Statistics

Descriptive statistics:

Statistics involving describing data. Involves summarizing and organizing data so they can be easily understood.

Inference statistics:

Complex set of procedures to draw conclusions over large populations with sample data.

Data

Numeric: wind speed, time duration, discrete etc.

Categorical: Car types, Binary, ordinal (ordered).

*------------------- 5/7/2024 Up to Page 5*

Exploratory Data Analysis

Data Structure:

* Rectangular (rows are records & columns are variable or features) and
* Non-Rectangular (spatial or graph)

In statistics we use mostly rectangular data:

**Estimates of Location**

An estimate of where most of the data is located (i.e., its central tendency)

|  |  |  |  |
| --- | --- | --- | --- |
| Key Term | Definition | Formula | Usage |
| Mean | Sum of all values/ number of values |  | average |
| Weighted Mean | Sum of all values times a weight / sum of weights |  | Some variables are intrinsically more variable than other and high variable observations are given lower weight. Ex: when taking average from multiple sensors giving lower weight for sensors that giving less accurate readings. |
| Median | The value such that one-half of data lies above and below |  | While calculating average household income in a city where bill gate lives the mean gives diff value where median gives right value no matter who is rich or not. |
| Percentile | The value such that P percentage of data lies below |  |  |
| Weighted Median | The value such that one half of the weighted sum lies above and below the sorted data. |  |  |
| Trimmed Mean | The average of all values after removing fixed number of extreme values | P smallest and largest values omitted | A trimmed means eliminate the influence of extreme values, EX: International diving the top score & bottom score from five judges are dropped and the final score is the average of scores from 3 remaining judges. This makes it difficult for a single judge to manipulate the scores. |

Note: Trimmed mean, Median and weighted median are robust to outliers.

Outlier: is a any value which is very distant from other values in data set and cause skewness.

*------------------- 5/8/2024 Up to Page 11*