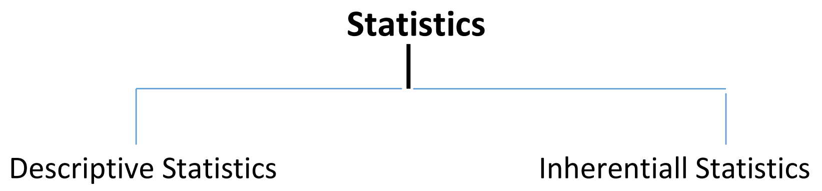

Statistics

Descriptive Statistics

Inherentiall Statistics

Descriptive Statistics

Aims to describe and summarize our data in a meaningful way. It provides a simple overview of the main characteristics of our data.

Key components

- Measures of central tendency
- Measures of dispersion
- Tables
- Charts

Measures of central tendency:

- Mean – defines sum of all observations divided number of observations.
- Median - When the values in a dataset are arranged in ascending order the median is the middle value.
- Mode – It refers to the value that appear most frequently in the data set.
    - If a dataset has one value that appears more frequently than any other it has one mode and is called unimodal

      Ex- 2, 3, 4, 4, 5, 6, 7

Measures of dispersion:

Describe how spread out the values in a dataset are measures of dispersion.

Three phases

- Variance & standard deviation
- Range
- Interquartile range

Variance & Standard deviation

The variance is a measure of variability. It is the average squared deviation from the mean. Standard deviation indicates the average distance b/n each data point and the mean.

Inferential Statistics

Refers to techniques that allow us to make inferences or predictions about a population based on a sample.

While descriptive statistics describe the data at hand, inferential statistics help generalize findings to a larger context or population. It involves estimation, hypothesis testing, regression and other techniques.

Key components

Population vs. Sample:

- **Population**: The entire group of individuals or data points.
- **Sample**: A subset of the population used to make inferences about the population.

Point Estimation: Point estimates are single values that serve as estimates of population parameters (e.g., estimating the population mean based on a sample mean).

Confidence Interval (CI): A range of values that is used to estimate the true population parameter, with a certain level of confidence (e.g., 95% confidence).

Hypothesis Testing: Hypothesis testing is used to test an assumption about a population parameter. It involves two hypotheses:

- **Null hypothesis (H_0)**: The default assumption, typically that there is no effect or difference.
- **Alternative hypothesis (H_1)**: The hypothesis that contradicts the null hypothesis.

t-test: A statistical test used to compare the means of two groups to determine if they are significantly different from each other.

Chi-Square Test: A test used to determine if there is a significant association between categorical variables.

Regression Analysis: Regression models help predict the relationship between variables. For example, simple linear regression models the relationship between two variables.

Frequency Distribution

Summary of how often each unique value or group of values appears in a dataset.

It's a fundamental concept used in descriptive statistics to organise data and gain insights about its distribution. It can be used for both categorical and numerical data, and they handling the detecting patterns and making decisions based on data.

Key components

Frequency: No of times a value or category appears in the dataset**.**

Relative Frequency: The proportion or percentage of times a value appears, calculated as:

Relative Frequency = Frequency of a value **/** Total number of observations

Cumulative frequency: The running total of frequencies up to a particular value or class. It's often used to calculate percentile ranks and understand the spread of data.

Class Intervals (for continuous data): In cases of continuous data, we group the data into intervals (or bins) to construct a grouped frequency distribution.