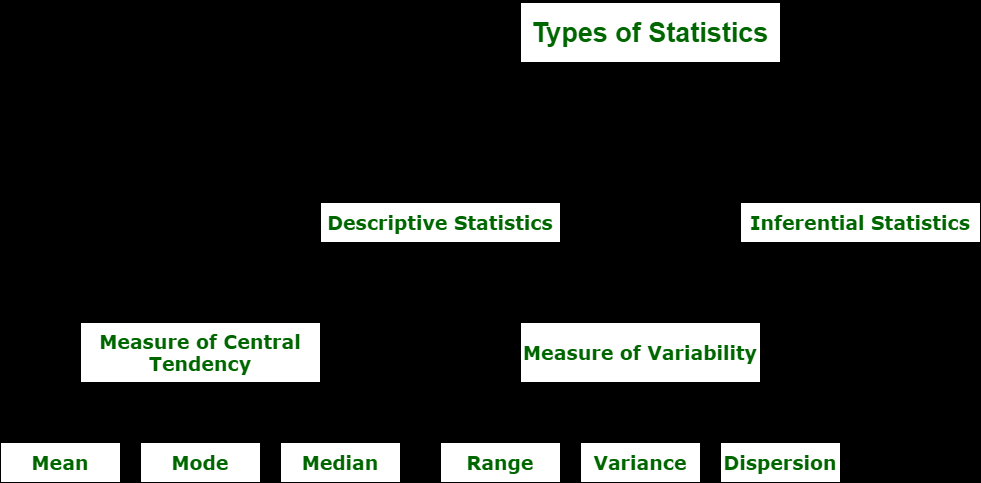
**STATISTICS**

# **Introduction of Statistics and its Types**

**Statistics simply means numerical data, and is field of math that generally deals with collection of data, tabulation, and interpretation of numerical data.**

* **Population –  
  It is actually a collection of set of individuals or events whose properties are to be analyzed.**
* **Sample –  
  It is the subset of a population.**

****

**Types of Statistics :**

**1. Descriptive Statistics :**

**Descriptive statistics uses data that provides a description of the population either through numerical calculation or graph or table.**

**(a). Measure of central tendency –**

**Measure of central tendency is also known as summary statistics that is used to represents the center point or a particular value of a data set or sample set.**

**Mean - Avg , Median - Middle , Mode - Most frequent.**

**b). Measure of Variability –**

**Measure of Variability is also known as measure of dispersion and used to describe variability in a sample or population.**

**Range, Variance, Standard deviation, Dispersion**

**2. Inferential Statistics :**

**Inferential Statistics makes inference and prediction about population based on a sample of data taken from population. It generalizes a large dataset and applies probabilities to draw a conclusion.**

**Types of inferential statistics –**

* **One sample test of difference/One sample hypothesis test**
* **Confidence Interval**
* **Contingency Tables and Chi-Square Statistic**
* **T-test or Anova**
* **Pearson Correlation**
* **Bi-variate Regression**
* **Multi-variate Regression**

**=====================================================================Probability**

**Probability refers to the extent of occurrence of events.**

**There are three approaches to the theory of probability:**

1. **Empirical Approach**
2. **Classical Approach**
3. **Axiomatic Approach**

**Axiomatic Approach. In this approach, we represent the probability in terms of sample space(S).**

* **Random Event :- If the repetition of an experiment occurs several times under similar conditions, if it does not produce the same outcome everytime but the outcome in a trial is one of the several possible outcomes, then such an experiment is called random event or a probabilistic event.**
* **Elementary Event – The elementary event refers to the outcome of each random event performed. Whenever the random event is performed, each associated outcome is known as elementary event.**
* **Sample Space – Sample Space refers to the set of all possible outcomes of a random event.Example, when a coin is tossed, the possible outcomes are head and tail.**
* **Event – An event refers to the subset of the sample space associated with a random event.**
* **Occurrence of an Event – An event associated with a random event is said to occur if any one of the elementary event belonging to it is an outcome.**
* **Sure Event – An event associated with a random event is said to be sure event if it always occurs whenever the random event is performed.**
* **Impossible Event – An event associated with a random event is said to be impossible event if it never occurs whenever the random event is performed.**
* **Compound Event – An event associated with a random event is said to be compound event if it is the disjoint union of two or more elementary events.**
* **Mutually Exclusive Events – Two or more events associated with a random event are said to be mutually exclusive events if any one of the event occurs, it prevents the occurrence of all other events.This means that no two or more events can occur simultaneously at the same time.**
* **Exhaustive Events – Two or more events associated with a random event are said to be exhaustive events if their union is the sample space.**

**+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Addition Rule in Probability**

**If A and B are two events in a probability experiment, then the probability that either one of the events will occur is: P(A or B)=P(A)+P(B)−P(A and B)**

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Multiplication Rule in Probability**

**If A and B are two independent events in a probability experiment, then the probability that both events occur simultaneously is:**

***P*(*A* and *B*)=*P*(*A*)⋅*P*(*B*)**

**In case of dependent events , the probability that both events occur simultaneously is:**

***P*(*A* and *B*)=*P*(*A*)⋅*P*(*B* | *A*)**

**============================================================**

**Population Mean And Sample Mean**

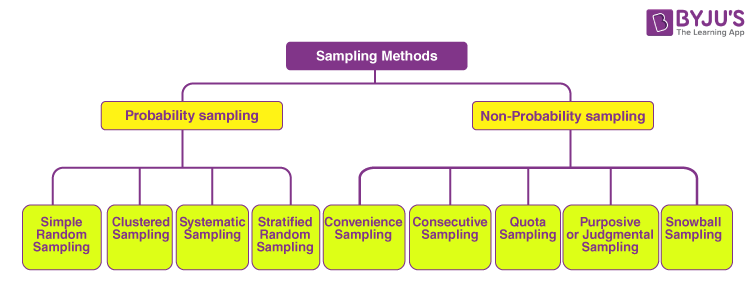
**Sample mean is the arithmetic mean of random sample values drawn from the population.**

**Population mean represents the actual mean of the whole population.**

**+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**What is Sampling Method And Its Types**

**In Statistics, the sampling method or sampling technique is the process of studying the population by gathering information and analyzing that data.**

****

1. **Probability Sampling**

**The probability sampling method utilizes some form of random selection. In this method, all the eligible individuals have a chance of selecting the sample from the whole sample space.**

