**Statistics for Data Science**

**Overview of Statistics**

**Statistics:**

The science of collecting, describing, and interpreting data is popularly known as Statistical leveraging in Data Science

Two areas of Statistics in Data Science:

**Descriptive statistics** – Methods of organizing, summarizing, and presenting data in an informative way

**Inferential statistics** – The methods used to determine something about a population on the basis of a sample

* Applications: https://en.wikipedia.org/wiki/List\_of\_fields\_of\_application\_of\_statistics
* **Descriptive statistics** are methods for organizing and summarizing data.
* For example, tables or graphs are used to organize data, and descriptive values such as the average score are used to summarize data.
* A descriptive value for a population is called a **parameter** and a descriptive value for a sample is called a **statistic**.
* Collect data
* e.g., Survey
* Present data
* e.g., Tables and graphs
* Summarize data
* e.g., Sample mean
* **Inferential statistics** are methods for using sample data to make general conclusions (inferences) about populations.
* Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.
* Estimation
* e.g., Estimate the population mean weight using the sample mean weight
* Hypothesis testing
* e.g., Test the claim that the population mean weight is 70 kg
* **Inference is the process of drawing conclusions or making decisions about a population based on sample results**

Basic Terminologies:

**Population**: A collection, or set, of individuals or objects or events whose properties are to be analyzed.

Two kinds of populations: *finite* or *infinite*.

**Sample**: A subset of the population.

**Variable**: A characteristic about each individual element of a population or sample.

**Data (singular)**: The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.

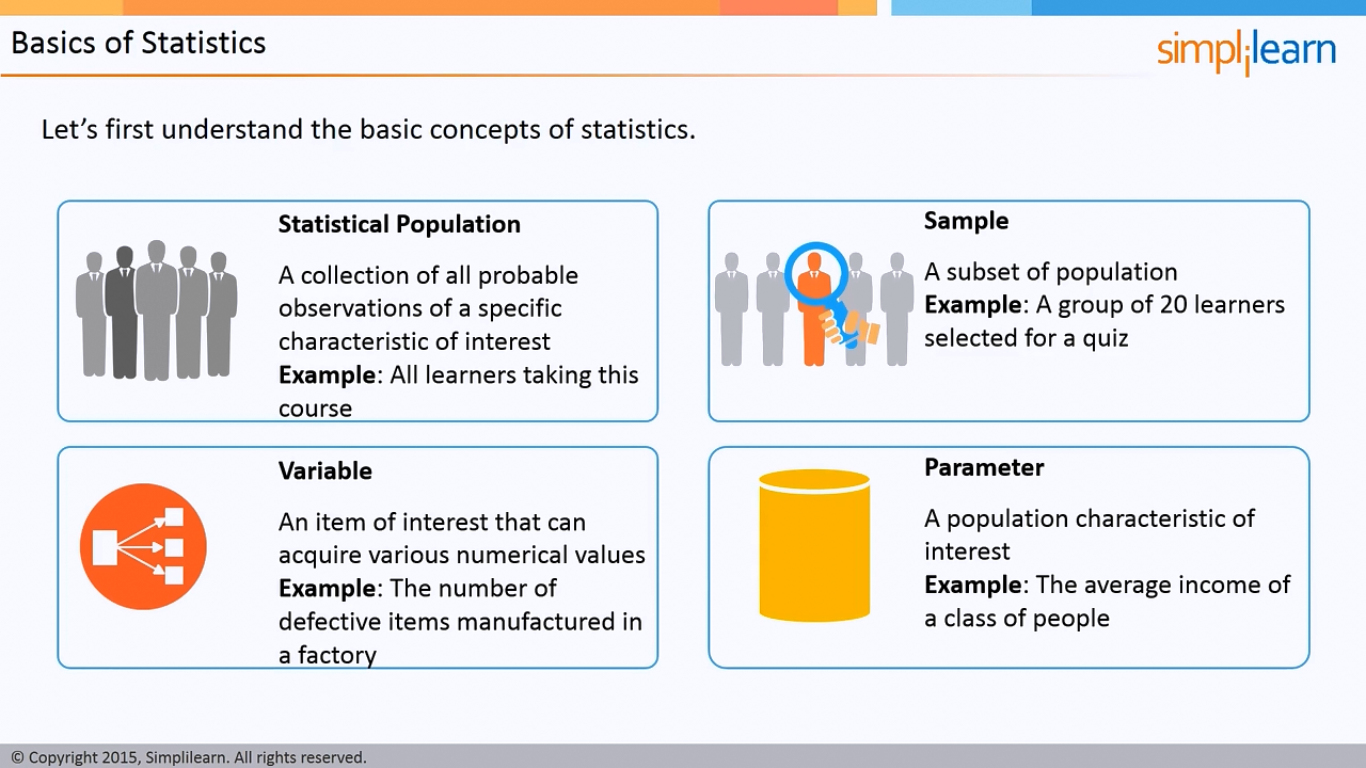
**Data (plural)**: The set of values collected for the variable from each of the elements belonging to the sample.

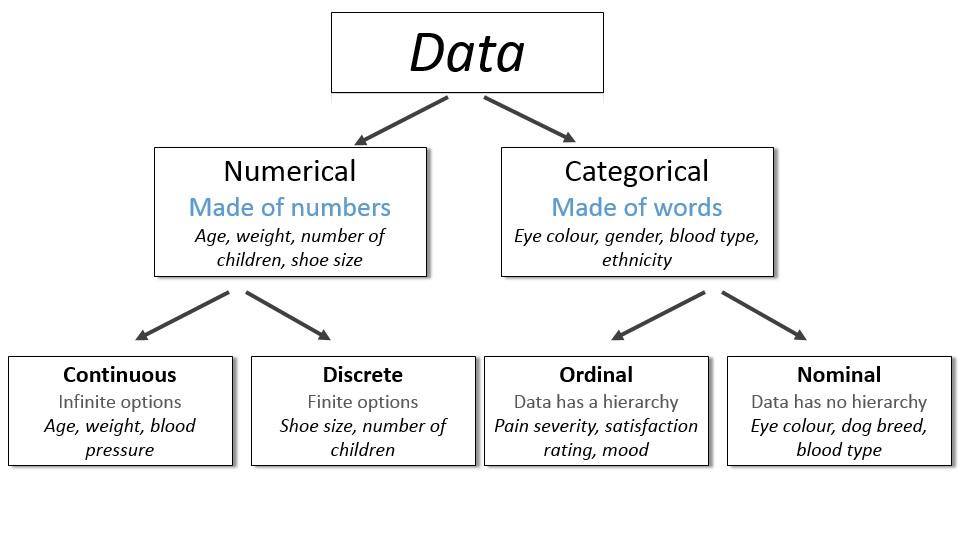
**Random Variable**:Variable are placeholder where you can store anything.It can number,or string,sentences.

**Experiment**: A planned activity whose results yield a set of data.

**Parameter**: A numerical value summarizing all the data of an entire population.

**Statistic**: A numerical value summarizing the sample data.





*Example*: Identify each of the following as examples of qualitative or numerical variables:

1. The temperature in Barrow, Alaska at 12:00 pm on any

given day.

2. The model of automobile.

3. Whether or not a 6 volt lantern battery is defective.

4. The weight of a lead pencil.

5. The length of time billed for a long distance telephone call.

6. The brand of cereal children eat for breakfast.

7. The type of book taken out of the library by an adult.

*Example*: Identify each of the following as examples of

1. nominal, (2) ordinal, (3) discrete, or (4) continuous variables:

* The length of time until a pain reliever begins to work.
* The number of chocolate chips in a cookie.
* The number of colors used in a statistics textbook.
* The brand of refrigerator in a home.
* The overall satisfaction rating of a new car.
* The number of files on a computer’s hard disk.
* The pH level of the water in a swimming pool.
* The number of staples in a stapler.

