
   - Find Q_1 at $\frac{1}{4}N$.
   - Find Q_3 at $\frac{3}{4}N$.
   - $IQR = Q_3 - Q_1$.

Comparison of Graphs

It is common to confuse a Histogram with a Bar Chart. Use this table as a quick guide:

Feature	Bar Chart	Histogram
Data Type	Qualitative / Discrete	Quantitative / Continuous
Gaps	There are gaps between bars	Bars are joined together
Width	Bar width is uniform	Width represents the class interval
Vertical Axis	Represents Frequency	Represents Frequency Density



Summary Checklist for Exams:

- **Mean:** Use for data without extreme outliers.
- **Median:** Best for "skewed" data (where there are very high or very low numbers).
- **Mode:** Use for "popularity" (e.g., most common shoe size).
- **Standard Deviation:** The gold standard for measuring how much scores vary from the average.