Descriptive Statistics

Descriptive statistics is a branch of statistics that focuses on summarizing and organizing data in a meaningful way. It provides a snapshot of the main features of a dataset through numerical measures, visualizations, and tabular representations, helping researchers and analysts understand the data briefly.

Measures of Central Tendency

These measures describe the central or typical value of a dataset, giving an idea of the average or most common data point.

- **Mean (Average):** The arithmetic mean is calculated by summing all data points and dividing by the total number of points. It provides a general idea of the data's central value but can be skewed by outliers.

$$\boldsymbol{Mean} = \frac{\sum \mathbf{x_i}}{\boldsymbol{N}}$$

Where x_i is the data points, and N is the number of points.

- **Median:** The median is the middle value in a dataset when the values are arranged in ascending or descending order. For datasets with an even number of observations, the median is the average of the two middle numbers.
- **Mode:** The mode is the most frequently occurring value in a dataset. A dataset can have one mode (unimodal), more than one mode (multimodal), or no mode at all.

Measures of Dispersion

These measures describe the spread or variability of the data, showing how much the data deviates from the central tendency.

- **Range:** The range is the difference between the maximum and minimum values in a dataset.

$$\boldsymbol{Range} = \boldsymbol{Max_{value}} - \boldsymbol{Min_{value}}$$

- **Variance:** Variance measures the average squared deviation of each data point from the mean. It provides an understanding of data spread but is expressed in squared units.

$$\boldsymbol{Variance}\ (\boldsymbol{\sigma^2}) = \frac{\sum(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^{\mathbf{2}}}{\boldsymbol{N}}$$

- **Standard Deviation (SD):** The standard deviation is the square root of variance and represents data spread in the same units as the data itself.

$$\boldsymbol{SD}(\boldsymbol{\sigma}) = \sqrt{\boldsymbol{Variance}}$$

- **Interquartile Range (IQR):** The IQR measures the spread of the middle 50% of data by subtracting the first quartile (Q1, 25th percentile) from the third quartile (Q3, 75th percentile).

$$\boldsymbol{IQR = Q3 - Q1}$$

Measures of Shape

These measures describe the distribution pattern or shape of the dataset.

- **Skewness:** Skewness quantifies the asymmetry of the data distribution. Positive skewness indicates a long tail on the right, while negative skewness indicates a long tail on the left.
    - Symmetrical data: Skewness ≈ 0
    - Right-skewed: Skewness > 0
    - Left-skewed: Skewness < 0
- **Kurtosis:** Kurtosis measures the "tailedness" of the data distribution. It indicates whether the data has heavy or light tails compared to a normal distribution.
    - Mesokurtic: Normal distribution (Kurtosis ≈ 0)
    - Leptokurtic: Heavy tails (Kurtosis > 0)
    - Platykurtic: Light tails (Kurtosis < 0)

Inferential Statistics

Inferential statistics is a branch of statistics that focuses on generalizing, predictions, or decisions about a population based on a sample of data. Unlike descriptive statistics, which merely summarizes data, inferential statistics uses probability theory to draw conclusions about a larger group based on observed data.

Inferential Statistics aims at:

- **Generalization:** Extend findings from a sample to the entire population.
- **Estimation:** Estimate population parameters (e.g., mean, proportion).
- **Hypothesis Testing:** Determine whether observed data supports a specific claim about the population.
- **Prediction:** Forecast future outcomes based on trends or patterns in the sample data.

1. Population vs. Sample

- **Population:** The entire group of individuals or items of interest.

  Example: All voters in a country.

- **Sample:** A subset of the population chosen for analysis.

  Example: 1,000 voters selected for a survey.

2. Parameters vs. Statistics

- **Parameter:** A numerical measure that describes a characteristic of the population (e.g., population mean).
- **Statistic:** A numerical measure that describes a characteristic of the sample (e.g., sample mean).

3. Sampling Methods

Choosing a representative sample is crucial for accurate inferences. Common sampling methods include:

- **Simple Random Sampling:** Every individual has an equal chance of being selected.
- **Stratified Sampling:** The population is divided into strata, and samples are drawn from each stratum.
- **Cluster Sampling:** The population is divided into clusters, and entire clusters are randomly selected.
- **Systematic Sampling:** Every k^{th} individual is selected after a random starting point.

4. Probability and Distributions

Inferential statistics relies heavily on probability theory and probability distributions.

