Statistics

- Statistics is the branch of mathematics where we collect, organize, analyse and represent the data for batter decision making. We apply statistics to different problems.

Population Sample

Deals with

Planning Designing

Data – Facts or pieces of information that can be measured.

- **Statistical Methods**

Descriptive Statistics → Mean (μ), Standard Deviation (σ)

Inferential Statistics → Conclusion

Descriptive Statistics :

Descriptive Statistics is a summary that describes or summarizes the collection of information/data.

It summarizes the sample data rather than learning about the population that sample data is representing.

Inferential Statistics :

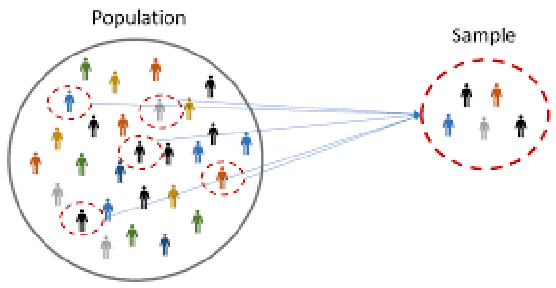
Inferential Statistics is the process of data analysis where we make the conclusions about population data using sample data.

Population (N) :

Population Data is the entire group that you want to draw conclusions about.

Sample (n) :

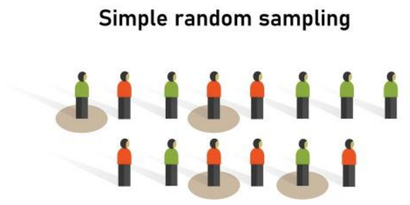
Sample is the group (part of population) from which you'll collect data.



Types of sampling :

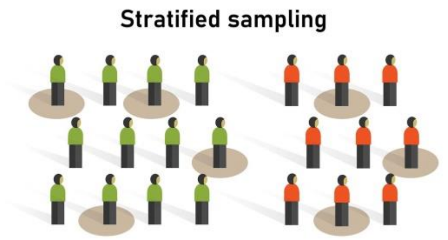
1. Simple Random Sampling

Simple Random Sampling is the process of sampling where every member of the population has equal chance of being selected.



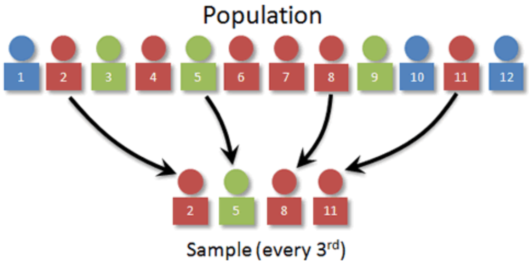
2. Stratified Sampling

Stratified sampling is a method of sampling where population (N) is split into non-overlapping group.



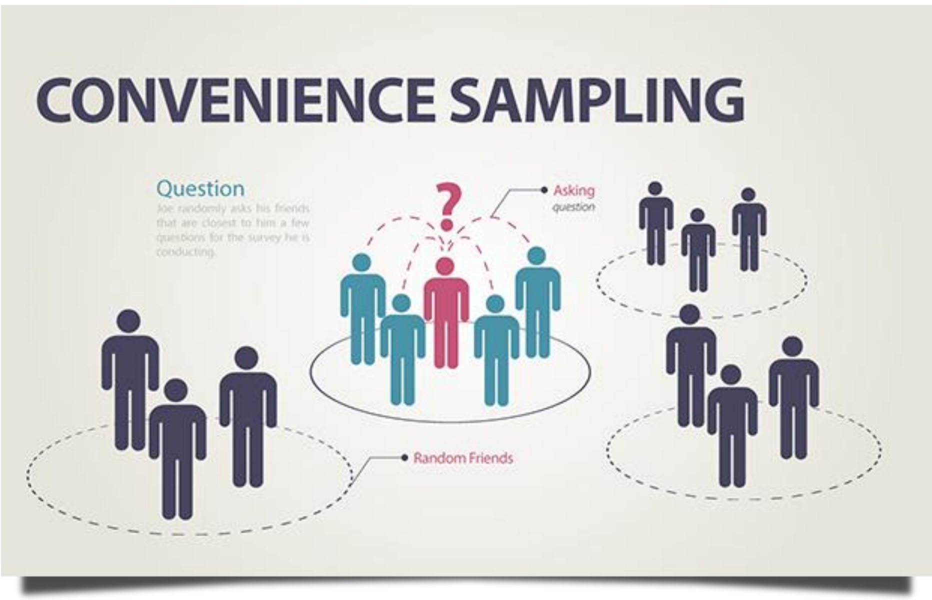
3. Systematic Sampling

Systematic sampling is **a probability sampling method where researchers select members from population at nth interval.**



4. Convenience Sampling

Convenience sampling is the process of taking sample data from those who has knowledge/expertise on the research area.



