Descriptive Statistics
&
Inferential Statistics

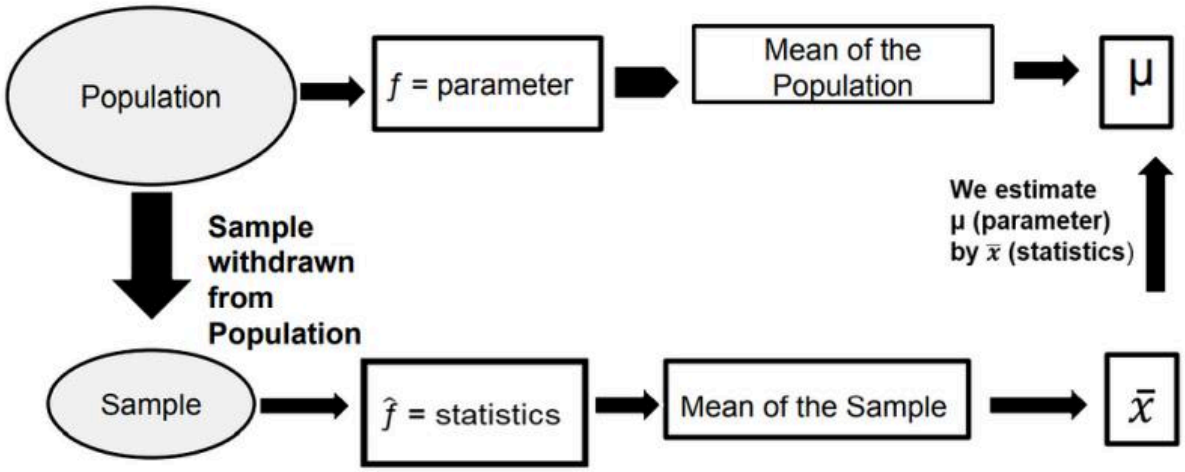
Prepared by
G.J.Rahul

Termilogies

Parameter: It is a measure that could be mean, median, variance, and many more for population data.

Statistic: It is a measure that could be mean, median, variance, and many more for sample data.

The following diagram, explains the co-relation between Parameter and Statistic in understandable manner:



Introduction

The first step of any data-related process is the collection of data. Once we have collected the data, what do we do with it? Data can be sorted, analyzed, and used in various methods and formats, depending on the project's needs. While analyzing a dataset, We use statistical methods to arrive at a conclusion.

What is Descriptive Statistics?

Descriptive statistics serves as the initial step in understanding and summarizing data. It involves organizing, visualizing, and summarizing raw data to create a coherent picture. The primary goal of descriptive statistics is to provide a clear and concise overview of the data's main features. This helps us identify patterns, trends, and characteristics within the data set without making broader inferences.

Key Aspects of Descriptive Statistics

- **Measures of Central Tendency:** Descriptive statistics include calculating the mean, median, and mode, which offer insights into the center of the data distribution.
- **Measures of Dispersion:** Variance, standard deviation, and range help us understand the spread or variability of the data.
- **Visualizations:** Creating graphs, histograms, bar charts, and pie charts visually represent the data's distribution and characteristics.

Types of Statistics

Aspect	Descriptive Statistics	Inferential Statistics
Purpose	Summarize and describe data	Draw conclusions or predictions
Data Sample	Analyzes the entire dataset	Analyzes a sample of the data
Examples	Mean, Median, Range, Variance	Hypothesis testing, Regression
Scope	Focuses on data characteristics	Makes inferences about populations
Goal	Provides insights and simplifies data	Generalizes findings to a larger population
Assumptions	No assumptions about populations	Requires assumptions about populations
Common Use Cases	Data visualization, data exploration	Scientific research, hypothesis testing

What is Inferential Statistics?

Inferential Statistics let's us make predictions and decisions about larger group based on smaller sample.

