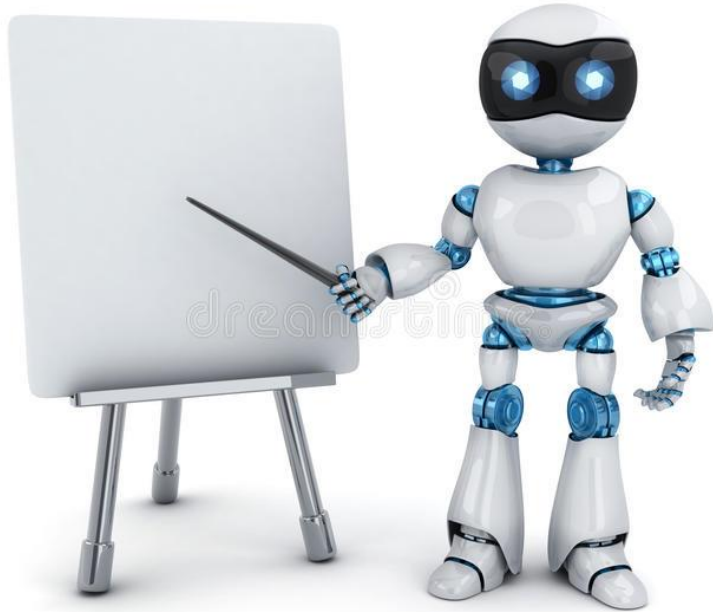




STATISTICS ESSENTIALS



Accredited by IABAC™



Overview of Statistics

Module - 1

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

- 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**.

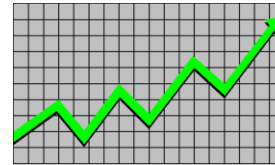
Collect data

e.g., Survey



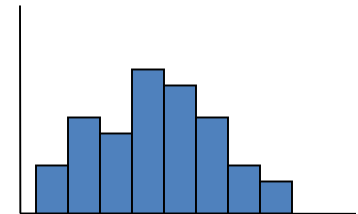
Present data

e.g., Tables and graphs



Summarize data

e.g., Sample mean = $\frac{\sum X_i}{n}$



- **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

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.

Let's first understand the basic concepts of statistics.



Statistical Population

A collection of all probable observations of a specific characteristic of interest

Example: All learners taking this course



Sample

A subset of population

Example: A group of 20 learners selected for a quiz



Variable

An item of interest that can acquire various numerical values

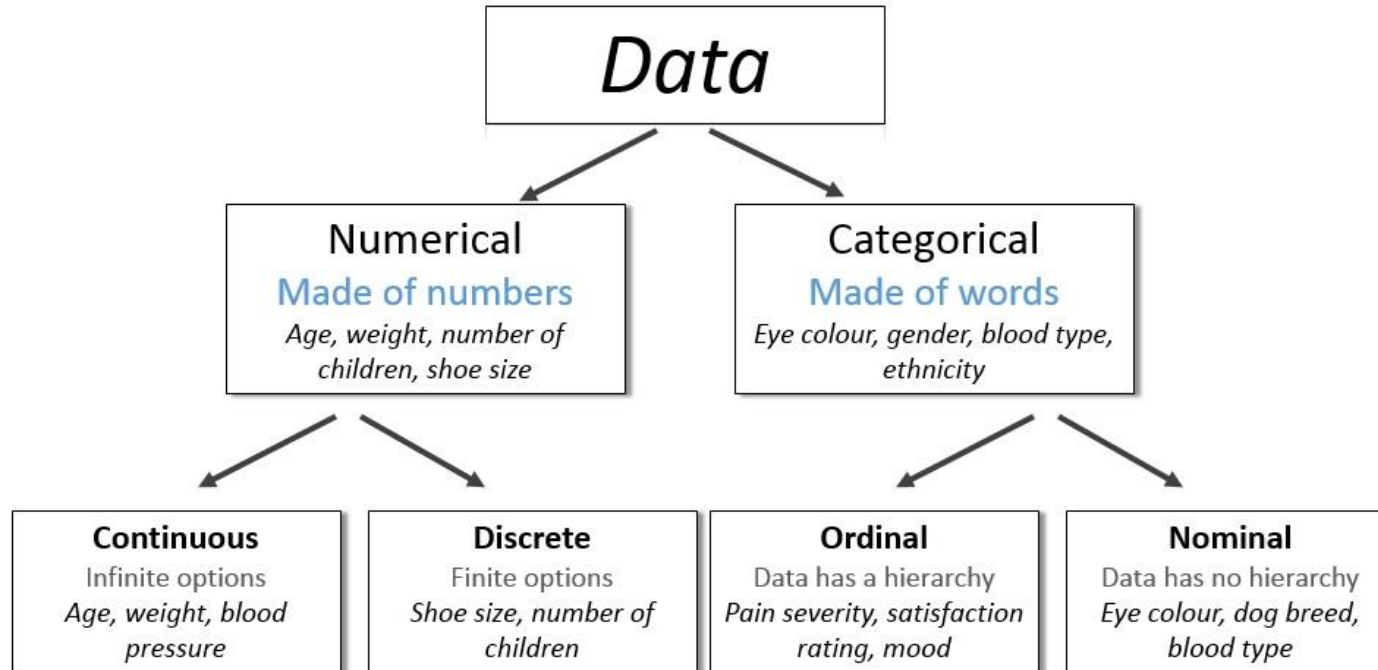
Example: The number of defective items manufactured in a factory



Parameter

A population characteristic of interest

Example: The average income of a class of people



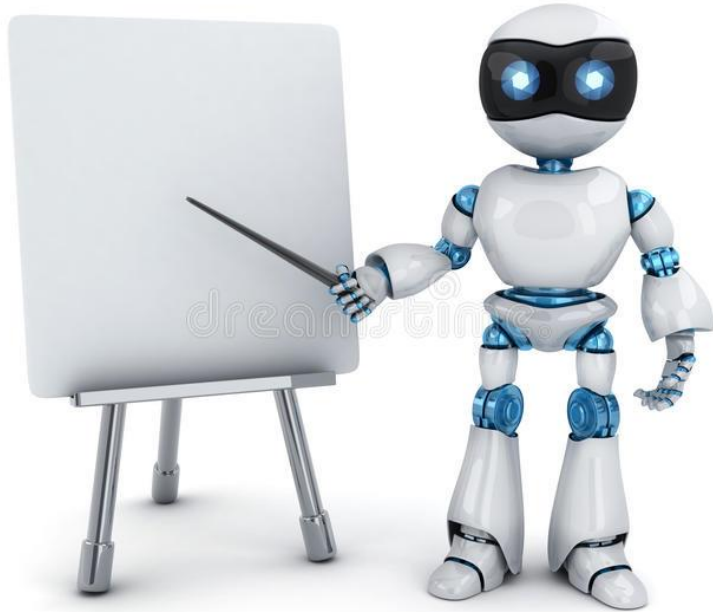
Example: Identify each of the following as examples of qualitative or numerical variables:

1. The temperature in Barrow, Alaska at 12:00 pm on any given day.
2. The make of automobile driven by each faculty member.
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 Data

Module - 2

- Collecting the data
- Presenting the data-->Visualization using Matplotlib and Seaborn.
- Summarizing the data-->Module 3

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

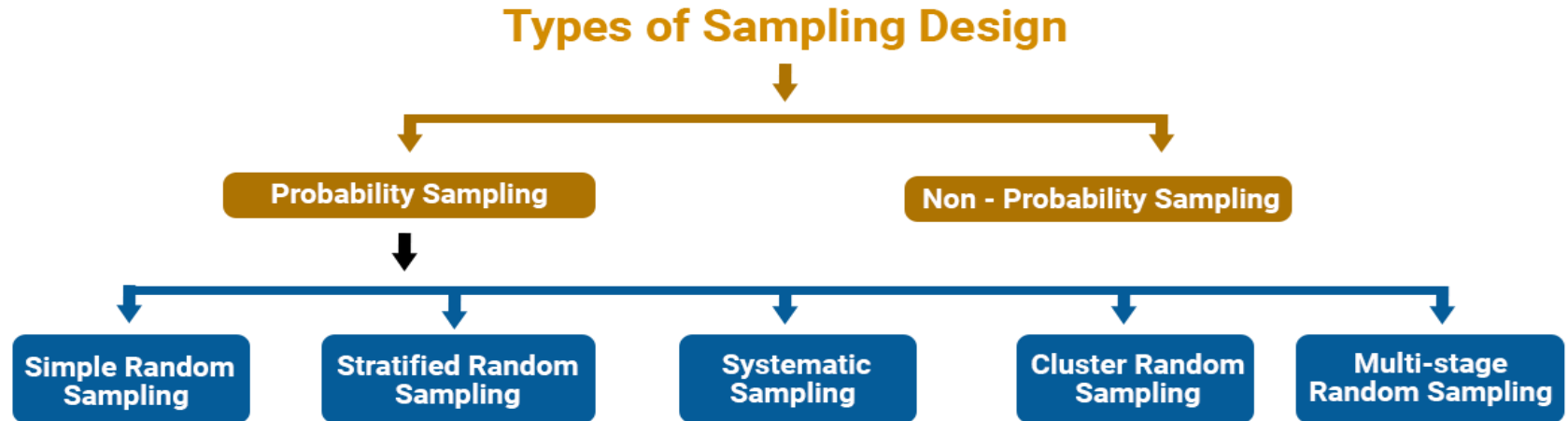
Step1:- Define the object or aim of the experiment.
i.e Estimate the average life of electronic component

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

1. **Experiment:** The investigator controls or modifies the environment and observes the effect on the variable under study.
1. **Survey:** Data are obtained by sampling some of the population of interest. The investigator does not modify the environment.
1. **Census:** A 100% survey. Every element of the population is listed. Seldom used: difficult and time-consuming to compile, and expensive.
1. **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.
1. **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.



1. **Simple Random sampling** :-each sample of the same size has an equal chance of being selected.
1. **Stratified Sampling** :-divide the population into groups called strata and then take a sample from each stratum.
1. **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.
1. **Systematic sampling** :-randomly select a starting point and take every n-th piece of data from a listing of the population.

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.

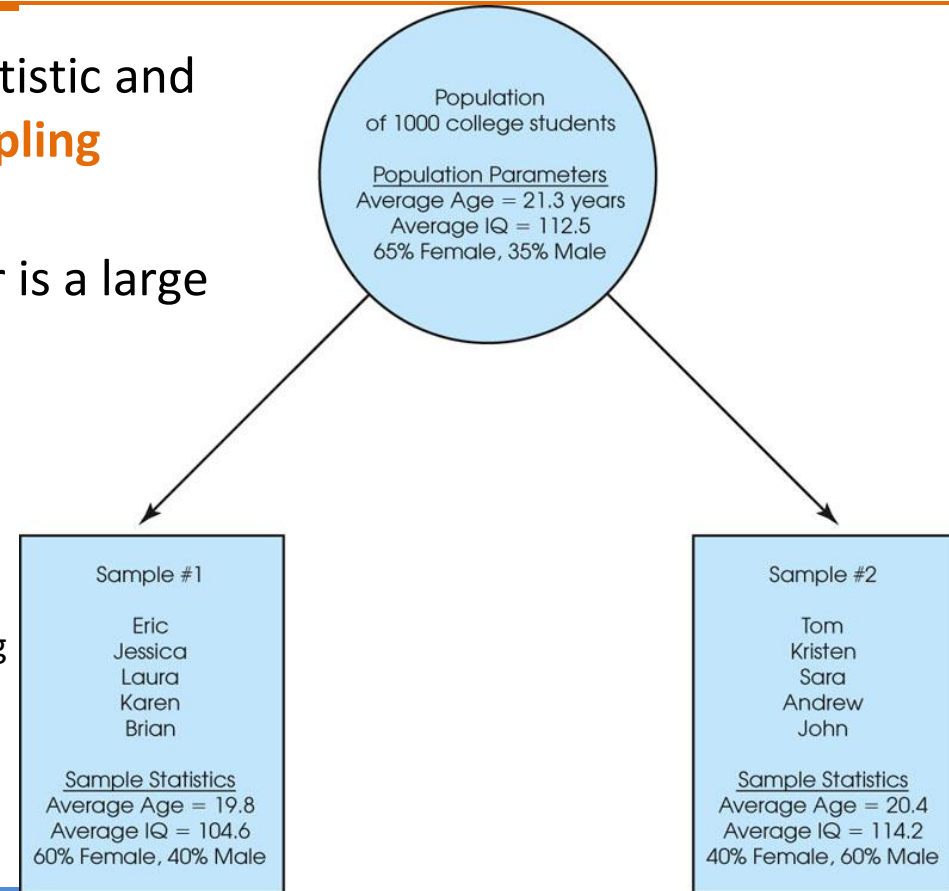
Each employee is numbered: 0001, 0002, 0003, etc. up to 2712.