Variables

A variable is a property that can take any value.

Quantitative Variable Qualitative/Categorical Variable

{Measured Numerically} {Based on some characteristics we
{Add, Subtract, Multiplication, Divide} can derive categorical value}

Quantitative Variables	Qualitative Variables
Take on numeric values	*Take on names or labels*
Number of students in a class	Eye color
Number of square feet in a house	Gender
Population size of a city	Breed of dog
Age of an individual	Level of Education
Height of an individual	Marital status

Quantitative Variable

Discrete Variable Continuous Variable

Whole Numbers A numeric **variable** that have an

Eg : infinite number of values between

Total number of children in the family any two values

2, 4, 5 Eg :

Height – 172.5, 163.9, 162.8

Variable Measurement Scale

There are 4 types of measured variables

1. **Nominal** : Categorical data

2. **Ordinal** : Order of the data matters but value doesn't.

3. **Interval** : Order matters, value also matters but natural zero is not present.

4. **Ratio** : Something measured on a ratio scale, has same properties as interval scale but with absolute zero point.

Eg :

Students (Marks)	Rank
100	1
96	2
57	3
85	4
44	5

Ordinal Data

Eg :

Temperatures

Farehheits

Interval

70 – 80   80 – 90   90 – 100   0

Zero doesn't make any useful meaning

Ratio Variable :

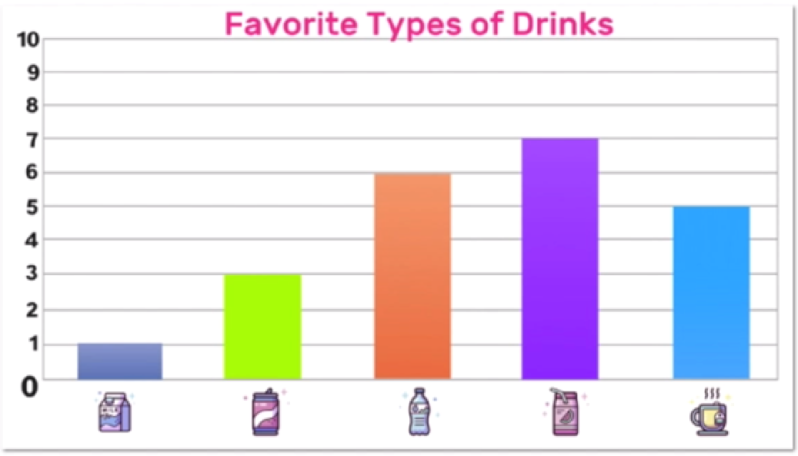
Provides more detailed informations. It includes true zero value.

Frequency Distribution :

Sample Dataset : Rose, Lily, Sunflower, Rose, Lily, Sunflower, Rose, Lily, Lily

Flower	Frequency	Cumulative Frequency
Rose	3	3
Lily	4	7
Sunflower	2	9

Bar-graph : (Categorical)

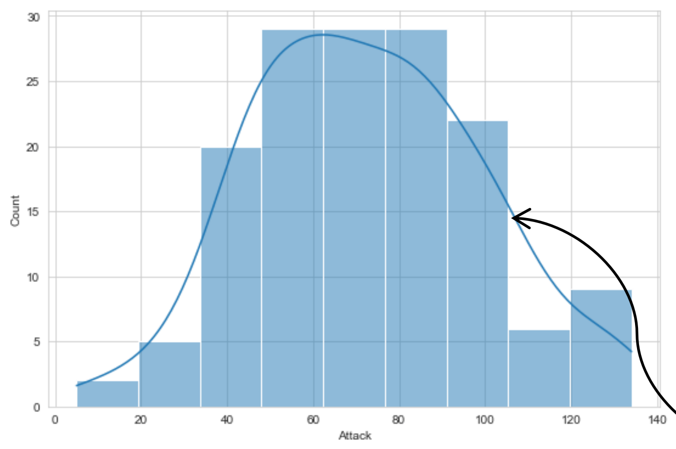


Histogram : (Continuous)

It is a graphical tool to summarize discrete or continuous data.