**Harnessing the Data**

Steps involved in Descriptive Statistics

* Collecting the data
* Presenting the data
* Summarizing the data

Making sense of data:

A sample which is drawn from the population should have same characteristics as the population.

Sampling can be:

* **with replacement**: a member of the population may be chosen more than once
* **without replacement**: a member of the population may be chosen only once (lottery ticket)

Collecting the data:

**Step1:-**Define the object or aim of the experiment.

i.e Estimate the average life of electronic component

**Step2:-**Define the variable and population of interest.

i.e usage,power rating,battery life etc

**Step3:-**Defining the data collection scheme and data measuring scheme.

i.e sampling procedure,sample size,data measuring device.

**Step4:-**Defining the appropriate descriptive and inferential analysis techniques

Methods used collect data:

Experiment: The investigator controls or modifies the environment and observes the effect on the variable under study.

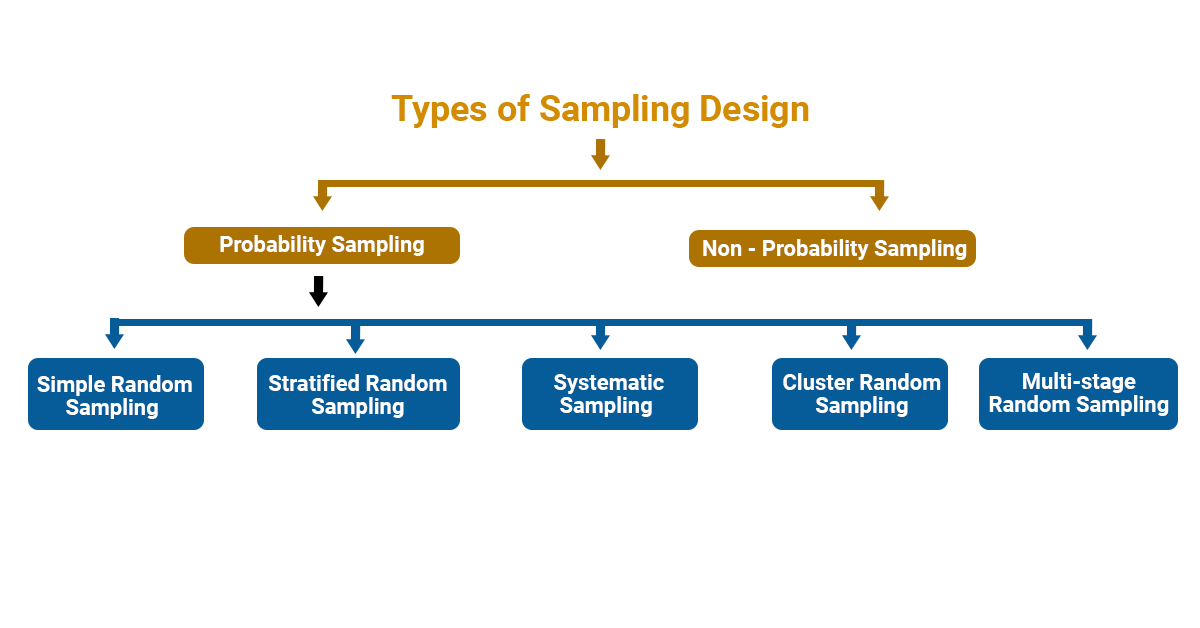
Survey: Data are obtained by sampling some of the population of interest. The investigator does not modify the environment.

Census: A 100% survey. Every element of the population is listed. Seldom used: difficult and time-consuming to compile, and expensive.

Judgment Samples: It is a non-probability sampling technique in which the sample members are chosen only on the basis of the researcher's knowledge and judgment.

Probability Samples: Samples in which the elements to be selected are drawn on the basis of probability. Each element in a population has a certain probability of being selected as part of the sample.

Types of Sampling



Probability Sampling:

**1. Simple Random sampling** :-each sample of the same size has an equal chance of being selected.

**2. Stratified Sampling** :-divide the population into groups called strata and then take a sample from each stratum.

**3. Cluster sampling** :-divide the population into strata and then randomly select some of the strata. All the members from these strata are in the cluster sample.

**4. Systematic sampling** :-randomly select a starting point and take every n-th piece of data from a listing of the population.

**5. Multistage Random** :- divide the population into clusters and select some clusters at the first stage. At each subsequent stage, you further divide up those selected clusters into smaller clusters, and repeat the process until you get the desired sample size.

Sampling Example:

*Example*: An employer is interested in the time it takes each employee to commute to work each morning. A random sample of 35 employees will be selected and their commuting time will be recorded.

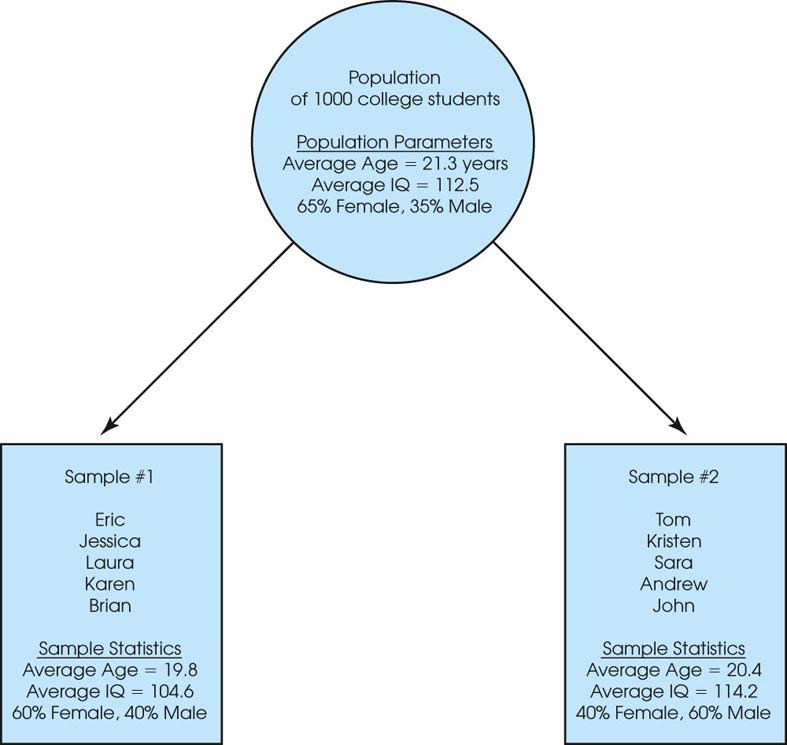
There are 2712 employees.

Each employee is numbered: 0001, 0002, 0003, etc. up to 2712.

Using four-digit random numbers, a sample is identified: 1315, 0987, 1125, etc.

Sampling Error:

* The discrepancy between a sample statistic and its population parameter is called **sampling error**.
* Defining and measuring sampling error is a large part of inferential statistics



**Figure 1.2**

A demonstration of sampling error. Two samples are selected from the same population. Notice that the sample statistics are different from one sample to another, and all of the sample statistics are different from the corresponding population parameters. The natural differences that exist, by chance, between a sample statistic and a population parameter are called sampling error.