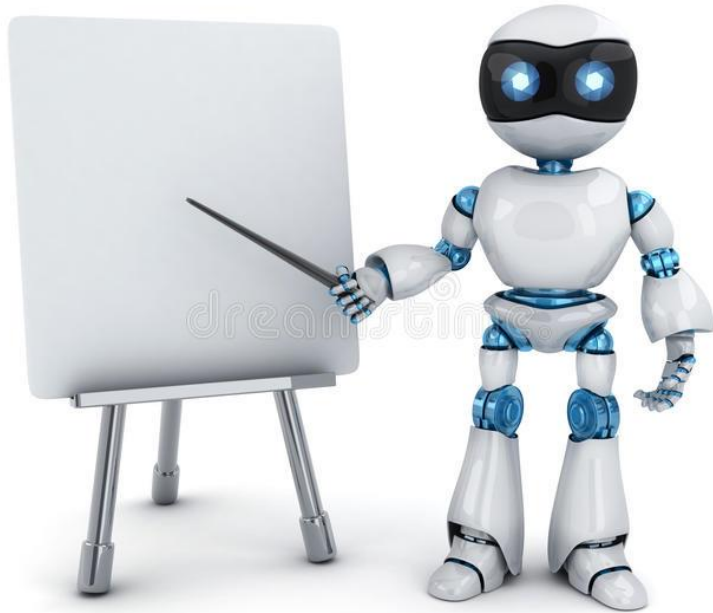
Using four-digit random numbers, a sample is identified: 1315, 0987, 1125, etc.

- 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

Module - 3

Measures of Central Tendencies

- Mean
- Median
- Mode

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

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

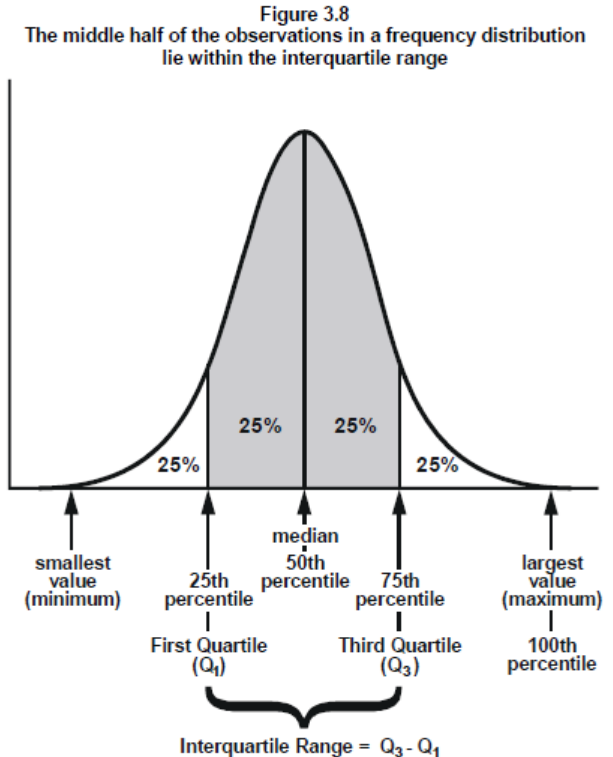
The mode is the number that occurs most often within a set of numbers.

Data Variability

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

Dataset 1	Dataset 2
20	11
21	16
22	19
25	23
26	25
29	32
33	39
34	46
38	52

The interquartile range is the middle half of the data.
Mathematically the interquartile range includes the 50% of data points that fall between Q_1 and Q_3 .



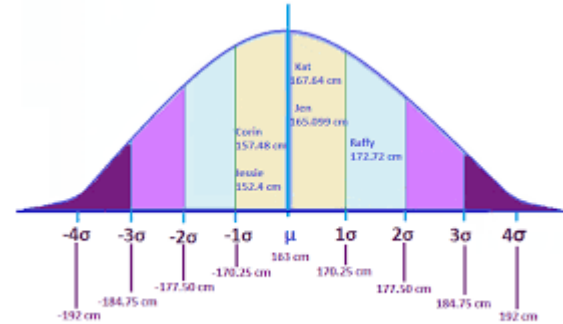
Variance is the average squared difference of the values from the mean. Unlike the previous measures of variability, the variance includes all values in the calculation by comparing each value to the mean.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

$$s^2 = \frac{\sum (X - M)^2}{N - 1}$$

Standard Deviation (σ)

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

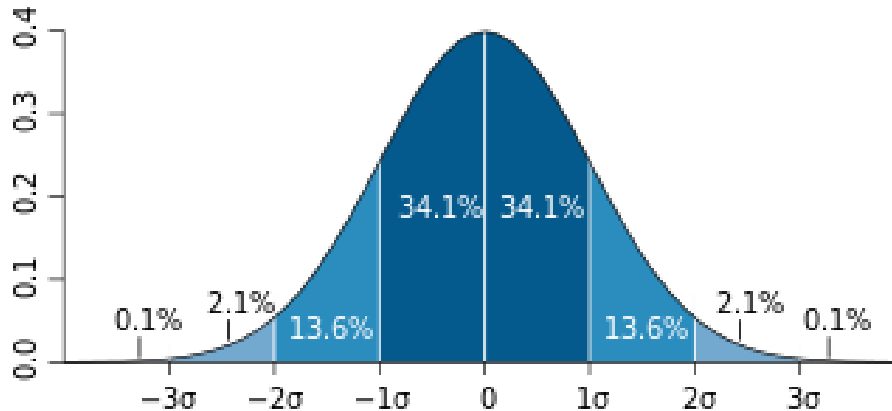


Standard Deviation (σ)