Eg:

Ages : {10, 12, 14, 18, 29, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}



Histogram

To construct histogram :

Bins/Bucket

If data set starts from zero.

Otherwise = (max – min) / bins

pdf : Smoothening of the histogram.

{Probability density function}

{Kernel Density Estimator}

<u>Measure of central Tendency :</u>

1. Mean 2. Median 3. Mode

Quantitative Data

Refers to the measure used to determine the centre of the distribution of the dataset.

{1,1,2,2,3,3,4,5,5,6,100}

<u>Mean</u> = (32+100)/11 = 12

<u>Median</u>

Sort the numbers

Odd = n Middle element

even = (n1 + n2)/2 Middle two element

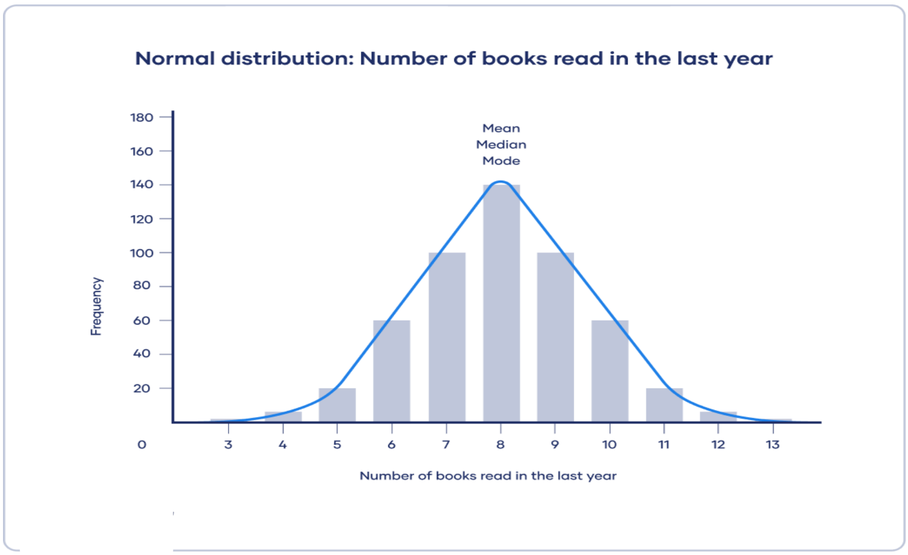
Median works well with outliers.

<u>Mode</u>

Most frequent element.

Eg :

{1,2,2,3,4,5,6,6,6,7,8,100,200,100} Mode = 6

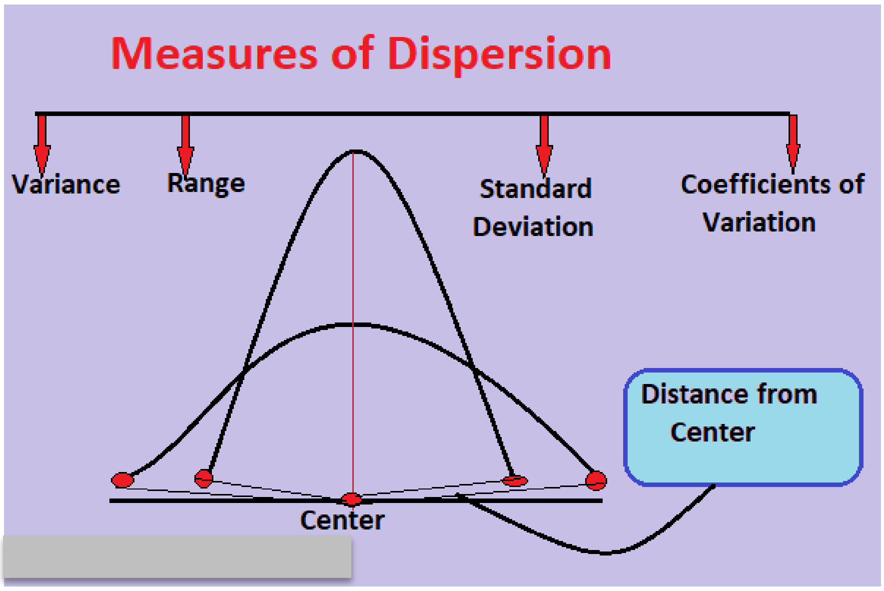


Measure of Dispersion :

