Statistics Essentials

# Module: 1

## Machine Learning Vs Artificial Intelligence:

**Machine learning** is the bottom-up approach where the analysis done with significance amount of training date.

**AI (Older System)** is designed to discover generic patterns and rules without having a lot of training data.

## Deep Learning:

The basics understanding of deep learning is basically how to design and architect neural network to achieve the goal.

## Graphical Model:

Graphical Models are **probabilistic model** where this model tries to find the probabilistic relation between data. This is mostly in the segment of **unsupervised learning**.

e.g.:

* Do the diagnostic for the patient by building a dependency graph by taking different report results of the patient.
* Finding “out of the ordinary” from the list of data. (Some unknown IP is trying to hack the system). From the unrelated data, need to derive certain things which were not planned or predicted before or identifying something unusual. As Graphical model will try to build the relation between all the data provided. So anything which is coming out of the scope will be identified with this algorithm.

## Reinforcement Learning:

Reinforcement Learning is an Active way of learning. In the Deep Learning, we usually prepare the dataset and train the dataset just one time. Once the training the dataset is done, we are ready for doing the prediction. In the case of Reinforcement learning, we need to train the system in active way where the machine will keep on learning with the real time information.

One of the best examples for the Reinforcement learning is “Driver Less Car”. In this example, system not only needs to learn from the existing training data, also it should get learn based on the real time data got during driving.

# Module: 2

Multiple factors which needs to act on for business problem and business understanding. This process is called “Cross Industry Standard Process for Data Mining Framework” 🡺 CRISP-DM

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Data Modelling
5. Model Evaluation
6. Model Deployment

## Business Understanding

Before applying any kind of data analysis, following understanding and analysis are important.

1. Understand the existing business model and Business Goal
2. What areas are giving more profit currently and what areas are already company invested?
3. What is the target needs to achieve
4. What kind of model can give better outcome with the given input?

## Data Understanding

* Collect Relevant data
* Describe Datasets
* Explore Data by plotting graphs
* Check data Quality

## Data Preparation

* Select Relevant Data
* Integrate Data Files
* Clean the Data
* Construct the data
* Format the data

## Data Modelling

Models should be

* Succinct
* Mathematically Sound
* Efficient
* Easy to use

Data modelling approach should be as below:

* Understand Family of the Model
* Choose appropriate algorithm from the chosen family of model

## Model Evaluation/Deployment

Modelling 🡪 Evaluation 🡪 Deployment 🡪 Modelling

It is cyclic process where evaluation and deployment of the model can be done. Based on the outcome, again remodelling can be done.

# Module: 7 (Additional Reference)

## Probabilities

<https://www.mathsisfun.com/data/probability.html>

<https://www.mathgoodies.com//lessons/vol6/addition_rules>

<https://www.mathgoodies.com//lessons/vol6/independent_events>

<https://www.calculatorsoup.com/calculators/discretemathematics/combinations.php>

## Mean and Median (The Basics)

Mean 🡪 Average of the number

Median 🡪 It is just the middle of the complete sorted number set.

Mode 🡪 It is the most repeated number in the complete number set.

When there is many out-liars, Median and Mode give the better result than Mean.

e.g. {3, 3, 3, 3, 1000} 🡺 in this data set, Mean will give a wrong value but Median and Mode will give closer values.

## Range and Mid Range

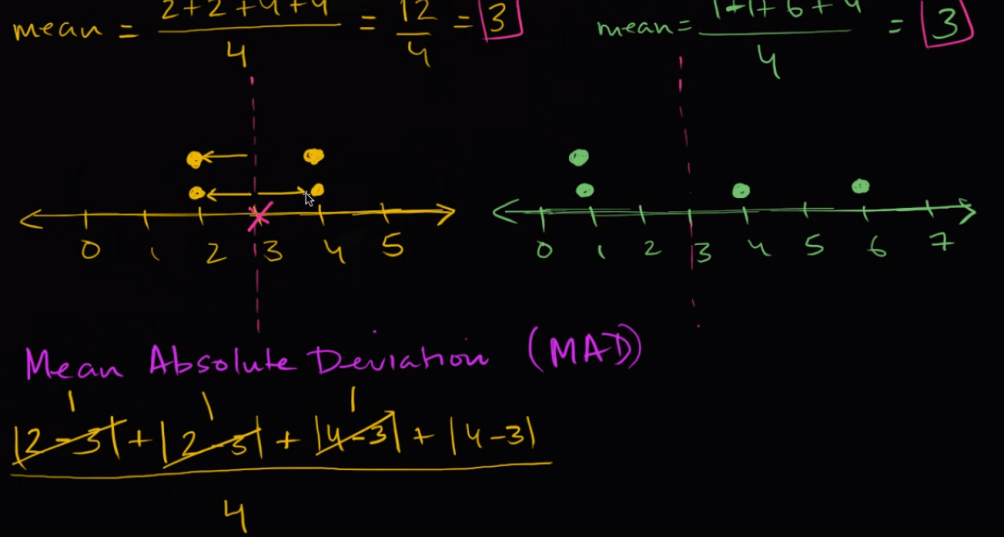


## Mean Absolute Deviation (MAD)

Average distance of each point from mean point is called Mean Absolute Deviation.