**Exploratory Analysis:**

**Central Tendency:** The property of data being concentrated in the centre.

**Measure of Central Tendency:**

Mean: The mean is the average of all numbers and is sometimes called the arithmetic mean.

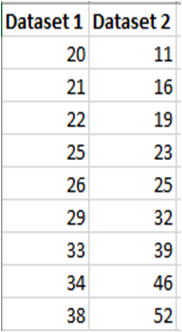
Median: The statistical median is the middle number in a sequence of numbers. To find the median, organize each number in order by size; the number in the middle is the median

Mode: The mode is the number that occurs most often within a set of numbers.

**Measure of Spread / Data Variability:**

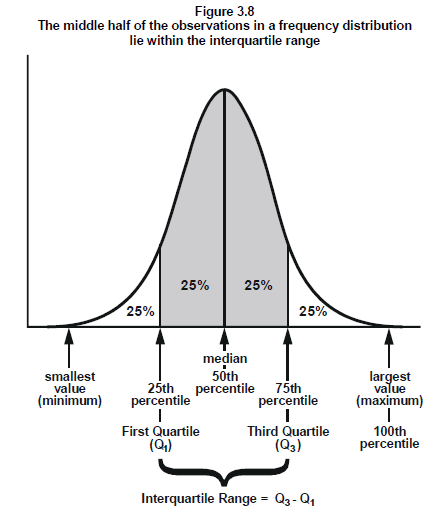
1. **Range:**

The range is the difference between the highest and lowest values within a set of numbers.



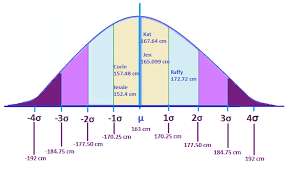
1. **Interquartile Range (IQR):** The interquartile range is the middle half of the data.

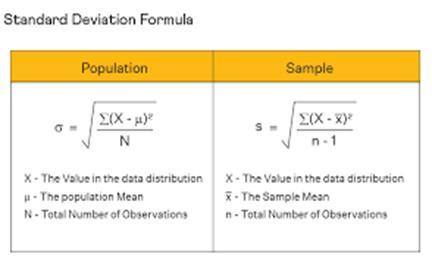
Mathematically the interquartile range includes the 50% of data points that fall between Q1 and Q3.



1. Standard Deviation:

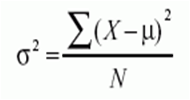
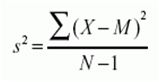
Standard Deviation (SD) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.





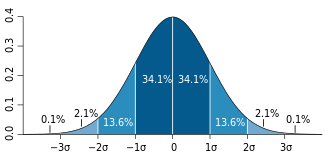
1. Variance:

Variance is the average squared difference of the values from the mean. Unlike the previous measures of variability, the variance includes all values in the calculation by comparing each value to the mean.

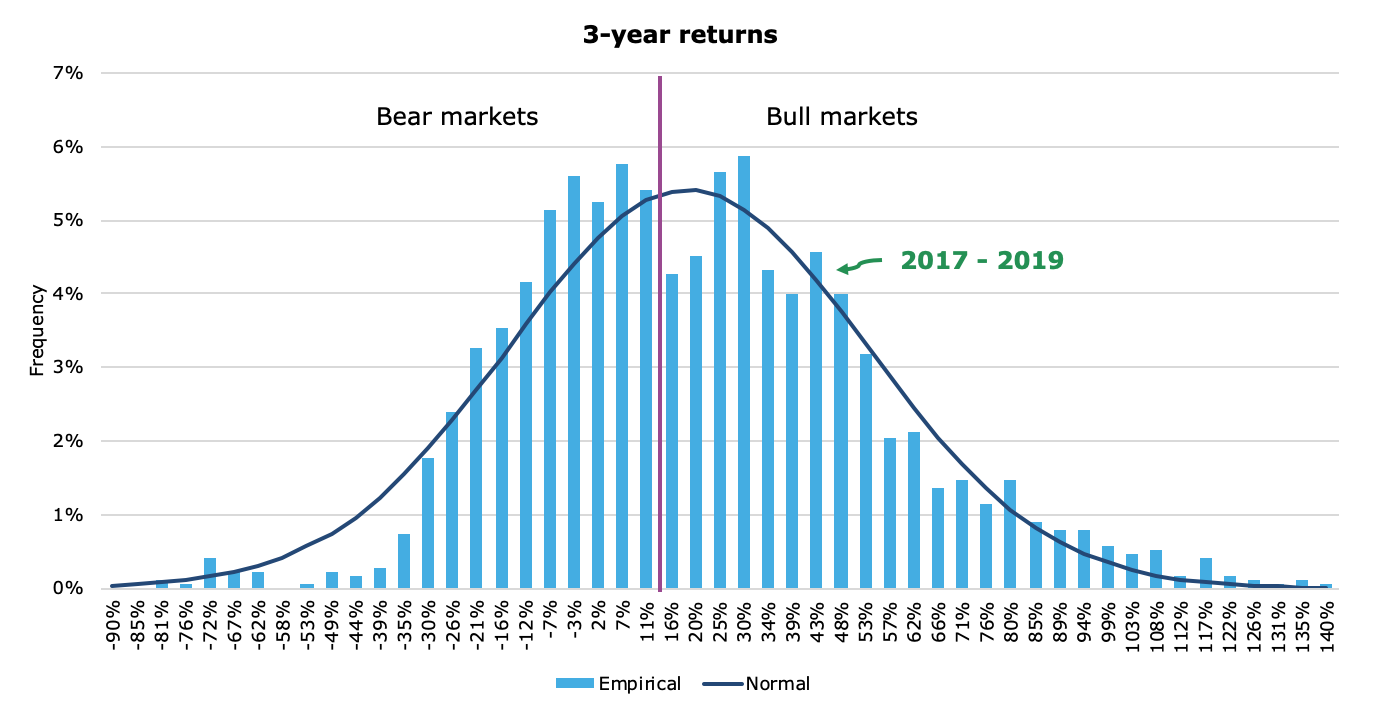
Percentile:

* A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.



Distribution:

The graphical representation of all observations is known as distribution

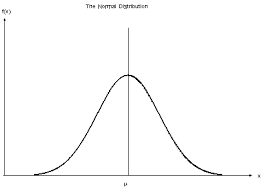


Normal Distribution, Mean, Variance

* The mean of a set of values is just the average of the values.
* Variance a measure of the width of a distribution. Specifically, the  variance is the mean squared deviation of points from the mean:
* The standard deviation is the square root of variance.
* The normal distribution is completed characterized by mean and variance.

Normal Distribution:

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



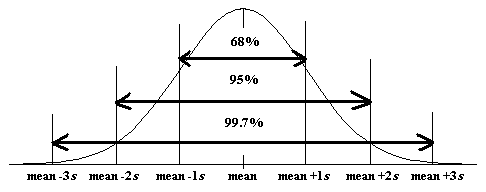
Properties of Normal Distribution

1. Empirical Rule
2. Distortion in Normal Distribution
3. Central Limit Theorem
4. Standard Normal Distribution
5. Outliers
6. QQ plot
7. Log,Sqrt,Boxcox transformation

Empirical Rule:

* The empirical rule states that for a normal distribution, nearly all of the data will fall within three standard deviations of the mean. The empirical rule can be broken down into three parts:
* 68% of data falls within the first standard deviation from the mean. (1 Sigma)
* 95% fall within two standard deviations. (2 Sigma)
* 99.7% fall within three standard deviations. (3 Sigma)

Any points lying after 3 sigma is outliers.