Measure of dispersion describes the spread of the data or its variation around the centre value. {Dispersion}

Spread

1. Variance
2. Standard Deviation



1. **Variance :**

Population Variance :

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

Sample Variance :

$$s^2 = \sum_{i=1}^{n} \frac{(x_{i-}\overline{x})^2}{n-1}$$

Spread is low means the elements present in the central region is more.

More variance : Data is more spread.

Variance = Spread = Dispersion = Is the extent to which distribution is stretched or squeezed.

2. **Standard Deviation** :

$\sigma = \sqrt{variance}$

Standard deviation shows how far the elements are from mean.

Percentiles and Quartiles : {find outliers}

Percentile is a value below which a certain percentage of observation lie.

Eg :

Dataset : {2,2,3,4,5,5,5,6,7,8,8,8,8,8,9,9,10,11,11,12}

What is the percentile range of 10 ?

n = 20

Percentile rank of x = $\frac{(\text{\# of values below x})}{x}$ * 100

$$= \frac{16}{20} * 100 = 80\%$$

What value exists at percentile ranking of 25% ?

Value = $\frac{\text{Percentile}}{100}(n+1)$

= $\frac{25}{100}(21)$ = 5.25 index

≈ $\frac{5+5}{2}$ = 5

5 Number Summary

1. Minimum
2. First Quartile
3. Median
4. Third Quartile
5. Maximum

Removing Outliers

{1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9,27}

Q1 = $\frac{25}{100}(20)$ = 5th index = 3

Q3 = $\frac{75}{100}(20)$ = 15 index = 8

Lower fence = Q1 – 1.5(IQR) = -4.5

Higher fence = Q3 + 1.5(IQR) = 15.5

Inter Quartile Range (IQR) = Q3 – Q1 = 8 – 3 = 5

Remaining data : {1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9}

Minimum : 1

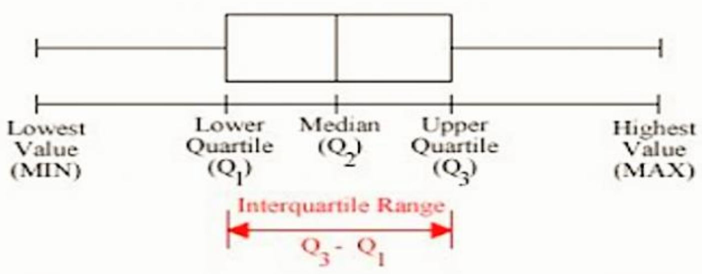
First Quartile : 3

Median : 5

Third Quartile : 8

Maximum : 9

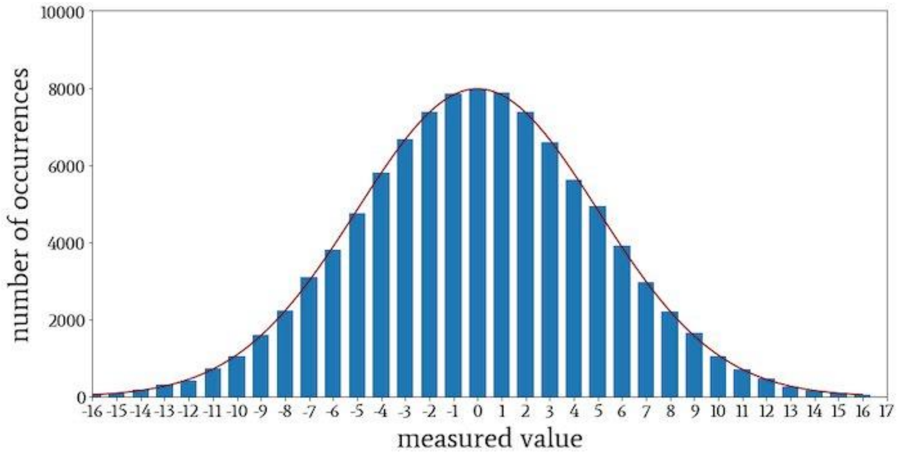
➡ **Boxplot**



➡ **Normal Distribution** :

A distribution is called normal distribution if we plot histogram with the data it'll be symmetric to the mean and dataset are more frequent towards the mean.