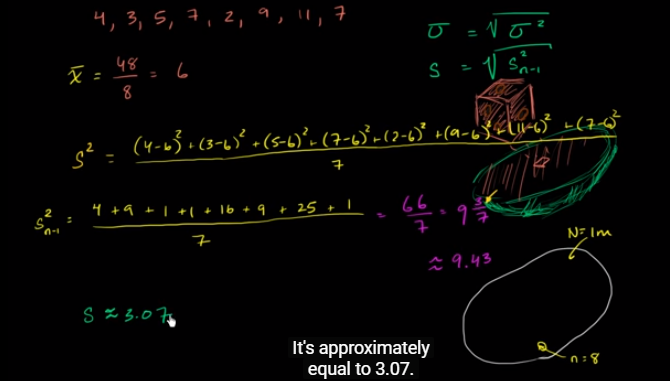
In below example, Mean of {2, 2, 4, 4} is 3. Now the MAD is

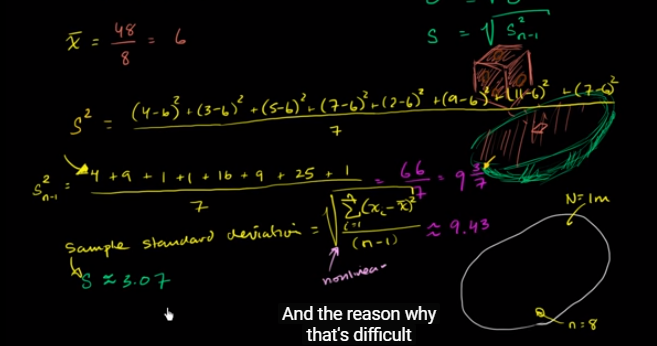
[ABS (2-3) + ABS (2-3) + ABS (4 – 3) + ABS (4 – 3)]/4

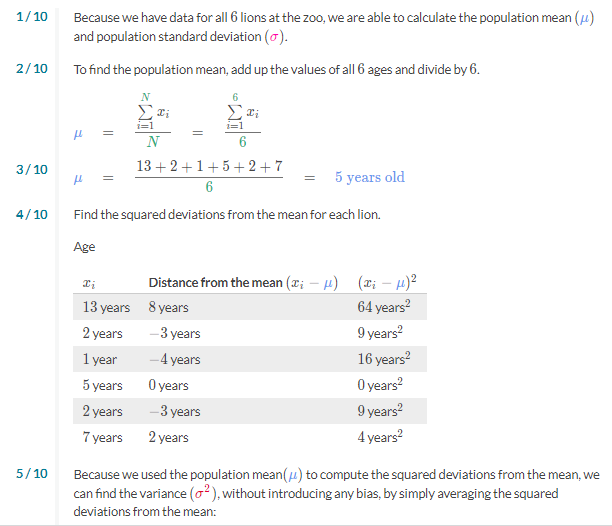


## Standard Deviation/Sample Standard Deviation

Difference between Standard deviation and Sample Standard Deviation is, Sample Standard deviation gives more appropriate result. In Sample standard deviation, in place of dividing by n, we divide it by n-1.

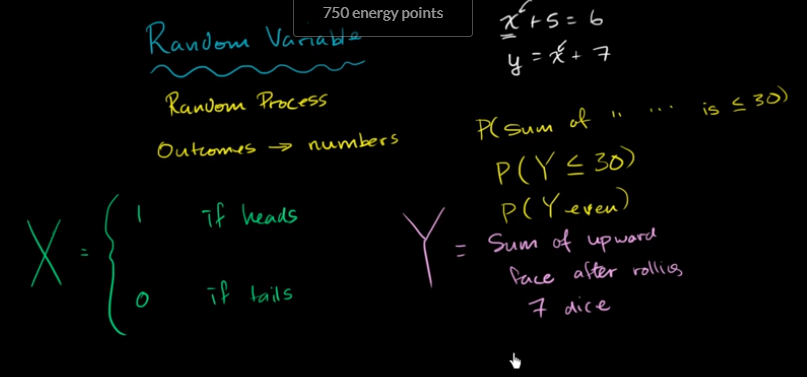






## Random Variable

Random variable is way of presenting the outcome of the probability.



Random variable is of 2 types

* Discrete 🡪 Separate Value
* Continuous 🡪 Any value in interval

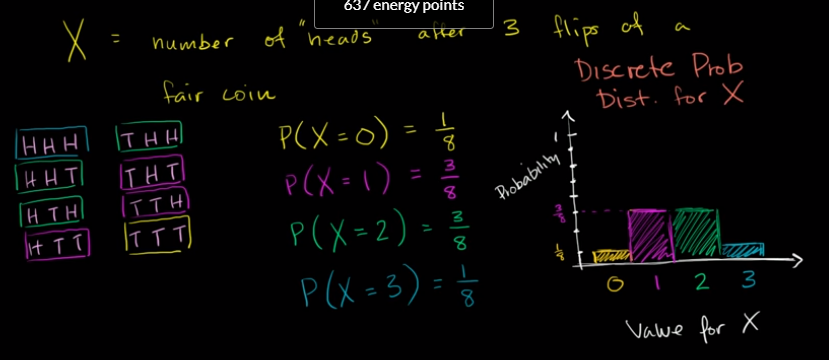
X = {1: Heads, 0: Tails 🡺 Discrete Random Variable

Y = **Exact** Mass of a Random Animal in the Zoo 🡪 Mass can vary from (~0 to ~5000 KG) 🡪 Hence this is not the Discrete Random Variable, it is continuous random variable.