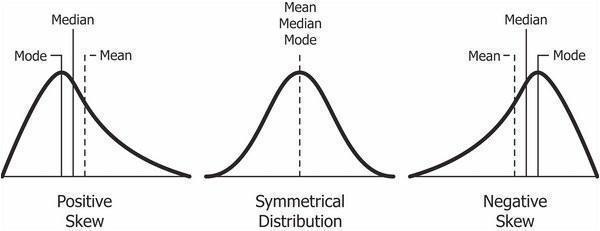
Distortion in Normal Distribution:

The distortion in normally distributed curves can be quantified in 2 ways

1. Skewness
2. Kurtosis

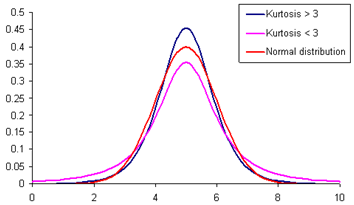
Skeweness in Normal Distribution

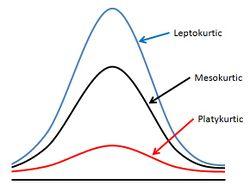
Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution



Kurtosis in Normal Distribution

In probability theory and statistics, kurtosis is a measure of the “peakedness" of the probability distribution of a real-valued random variable.





How much Skewness & Kurtosis

* If the skewness is between -0.5 and 0.5, the data are fairly symmetrical. If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data is moderately skewed.
* If the skewness is greater than 1or less than -1, the data is highly skewed.
* A standard normal distribution has kurtosis of 3 and is recognized as mesokurtic. An increased kurtosis (>3) can be visualized as a thin “bell” with a high peak whereas a decreased kurtosis corresponds to a broadening of the peak and “thickening” of the tails.

Which is best – Mean / Median / Mode?

* When you have a symmetrical distribution for continuous data, the mean, median, and mode are equal. In this case, analysts tend to use the mean because it includes all of the data in the calculations. However, if you have a skewed distribution, the median is often the best measure of central tendency.
* When you have ordinal ,categorical,count(discrete), the median or mode is usually the best choice. For categorical data, you have to use the mode.

Correcting the distortion in Normal Distribution

Transformation is nothing but taking a mathematical function and applying it to the data.

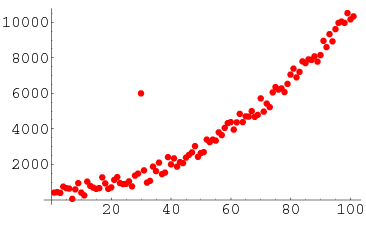
1. Log Transformation [Each data point is replaced with log(x) to obtain ND]
2. Square-Root Transformation [Each data point is replaced by its square root]
3. Reciprocal Transformation [ It takes the inverse of x ie., 1/x]
4. Box-Cox Transformation [Transformation of non-normal dependent variables to normal shape]

**Reason:** To transform the data to either reduce the skewness or to normalize the data or simply making the data easier to understand.

**Outliers:**

An outlier is an observation point that is distant 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 data set.

<https://tribe.datamites.com/posts/outliers>

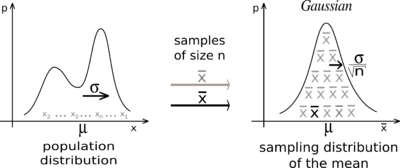


Central Limit Theorem:

The central limit theorem states that the distribution of sample means approximates a normal distribution as the sample size gets larger (assuming that all samples are identical in size), regardless of population distribution shape.

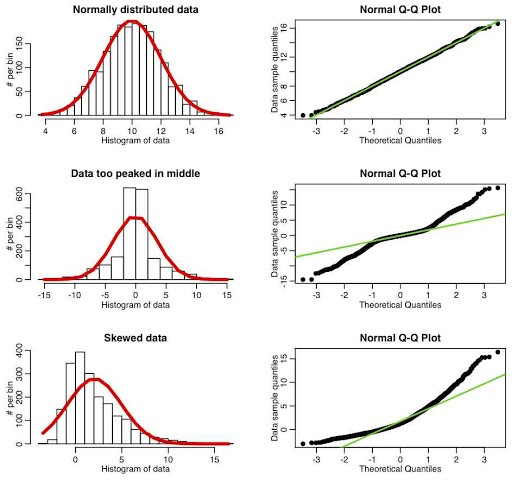
**CLT in one sentence "Even if I'm not normal, the average is normal"**

When collecting means of the samples from any distribution, the no of samples taken for calculating the mean should be greater or equal to 30.



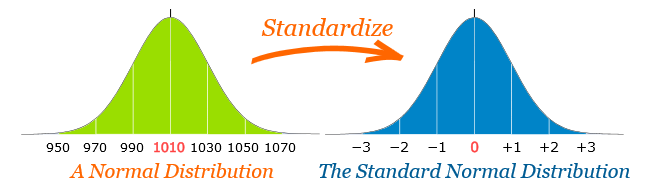
**Q-Q Plot**

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution / Normal distribution or not.