1. **Non-probability Sampling**

**The non-probability sampling method is a technique in which the researcher selects the sample based on subjective judgment rather than the random selection. In this method, not all the members of the population have a chance to participate in the study.**

**=====================================================================**

**What is Variables And Its Types?**

**Variable is an algebraic term that denotes the unknown value that is not a fixed value which is in numerical format.**

#### **Discrete Variable:**

#### **Continuous Variables**

#### **Independent Variables**

#### **Dependent Variables**

#### **Categorical Variables**

**=====================================================================**

**Variable Measurement Scales**

**Nominal**

**Nominal scales assign numbers as labels to identify objects or classes of objects.**

**Ordinal**

**Ordinal scales build upon nominal scales by assigning numbers to objects to reflect a rank ordering on an attribute in question.**

**Interval**

**Interval scales build upon ordinal scale variables. In an interval scale, numbers are**

**assigned to objects such that the differences (but not ratios) between the numbers can be meaningfully interpreted.**

**Ratio**

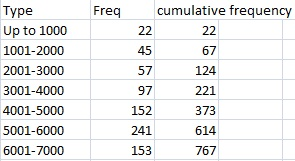
**Ratio scales have all the attributes of interval scale variables and one additional attribute:ratio scales include an absolute “zero” point.**

**====================================================================**

**Frequency Distribution And Cumulative Frequency**

**Relative frequency is the fraction or proportion times an answer occurs in the data set. Relative frequencies can be written in decimal, fraction or percents.**

**Cumulative frequency is the collection of all previous frequencies together. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the present given row or column.**

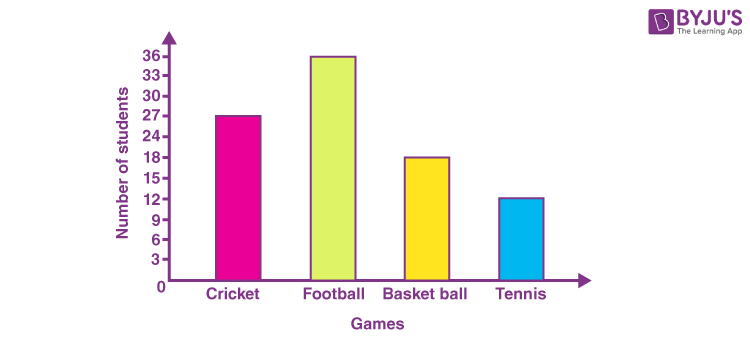
****

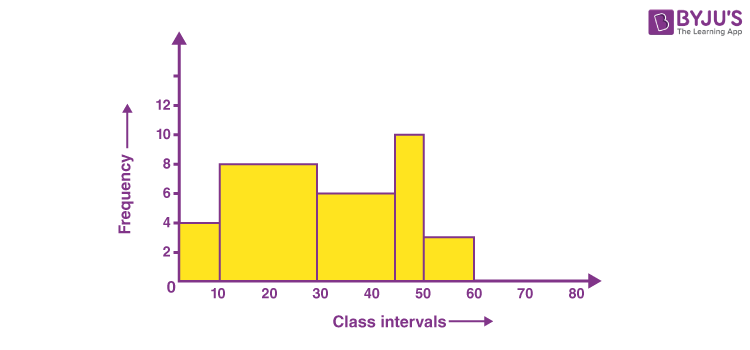
**=====================================================================**

**Histograms**

A **histogram** is a graphical representation of a grouped frequency distribution with continuous classes.

| **Histogram** | **Bar Graph** |
| --- | --- |
| **It is a two-dimensional figure** | **It is a one-dimensional figure** |
| **The frequency is shown by the area of each rectangle** | **The height shows the frequency and the width has no significance.** |
| **It shows rectangles touching each other** | **It consists of rectangles separated from each other with equal spaces.** |

****

****

**=====================================================================**

**Percentiles And Quantiles**

**Quartiles and percentiles are a measures of variation, which describes how spread out the data is.**

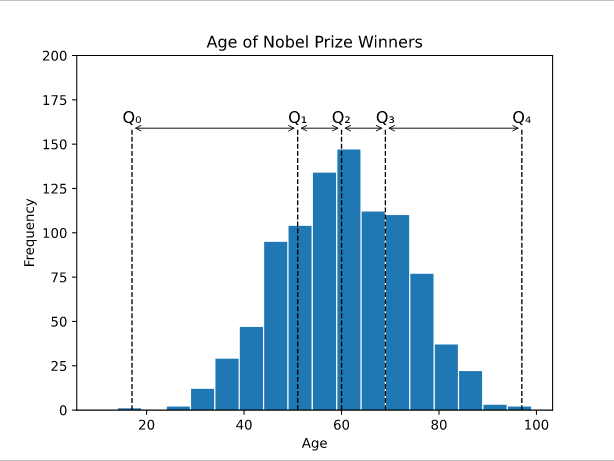
**Quartiles are values that separate the data into four equal parts.**

**import numpy**

**values = [13,21,21,40,42,48,55,72]**

**x = numpy.quantile(values, [0,0.25,0.5,0.75,1])**

**print(x)**

****

**Percentiles are values that separate the data into 100 equal parts.**

**Method to find the 65th percentile of the values 13, 21, 21, 40, 42, 48, 55, 72:**

**import numpy**

**values = [13,21,21,40,42,48,55,72]**

**x = numpy.percentile(values, 65)**

**print(x)**

**+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Five Number Summary**

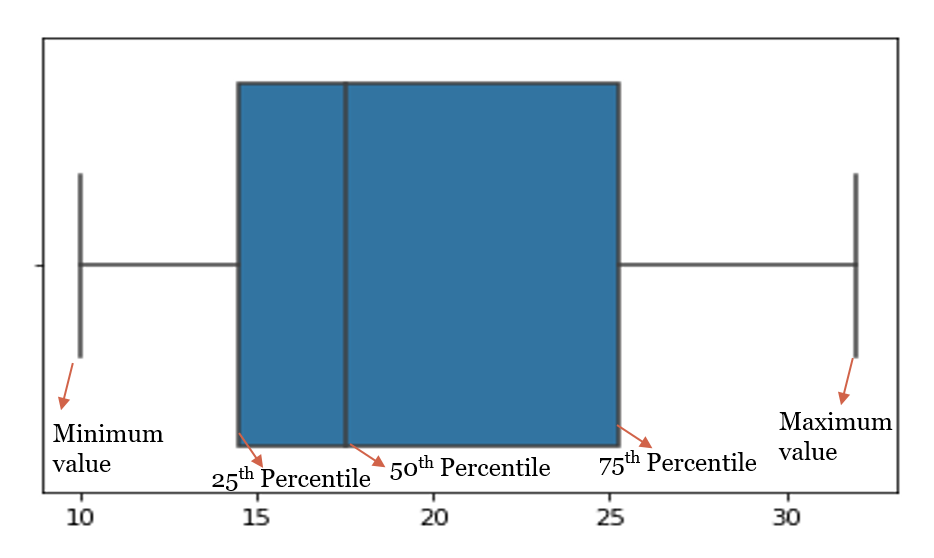
**Five number summary is a part of descriptive statistics and consists of five values and all these values will help us to describe the data.**