**Exact** winning time in 100M 2016 Olympics 🡪 this is continuous

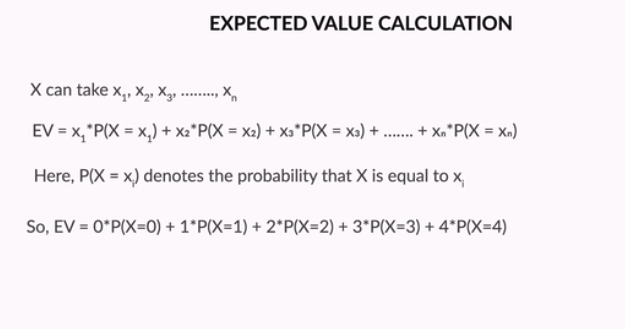
Winning time with nearest 10 in 100M 2016 Olympics 🡪 This Discrete.

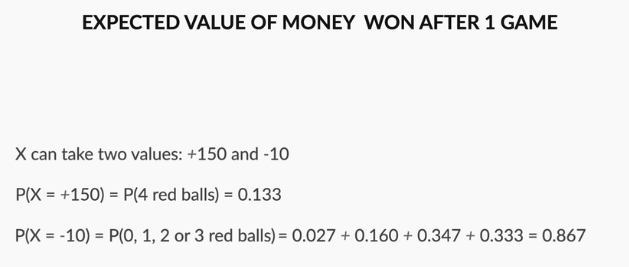
X = 0 🡺 Probability of 0 Heads X= 1 🡪 Probability of 1 Heads

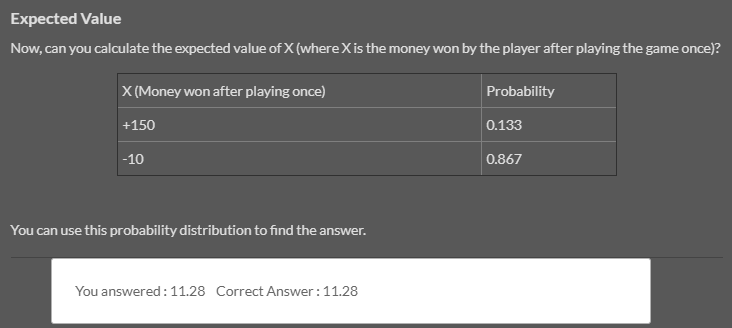


# Module: 3 (Inferential Statistics)

## Basics of Probability



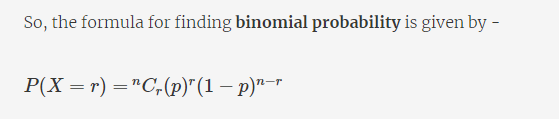


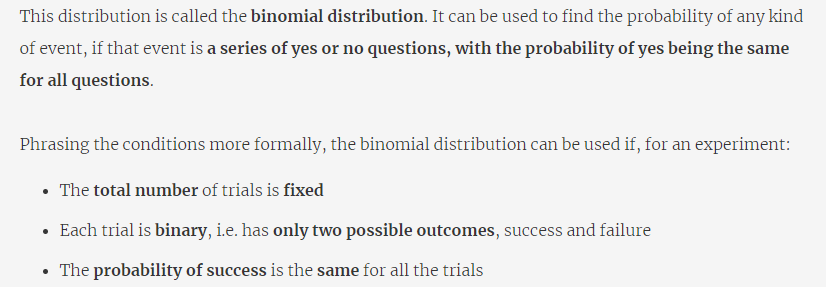


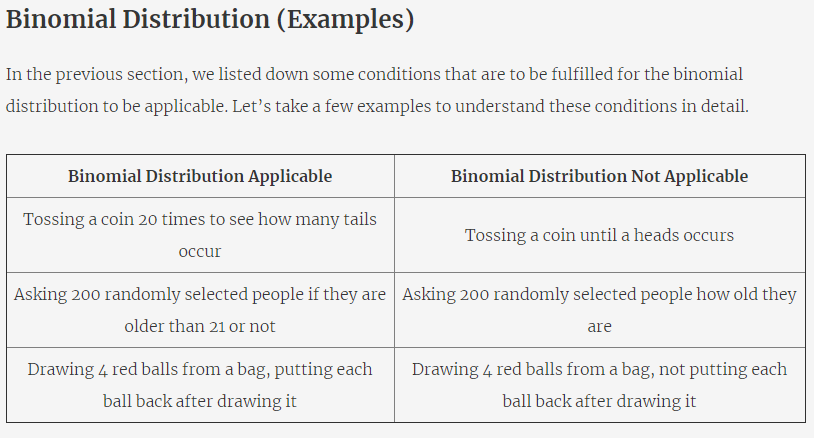


## Binomial Probability Distribution

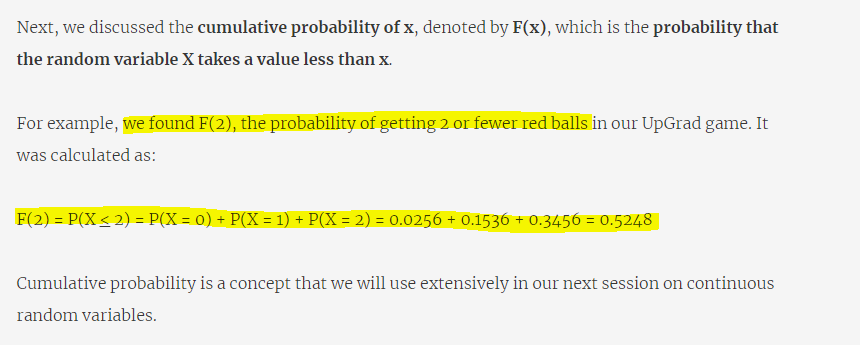
|  |
| --- |
| * Total Number of Balls n * Total Number of red balls to get is r times. |