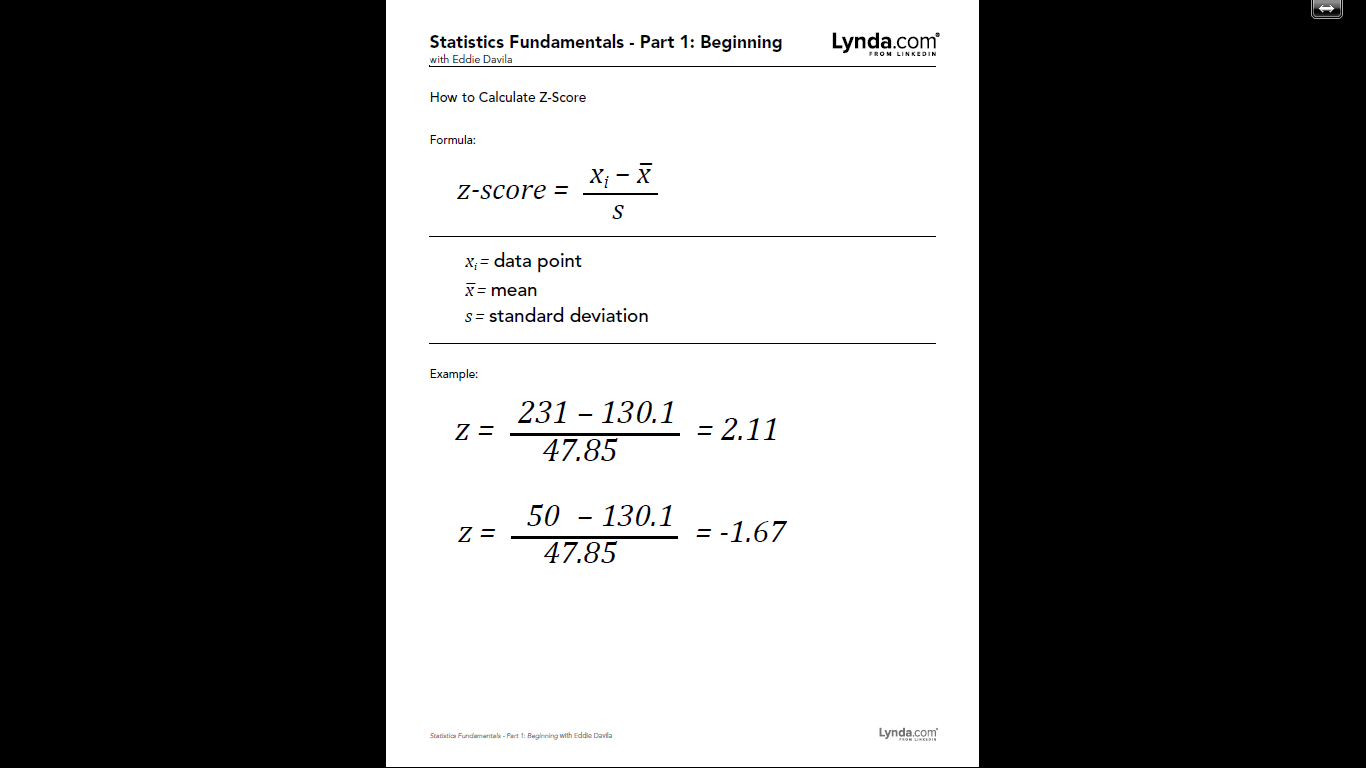
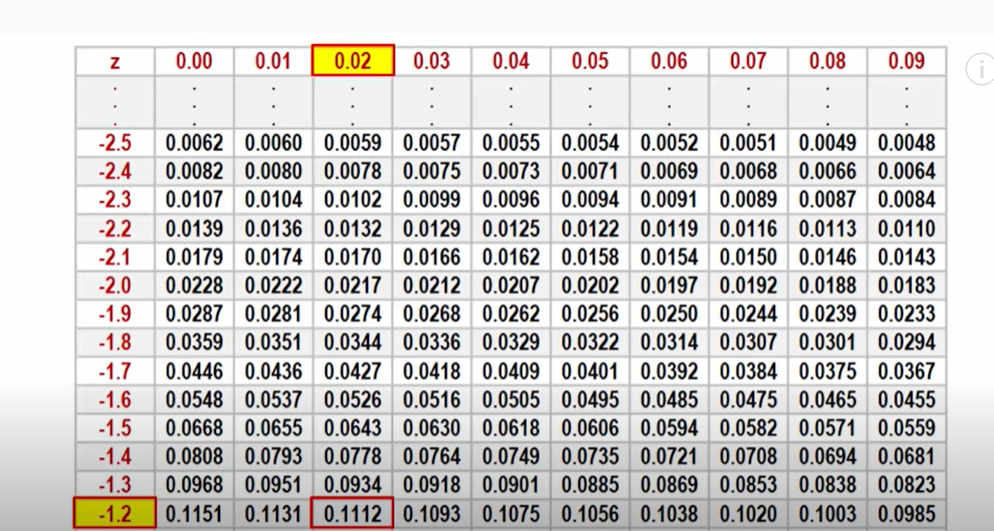
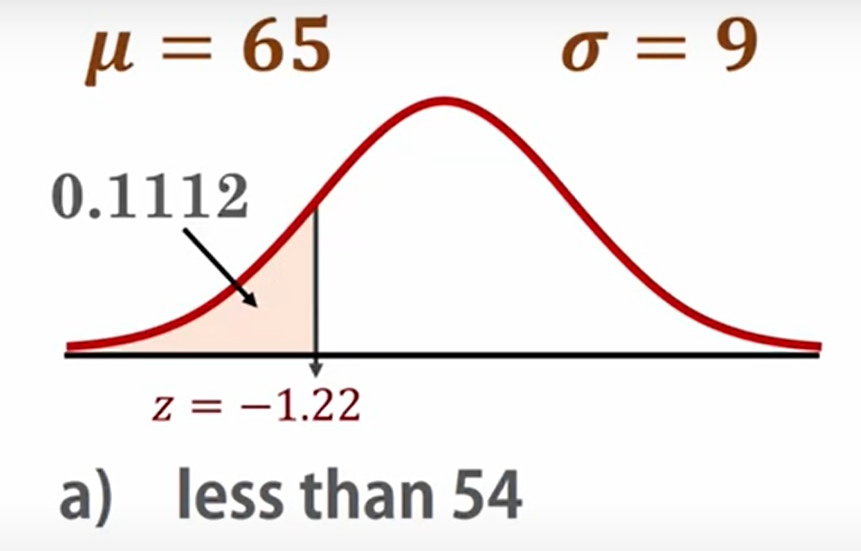
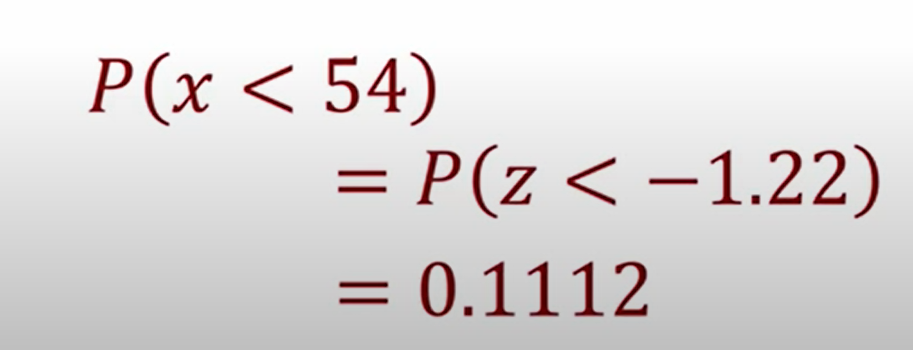


Standard Normal Distribution:

The standard normal distribution is a special case of the normal distribution. It is the distribution that occurs when a normal random variable has a mean of zero and a standard deviation of one.



Z-Score / Z-Value / Standard Score

* A z-score (aka, a standard score) indicates how many standard deviations an element is above or below from the mean. A z-score can be calculated from the following formula.
* z = (X - μ) / σ
* 
* 
* 
* 
* So on this normal curve, for z=-1.22, the area on the left is 0.1112 as we saw in the z-score table. Therefore, the probability that x < 54 is 11.12%.
* Z-score is the powerful tool to find the probability of distribution using Z-score table.

**Outliers:**

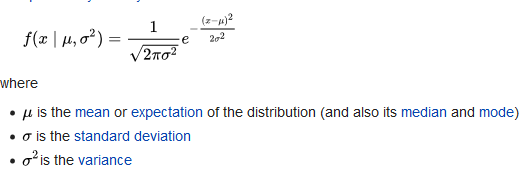
* Z-scores generally range from -3.0 to +3.0.
* - For bell shaped distributions, the empirical rule says 99.7% of all the data values have z-scores between -3.0 and +3.0.
* - We consider any z-score that is either less than -3.0 or greater than +3.0 to be an **outlier**.

**Probability Density Function (PDF):**

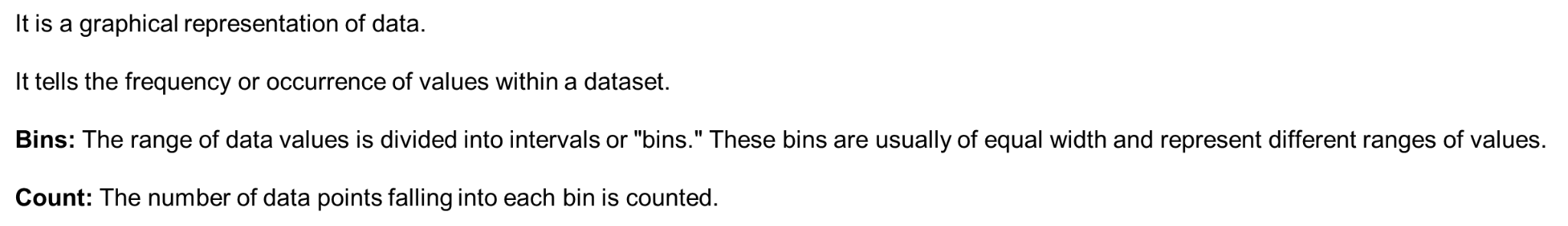
Probability density is the relationship between observations and their probability.