* **The minimum value (the lowest value)**
* **25th Percentile or Q1**
* **50th Percentile or Q2 or Median**
* **75th Percentile or Q3**
* **Maximum Value (the highest value)**

**=====================================================================**

**BOX PLOT**

**Boxplots are the graphical representation of the distribution of the data using Five Number summary values. It is one of the most efficient ways to detect outliers in our dataset.**

****

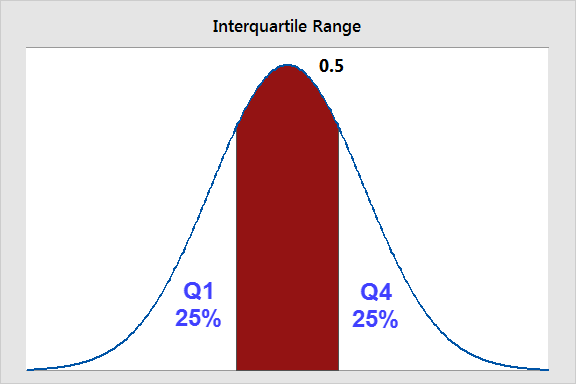
**n statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the dataset. An outlier can cause serious problems in statistical analyses.**

**==================================================================**

**Inter Quartile Range(IQR)**

**Interquartile range = Upper Quartile – Lower Quartile = Q­3 – Q­1**

**The interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles. First Quartile is denoted by Q1 known as the lower quartile, the second Quartile is denoted by Q2 and the third Quartile is denoted by Q3 known as the upper quartile.**

****

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Effect Of Outliers And Its Removal**

**An Outlier is that observation which is significantly different from all other observations.**

**When we should remove outliers**

**I believe that the dropping outlier is always a harsh step and should be taken only in extreme conditions when we’re very sure that the outlier is due to a measurement error, which we generally do not know while doing analysis.**

**Sometimes outliers indicate a mistake in data collection. Other times, though, they can influence a data set, so it’s important to keep them to better understand the dataset in the big picture.**

**================================================================**

## **Probability Density Function?**

**The Probability Density Function(PDF) defines the probability function representing the density of a continuous random variable lying between a specific range of values.**

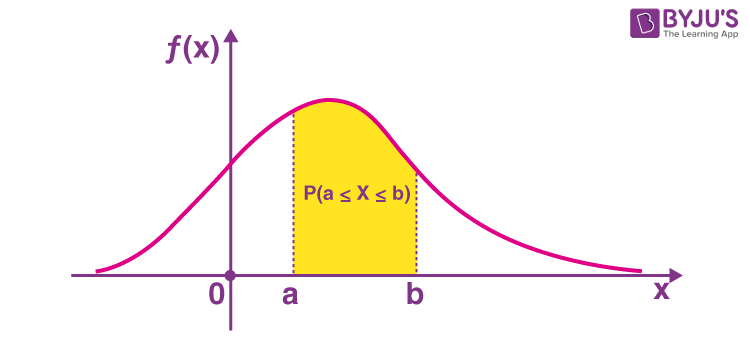
**The different types of variables**

**Discrete Variable: A variable that can only take on a certain finite value within a specific range is called a discrete variable. e.g., a sum of two dice.**

**Continuous Variable: A continuous random variable can take on infinite different values within a range of values, e.g., amount of rainfall occurring in a month.**

### Probability Density Function Graph

The probability density function is defined as an integral of the density of the variable density over a given range.

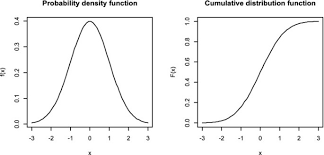
****

**Eg.**

* **In Statistics, it is used to calculate the probabilities associated with the random variables.**
* **The probability density function is used in modelling the annual data of atmospheric NOx temporal concentration**
* **It is used to model the diesel engine combustion.**

**===================================================================**

**Cumulative Distribution Function**

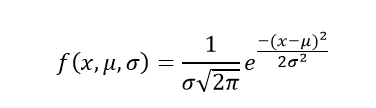
****

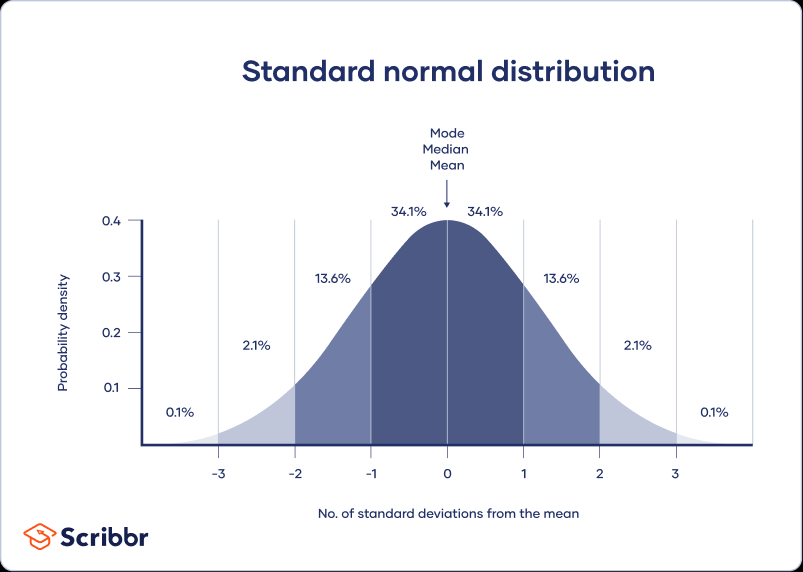
**===================================================================**