### Cumulative Probability



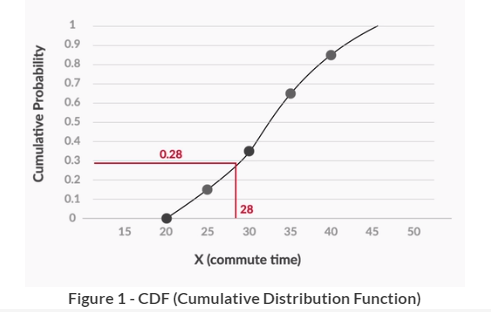
## Continuous Probability Distribution

### Probability Distribution

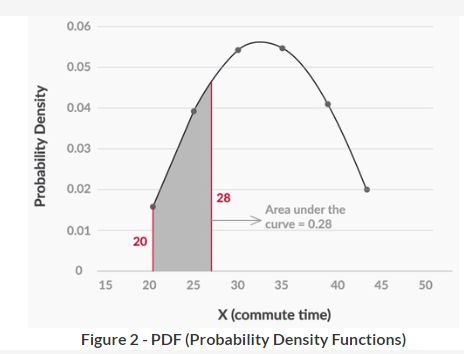
Below is the data of the Probability and cumulative probability value in the chart.



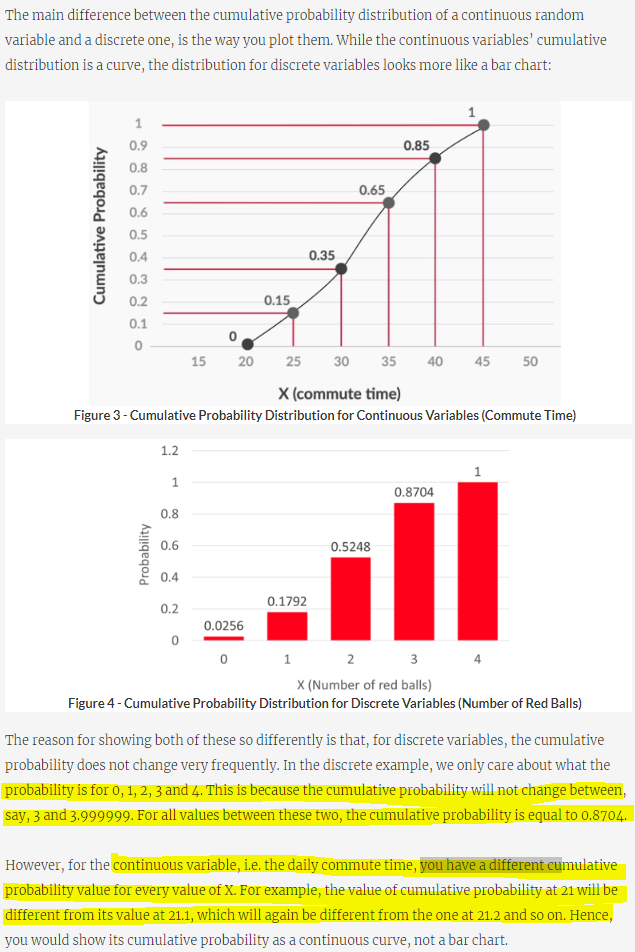
If we put the graph of “Commute Time” to “Cumulative Probability” then we will get a increasing sequence of a Graph which is called CDF (Cumulative Distribution Function)

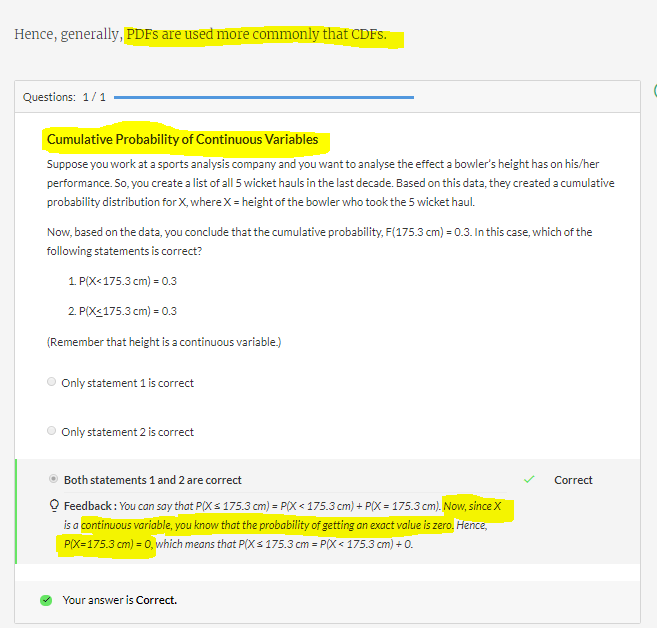


If we put the Graph of “Commute Time” to “Probability” then to get the cumulative probability by PDF (Probability Density Function)

