Statistics

Descriptive Statistics

Descriptive Statistics is the branch of statistics that describes and summarizes data

Types of variables:-
Categorical variables typically have a small number of possible values that are called categories or classes
Numeric variables have numbers as values.
Count: the number of values that a variable consists of
Minimum: the smallest value of a numeric variable
Maximum: the largest value of a numeric variable

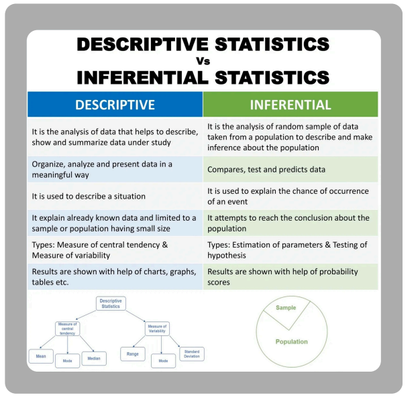
Measures of Central Tendancy: Descriptive statistics that attempt to represent a variable by its central value.
- **Mean:** add the values of the numeric variable and divide by the count
- **Median:** middle value in a sorted list of values of a numeric variable.
- **Mode:** value with the highest count

Measure of Variability: Descriptive statistics that represent how distant the typical values of a numeric variable are from mean
- **Variance:** for a sample ($s2$) represents how distant the values are from the mean.
- **Standard Deviation:** (s) for a sample is the square root of variance
- **Range:** the largest difference between any two values of a numeric variable, Range = Maximum - Minimum

Inferential Statistics

Inferential statistics are commonly used when the user needs to conclude for the whole population at hand, using different types of tests. It is a tool used to understand patterns by taking and evaluating a sample and drawing the requisite conclusions about a large population. We can generalize and forecast other relationships within the entire population using variables and the relations between them from the study, regardless of the degree.

Methods commonly used:
- Regression Analysis
- Correlation Analysis
- Statistical Significance
- Hypothesis Testing
- Variance/Co-Variance Analysis



Scales of Measurement

Nominal

A nominal scale is the 1st level of measurement scale in which the numbers serve as "tags" or "labels" to classify or identify the objects. A nominal scale usually deals with the non-numeric variables or the numbers that do not have any value.
Characteristics of Nominal Scale
- A nominal scale variable is classified into two or more categories. In this measurement mechanism, the answer should fall into either of the classes.
- It is qualitative. The numbers are used here to identify the objects.
- The numbers don't define the object characteristics. The only permissible aspect of numbers in the nominal scale is "counting."

Example:
What is your gender?
M- Male
F- Female
Here, the variables are used as tags, and the answer to this question should be either M or F.

Ordinal

The ordinal scale is the 2nd level of measurement that reports the ordering and ranking of data without establishing the degree of variation between them. Ordinal represents the "order." Ordinal data is known as qualitative data or categorical data. It can be grouped, named and also ranked.
Characteristics of the Ordinal Scale
- The ordinal scale shows the relative ranking of the variables
- It identifies and describes the magnitude of a variable
- Along with the information provided by the nominal scale, ordinal scales give the rankings of those variables
- The interval properties are not known
- The surveyors can quickly analyse the degree of agreement concerning the identified order of variables

Examples:
- Ranking of school students – 1st, 2nd, 3rd, etc.
- Ratings in restaurants
- Evaluating the frequency of occurrences
  - Very often
  - Not often
  - Not at all
- Assessing the degree of agreement
  - Totally agree
  - Agree
  - Neutral
  - Disagree
  - Totally disagree

Interval

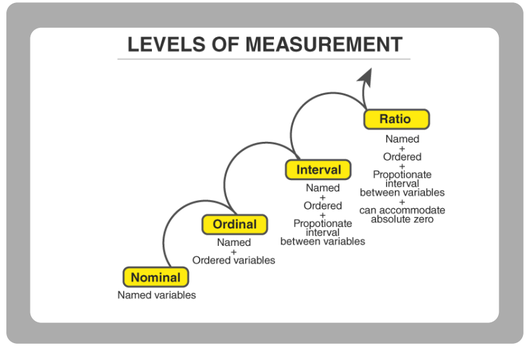
The interval scale is the 3rd level of measurement scale. It is defined as a quantitative measurement scale in which the difference between the two variables is meaningful. In other words, the variables are measured in an exact manner, not as in a relative way in which the presence of zero is arbitrary.
Characteristics of Interval Scale:
- The interval scale is quantitative as it can quantify the difference between the values
- It allows calculating the mean and median of the variables
- To understand the difference between the variables, you can subtract the values between the variables
- The interval scale is the preferred scale in Statistics as it helps to assign any numerical values to arbitrary assessment such as feelings, calendar types, etc.

