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*

[Refer Estimates of Location in Python Notebooks]

*------------------- 5/9/2024 Up to Page 12*

**Estimates of Variability**

Measures whether the data values are tightly clustered or spread out.

At the heart of the statistics lies variability:

* Measuring it
* Reducing it
* Distinguishing random from real variability
* Identifying the various sources of real variability
* Make decision out of it in presence.

|  |  |
| --- | --- |
| Key Terms | Definition |
| Deviation | The difference between observed variable from the mean or estimated location |
| Variance | The sum of squared deviation from mean divided by n -1 (n is the number of data values) |
| Standard Deviation | Square root of variance |
| Mean Absolute Deviation (Manhattan Norm, l1 – norm) | Mean of absolute values of the deviation from the mean |
| Median Absolute Deviation from Median | Median of absolute values of the deviation from the median |
| Percentile | The value such that p percent of values take this value or less and (100-p) values take this value or more |
| Interquartile range | The difference between percentile 75th percentile and 25th percentile |

*------------------- 5/10/2024 Up to Page 14 (Additional notes pending)*