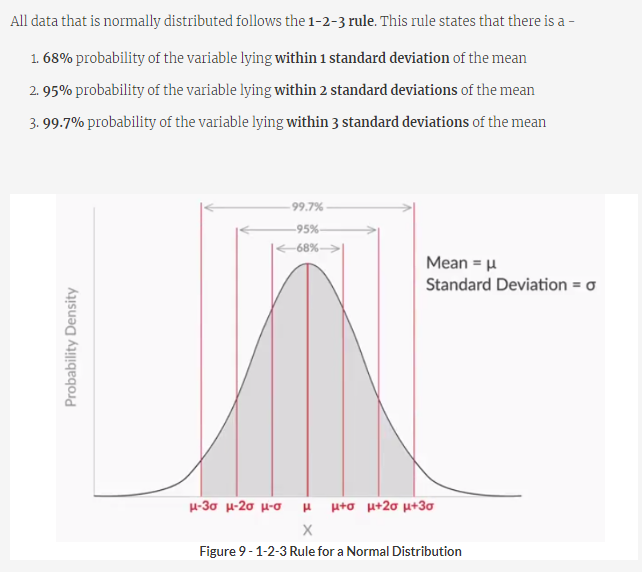


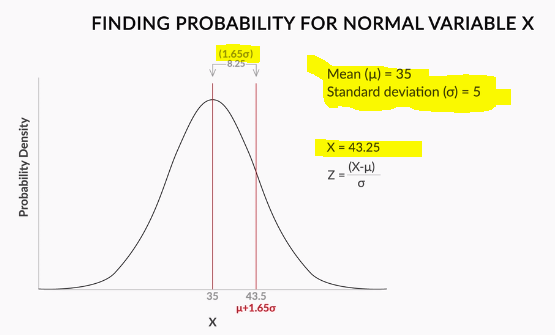
### Difference between Probability Distribution and Continuous Probability