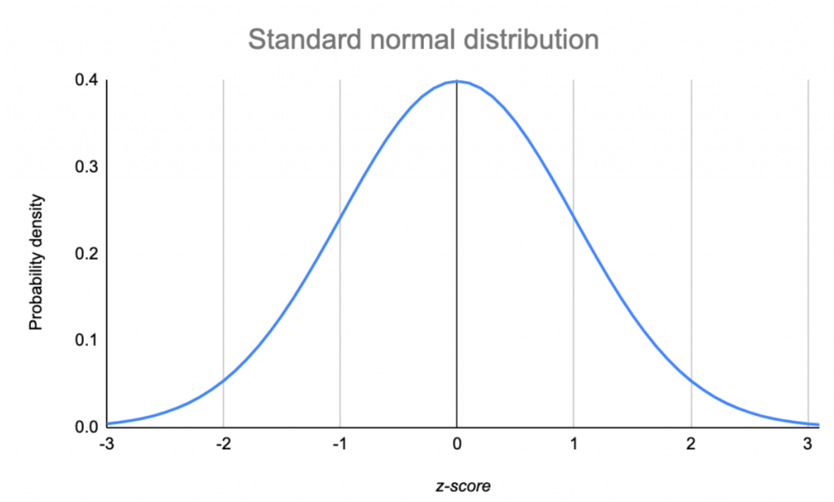
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



➡ Standard Normal Distribution :

A distribution is called standard normal distribution where mean will be zero and standard deviation will be one. And most of the data will lie in between --3σ to $+3\sigma$.

If the dataset will be in different unit ML algorithm will take more time. So we will apply SND on these dataset to bring them to the same scale.



➡ Standardization :

In standardization we scale down the value, where the mean will be zero and standard deviation will be one.

➡ Normalization :

In normalization we try to convert a dataset in between a range.

1. **MinMax Scaler** :

With the help of MinMax scaler we convert the data which will range in between 0 to 1.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Covariance :

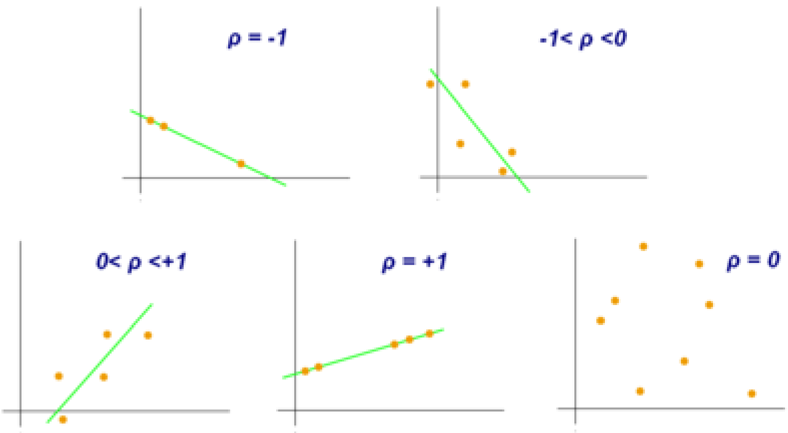
Covariance helps us to find-out the direction of relationship.

$$\text{Cov}\ (X, Y) = \frac{\sum (X_i - \overline{X})(Y_j - \overline{Y})}{n}$$

Pearson Correlation Coefficient :

Pearson Correlation coefficient measures the strength and relationship between two variables. The value ranges in between -1 to +1. The values more towards +1 they are more positively correlated and the values more towards -1 the more negatively correlated they are.

