Statistics

Descriptive Vs Inferential

Descriptive

- **organizing, summarizing, and visualizing data**. Doesn't make predictions.
    - Organizing and summarizing data using number & graphs

1. Data Summary:
    1. **Bar Graphs** - Compare categories easily.
    2. **Histograms** - Show frequency distributions.
    3. **Pie Charts** - Represent proportions in a cool, circular way.
2. Measures of Central Tendency:
    1. **Mean(Average)** - The most common way to summarize data.
    2. **Mode** - The most frequently occurring value.
    3. **Median** - The middle value when data is sorted.
3. Measures of Variability:
    1. **Range** - Difference between the largest and smallest values.
    2. **Variance** - How much the data points deviate from the mean.
    3. **Standard deviation** - A measure of data spread → lower values mean data is clustered, higher values mean it's more spread out.

Inferential

- Predicting the Future(استغفر الله العظيم يارب), instead of just describing the data, they help us **make conclusions** and **predictions**.

1. Using sample data to make an inference or draw a conclusion.
2. Uses probability(the core) to determine how confident we can be that the conclusions we make are correct .

Feature	Descriptive	Inferential
Purpose	Summarizes and organizes data	Makes predictions
Focus	Past & present data	Future trends & conclusions التنبؤ بالغيب واليعوذ بالله
Methods	Graphs, averages, dispersion	Probability, sampling
E.x	The average test score is 75	Students who study 2+ hours daily score higher

---

Mean, Mode, Median, Range

1. **Mean (Average):**
   - Sum of all values ÷ Number of values
   - **Sensitive to outliers** (extreme values can skew it)
2. **Mode:**
   - The **value that appears most often** in the dataset
   - Dataset can have **one mode (unimodal), two modes (bimodal), or more (multimodal)**
   - If **no number repeats**, there's **no mode**
3. **Median:**
   - Arrange data **in order**, then pick the middle number
   - If even numbers, take the **average of the two middle values**
   - **Not affected by outliers**, making it a good choice for skewed data
4. **Range(Spread of data):**
   - Highest value - Lowest value
   - Shows how **spread out** the data

When to use what?

- **Mean:** when data is balanced (no extreme values)
- **Median:** when data have outliers
- **Mode** when looking for **most common occurrences**
- **Range** to check **variability** in data

Variance

- Variance (**σ^2 for population, s^2 for sample**) measures **how spread out the data is** from the **mean**. A **higher variance** means data points are more spread out, while a **lower variance** means they are closer to the mean.
- **Steps to Calculate Variance:**
  1. Find the Mean (Average): $\sum X/N$
  2. Subtract the Mean from Each Value: $(X - \mu)$
  3. Square Each Deviation: $(X - \mu)^2$
  4. Find the Average of Squared Deviations
     - For Population Variance (σ^2): $\sum(X - \mu)^2/N$
     - For Sample Variance (s^2): $\sum(X - \mu)^2/N - 1$
       - $N - 1$ to correct bias
- **Keys:**

    - **Use σ^2 for populations , s^2 for samples**
    - **Variance is always non-negative**

> **The square root of variance = Standard Deviation (σ or s)**

Interquartile range (IQR) & Detect outliers

- Measures the **spread of the middle 50% of data**. It's useful for detecting **outliers** and understanding variability without being affected by extreme values.
- **Steps to find IQR:**
    1. **Arrange the data in Ascending Order**
    2. **Find Q2: (Middle Quartile == 50th Precentile)**
    3. **Find Q1: (First Quartile == 25th Precentile)**
    4. **Find Q3: (Third Quartile == 75th Precentile)**
    5. **Compute the IQR:** $(Q_3 - Q_1)$
- **How to find the outlier:**
    - **Outlier** is any value that is **too far from the rest of the data**
        1. **Calculate the range:** $[Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR]$
        2. **Identify Outliers**
            - Any value **less than** the **Lower Bound** is an **outlier**
            - Any value **greater than** the **Upper Bound** is an **outlier**