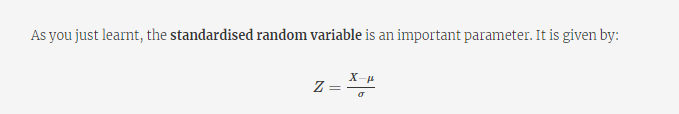


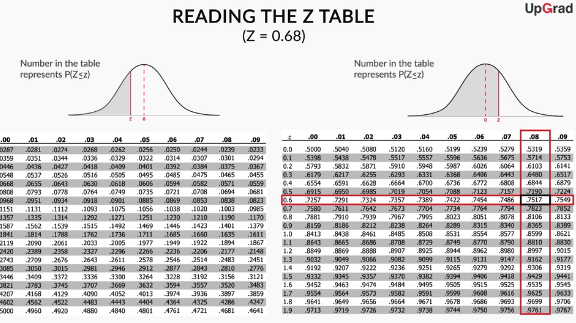


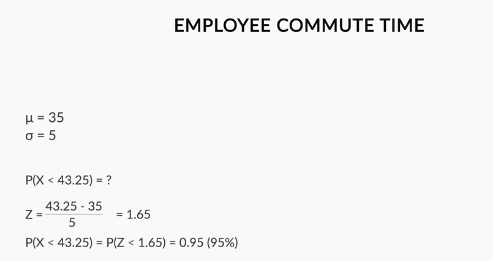
### Normal Distribution





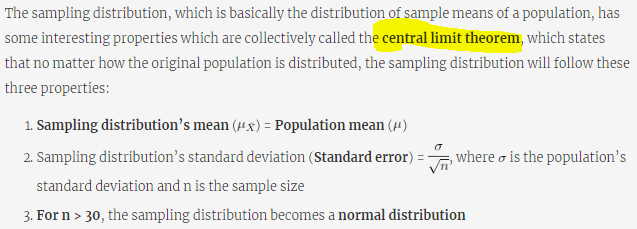


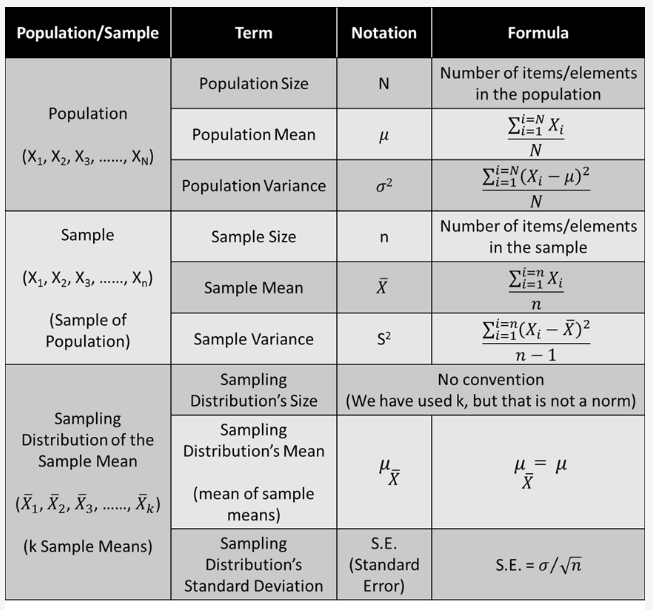




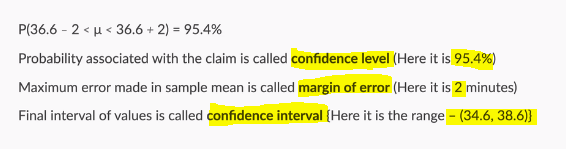


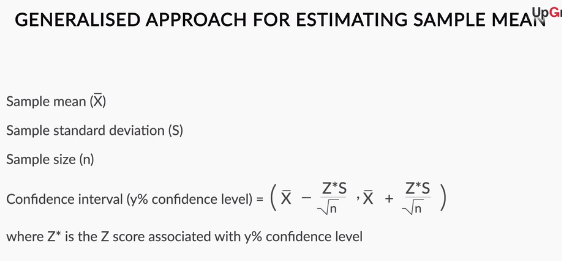
## Central Limit Theorem

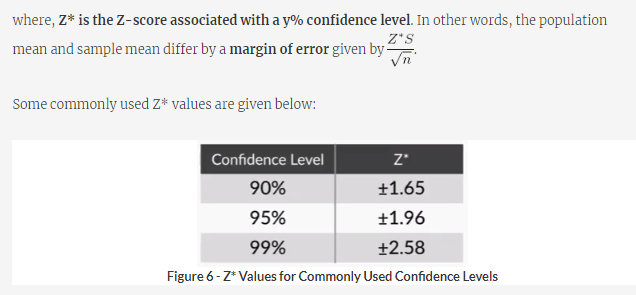




|  |  |
| --- | --- |
|  |  |



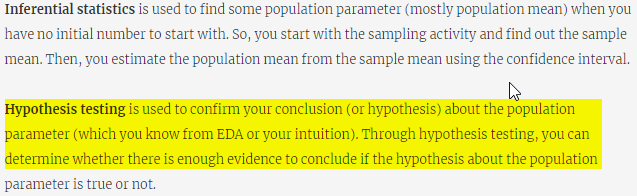


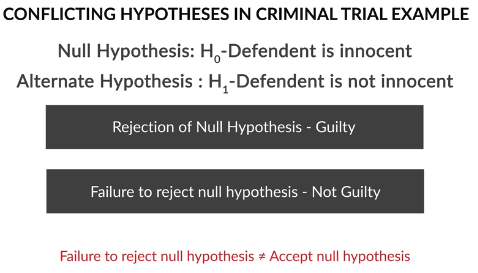


### One Example of CLT:

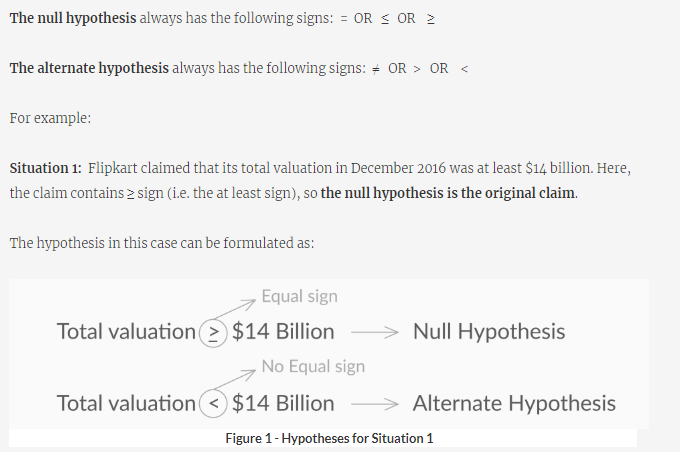
|  |
| --- |
| **Question:** To find Population Mean |
| Fix the Confidence Level to 99%. |

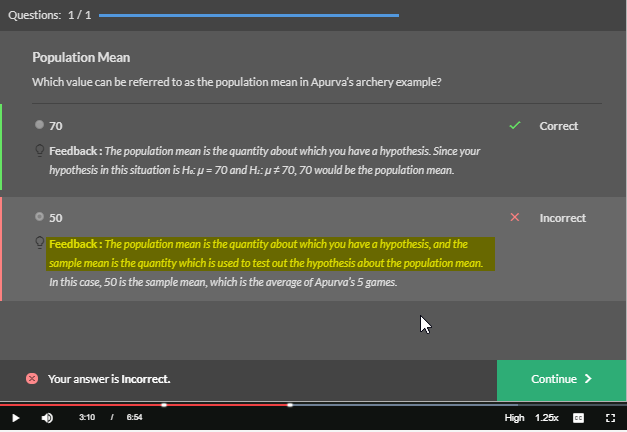
# Module: 4 (Concept of Hypothesis Testing-1)

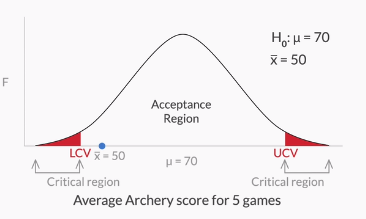


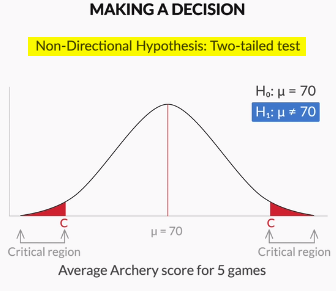




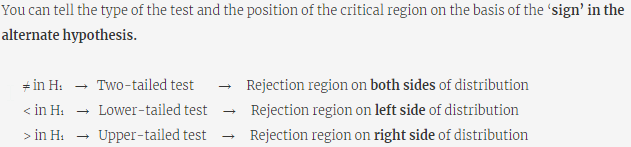








|  |  |
| --- | --- |
|  |  |



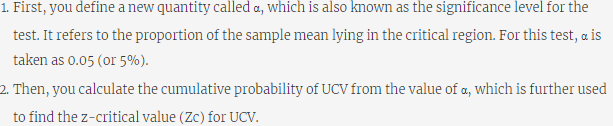
**Important Note:**

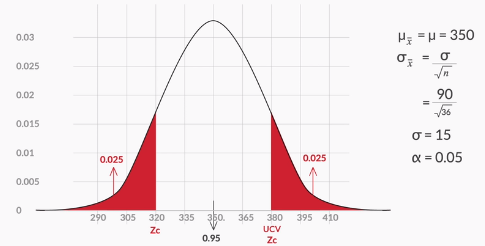
NULL Hypothesis is the Initial Mean value which was estimated before and claimed to be true. By Sample Mean we try to find out of the NULL Hypothesis is true or false.