Examples:
- Net Promoter Score (NPS)
- Bipolar Matrix Table

Ratio

The ratio scale is the 4th level of measurement scale, which is quantitative. It is a type of variable measurement scale. It allows researchers to compare the differences or intervals. The ratio scale has a unique feature. It possesses the character of the origin or zero points.
Characteristics of Ratio Scale:
- Ratio scale has a feature of absolute zero
- It doesn't have negative numbers, because of its zero-point feature
- It affords unique opportunities for statistical analysis. The variables can be orderly added, subtracted, multiplied, divided. Mean, median, and mode can be calculated using the ratio scale.
- Ratio scale has unique and useful properties. One such feature is that it allows unit conversions like kilogram – calories, gram – calories, etc.

Examples:
What is your weight in Kgs?
- Less than 55 kgs
- 55 – 75 kgs
- 76 – 85 kgs
- 86 – 95 kgs
- More than 95 kgs



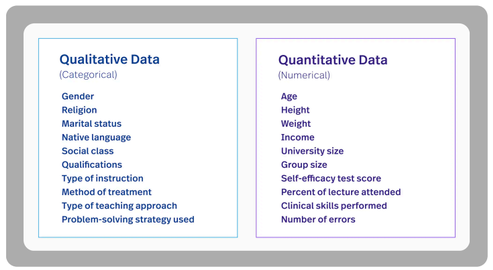
Data

Qualitative

Qualitative data:
- Is interpretation-based, descriptive, and relating to language.
- Qualitative data can help us to understand why, how, or what happened behind certain behaviors.
- Subjective and unique.
- Qualitative research methods are interviewing and observing.
- Analyzed by grouping the data into categories and themes.

Quantitative

Quantitative data:
- Is numbers-based, countable, or measurable.
- Tells us how many, how much, or how often in calculations.
- Fixed and universal.
- Quantitative research methods are measuring and counting.
- Analyzed using statistical analysis.

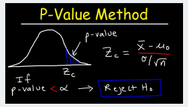


Hypothesis Testing

- **Null hypothesis (H0):** The null hypothesis is the starting assumption in statistics. It says there is no relationship between groups. For Example A company claims its average production is 50 units per day then here:
  - Null Hypothesis: H_0: The mean number of daily visits (μ) = 50.
- **Alternative hypothesis (H1):** The alternative hypothesis is the opposite of the null hypothesis it suggests there is a difference between groups. like The company's production is not equal to 50 units per day then the alternative hypothesis would be:
  - H_1: The mean number of daily visits (μ) ≠ 50.

P-Value

P-value: When analyzing data the p-value tells you the likelihood of seeing your result if the null hypothesis is true. If your P-value is less than the chosen significance level then you reject the null hypothesis otherwise accept it.



Confidence Interval

Confidence Interval: A range of values that estimates where the true population value is likely to fall. Instead of just saying The average height of students is 165 cm a confidence interval allow us to say We are 95% confident that the true average height is between 160 cm and 170 cm.
Confidence level tells us how sure we are that the true value is within a calculated range. If we have to repeat the sampling process many times we expect that a certain percentage of those intervals will include the true value.
- 90% Confidence: 90% of intervals would include the true population value.
- 95% Confidence: 95% of intervals would include the true value which is commonly used in data science.
- 99% Confidence: 99% of intervals would include the true value but the intervals would be wider.

Regression Analysis

Regression: A statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between a dependent variable and one or more independent variables.
Linear regression is the most common form of this technique. Also called simple regression or ordinary least squares (OLS), linear regression establishes the linear relationship between two variables.
Linear regression is graphically depicted using a straight line of best fit with the slope defining how the change in one variable impacts a change in the other. The y-intercept of a linear regression relationship represents the value of the dependent variable when the value of the independent variable is zero.