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

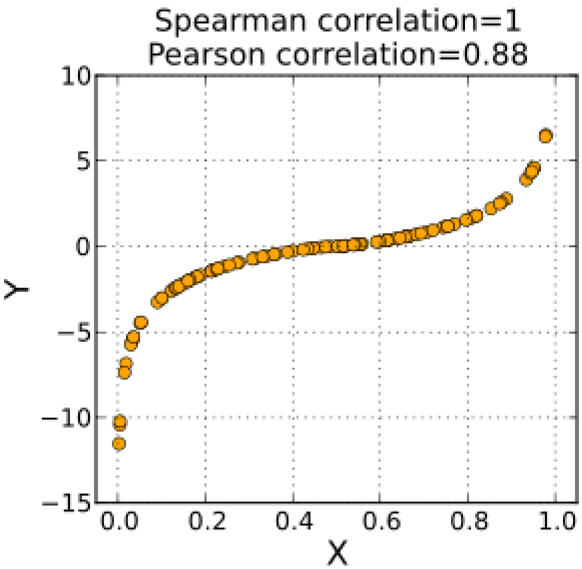


Spearman's rank correlation coefficient :

Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

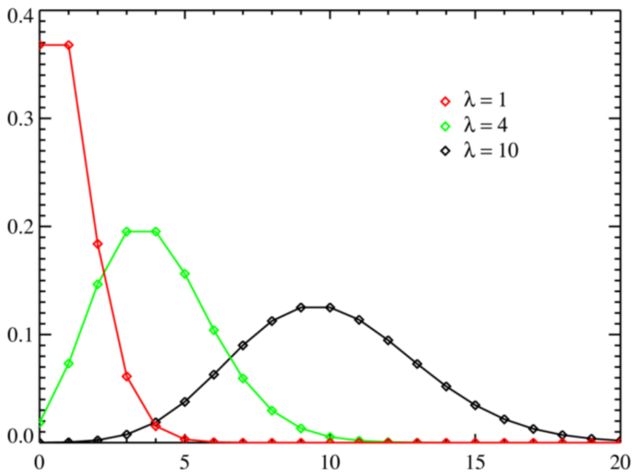
$$r_s = \rho(r_x, r_y) = \frac{covariance(r_x, r_y)}{\sigma_{ry} * \sigma_{ry}}$$

$$= \rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$



Poisson Distribution :

Poisson distribution is a probability distribution which gives the probability of an event can occur in a given interval of time.



3 conditions of Poisson distribution :

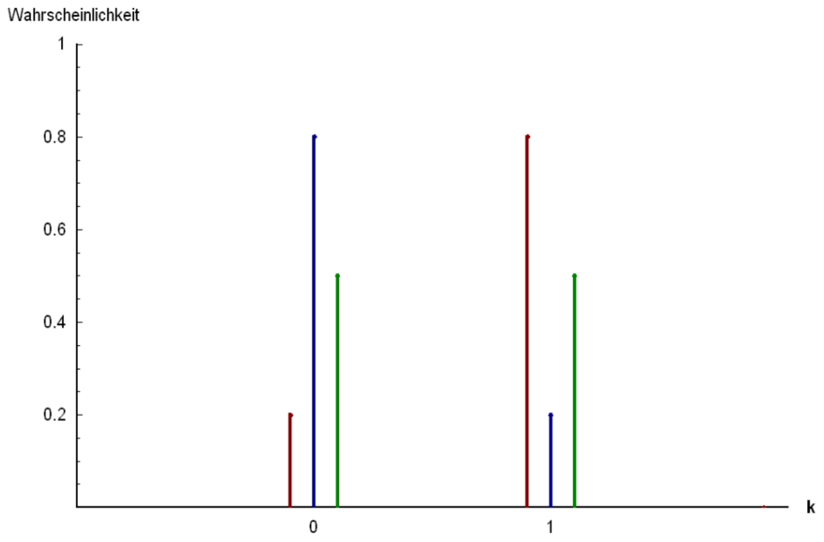
- **Events are independent of each other.** The occurrence of one event does not affect the probability of another event will occur.
- The average rate (events per time period) is constant.
- Two events cannot occur at the same time.

$$f(x) = \frac{\lambda^x}{x!}e^{-\lambda}$$

Bernoulli Distribution :

Bernoulli distribution is a discrete probability distribution which takes the value 1 with probability p and the value 0 with probability q = 1-p. It is a set of possible outcomes of any single experiment.

$$f(x) = \begin{cases} p^x * (1-p)^{1-x} & \text{if } x = 0,1 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$$