Standard Deviation Formula

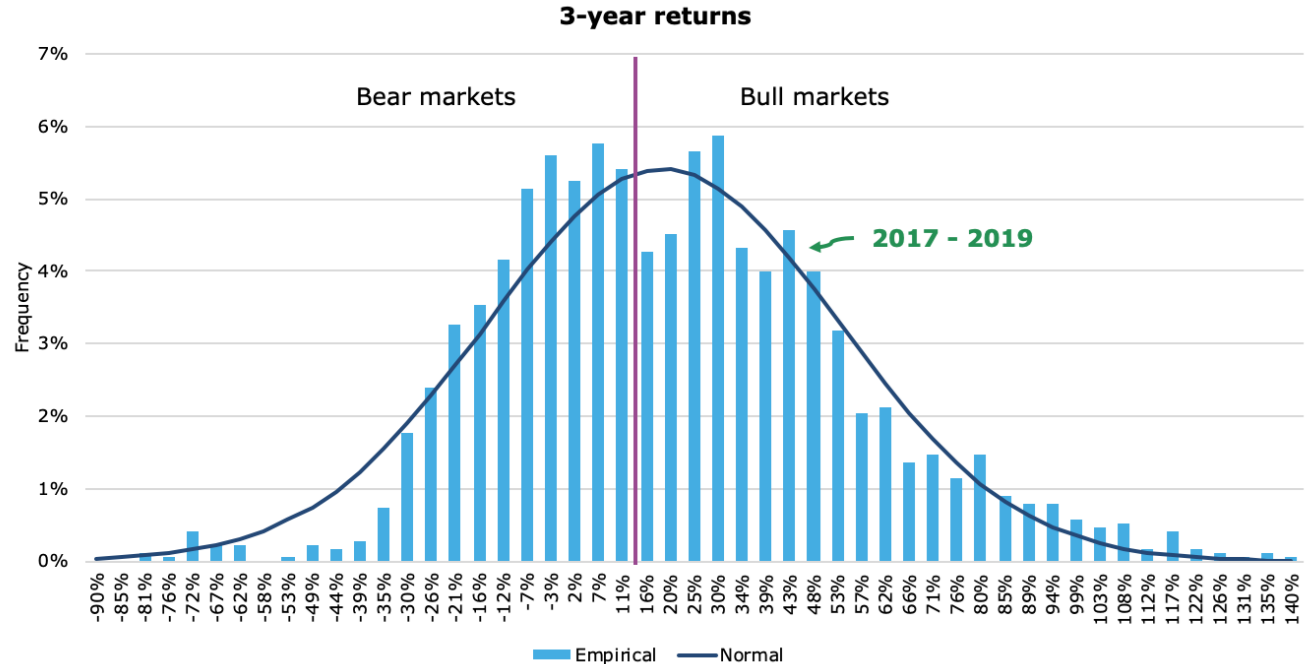
Population	Sample
$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$ <p>X - The Value in the data distribution μ - The population Mean N - Total Number of Observations</p>	$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n - 1}}$ <p>X - The Value in the data distribution \bar{x} - The Sample Mean n - Total Number of Observations</p>

- 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.



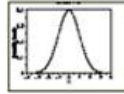
Python Implementation

The graphical representation of all observations is known as distribution

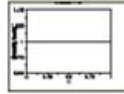


Types Of Distributions

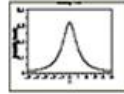
Continuous Distribution



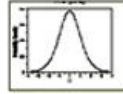
Normal Distribution



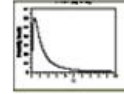
Uniform Distribution



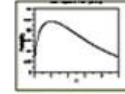
Cauchy Distribution



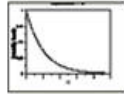
T Distribution



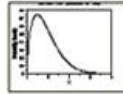
F Distribution



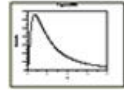
Chi-Square Distribution



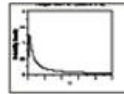
Exponential Distribution



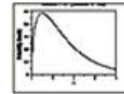
Weibull Distribution



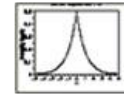
Lognormal Distribution



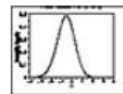
Bimbaum
Saunders
(Fatigue Life)
Distribution



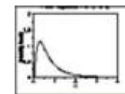
Gamma Distribution



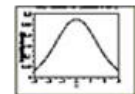
Double
Exponential
Distribution



Power Normal
Distribution

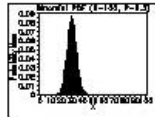


POwer
Lognormal
Distribution

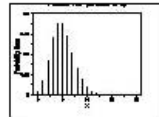


Tukey-Lambda
Distribution

Discrete Distribution



Binomial
Distribution



Poisson
Distribution

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

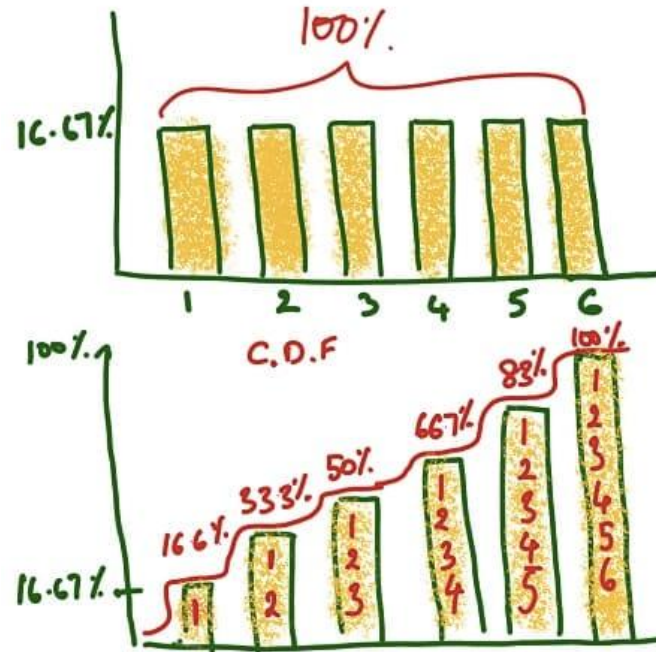
$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where

- μ is the mean or expectation of the distribution (and also its median and mode)
- σ is the standard deviation
- σ^2 is the variance

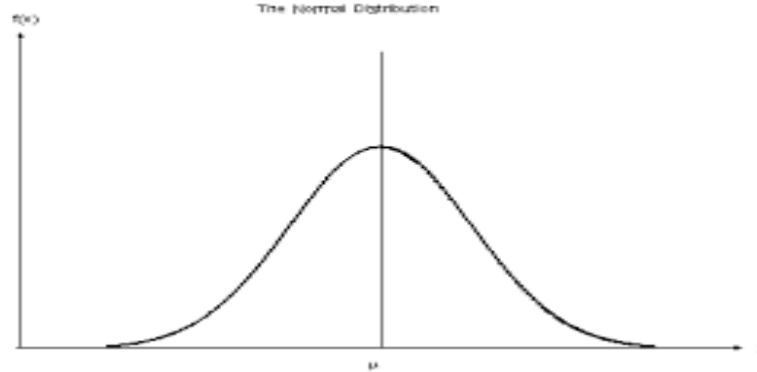
Cumulative Density Function

It provides a shortcut for calculating many probabilities at once. We integrate the **pdf** function to get the cumulative probability.



