Exploratory Data Analysis

- Rectangular data is the basic data structure for statistical and machine learning models

- A column within a table is commonly referred to as feature.

- A row within a table is commonly referred to as a record.

Mean

Formula for the mean is pretty simple:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

but if the values are sorted, and we want to trim the smallest and the largest values, we can use trimmed mean:

$$\bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n-2p}$$

Therefore eliminating the influence of extreme values.

There is also weighted mean which is used when there are values that are much more variable only coming from a particular resource.

For example, if we are taking the average from multiple sensors, and if one sensor is particularly variable, we may give the values collected from that sensor less weight.

Also, sometimes data does not represent the different groups we're interested in equally. To correct that, we give a higher weight to the values from the group that were under represented.

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

Outliers

Median is referred to as robust estimate of location since it is not influenced by outliers (extreme cases).

Outliers are not inheently invalid or erroneous. In contrast, outliers are sometimes informative in anomaly detection.

Still, outliers are often result of data errors such as mixing units (km vs m) or bad readings from the source.

When outliers are result of a bad data, mean will result in a poor estimate while the median will still be valid.

The median is not the only robust estimate of location. A trimmed mean can also be used to avoid the influence of outliers. It can be thought of as a compromise between the mean and the median.

Location in statistics is used to describe the central tendency of a dataset.