Starting with Statistics

Contents

Descriptive Statistics Vs Inferential Statistics

Statistics is the science of collecting, analyzing, and interpreting data. There are two main categories of statistics, Descriptive and Inferential Statistics.

1. **Descriptive Statistics:**

   Organizing and summarizing data using numbers and graphs.

   - **Graph Representation:**
     - Bar Graph, Histogram, Pie Chart, etc…
     - Shape of Graph and Skewness
   - **Numbers as Measurements:**
     - Measure of Central Tendency: mean, median, & mode
     - Measure of Variability: range, variance, & standard deviation

2. **Inferential Statistics:**
   Using sample data to make an inference – finding out something that you can find out indirectly from what you already know. Uses probability to prove conclusions are correct

Mean, Median, Mode, and Range

The **mean** (or average) is the sum of all values in a dataset divided by the number of values. It represents the central value of the data.

The **median** is the middle value when the numbers are arranged in ascending order. If there is an odd number of values, the median is the middle one. If there is an even number, the median is the average of the two middle values.

The **mode** is the number that appears most frequently in a dataset. A dataset can have no mode, one mode, or multiple modes.

The **range** is the difference between the highest and lowest values in a dataset. It measures the spread of the data.

Variance

Variance is a measure of how much the data values spread out from the mean. A **low variance** means the data points are close to the mean, while a **high variance** indicates that the data points are more spread out.

- **Population Variance** (σ^2): Used when calculating variance for an entire population.
- **Sample Variance** (s^2): Used when calculating variance from a sample of the population.

To calculate variance:

1. first find the **mean** of the data.
2. Then, subtract the mean from each value and square the differences.
3. Next, find the **average of these squared differences**.
    - For **population variance (σ^2)**, divide by the total number of values N,
    - while for **sample variance (s^2)**, divide by $n-1$.

Standard Deviation

The **standard deviation** is a measure of how spread out the data points are from the mean. It is the **square root of the variance**, making it easier to interpret since it has the same units as the original data. A **higher standard deviation** means more spread, while a **lower standard deviation** means the data is closer to the mean.

- **Population Standard Deviation (σ)**: Used when calculating for an entire population.
- **Sample Standard Deviation (s)**: Used when calculating from a sample of the population.

To Calculate Standard Deviation:

1. Find the Mean
2. Subtract the Mean from Each Value
3. Square Each Difference
4. Find the Variance
5. Take the Square Root of the Variance

Interquartile Range (IQR) & Outliers

The **Interquartile Range (IQR)** measures the spread of the middle 50% of the data. It helps in identifying data variability while reducing the effect of extreme values (outliers).

How to Calculate IQR:

1. **Order the Data**: Arrange the values in ascending order.
2. **Find Q2 (Second Quartile)**: median of all values
3. **Find Q1 (First Quartile)**: The **median** of the lower half of the data (25th percentile).
4. **Find Q3 (Third Quartile)**: The **median** of the upper half of the data (75th percentile).
5. **Calculate IQR:** IQR = Q3 − Q1

Outliers

Outliers are extreme values that are significantly different from the rest of the data. They are identified using the **1.5 × IQR rule**.

How to Identify Outliers:

1. **Find the Lower Bound**: Lower Bound = Q1 − (1.5×IQR)
   Find the Upper Bound: Upper Bound=Q3+(1.5×IQR
2. **Any values below the Lower Bound or above the Upper Bound are outliers.**