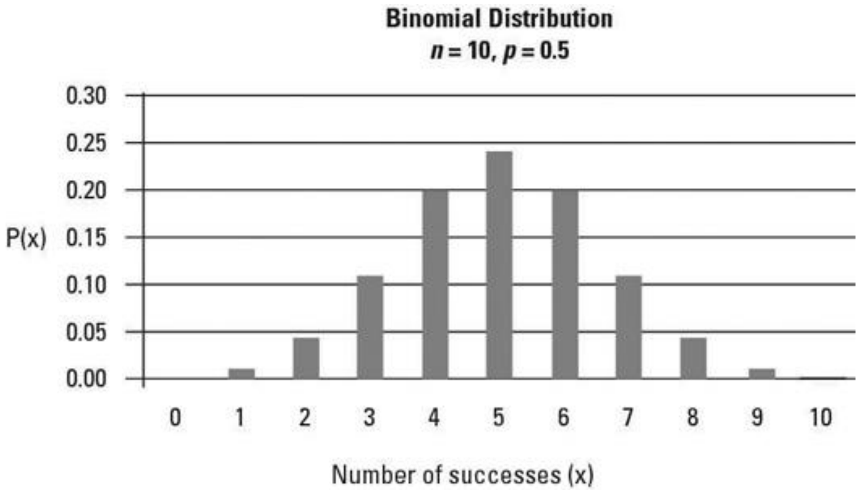


Binomial Distribution :

Binomial Distribution is a discrete probability distribution which represents number of success in n independent experiments.

Binomial Distribution Formula

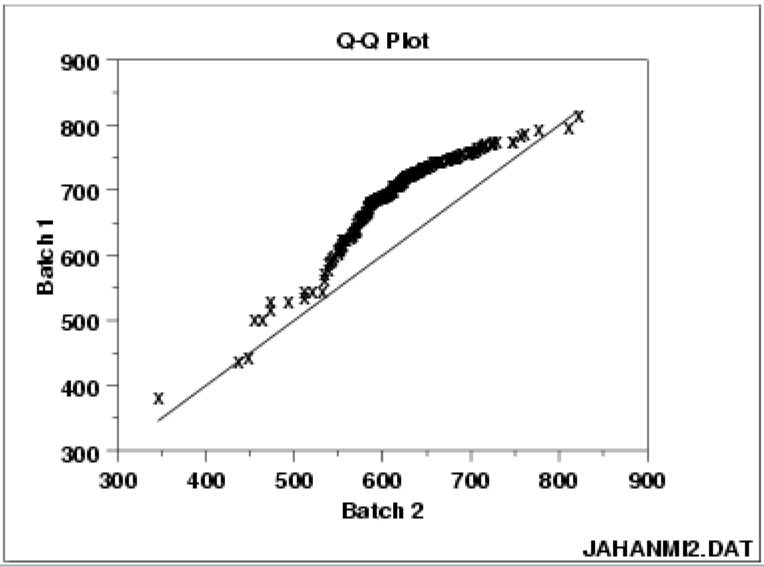
$$P(X) = {}_nC_x\, p^x (1 - p)^{n-x}$$



Q-Q plot :

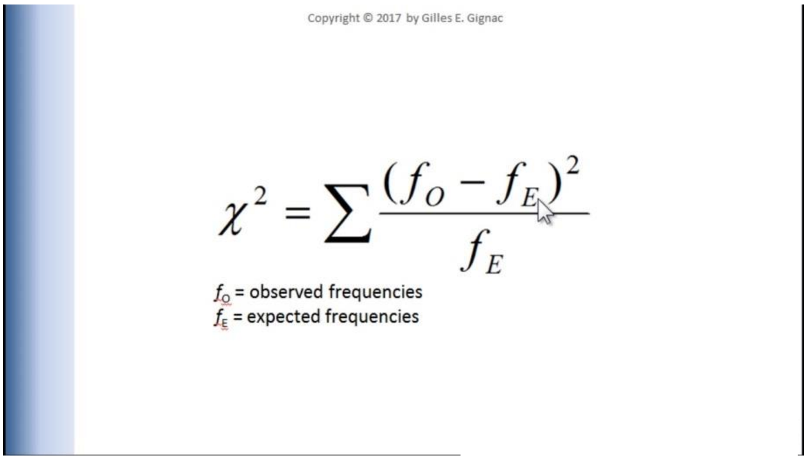
Q-Q plot is a probability plot for compare and analyze two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then poins on Q-Q plot perfectly lie in straight line y=x.

With the help of Q-Q plot we can identify a distribution is Gaussian or not.