Key Aspects of Inferential Statistics

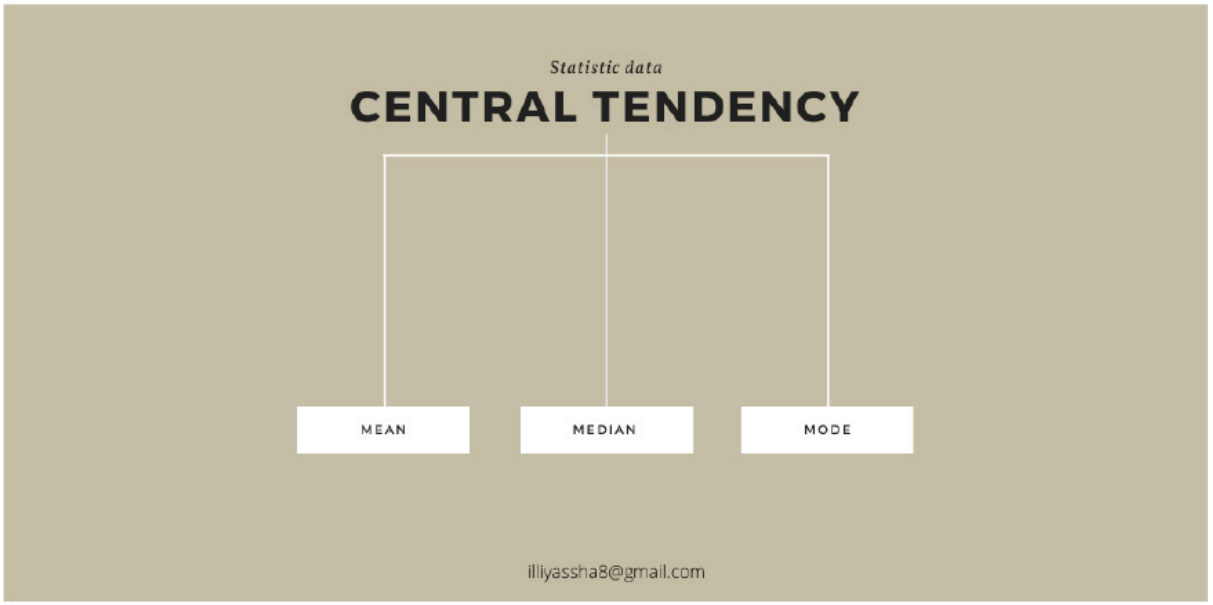
Sampling Techniques: Inferential statistics relies on carefully selecting representative samples from a population to make valid inferences.

Hypothesis Testing: This process involves setting up hypotheses about population characteristics and using sample data to determine if these hypotheses are statistically significant.

Confidence Intervals: These provide a range of values within which we're confident a population parameter lies based on sample data.

Regression Analysis: Inferential statistics also encompass techniques like regression analysis to model relationships between variables and predict outcomes.

Mean Median Mode



Mean:
The “Mean” is the average of the data. The average can be identified by summing up all the numbers and then dividing them by the number of observations.

Example:

Data – 10,20,30,40,200

Mean = [10+20+30+40+200] / 5
Mean = 60

Median:
The median can be identified by ordering the data, splitting it into two equal parts, and then finding the number in the middle. It is the best way to find the center of the data.

Example:
Odd number of Data – 10,20,30,40,50
Median is 30.
Even the number of data – 10,20,30,40,50,60

Find the middle 2 data and take the mean of those two values.

30 and 40 are middle values.

Now, add them and divide the result by 2
30+40 / 2 =35
Median is 35

Mode:
The mode of the data is the most frequently occurring data or elements in a dataset. If an element occurs the highest number of times, it is the mode of that data. If no number in the data is repeated, then that data has no mode. There can be more than one mode in a dataset if two values have the same frequency, which is also the highest frequency.

Example:
Data – 1,3,4,6,7,3,3,5,10, 3
Mode is 3, because 3 has the highest frequency (4 times)

Inter Quartile Range (IQR)

- Quartiles are special percentiles.
- 1st Quartile Q1 is the same as the 25th percentile.
- 2nd Quartile Q2 is the same as 50th percentile.
- 3rd Quratile Q3 is same as 75th percentile

Steps to find quartile and percentile:

- The data should sorted and ordered from the smallest to the largest.
- For Quartiles, ordered data is divided into 4 equal parts.
- For Percentiles, ordered data is divided into 100 equal parts.