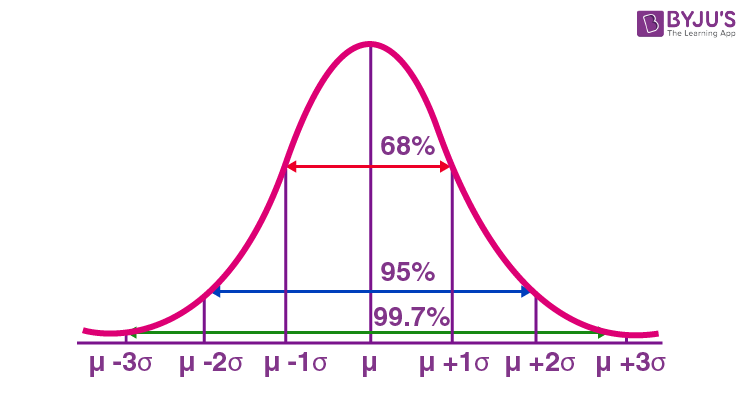
**Normal distribution**

**Normal distribution, also known as the Gaussian distribution, is a probability distribution function that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.**

* **The mean, median and mode are exactly the same.**
* **The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.**
* **The distribution can be described by two values: the mean and the standard deviation.**

****

****

****

**=====================================================================**

**Z-SCORE /Standardization**

**Z-score is also known as standard score gives us an idea of how far a data point is from the mean. It indicates how many standard deviations an element is from the mean. Hence, Z-Score is measured in terms of standard deviation from the mean.**

**A z-score can be calculated using the following formula.**

#### **z = (X – μ) / σ**

**where,**

**z = Z-Score,**

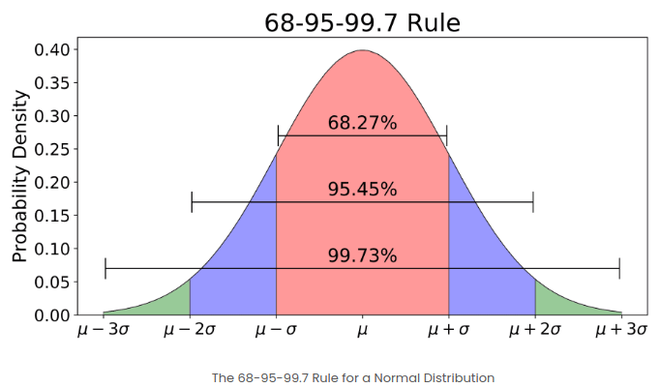
**X = The value of the element,**

**μ = The population mean, and**

**σ = The population standard deviation**

### **Interpretation of Z-score**

* **An element having a z-score less than 0 represents that the element is less than the mean.**
* **An element having a z-score greater than 0 represents that the element is greater than the mean.**
* **An element having a z-score equal to 0 represents that the element is equal to the mean.**
* **An element having a z-score equal to 1 represents that the element is 1 standard deviation greater than the mean; a z-score equal to 2, 2 standard deviations greater than the mean, and so on.**
* **An element having a z-score equal to -1 represents that the element is 1 standard deviation less than the mean; a z-score equal to -2, 2 standard deviations less than the mean, and so on.**
* **If the number of elements in a given set is large, then about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2; about 99% have a z-score between -3 and 3. This is known as the Empirical Rule or the 68-95-99.7 Rule and can be demonstrated in the image below**

****

**=====================================================================**

**Standardization Vs Normalization**

**Feature scaling techniques**

**1. Normalization [0, 1]**

**Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:**

**X\_new = (X - X\_min)/(X\_max - X\_min)**

**This scales the range to [0, 1]**

**Normalization is useful when there are no outliers as it cannot cope up with them. Usually, we would scale age and not incomes because only a few people have high incomes but the age is close to uniform.**