Normal Distribution

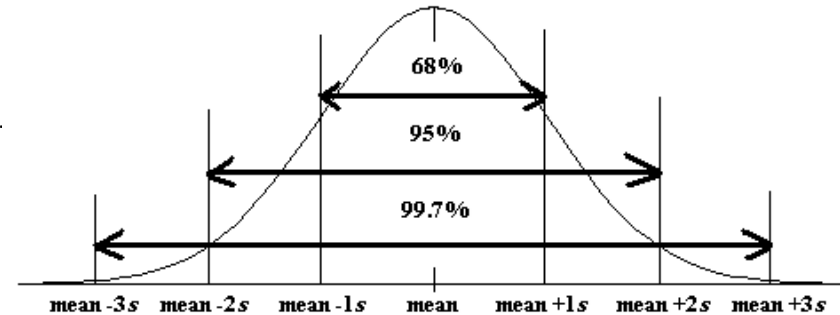
Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.



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

- 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
- 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.

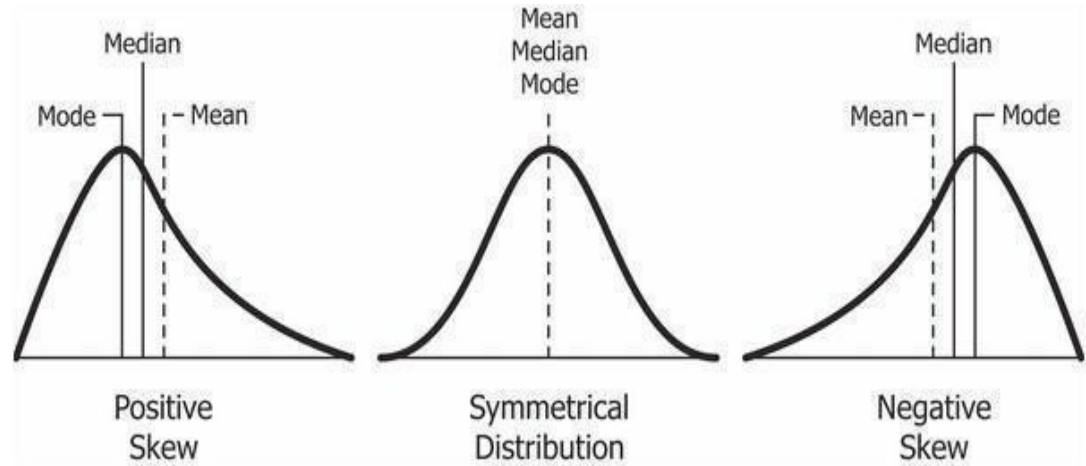


The distortion in normally distributed curves can be quantified in 2 ways

1. Skewness
2. Kurtosis

Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left

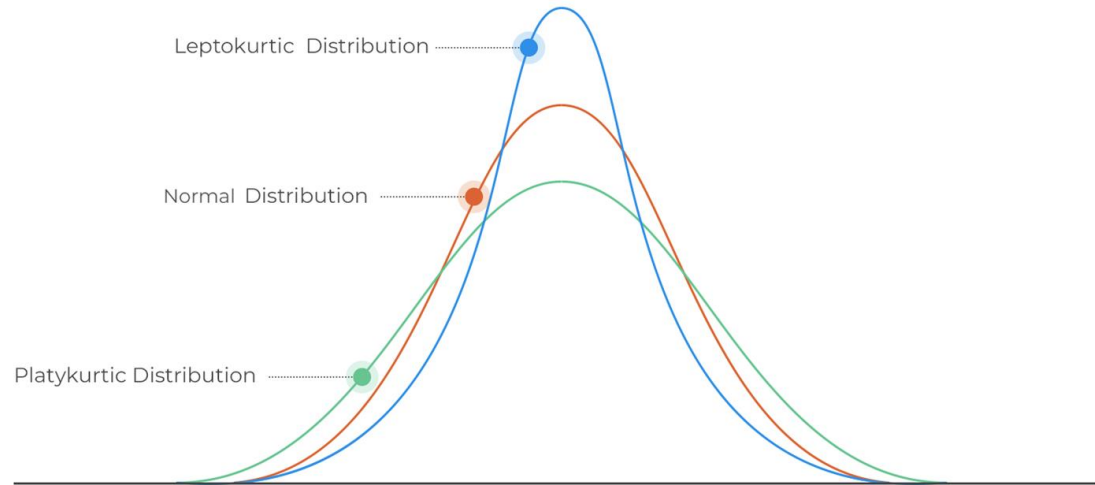
or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution



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



Kurtosis

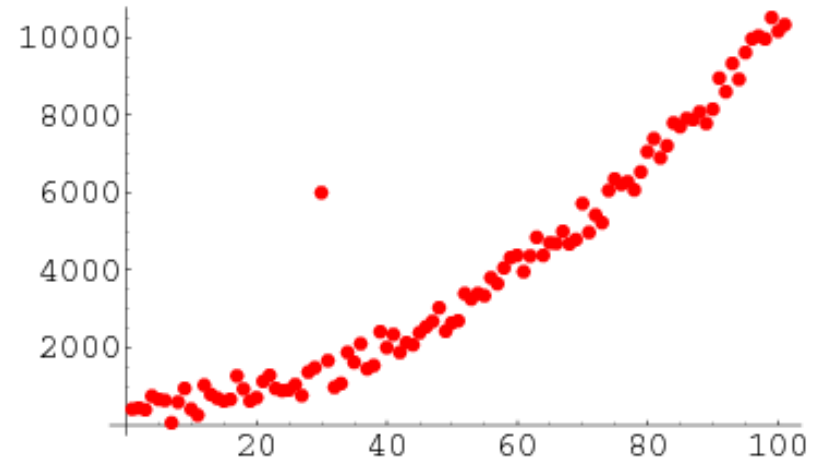


- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical. If the skewness is between -1 and -0.5 or between 0.5 and 1, the data is moderately skewed.
- If the skewness is greater than 1 or less than -1, the data is highly skewed.
- A standard normal distribution has kurtosis of 3 and 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.

- 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.

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.

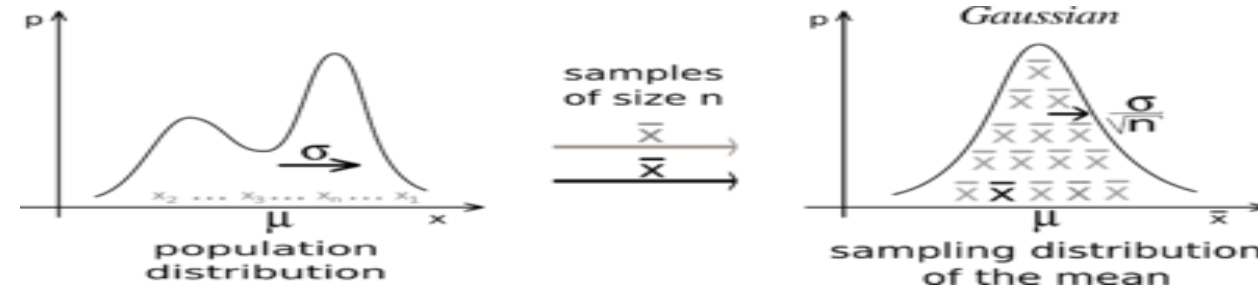
<https://tribe.datamites.com/posts/outliers>



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.