The Inter Quartile Range is the difference between the third quartile (Q3) and the first quartile (Q1)

$$IQR = Q3 - Q1$$

Range

The range is the difference between the largest and the smallest value in the data.

Bivariate analysis often involves techniques such as:

Scatter Plots: These visualizations showcase the relationship between two variables, with each data point plotted on the graph.
Correlation: Calculating correlation coefficients helps you quantify the strength and direction of the relationship between variables.
Regression Analysis: This technique allows you to model the relationship between variables, predicting the outcome of one based on the other.

Apart from the above-mentioned Descriptive statistics, the following are commonly used:

1. Categorical Variables: These are variables that represent distinct groups or categories. Analysis often involves graphical representations and contingency tables to understand the relationships between categories.
2. Contingency Tables: These tables are used to display the frequency distribution of categorical variables. They help in analyzing the relationship between different categorical variables in multivariate data.
3. Box Plot: A graphical representation that shows the distribution of a dataset through its quartiles. It highlights the median, quartiles, and extreme values, providing a clear picture of the data's spread and potential outliers.
4. Graphical Representation: This involves using visual tools like box plots, histograms, and scatter plots to summarize and analyze data, making it easier to identify patterns, trends, and extreme values in both univariate and multivariate datasets.
5. Extreme Values: These are the data points that are significantly higher or lower than the majority of the data. They can heavily influence the mean and standard deviation and are often highlighted in box plots and other graphical representations

Summary

`Can Descriptive Statistics be used to make inferences or predictions?`

Descriptive statistics summarize the data you have. They use measures like mean, median, and standard deviation to give you a general idea of what the data looks like. This process often involves exploratory data analysis, where open exploration of the data can reveal patterns and insights. For instance, calculating mean scores is a common part of this analysis.

Inferential statistics use your data to conclude a larger population. This allows you to make predictions about things you haven't observed yet. Here, you would identify the dependent and independent variables in your study, which are crucial for making these inferences.

Think of it like this:
Descriptive statistics describe your apartment, while inferential statistics use the features of your apartment to guess about the entire apartment building.

So, while descriptive statistics can't directly predict the future, they can help you understand the data and prepare it for inferential statistics, which can then be used for predictions. Summary statistics from your exploratory data analysis can provide the foundation for these predictive models.

References:
Descriptive Statistics Definitions, Types, Examples | Analytics Vidhya

Inferences Statistics

Key Concepts in Inferential Statistics

Population and Sample

Population: The entire set of individuals or items of interest in a study.

Sample: A subset of the population used to make inferences about the population. Sampling is essential because it is often impractical to collect data from the entire population.

Parameters and Statistics

Parameter: A numerical value that describes a characteristic of a population (e.g., population mean).

Statistic: A numerical value that describes a characteristic of a sample (e.g., sample mean).

Hypothesis Testing

A method for testing a hypothesis about a parameter in a population using data measured in a sample.

Null Hypothesis (H0): The assumption that there is no effect or difference.

Alternative Hypothesis (H1): The assumption that there is an effect or difference.

P-value: The probability of obtaining a result at least as extreme as the one observed, assuming the null hypothesis is true.

Significance Level (α): The threshold for rejecting the null hypothesis, commonly set at 0.05.

Confidence Intervals

A range of values, derived from a sample, that is likely to contain the population parameter.

For example, a 95% confidence interval means we are 95% confident that the interval contains the population parameter.

Regression Analysis

A statistical technique used to model and analyze the relationships between variables.

Simple Linear Regression:
Examines the relationship between two variables.

Multiple Regression:
Examines the relationship between multiple independent variables and a dependent variable.

ANOVA (Analysis of Variance):
A statistical method used to compare means across three or more groups to determine if at least one mean is different from the others.

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