**It is used when features are of different scales.**

**2. Standardization (Mean -0, Standard D - 1)**

**Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.**

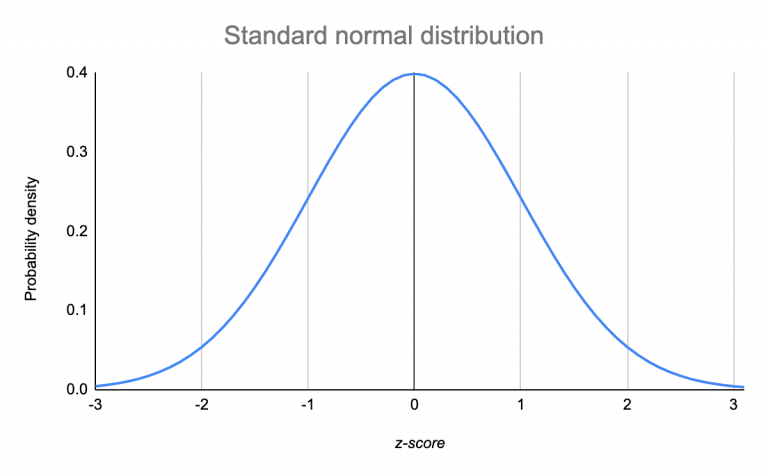
**X\_new = (X - mean)/Std**

**It is used when we want to ensure zero mean and unit standard deviation.**

**==========================================================**

**Standard Normal Distribution(m=0,sd=0, no m see in image)**

**The standard normal distribution, also called the *z*-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1.**

****

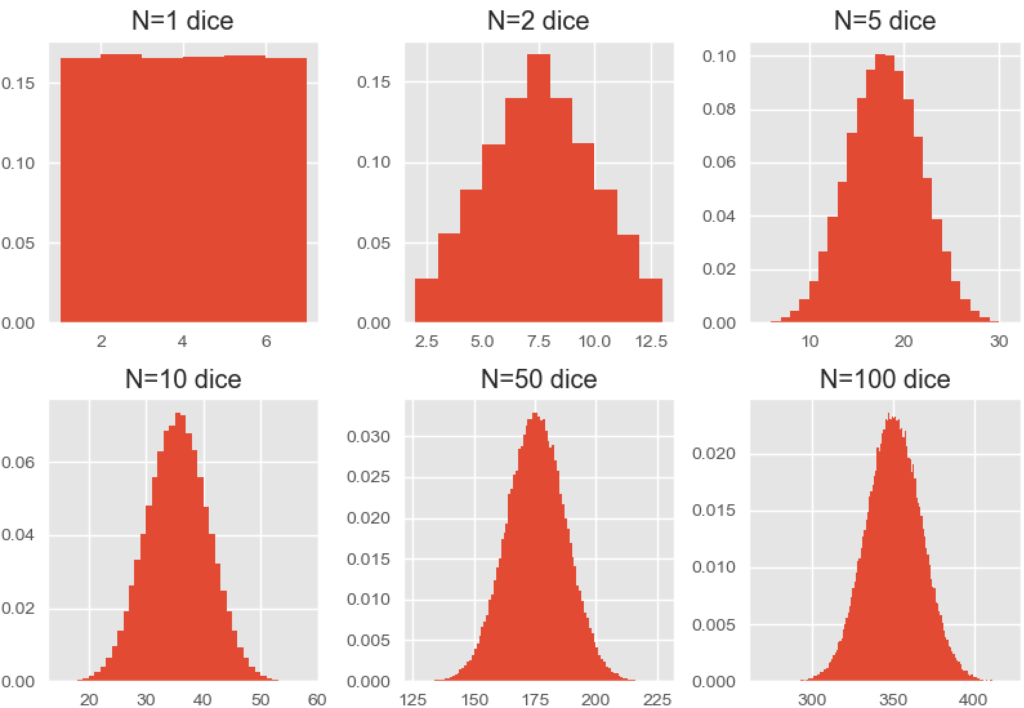
**When you standardize a normal distribution, the mean becomes 0 and the standard deviation becomes 1. This allows you to easily calculate the probability of certain values occurring in your distribution, or to compare data sets with different means and standard deviations.**

**===========================================================**

**Central Limit Theorem**

**The central limit theorem states that whenever a random sample of size n is taken from any distribution with mean and variance, then the sample mean will be approximately normally distributed with mean and variance. The larger the value of the sample size, the better the approximation to the normal.**

**Here’s what the Central Limit Theorem is saying, graphically. The picture below shows one of the simplest types of test: rolling a fair die. The more times you roll the die, the more likely the shape of the distribution of the means tends to look like a normal distribution graph.**

****

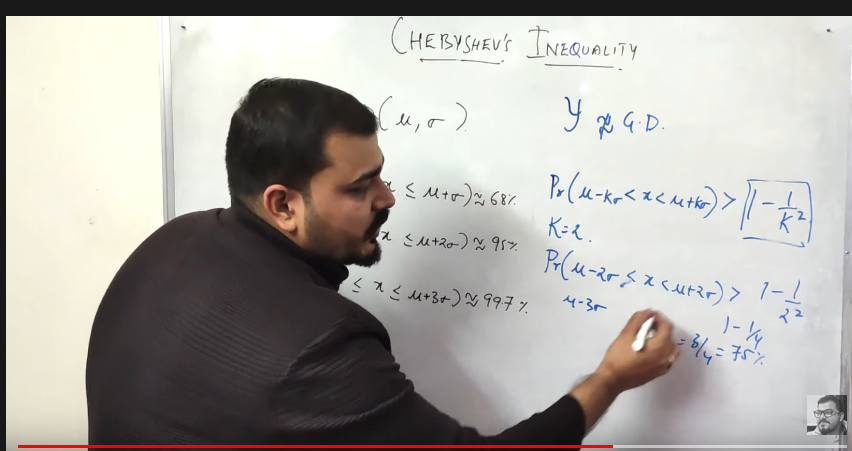
**=====================================================================**

**Chebyshevs Inequality**

**For a random variable which doesn’t follow normal distribution or GD.**

**P(mue - SD < X <mue + SD) > 1-1/k2.**

**Theorem states that no more than 1 / *k*2 of the distribution's values will be more than k standard deviations away from the mean. Looked at another way, 1 - (1 / *k*2) of the distribution's values will lie within *k* standard deviations of the mean.**

****

**For k=2 and k=3**

**, the theorem states that the proportion within 2 and 3 deviations will be at least**

**1−1/22=75% and 1−1/32≈89%**

**====================================================================**

**Covariance**

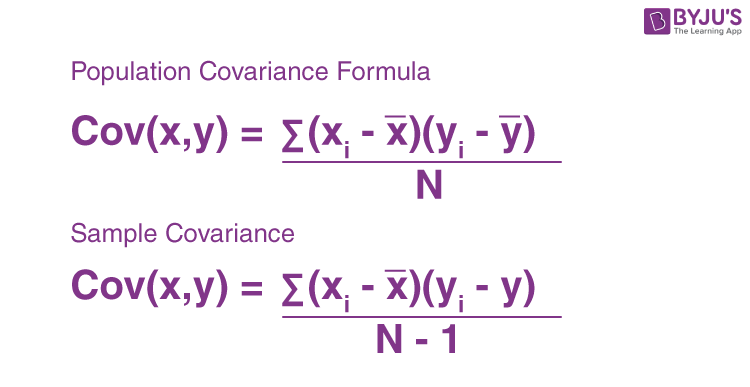
**Covariance is a measure of the relationship between two random variables and to what extent, they change together.**

### **Positive Covariance**

**If the covariance for any two variables is positive, that means, both the variables move in the same direction. Here, the variables show similar behaviour.**

### **Negative Covariance**

**If the covariance for any two variables is negative, that means, both the variables move in the opposite direction.**

****

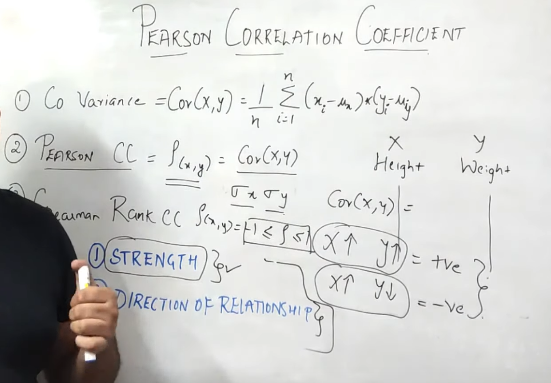
**================================================================**

