Statistics

- Statistics is a branch of mathematics that deals with collection, analyzing and interpreting large amount of data.

Why is Statistics important?

- Statistics allows us to derive knowledge from large dataset and this knowledge can then be used to make predictions, decisions and classifications etc.

Where is Statistics used?

Statistics are used in various fields, some are:

- Medical Research
- Stock Market
- Sales Projection
- Weather forecasting

Sampling -

Sampling is the process of collecting data to perform analysis on.

Sample vs population:-

Population is the entire dataset such as the whole population of a country. Sample is subset of that population which is analyzed to make inference.

Sample frame -

A Sampling frame is a list from which sample is selected, such as Citizen Register for a country or Employee List for a company etc.

- Sample error is an error that leads to our sample not accurately representing our population.

Non Sampling Error occurs due to poor sample design, inaccurate measurements, bias in data collection etc.

Random Sampling is the process of selecting a subset/sample for a population in such a way that every data point is equally likely to be included in the sample.

Stratified Sampling :- is the process of dividing your sample into layers of groups and then perform random sample for each group

Systematic Sampling is the process of selecting your sample by picking every k^{th} element in your population. You don't need a list for this.

Central Tendencies is used to indicate where does the middle or center of the distribution of our data lies.

- Mean
- Median
- Mode.

Mode is used to indicate the most frequent data point, in other words the one which occurs most

Median is the middle of the data. If the data is arranged in ascending order then the data element which occurs right at the center is the median.

Mean is the average of the data.

Trimmed Mean is used to deal with the outliers by trimming or removing some data from both end so as to get rid of outliers.

Weighted Mean is used when certain values are supposed to count more in some context.
E.g. Calculating average grade of a student based on their grade distribution

Variation in statistics is used to show how data is dispersed or spread out. Several measure of variation are used in statistics.

Range is the difference b/w the highest and the lowest values in our dataset.

Percentile are score that are used to describe a value below which some observation fall.
E.g. If X is at 70th Percentile it means 70% of other data points from our sample are below X.

Quartiles are used to break data into 4 parts so as to better find the spread of data in a way that is less influenced by outliers.

Interquartile Range (IQR) is the difference between the lower and upper quartile. This gives us a better idea of a range of data.

Correlation is a term that is a measure of the strength of a linear relationship between two quantitative variables

- Correlation allows us to check whether one variable has any influence on other variable

Positive Correlation is a term that used to describe a positive linear relationship between two quantitative variable.

No Correlation is a term used to describe no linear relationship between two quantitative variable.

Negative correlation is a term that is used to describe the strength of a negative linear relationship between two quantitative variables

- (Increase in one quantity means decrease in another)

Standard Variance measure how far a set of numbers are spread out from their average value.

Standard Deviation is used to express the magnitude by which the members of a group diff. from the mean value for the group.

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

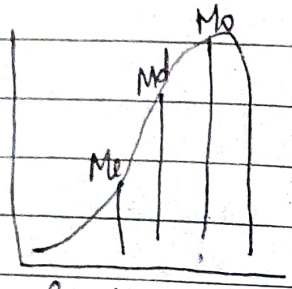
Standard Variance

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

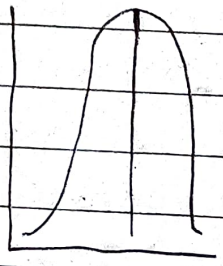
Standard Deviation

$\bar{x}$ = mean

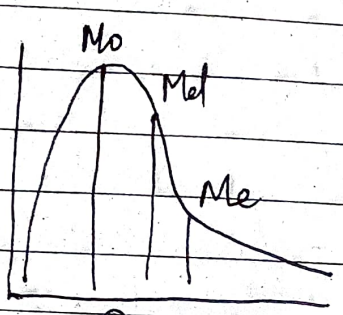
Normal Distribution is a term that is used to describe a distribution which when plotted gives us a shape of bell curve. It has mean of zero and standard deviation of 1.