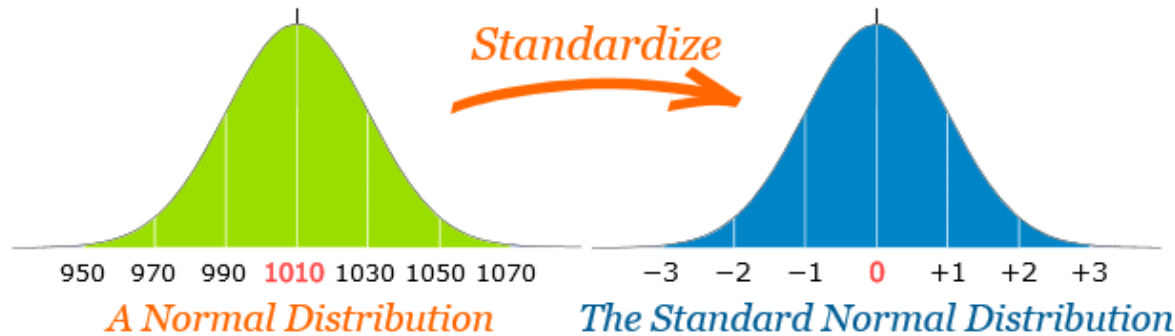
CLT in one sentence "Even if I'm not normal, the average is normal"

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.



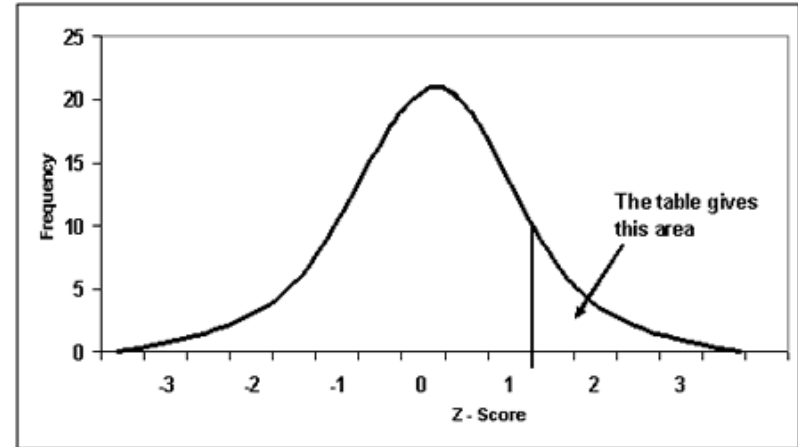
Standard Normal Distribution

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.



- A z-score (aka, a standard score) indicates how many standard deviations an element is from the mean. A z-score can be calculated from the following formula.

- $z = (X - \mu) / \sigma$



Calculating Z-Score

Formula:

$$z\text{-score} = \frac{x_i - \bar{x}}{s}$$

x_i = data point

\bar{x} = mean

s = standard deviation

Example:

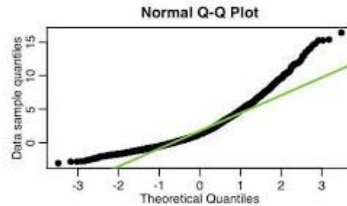
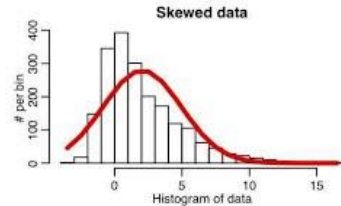
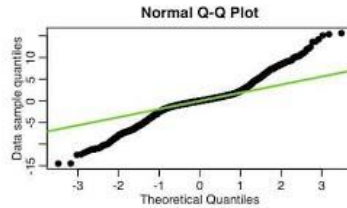
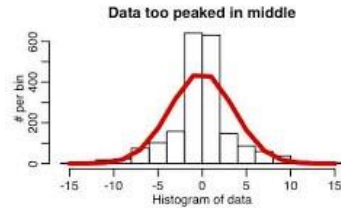
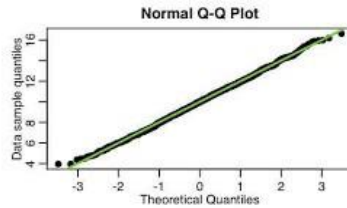
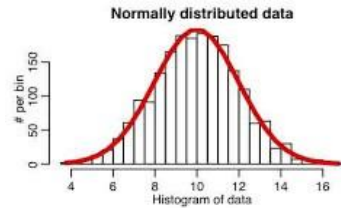
$$z = \frac{231 - 130.1}{47.85} = 2.11$$

$$z = \frac{50 - 130.1}{47.85} = -1.67$$

Numerical Question

Q-Q Plots

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution etc.



Correcting distortion In Normal Distribution

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

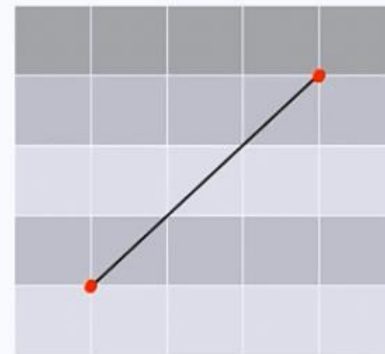
1. Log Transformation
2. Square-Root Transformation
3. Reciprocal Transformation
4. Box-Cox Transformation

- Euclidean Distance
- Manhattan Distance
- Minkowski Distance

Euclidean Distance

It is a classical method to calculate the distance between two objects X and Y in the Euclidean space (1- or 2- or n- dimension space). This distance can be calculated by traveling along the line, connecting the points.