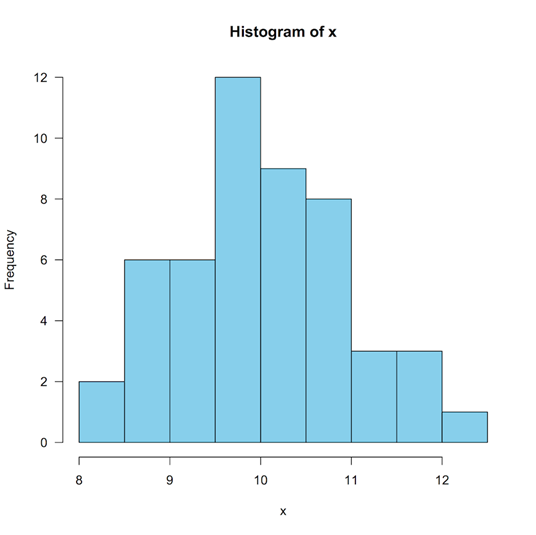
The overall shape of the probability density is referred to as a probability distribution, and the calculation of probabilities for specific outcomes of a random variable is performed by a probability density function, or PDF for short.

The probability density function for Normal distribution is given as

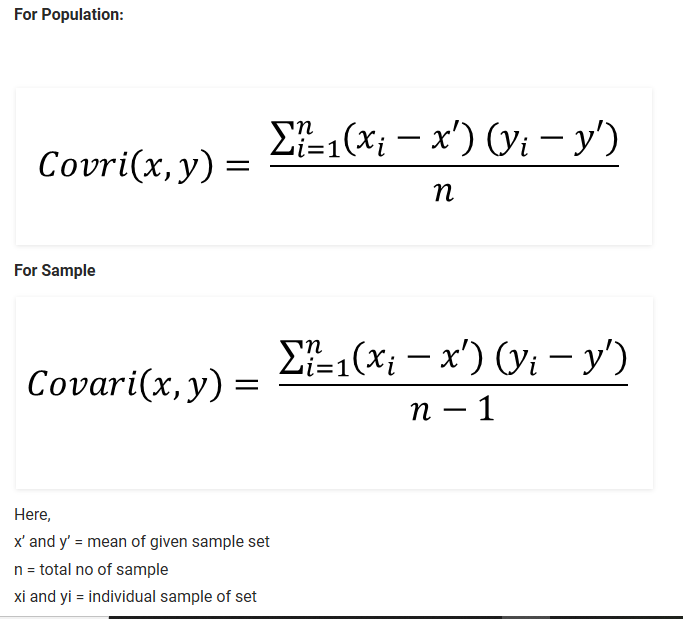


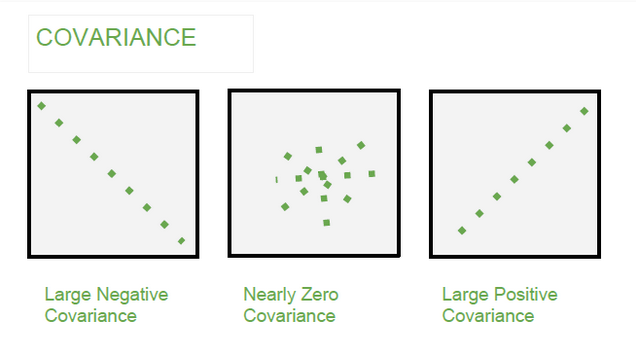
Histogram:





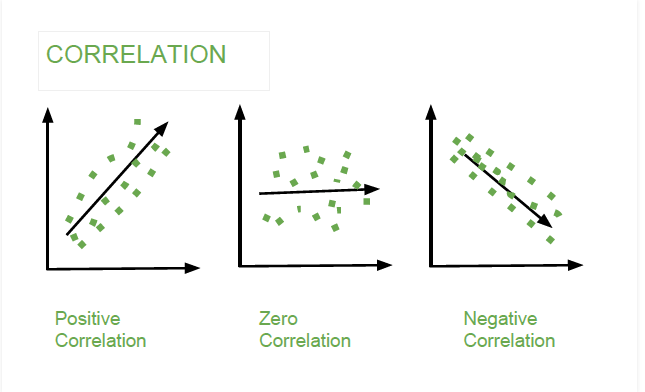
Covariance:

* It is the relationship between a pair of random variables where change in one variable causes change in another variable.
* It can take any value between -∞ to +∞, where the negative value represents the negative relationship whereas a positive value represents the positive relationship. 



Correlation:

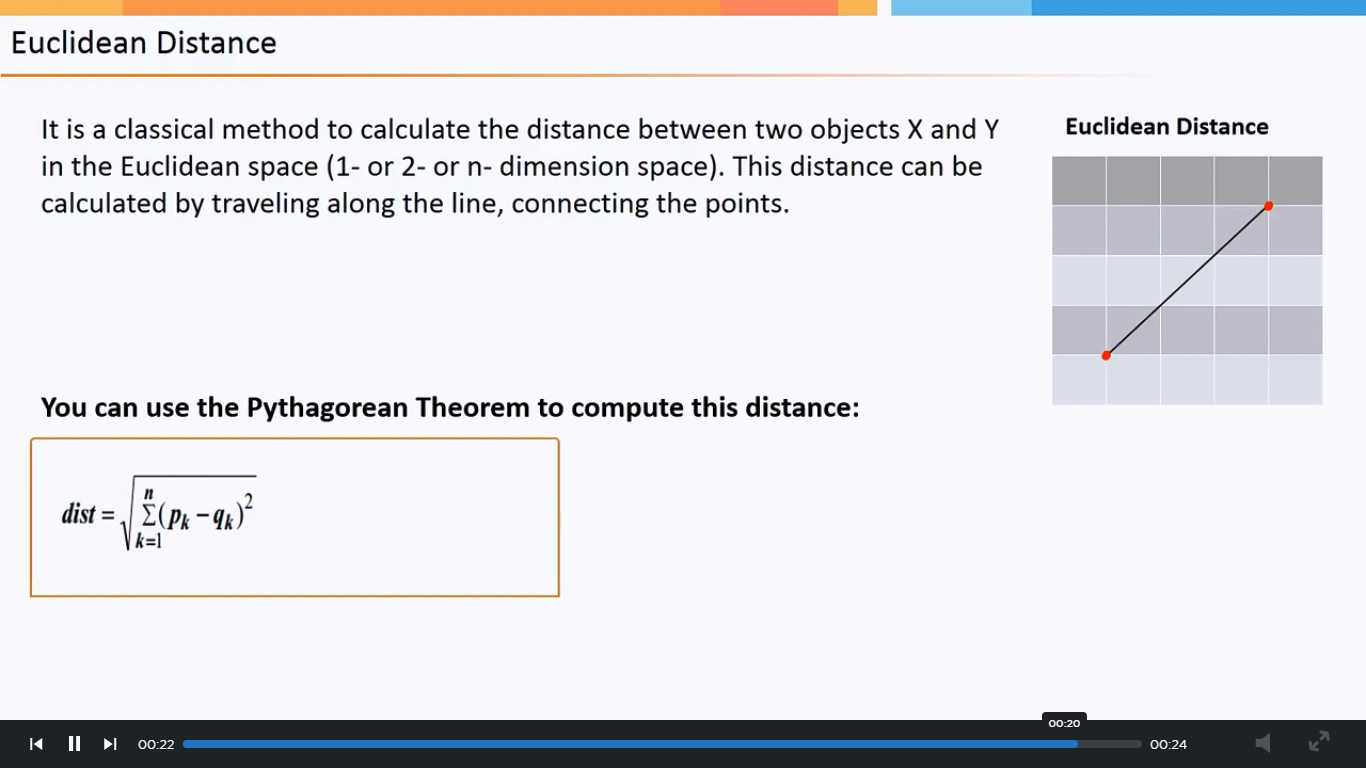
* It is the scaled version of Covariance.
* Correlation is a step ahead of covariance as it quantifies the relationship between two random variables. In simple terms, it is a unit measure of how these variables change concerning each other (normalized covariance value).
* 