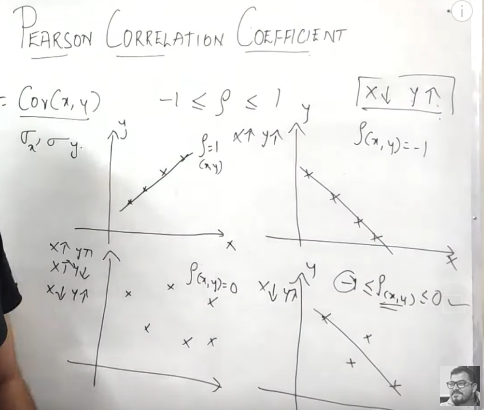
**Pearson Correlation Coefficient(cov(x,y)/SD(x),SD(y))**

**The value of the Pearson correlation coefficient product is between -1 to +1. When the correlation coefficient comes down to zero, then the data is said to be not related. While, if we are getting the value of +1, then the data are positively correlated and -1 has a negative correlation.**

**Pearson Corelation tell along with covariance(just direction)**

1. **Strength/magnitude.**
2. **Direction of relation.**

****

****

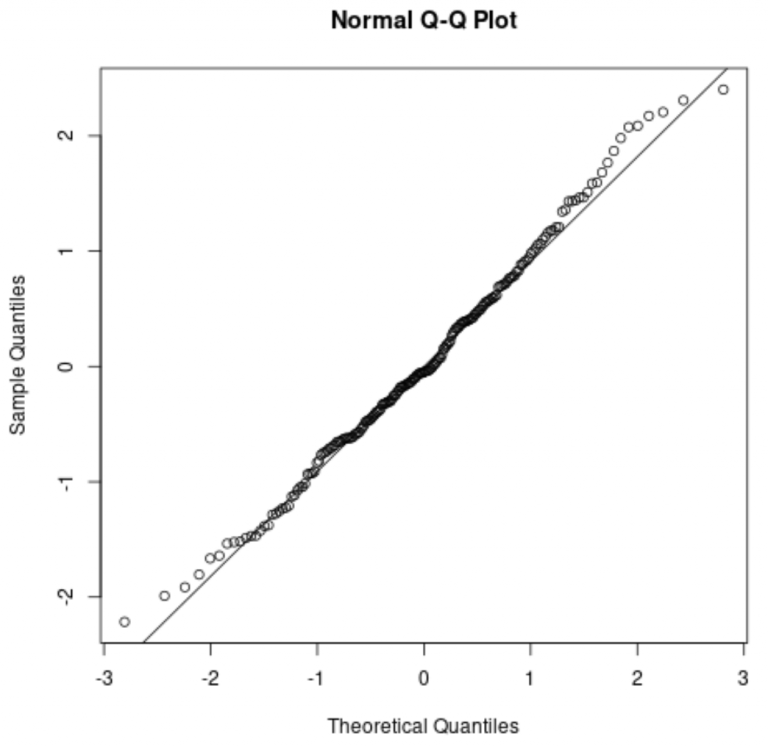
**===============================================================**

**QQ plot Check data is Normally Distributed**

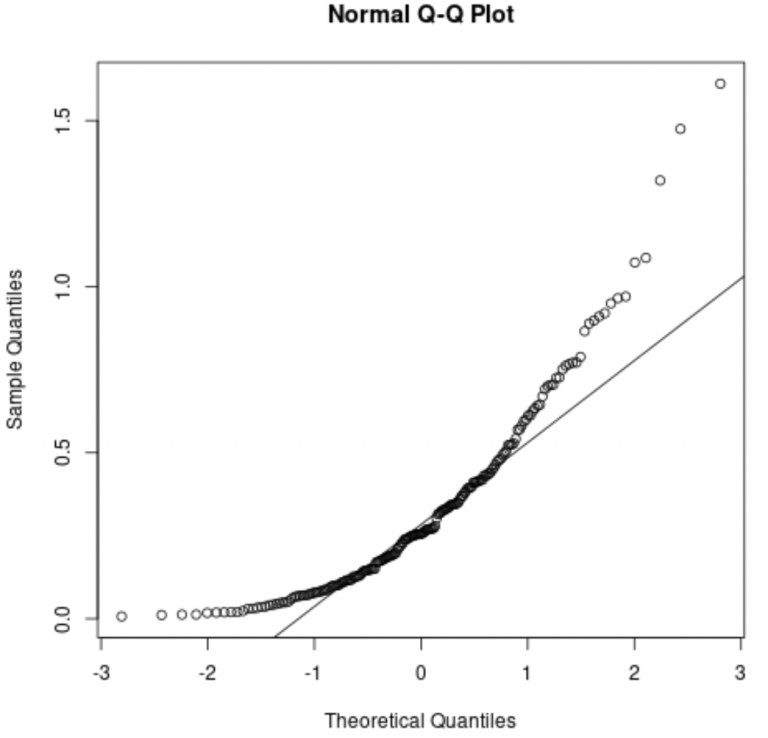
**This type of plot is used to determine whether or not a set of data follows a normal distribution.**

**If the data is normally distributed, the points in a Q-Q plot will lie on a straight diagonal line.**

**the more the points in the plot deviate significantly from a straight diagonal line, the less likely the set of data follows a normal distribution.**

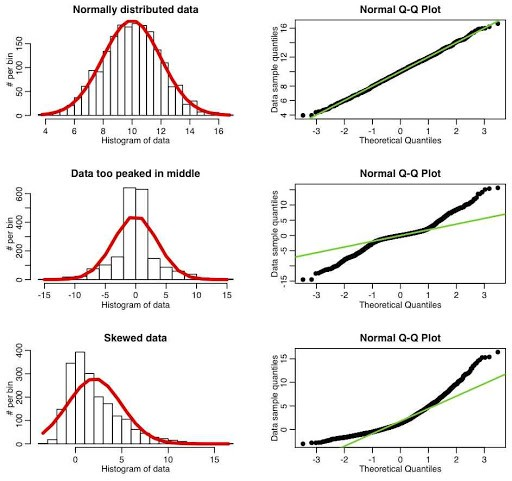
****

**We can see that the points lie mostly along the straight diagonal line with some minor deviations along each of the tails. Based on this plot, we could safely assume that this set of data is normally distributed.**

****

**We can see that the points deviate significantly from the straight diagonal line. This is a clear indication that the set of data is not normally distributed.**

**This should make sense considering we specified that the data should follow an exponential distribution.**

****

**====================================================================**

**Bernoulli Distribution And Binomial**

**The Bernoulli distribution represents the success or failure of a single Bernoulli trial. The Binomial Distribution represents the number of successes and failures in n independent Bernoulli trials for some given value of n.**