Euclidean Distance



You can use the Pythagorean Theorem to compute this distance:

$$dist = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

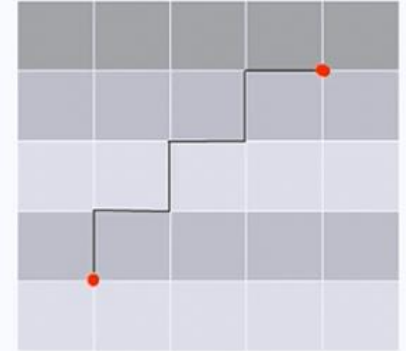
Manhattan Distance

It is similar to Euclidean Distance, but the distance (for example, two points, separated by building blocks in a city) is calculated by traversing vertical and horizontal lines in the grid-based system.

You can use the following formula to compute this distance:

$$d_t = |x_2 - x_1| + |y_2 - y_1|$$

Manhattan Distance



It is a metric on the Euclidean space and can be considered as a generalization of both the Euclidean and Manhattan distances.

You can use the following formula to compute this distance:

$$dist = \left(\sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

When $r = 1$; it computes the Manhattan distance.

When $r = 2$; it computes the Euclidean distance.

When $r = \infty$; it computes Supremum.

- 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 -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.

For Population:

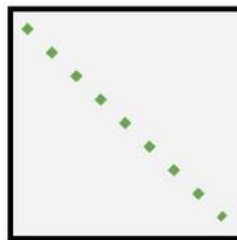
$$Covari(x, y) = \frac{\sum_{i=1}^n (x_i - x') (y_i - y')}{n}$$

For Sample

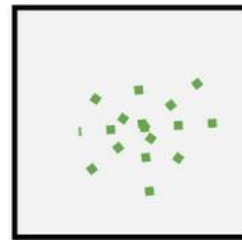
$$Covari(x, y) = \frac{\sum_{i=1}^n (x_i - x') (y_i - y')}{n - 1}$$

Here,
x' and y' = mean of given sample set
n = total no of sample
xi and yi = individual sample of set

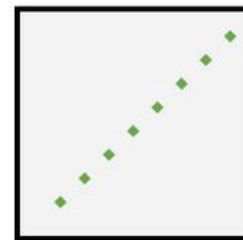
COVARIANCE



Large Negative
Covariance



Nearly Zero
Covariance



Large Positive
Covariance

- It is the scaled version of Covariance.
- This is a dimensionless metric.

Formula –

$$\text{Corr}(x, y) = \frac{\sum_{i=1}^n (x_i - x') (y_i - y')}{\sqrt{\sum_{i=1}^n (x_i - x')^2 \sum_{i=1}^n (y_i - y')^2}}$$

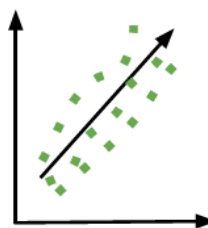
Here,

x' and y' = mean of given sample set

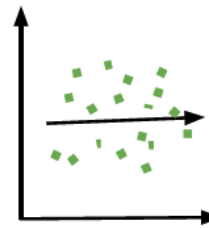
n = total no of sample

x_i and y_i = individual sample of set

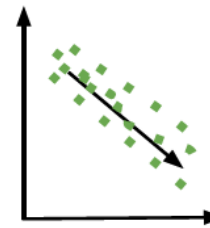
CORRELATION



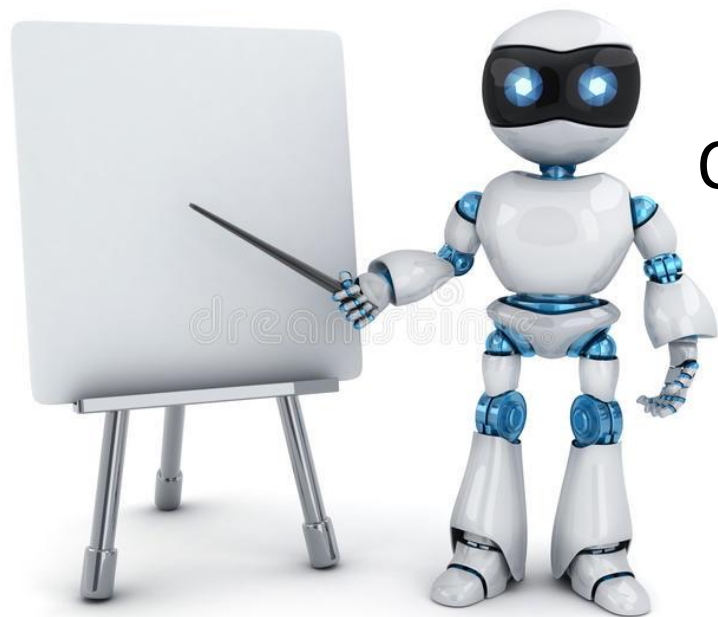
Positive
Correlation



Zero
Correlation



Negative
Correlation



Hypothesis Testing & Other computational Techniques

Module - 4

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

- 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 losses team had under his captaincy. We can test this statement based on all the match data.

- When a hypothesis specifies an exact value of the parameter, it is a simple hypothesis.

Ex:- Motorcycle company claiming that a certain model gives an average mileage of 100 km per liter; this is a case of simple hypothesis.

- If the hypothesis specifies a range of values then it is called a composite hypothesis.

Ex:-Average age of students in a class is greater than 20. This statement is a composite 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.

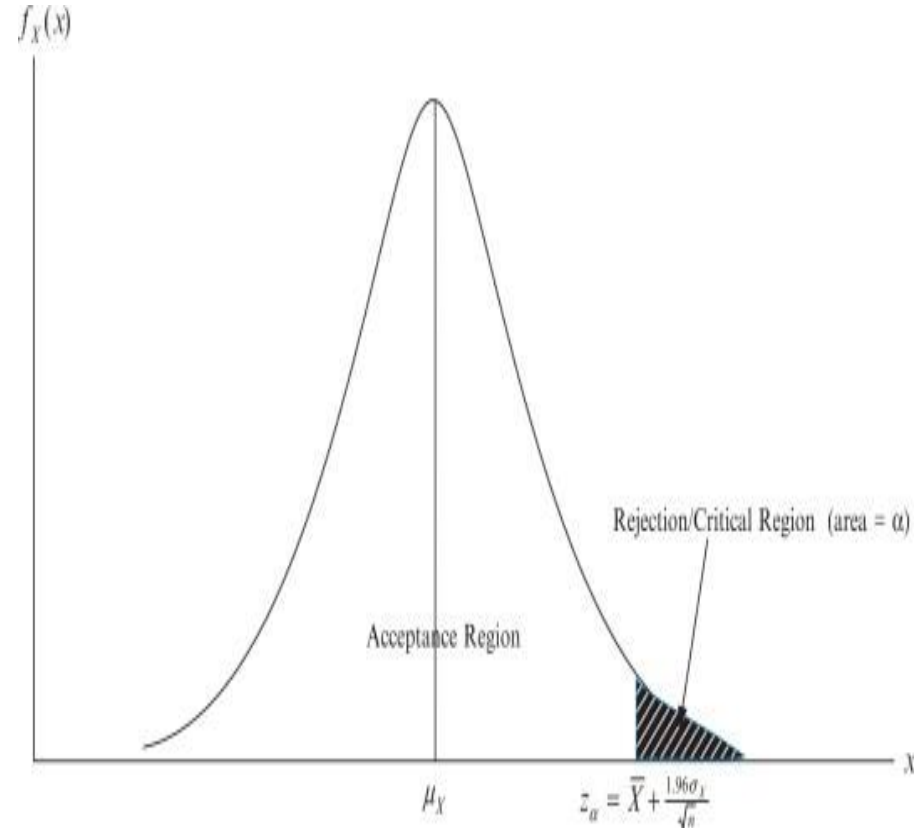
- The alternative hypothesis complements the null hypothesis.
- It is opposite of the null hypothesis such that both alternate and null hypotheses together cover all the possible values of the population parameter.