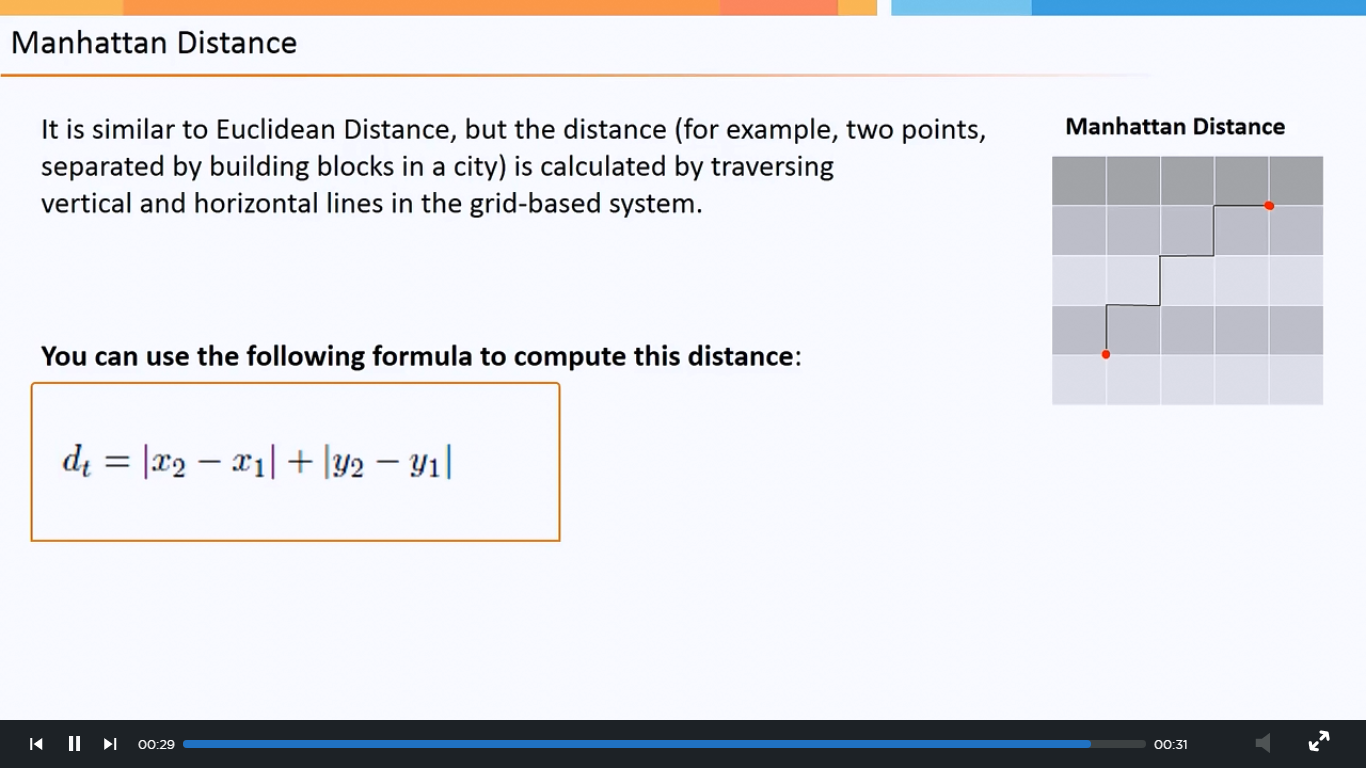
**Statistical Distance:**

* Euclidean Distance
* Manhattan Distance
* Minkowski Distance

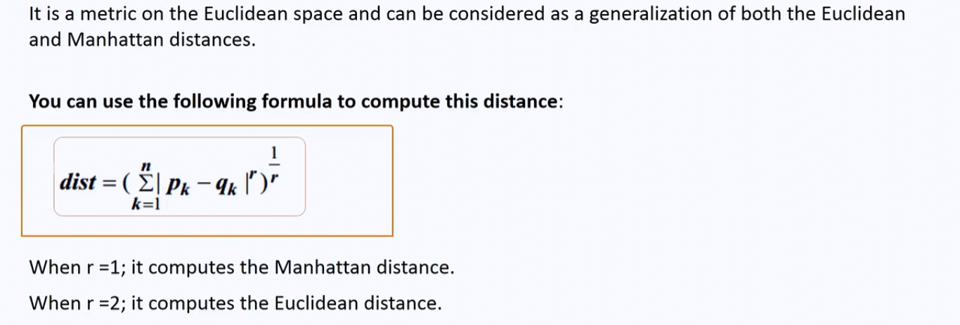
Euclidean Distance:



Manhattan Distance:



Minkowski Distance:



**Hypothesis Testing:**

Hypothesis:

* Hypothesis is a statement, assumption or claim about the value of the parameter (mean, variance, median etc.).
* A hypothesis is an educated guess about something in the world around you. It should be testable, either by experiment or observation.

Ex:-if we make a statement that “Dhoni is the best Indian Captain ever.” This is an assumption that we are making based on the average wins and loses team had under his captaincy. We can test this statement based on all the match data.

Hypothesis Testing:

Comparing and analysing the relationships

* Does the treatment with new drug help more patients than the standard treatment with old drug?
* Which of these four methods is the most efficient way of teaching machine learning?

Types of Hypothesis Testing:

* When a hypothesis specifies an exact value of parameter, it is simple hypothesis. For eg., Motor cycle company claims that a certain model gives an average mileage of 100km per litre, this is a case of simple hypothesis.
* If a hypothesis specifies a range of values then it is called a composite hypothesis. For eg., Average age of students in a class is greater than 20. This statement is a composite hypothesis.

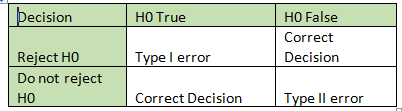
Null Hypothesis: The null hypothesis is the hypothesis to be tested for possible rejection under the assumption that it is true. The concept of the null is similar to innocent until proven guilty.

Alternate Hypothesis: The alternative hypothesis complements the Null hypothesis. It is opposite of the null hypothesis such that both Alternate and null hypothesis together cover all the possible values of the population parameter.