- **Normal Distribution:** A bell-shaped curve representing many natural phenomena. It is central to many inferential techniques.
- **Sampling Distribution:** The probability distribution of a statistic (e.g., sample mean) across many samples.

Techniques in Inferential Statistics

1. Estimation

Estimation involves calculating sample statistics to estimate population parameters.

- **Point Estimation:** Provides a single value estimate for a parameter. Example: Sample mean as an estimate of population mean (μ).
- **Interval Estimation (Confidence Intervals):** Provides a range of values within which the parameter is likely to lie, with a given confidence level (e.g., 95% confidence interval).

2. Hypothesis Testing

Hypothesis testing evaluates whether a claim about a population parameter is supported by sample data.

- **Null Hypothesis (H_0):** Assumes no effect or no difference (e.g., $\mu = 50$).
- **Alternative Hypothesis (H_a):** Suggests an effect or difference (e.g., $\mu/50$).

Probability and Frequency Distribution

Probability

Probability is a measure of the likelihood of a specific event occurring, expressed as a value between 0 and 1. A probability of 0 means the event cannot happen, while a probability of 1 means the event is certain.

- **Experiment:** Any process or activity that produces outcomes. For example, rolling a die or flipping a coin.
- **Outcome:** A possible result of an experiment. Example: Rolling a 4 on a die.
- **Event:** A collection of one or more outcomes. Example: Rolling an even number on a die.
- **Sample Space (S):** The set of all possible outcomes of an experiment. Example: For a six-sided die,

S = {1,2,3,4,5,6}.

$$\boldsymbol{P(E) = \frac{Number\ of\ Favourable\ Outcomes}{Total\ Number\ of\ Outcomes}}$$

Where *P(E)* is the probability of event *E*.

Types of Probability

1. **Classical Probability:** Based on equally likely outcomes. Example: The probability of rolling a 3 on a fair die is P(3) = 16P(3) = \frac{1}{6}P(3)=61.
2. **Empirical Probability:** Based on experimental or observed data. Example: If a die is rolled 100 times and a 3 appears 20 times, P(3) = 20100=0.
3. **Subjective Probability:** Based on personal judgment or experience. Example: Estimating the probability of rain tomorrow as 70%.

Rules of Probability

1. **Addition Rule:** For mutually exclusive events (A and B cannot happen simultaneously):

$$P(A\ or\ B) = P(A) + P(B)$$

2. **Multiplication Rule:** For independent events (A and B do not influence each other):

$$P(A\ and\ B) = P(A)\ \times P(B)$$

3. **Complement Rule:** The probability of the complement of an event E (not E) is:

$$P(not\ E) = 1 - P(E)$$

Frequency Distribution

Frequency distribution is a statistical technique that organizes raw data into categories or intervals, making it easier to analyse patterns and trends.

- **Class Interval:** A range of values into which data is grouped. Example: 0–10, 11–20.
- **Frequency (f):** The number of data points falling within a class interval.
- **Class Width:** The difference between the upper and lower boundaries of a class interval.

$$\boldsymbol{Class\ Width = Upper\ Bound - Lower\ Bound}$$

- **Relative Frequency:** The proportion of data points in each class relative to the total number of data points.

$$\boldsymbol{Relative\ Frequency = \frac{Frequency\ of\ Class}{Total\ Frequency}}$$

- **Cumulative Frequency:** The running total of frequencies up to a certain class.

Visual Representations of Frequency Distribution

1. **Histogram:** A bar graph where the x-axis represents class intervals and the y-axis represents frequencies.
2. **Frequency Polygon:** A line graph connecting midpoints of class intervals at their corresponding frequencies.
3. **Ogive (Cumulative Frequency Graph):** A line graph showing cumulative frequencies.

Relationship Between Probability and Frequency Distribution

- **Empirical Probability:** Probability can be estimated from frequency distribution by dividing the frequency of an event by the total number of observations.

$$\boldsymbol{P(E) = \frac{f}{N}}$$

Where *f* is the frequency of the event, and *N* is the total frequency.

Probability Distribution: If we map the probabilities of different outcomes (events), the result is a probability distribution. For example:

- Discrete Probability Distribution: Probabilities for discrete variables (e.g., rolling a die).
- Continuous Probability Distribution: Probabilities for continuous variables (e.g., heights of individuals).