**A product can be defective or non-defective, etc. These types of independent trials which have only two possible outcomes are known as Bernoulli trials.**

**For the trials to be categorized as Bernoulli trials it must satisfy these conditions:**

* **A number of trials should be finite.**
* **The trials must be independent.**
* **Each trial should have exactly two outcomes: success or failure.**

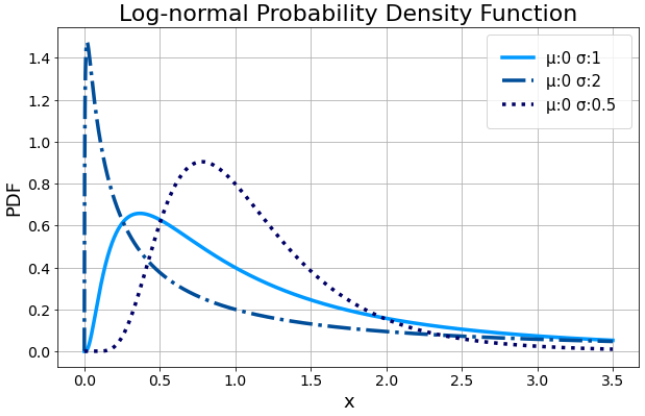
**It is the collection of Bernoulli trials for the same event, i.e., it contains more than 1 Bernoulli event for the same scenario for which the Bernoulli trial is calculated.**

**The binomial distribution is used when there are exactly two mutually exclusive outcomes of a trial.**

**====================================================================**

**Log Normal Distribution**

**The log-normal distribution is a right skewed continuous probability distribution, meaning it has a long tail towards the right. It is used for modelling various natural phenomena such as income distributions**

****

**===============================================================**

**Confidence Interval In statistics**

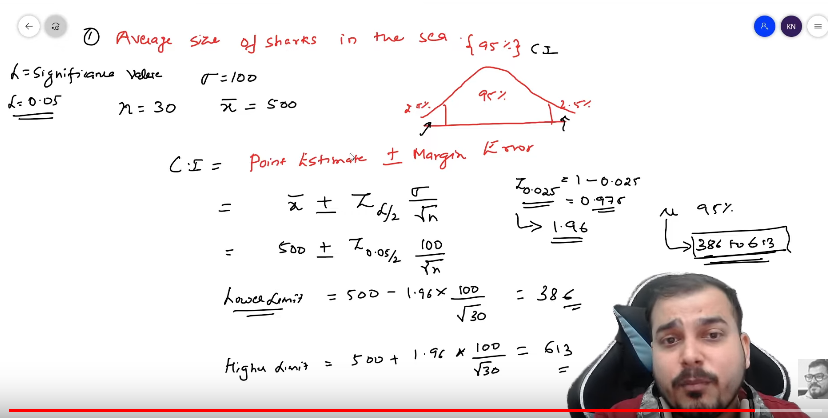
**The confidence interval is the range of values that you expect your estimate to fall between a certain percentage of the time if you run your experiment again or re-sample the population in the same way.**

**The confidence level represents the proportion of acceptable confidence intervals that contain the true value of the unknown parameter. Like when you estimate size of sharks in sea and get value 200 and 400. You tel size is between 200 and 400.**

**A confidence interval is the mean of your estimate plus and minus the variation in that estimate.**

**The confidence interval is based on the mean and standard deviation. Thus, the formula to find CI is**

**X̄ ± Zα/2 × [ σ / √n ]**

****

**================================================================**

All Transformation Techniques

**================================================================**

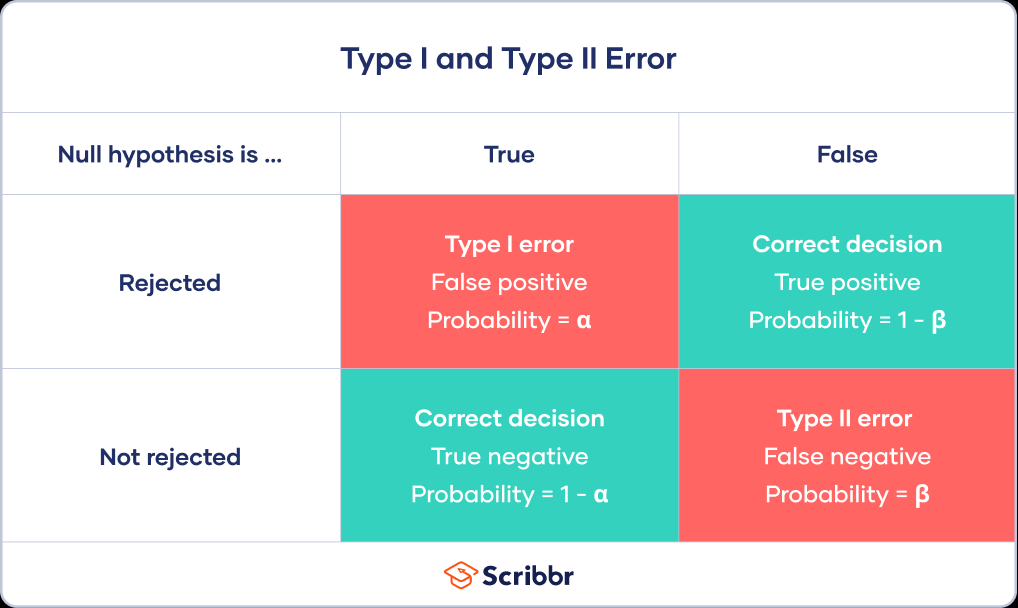
**Type 1 And Type 2 error**

**In statistics, a Type I error is a false positive conclusion, while a Type II error is a false negative conclusion.**

* **Type I error (false positive): the test result says you have coronavirus, but you actually don’t.**
* **Type II error (false negative): the test result says you don’t have coronavirus, but you actually do.**

**A Type I error means rejecting the null hypothesis when it’s actually true.**

**A Type II error means not rejecting the null hypothesis when it’s actually false.**

****

**==============================================================**

**Hypothesis Testing, p value[0-1]/Significance value**

**Hypothesis testing can be defined as the statistical framework which can be used to test whether the claim made about anything is true or otherwise.**

**The null hypothesis represents the default state of belief in the real worl.**

**The alternate hypothesis represents something different and unexpected.**

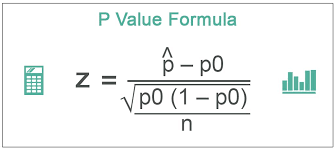
**The *p*-value is a number, calculated from a statistical test, that describes how likely you to found a particular set of observations if the null hypothesis were true.**

