Desc. Stats

Wednesday, May 19, 2021 5:40 PM

Univariate Analysis

- Central Tendency :

  In statistics, a **central tendency** (or measure of **central tendency**) is a **central** or typical value for a probability distribution. It may also be called a center or location of the distribution.

  Measures Of Central Tendency :

- Mean : Use when outliers are not present. (**sum(n)/n**)
- Median : Outliers doesn't affect . For Odd No of Values : **(nth + 1) /2**
  For Even No of Values : (**nth/2 + (nth+1)/2) / 2**
- Mode : Use on Categorical Data, Outliers doesn't affect. Bimodal If there are two modes.

Note :

Mean is used because it has less Time Complexity (**n**), While Median has high Time Complexity { **o(n.log(n))** }.

Measures of Spread :

Measures of spread describe how similar or varied the set of observed values are for a particular variable (data item). Measures of spread include the range, quartiles and the interquartile range, variance and standard deviation.

- Range : Difference Between Highest and Lowest Values. (Affected by Outliers)
- IQR : Difference Between Q3(75 percentile) and Q1(25 percentile). (Not Affected by Outliers)
- Variance : On average, how far away a data point from mean. (σ^2) tells us how data points are spread out. It is the average of the distances from each data point to the mean, squared.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(xi - \mu)^2}{N}$$

Where μ is mean
It affected by outliers.

- Standard Deviation :
  - $$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N}}$$

It affect by Outliers

- Median Absolute Deviation :
  For a univariate data set X_1, X_2, ..., X_n, the MAD is defined as the median of the absolute deviations from the data's median

$$\mathrm{MAD} = \mathrm{median}(|X_i - \tilde{X}|)$$

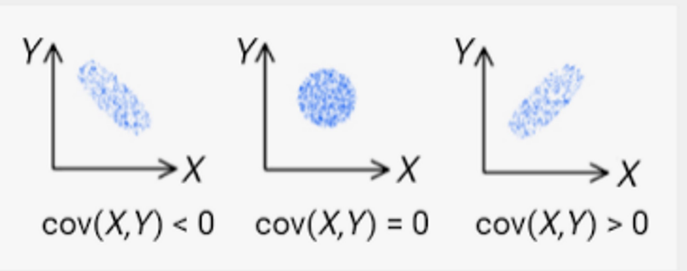
$$\mathrm{MAD} = \mathrm{median}(|X_i - \bar{X}|)$$

that is, starting with the residuals (deviations) from the data's median, the MAD is the median of their absolute values. And X is mean
Outliers are relatively affect less as compared to Std.

Bivariate Analysis :

❖ Measures of Relationship :

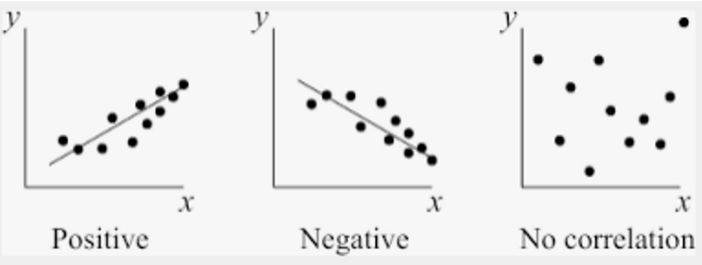
❖ Covariance :
covariance is a measure of the joint variability of two random variables. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values (that is, the variables tend to show similar behavior), the covariance is positive.[2] In the opposite case, when the greater values of one variable mainly correspond to the lesser values of the other, (that is, the variables tend to show opposite behavior), the covariance is negative. The sign of the covariance therefore shows the tendency in the linear relationship between the variables.



$$\mathrm{Cov}(x,y) = \frac{\Sigma\,(x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

❖ Corelation : **Correlation** is a **statistical** measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate).

$1 <= \rho <= 1$



$$\mathbf{COR(X,\ Y)} = \frac{\mathbf{COV(X,\ Y)}}{\sqrt{\mathbf{VAR(X)VAR(Y)}}}$$