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.

Estimates of Location

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 underrepresented.

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

Outliers

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

Outliers are not inherently 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.

Estimates of Variability

- Interquartile range is the difference between the 75th percentile and the 25th percentile. Also known IQR.

Mean absolute deviation

Let's say we have a dataset $\{1, 4, 4\}$. The deviations from the mean are the differences: $1 - 3 = -2$ $4 - 3 = 1$ $4 - 3 = 1$

But we can't estimate the variability with only these values. In fact, the sum of the deviations from the mean is precisely zero.

To solve this, we take the absolute value of the deviations when we are using mean absolute deviation.

- $MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$ where $\bar{x}$ is the sample mean.

Variance and Standart Deviation

The best known estimates of variability are the variance and the standard deviation

- Variance $= s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

- Standard deviation $= s = \sqrt{\text{Variance}}$

You might wonder, "Why are we dividing by n-1". That's because, if you have a dataset consist of 10 elements, sample of that dataset is naturally smaller. And everything becomes much more closer to mean because of the small size.

It causes underestimation of the variance, so we use (n-1) instead of n to make the variance bigger.

Why (-1). because we only lose one degree of freedom when we estimate the variance of a sample, that is the mean of the sample.

- Median absolute deviation = median $(|x_1 - m|, |x_2 - m|, \ldots, |x_N - m|)$

- $A = \{2, 4, 5, 6, 10\} \rightarrow$ median $= 5$

$A' = \{|2-5|, |4-5|, |5-5|, |6-5|, |10-5|\} = \{0, 1, 1, 3, 5\}$

Median absolute deviation of A is the median of $A' \rightarrow 1$

Estimates Based on Percentile

Statistics based on sorted data are referred to as order statistics.

The most basic measure is the range: the difference between the largest and smallest numbers.

In a dataset, P th percentile means that at least (100-P) percent of the values take on this value or more.

Also, quantile is essentially the same as percentile, with quantiles indexed by fractions (so the .8 quantile is the same as the 80th percentile)

A common measurment of variability is the difference between 25th percentile and the 75th percentile, called the interquartile range or IQR

- $\{3,1,5,3,6,7,2,9\}$ → we sort these → $\{1,2,3,3,5,6,7,9\}$ → 25th percentile is at 2,5 and the 75th percentile 6,5 so the interquartile range is $6.5 - 2.5 = 4$.

Exploring the Data Distribution

- Location and variability are referred to as the first and second moments of a distribution.

- The third and fourth moments are skewness and kurtosis.

- Skewness refers to whether the data is skewed to larger or smaller values. (i.e if the skewness > 0, the data is skewed to the right)

- Kurtosis indicates the propensity of the data to have extrem values. (i.e high kurtosis means more data in the tails, low kurtosis means fewer extreme values)

- Skewness and kurtosis are generally discovered through visualizations such as plots.

Exploring Binary and Categorical Data

- When we have numeric values, there are often too many unique values to easily visualize, so we group numbers into bins.

- When we create bins, we actually create an ordered factor, which is naturally ordered. And we can create a histogram with this categorized data.

- Like with bins, we can use something similar with categorical data which is bar chart.

Mode, is the value, or values in case of a tie, that appears most often in the data.

Expected Value

- A special type of categorical data is data in which the categories represent or can be mapped to discrete values on the same scale.

- For example, let's think of a service that has three different subscription tiers.

For the sake of our scenario, this service offers free webinars. And after these webinars; %5 of the attendees will sign up for the $300 service, 15% will sign up for the $50 subscription and %80 will not sign up for anything.

We can find the EV by basically estimating the weighted mean of this dataset.

$$EV = (0,05).(300) + (0,15)(50) + (0,80).(0) = 22,5$$

- EV is a fundamental concept in business valuation and capital budgeting.

Correlation

- $r = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x s_y}$ (Pearson's correlation coefficient)

- The correlation coefficient always lies between +1 and -1, which respectively means perfect positive correlation and perfect negative correlation.

... More on the related notebook

Exploring Two or More Variables

- Estimators like mean and variance can only look at one variable at a time, this is called univariate analysis.

- Additionally, correlation analysis is an important method that compares two variables, this is called bivariate analysis.

- In this section, we will look at additional estimates and plots, and at more than two variables, that is called multivariate analysis.

... More on the related notebook