Hypothesis Testing – Case Discussion

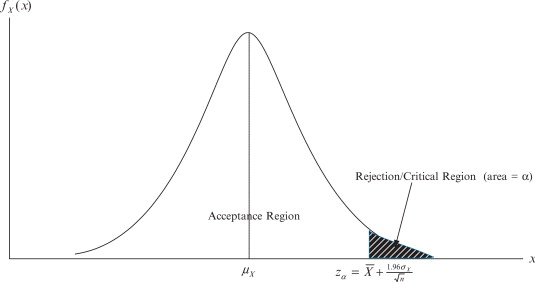
* Consider a court of law; the null hypothesis is that the defendant is innocent
* We require evidence to reject the null hypothesis (convict)
* When we collect evidence and try to reject null hypothesis, there are 2 errors that could potentially occur: Type 1 and Type 2 errors.

Type I Error and Type II Error



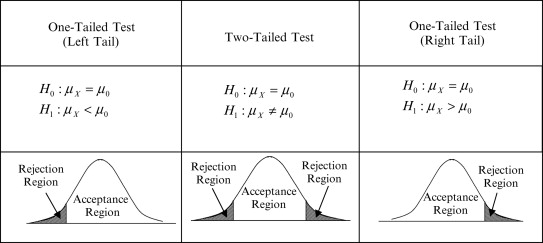
Critical Region:

* The critical region is that region in the sample space in which if the calculated value lies then we reject the null hypothesis.
* The critical region lies in one tail or two tails on the probability distribution curve according to the alternative hypothesis.
* The value of critical region is denoted by α.
* It is known as level of significance. i.e what is passing criteria of test.



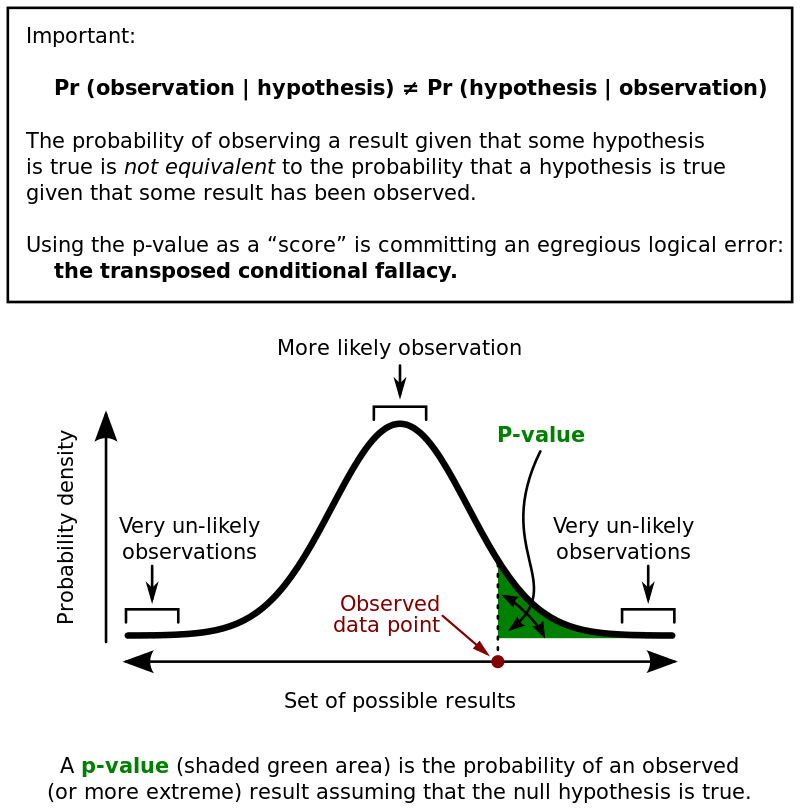
Three Cases of Critical Region Arise:

* If the alternate hypothesis gives the alternate in both directions (less than and greater than) of the value of the parameter specified in null hypothesis, it is called Two tailed test.
* Here according to H1, mean can be greater than or less than 100. This is an example of Two tailed test.
* e.g. if H0: mean= 100 H1: mean not equal to 100
* If the alternate hypothesis gives the alternate in only one direction (either less than or greater than) of the value of the parameter specified in null hypothesis, it is called One tailed test.
* Similarly, if H0: mean>=100 then H1: mean< 100
* Here, mean is less than 100, it is called One tailed test.



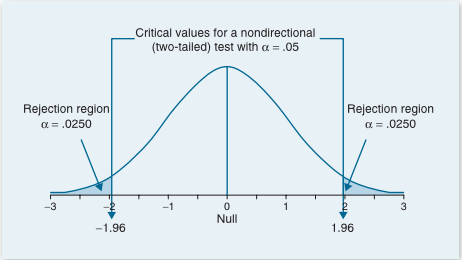
P-Value:

* Since we know that passing value of a test can be 1% or 5%.But we must also know what are the test score and this test score is known as p- value.
* Technically p-value (probability value) is the smallest level of significance at which a null hypothesis can be rejected.
* If p-value is greater than alpha, we do not reject the null hypothesis.
* If p-value is smaller than alpha, we reject the null hypothesis.



Process of Hypothesis Testing:

1. Set up Null hypothesis and Alternate hypothesis.
2. Decide the level of significance.(1% or 5%)
3. Select the test as per requirement.
4. Calculate the p-value.
5. If p-value less than level of significance, reject the null hypothesis.
6. If p-value more than level of significance, accept the null hypothesis.



Types of Hypothesis Tests:

* Parametric Tests:-Those test which considers the shape of distribution of sample.
* Non – Parametric Tests:-Those test which do not considers the shape of distribution of sample.

Parametric Hypothesis Test:

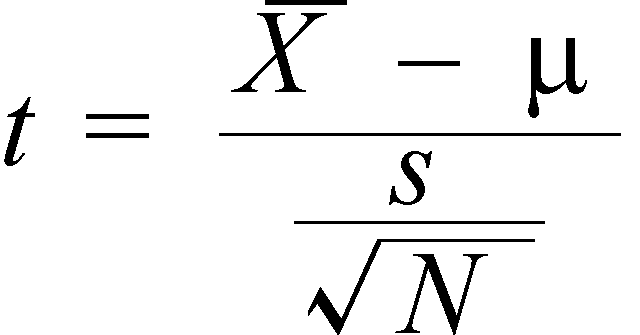
1. Z test
2. T/Student’s T test
3. Paired t Test
4. One Way ANOVA

Non-Parametric Hypothesis Test:

1. Chi Square Test
2. Mann-Whitney Test
3. Wilcoxon Signed-Rank Test
4. Kruskal-Wallis Test
5. Friedman’s ANOVA

T-Test:

* A t-test is an analysis of two populations means through the use of statistical examination; a t-test with two samples is commonly used with small sample sizes, testing the difference between the samples when the variances of two normal distributions are not known.



X -> Mean of sample set

μ -> Mean of Population

S -> Standard deviation of sample

N -> Sample size

Paired T-Test:

The paired t-test is performed when the samples typically consist of matched pairs of similar units, or when there are cases of repeated measures.

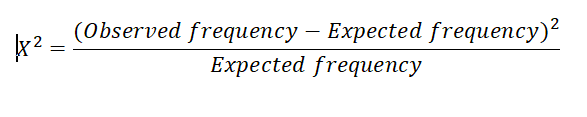
For example, there may be instances of the same patients being tested repeatedly—before and after receiving a particular treatment. In such cases, each patient is being used as a control sample against themselves.

One-Way Anova:

The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) group.

Chi-Square Test:

Chi-square test is used for categorical features in a dataset. We calculate Chi-square between each feature and the target and select the desired number of features with best Chi-square scores. It determines if the association between two categorical variables of the sample would reflect their real association in the population. Chi- square score is given by



Observed frequency = No. of observations of class

Expected frequency = No. of expected observations of class if there was no relationship between the feature and the target.