By taking initial Mean, we derive the range of the Mean Value in the acceptance Region. Then we try to find out if the sample mean is coming under acceptance region or in critical section.

If sample Mean is coming in “Acceptance Region” then Null Hypothesis is true and can’t be rejected. If sample mean is coming under “Critical Section” then Null Hypothesis could have rejected. That means the Initial Mean may not be correct.

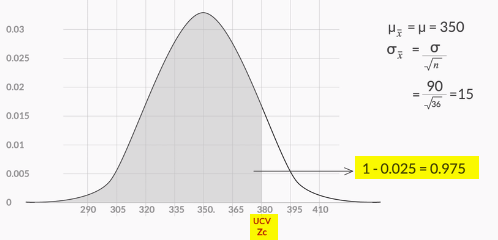
## Critical Value Method

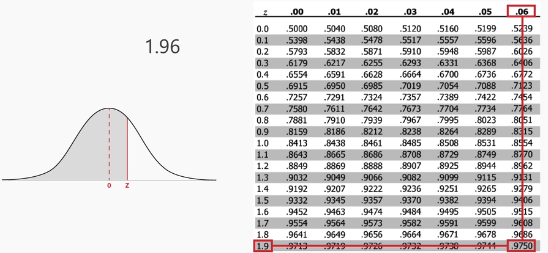


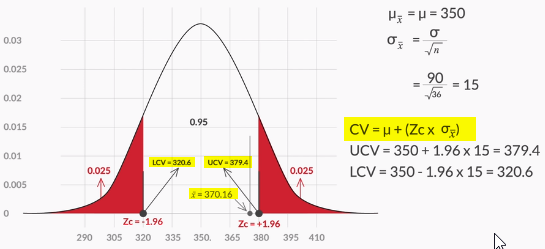


With the given information find following things:

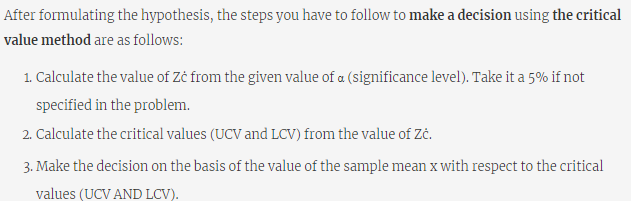
1. Find the Standard Error by taking the sample standard deviation/Sqrt(SampleSize)
2. Find the Z Score for the Critical Section.

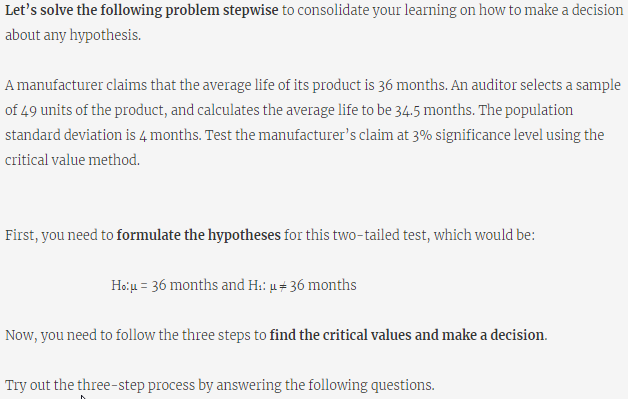


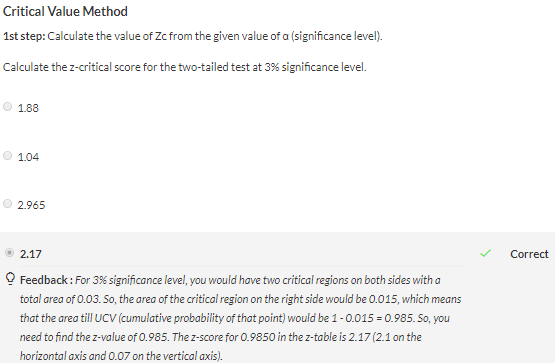


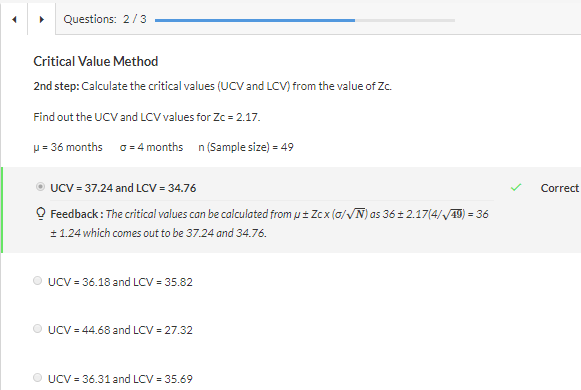


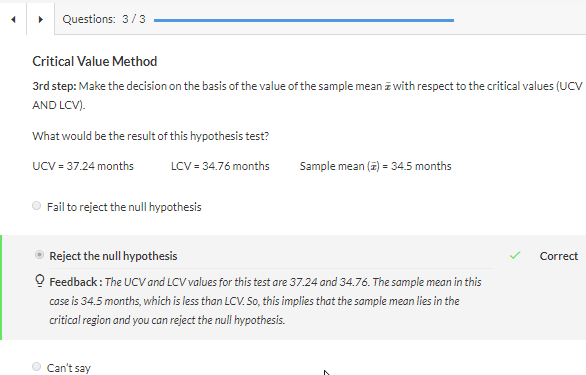
Now from above formula, the LCV and UCV calculation is done to find if the sample mean is coming under “Acceptance Region” or “Critical Section”