- Consider a court of law; the null hypothesis is that the defendant is innocent.
- We require evidence to reject the null hypothesis (convict).
- If we require little evidence, then we would increase the percentage of innocent people convicted (type I errors); however we would also increase the percentage of guilty people convicted (correctly rejecting the null).
- If we require a lot of evidence, then we increase the percentage of innocent people let free (correctly accepting the null) while we would also increase the percentage of guilty people let free (type II errors).

Type I and Type II Error




Decision	H0 True	H0 False
Reject H0	Type I error	Correct Decision
Do not reject H0	Correct Decision	Type II error

- 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.



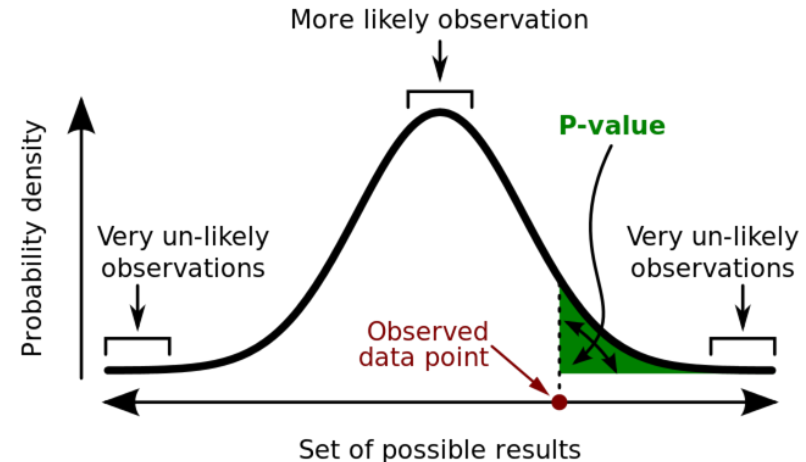
Three Cases Of Critical Region Arise

- 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 H_1 , mean can be greater than or less than 100. This is an example of Two tailed test.
- e.g. if H_0 : mean = 100 H_1 : 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 H_0 : mean ≥ 100 then H_1 : mean < 100
- Here, mean is less than 100, it is called One tailed test.

One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X < \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X \neq \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X > \mu_0$
		

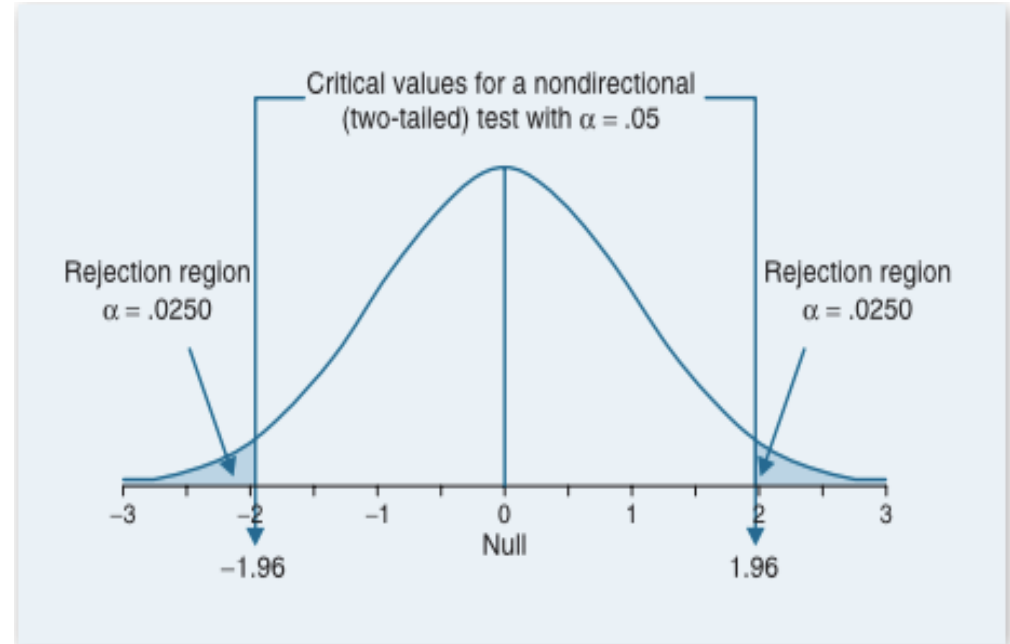
Typical level of significance may be 1% or 5%. But we must also compute the test score, which is known as p-value.

- Technically, p-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.



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

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.



- **Parametric Tests:-** Those which consider the shape of distribution of sample
- **Non-Parametric Tests:-** Those which are independent of the sample distribution

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

Hypothesis Tests:-Non Parametric

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

- 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

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

\bar{X} -> Mean of sample set

μ -> Mean of Population

s -> Standard deviation of sample

N -> Sample size

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.

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.

In ANOVA we compute a **single statistic** (an F-statistic) that compares variance **between groups** with **variance within each group**.

$$F = \frac{VAR_{between}}{VAR_{within}}$$

- The higher the **F-value** is, the less probable is the null hypothesis that the samples all come from the same population.
- We can look up the F-statistic value in a cumulative **F-distribution** (similar to the other statistics) to get the p-value.
- ANOVA tests can be much more complicated, with multiple dependent variables, hierarchies of variables, correlated measurements etc.

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

$$\chi^2 = \frac{(\text{Observed frequency} - \text{Expected frequency})^2}{\text{Expected frequency}}$$

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.

Python Implementation



Thank
You.