Chi-square test :

Chi-square test claims about population proportion.
It is a non-parametric test that is performed on categorical variable.



Hypothesis Testing :

Hypothesis testing is a method used to decide whether the data to draw the inference about the parameter is true or not.

z-test

We go for t-test if

i. We know the population std deviation. or

ii. We don’t know the population std deviation but our sample size is greater than 30 (n>30).

$$\text{z-statistic} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

<u>t-test</u>

We go for t-test if

i. We don’t know the population std deviation.

ii. Our sample size is less than 30 (n<30).

iii. Sample std deviation is given.

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$$

<u>Annova Test</u> :

An ANNOVA(analysis of variance) test is a statistical test used to determine the statistical difference between two or more categorical groups by calculating mean by using variance.

$$SS_{total} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - \overline{x})^2$$

$$SS_{between} = \sum_{j=1}^{p} n_j (\overline{x}_j - \overline{x})^2$$

$$SS_{within} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - \overline{x}_j)^2$$

There are different types of ANOVA tests

i. One way

ii. Two way

The difference between these two types depends on the number of independent variables in the dataset.

A one-way ANOVA has one categorical independent variable and a normally distributed continuous variable.

A two-way ANOVA has two or more categorical independent variables and a normally distributed continuous dependent variable.

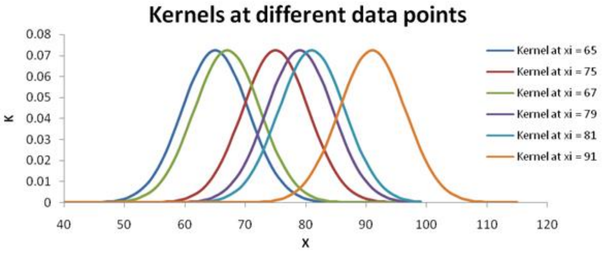
1 sample t-test and 2 sample t-test :

A **one-sample t-test** is used to compare a single population to a standard value (for example, to determine whether the average lifespan of a specific town is different from the country average).

A **paired t-test** is used to compare a single population before and after some experimental intervention or at two different points in time (for example, measuring student performance on a test before and after being taught the material).

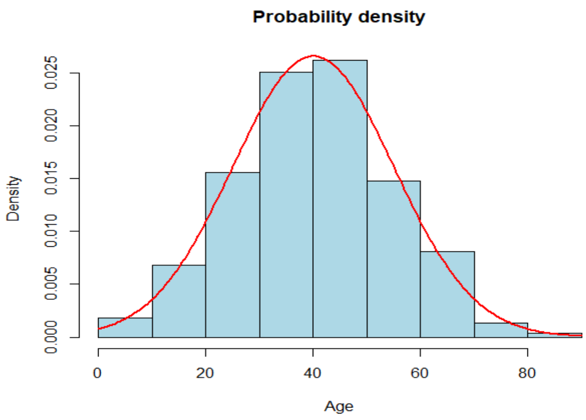
Kernel density estimation :

Kernel density estimation (KDE) is a smoothing process where inferences about population are made based on finite sample value and main aim of KDE is to estimate probability density function for the given dataset.



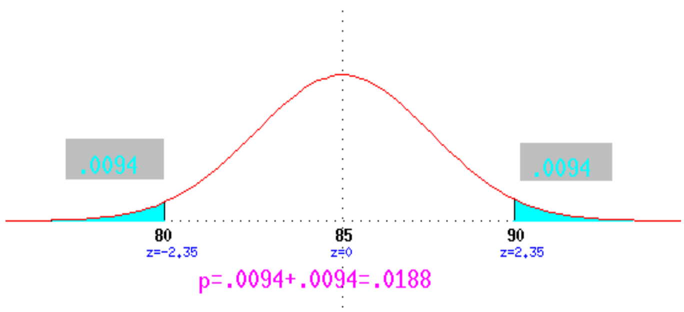
Probability density function :

Probability density function (PDF) is nothing but a probability function which represents the density of continuous random variable lying in between a range of values.



P-value :

P-value always represents the significance level. It tells you how many values are not contributing out of whole experiments. (in general words p-value tells you how many experiments are going to fail out of 100)



Bell curve :

A bell curve is a graph which describes the normal distribution and has a shape similar to a bell.

The top of the curve shows mean, median and mode of the dataset.