~~positive~~ Negatively Skewed

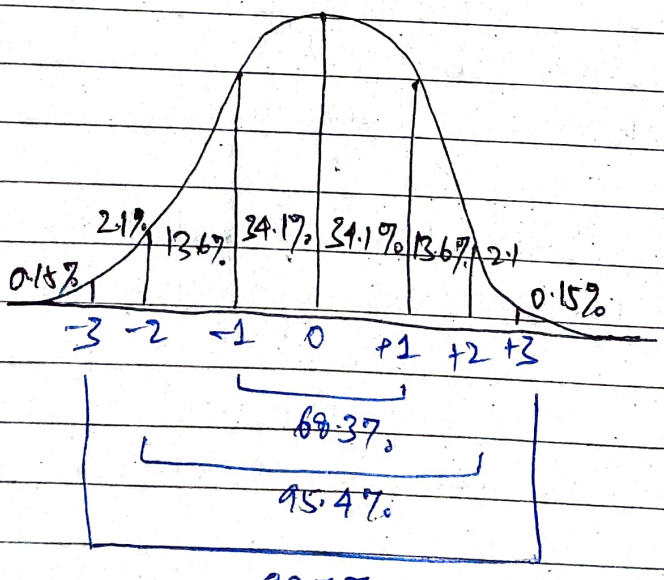


(No Skew)



(Positively Skewed)

Empirical Rule is used to remember the percentage of values that lies within a band around the mean in a normal distribution with a width of two, four and six standard deviation.



Z score is a measure of how many standard deviations below or above the population mean a raw score. Z score can be placed on a normal distribution curve.

$$Z = \frac{x - \mu}{\sigma}$$

(score) (Mean) (S.D.)

- Linear Regression is basic and commonly used type of predictive analysis.

It is used to create a model that can predict a dependent variable using an independent variable.

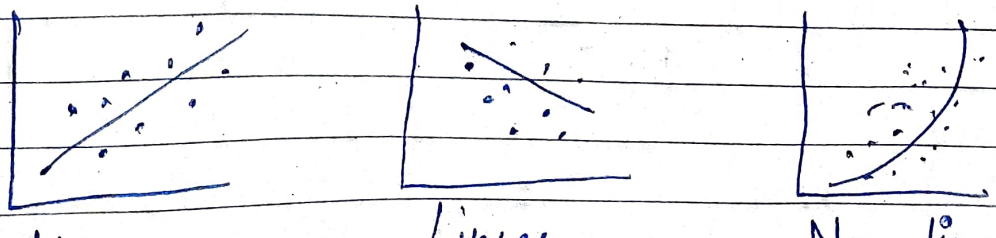
$$Y = \beta_0 + B_1 X_i + \varepsilon_i$$

Y: dependent Variable
$\beta_0 + B_1 X_i$: Linear Component
ε_i: Random Error component
X_i: Independent Variable

β_0 = Population Y intercept
β_1 - Population Slope Coefficient

Simple linear Regression is useful for finding relationship between two continuous variables.

One is predictor or independent variable and the other is the response or dependent variable.



Linear Linear Non-linear.

In general the data doesn't fall exactly on the line, So the regression equation should include an implicit error term

The fitted values (predicted values) are typically denoted by $\hat{Y}$ (Y-hat)

$$Y_i = b_0 + b_1 X_i + e_i$$

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_i$$

If $b1 > 0$, then x (predictor) and y (target) have positive relationship ie., an increase in x will increase Y.

$b1 > 0$ — Positive Relationship

If $b1 < 0$, then X (predictor) and Y (Target) have negative relationship ie an increase in X will decrease Y

$b1 < 0$ - Negative Relationship.

What are Residuals?

The difference between the observed value ~~and~~ of the dependent variable (Y) and the predicted ($\hat{y}$) is called Residuals. Each data point has 1 residual.

$$RSS = \sum_{k=1}^{n} (\text{Actual} - \text{Predicted})^2$$

The line with the lowest value of the residual sum of square would be the best fit line.