Pre-Read: Analyzing Sachin Tendulkar's ODI Career 2

---

1. Descriptive Statistics

Descriptive statistics are methods used to summarize and describe data. Instead of looking at raw data values, we use numerical and graphical summaries to capture the "big picture."

- What it does: Provides measures such as averages, spread, and shape of the data.
- Why it matters: It's the first step in data analysis—helping us understand patterns, detect anomalies, and prepare for deeper inferential statistics.
- Examples: Mean test score of a class, spread of salaries in a company, or percentage of customers preferring a product.

---

2. Measures of Central Tendency

These describe the center or typical value of a dataset.

- Mean (Average): Sum of values ÷ Number of values. Sensitive to outliers.
- Median: Middle value when data is ordered. Robust against outliers.
- Mode: Most frequently occurring value. Useful for categorical data.

Why important?

- Gives a single "representative" value.

- Different measures are useful in different contexts (e.g., median salary is better than mean when there are a few billionaires).

---

3. Measures of Spread (Variability)

Spread tells us how much the data values differ from each other and from the central tendency.

- Range: Difference between max and min. Quick but sensitive to outliers.
- Variance: Average of squared deviations from the mean.
- Standard Deviation (SD): Square root of variance. Expressed in the same units as the data.
- Interquartile Range (IQR): Spread of the middle 50% of data (Q3 – Q1).

Why important?

- Central tendency alone can be misleading (two classes can have the same mean score but one may have students tightly clustered while another has wide variation).
- Spread measures help in understanding risk, consistency, and reliability.

---

4. Normal Distribution

A special probability distribution shaped like a bell curve.

- Properties: Symmetrical, unimodal, mean = median = mode, defined by mean (μ) and standard deviation (σ).
- Empirical Rule (68–95–99.7):

- ~68% of data lies within 1 SD of mean
- ~95% within 2 SDs
- ~99.7% within 3 SDs

Why important?

- Many natural and business phenomena (heights, test scores, stock returns) follow a normal distribution.
- Underpins hypothesis testing, confidence intervals, z-scores, and much of inferential statistics.

---

5. Random Variables & Probability Distribution Functions (PDFs)

A random variable (RV) assigns numerical values to outcomes of a random experiment.

- Discrete RV: Takes countable values (e.g., number of goals scored).
- Continuous RV: Takes infinitely many values (e.g., time taken to run 100m).

Probability Distribution Function (PDF): Describes how probabilities are distributed over values of a random variable.

- For discrete RVs → Probability Mass Function (PMF): $P(X=x)$P(X = x)P(X=x).
- For continuous RVs → Probability Density Function (PDF): Probabilities from areas under the curve, not point values.

Why important?

- Random variables and PDFs form the foundation of probability modeling.

- They allow us to calculate probabilities, expectations, and risks in uncertain situations (e.g., predicting demand, modeling investment returns, or assessing medical test outcomes).