Descriptive Statistics UNIT 2

Dr. Praveen Barapatre

Importance of Data Inspection

- Examining individual data values is crucial for understanding the data.
- This initial inspection helps identify:
  - The mix of data types (numerical, categorical, etc.)
  - Duplicate values
  - Missing or incorrectly coded data
  - Zero values and unexpected values

Role of Descriptive Statistics

- Summarize the data in a manageable way.
- Provide a concise overview of the data's key characteristics.

Complementary Visualizations:

- Simple graphs like line charts, histograms, and scatter plots enhance understanding of the data.
- Descriptive statistics and visualizations work together to effectively communicate the data's story.

Descriptive Statistics

- Count: The total number of data points (n).
- Top m, Bottom m: Identifies the highest (Top) or lowest (Bottom) m values in the ordered data set.
  - SQL can provide these using TOP or specific functions like FIRST() and LAST().
- Variety (Diversity): The number of unique values in the data set. Not to be confused with information diversity measures.
  - Achieved using DISTINCT in SQL.

- Majority: The most frequent data value(s). Often used for specific subsets of data, not the entire dataset unless a value makes up over 50%.
- Minority: The least frequent data value(s). Similar usage to majority, often for subsets of data.
- Maximum (Max): The highest value in the data set (can have duplicates).
  - Found using MAX() in SQL.
- Minimum (Min): The lowest value in the data set (can have duplicates).
  - Found using MIN() in SQL.

- Sum: The total obtained by adding all data values together.
  - Calculated using SUM() in SQL.
- Average: The arithmetic mean, which is the sum of all values divided by the number of values (n).
  - Achieved using AVG() in SQL.

Key Points

- This section emphasizes the importance of initial data analysis and basic descriptive statistics.
- Many statistical software packages and SQL databases offer functions to compute these statistics.

Central Tendency

- In statistics, central tendency refers to a collection of measures that summarize the "center" or typical value of a data set. These measures aim to represent a large set of data with a single number. There are three main types of measures of central tendency:
- **Mean (Arithmetic Mean):** This is the most common and well-known measure. It's calculated by adding all the values in the data set and then dividing by the number of values.

Central Tendency

- **Median:** The median is the "middle" value when the data is arranged in ascending or descending order. If you have an even number of data points, the median is the average of the two middle values.
- **Mode:** The mode is the most frequent value in the data set. It can be useful for identifying the most common category in categorical data.

The choice of which measure of central tendency to use depends on the characteristics of your data set. For example, the mean can be sensitive to outliers (extreme values), while the median is not. The mode is most appropriate for categorical data.

Central Tendency

Data with Weights or Frequencies:

- The text introduces $\{f_i\}$which represents weights, frequencies, or probabilities associated with each data point $\{x_i\}$.
- The total weight/frequency (N) is the sum of all individual weights/frequencies.
- Relative frequencies $\{p_i\}$are obtained by dividing individual frequencies by the total (N). They sum to 1 and can be interpreted as probabilities.

Types of Averages for Different Data:

- **Arithmetic Mean (most common):** Best suited for untransformed, continuous data (e.g., heights, weights).
- **Mode:** Useful for nominal data (categorical data with no intrinsic order, e.g., hair color). It identifies the most frequent value(s).
- **Harmonic Mean:** More appropriate for ratio data (data with a meaningful zero point, e.g., speeds, rates).
- **Geometric Mean:** Used for percentage growth and rate data.