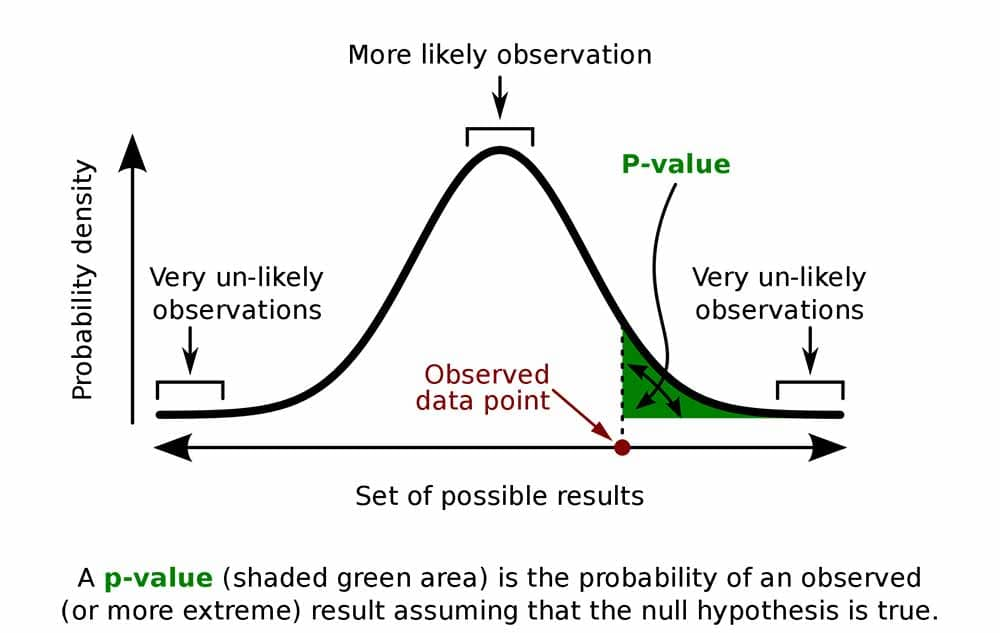
**P Value - Is the probability for the Null Hypothesis to be true.The smaller the *p*-value, the more likely you are to reject the null hypothesis.**

**If p = 0.1 , 10 out of 100, Null hypothesis wil be true. If p=0.001 1 out of 100.**

**If p value less than 0.05 then , I will reject NULL Hypothesis.**

* **Null hypothesis (*H*0): He is inocent.**
* **Alternative hypothesis ( *H*1): He is guilty.**

****

****

* **P≤ 0.05; Hypothesis rejected**
* **P>.05; Hypothesis Accepted**

**===============================================================**

**Steps For Hypothesis Testing**

1. **State your research hypothesis as a null hypothesis and alternate hypothesis (Ho) and (Ha or H1).**
2. **Collect data in a way designed to test the hypothesis.**
3. **Perform an appropriate statistical test.**
4. **Decide whether to reject or fail to reject your null hypothesis.**
5. **Present the findings in your result**[**s**](https://www.scribbr.com/dissertation/results/) **and discussion section.**

**===================================================================**

**T test(Apply when you have one numeral variable)**

**Corelation test - When we have two numeral variable.**

**A t-test is a statistical test that is used to compare the means of two groups.It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.**

**A t-test can only be used when comparing the means of two groups (a.k.a. pairwise comparison). If you want to compare more than two groups, or if you want to do multiple pairwise comparisons, use an ANOVA test or a post-hoc test.**

**The t-test is a parametric test of difference, meaning that it makes the same assumptions about your data as other parametric tests. The t-test assumes your data:**

1. **are independent**
2. **are (approximately) normally distributed.**
3. **have a similar amount of** [**vari**](https://www.scribbr.com/statistics/variance/)**a**[**nce**](https://www.scribbr.com/statistics/variance/) **within each group being compared (a.k.a. homogeneity of variance)**

### **One-sample, two-sample, or paired t-test?**

* **If the groups come from a single population (e.g. measuring before and after an experimental treatment), perform a paired t-test.**
* **If the groups come from two different populations (e.g. two different species, or people from two separate cities), perform a two-sample t-test (a.k.a. independent t-test).**
* **If there is one group being compared against a standard value (e.g. comparing the acidity of a liquid to a neutral pH of 7), perform a one-sample t-test.**

**====================================================================**

**z test**

**A z test is a test that is used to check if the means of two populations are different or not provided the data follows a normal distribution.**

**A z test is conducted on a population that follows a normal distribution with independent data points and has a sample size that is greater than or equal to 30.**

**=================================================================**

**Annova test**

**ANOVA test, in its simplest form, is used to check whether the** [**means**](https://www.cuemath.com/data/mean/) **of three or more populations are equal or not. The ANOVA test applies when there are more than two independent groups.**

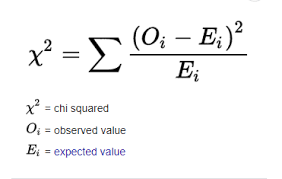
**====================================================================**

**Chisquare test(A chi-square test is a statistical test used to compare observed results with expected results.)**

**This test used when we have two categorical features.**

**One Sample categorial test - When we have one categorial feature,**

**The Chi-Square test is a statistical procedure for determining the difference between observed and expected data. This test can also be used to determine whether it correlates to the categorical variables in our data.**

****

**The degrees of freedom in a** [**statistical**](https://www.simplilearn.com/tutorials/machine-learning-tutorial/statistics-for-machine-learning) **calculation represent the number of variables that can vary in a calculation.**

**===================================================================**

| **Z Test** | **T-Test** |
| --- | --- |
| **A z test is a statistical test that is used to check if the means of two data sets are different when the population variance is known.** | **A** [**t-test**](https://www.cuemath.com/t-test-formula/) **is used to check if the means of two data sets are different when the population variance is not known.** |
| **The sample size is greater than or equal to 30.** | **The sample size is lesser than 30.** |
| **The** [**data**](https://www.cuemath.com/data/) **follows a normal distribution.** | **The data follows a student-t distribution.** |
| **The one-sample z test statistic is given by ¯¯¯x−μσ√n** |  |

**====================================================================**