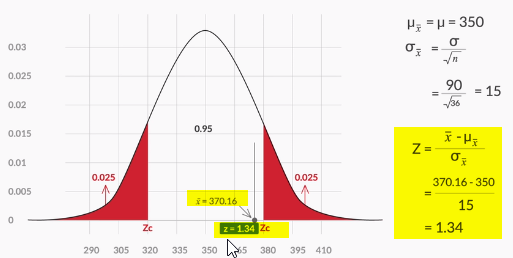


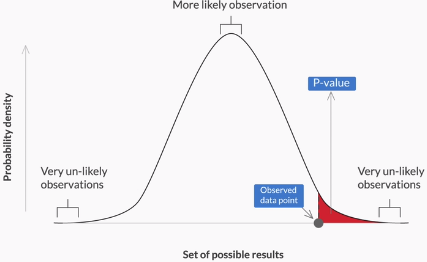






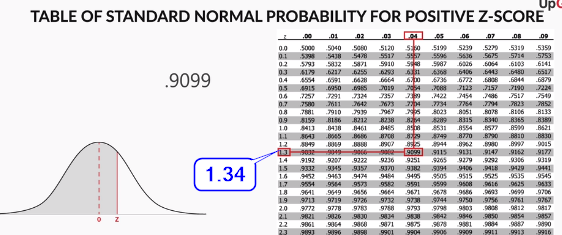
## P-Value Method

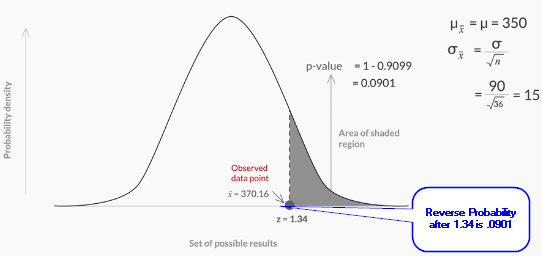


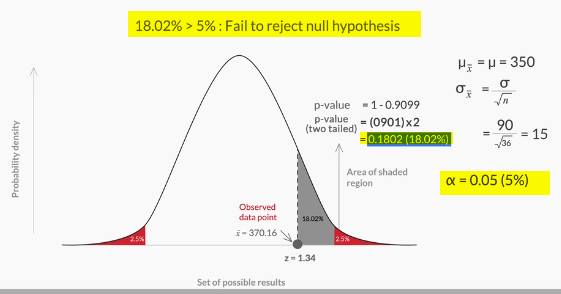


**Larger the P-Value, Observed-Data Point is near to mean and Hypothesis can be accepted.**

**If P-Value is less, then Observed-Data Point is far from mean and Hypothesis can be rejected.**

****

****

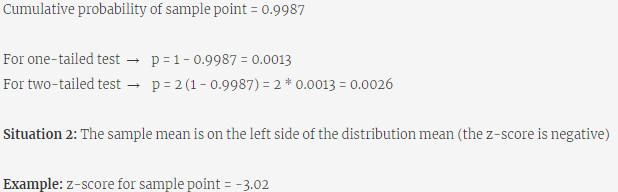
****

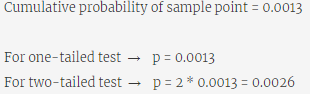
**Alfa is 0.05 🡺 Critical Section is in 5% Region.**

**Now we derive Sample mean is in 1802% Region. Hence Null Hypothesis can’t be rejected.**

### Another Example:

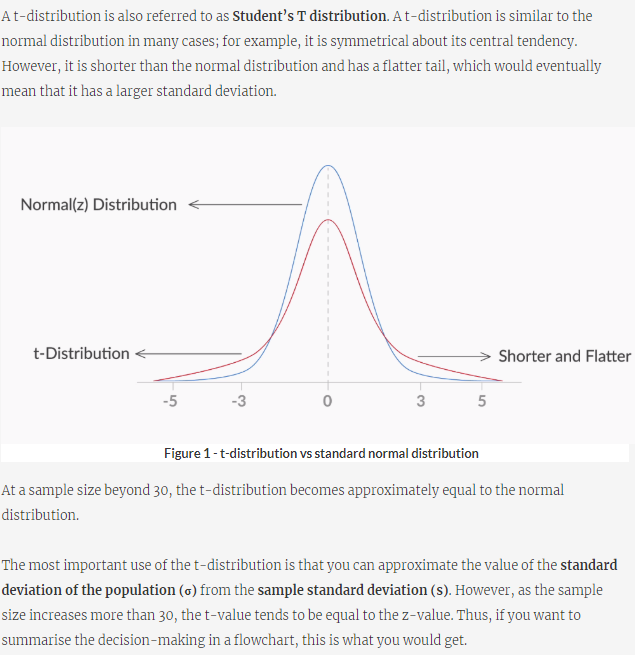


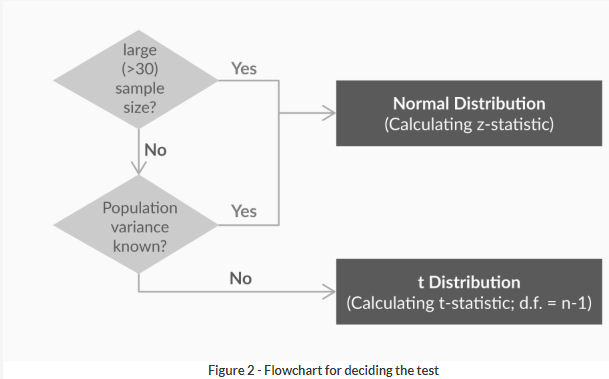


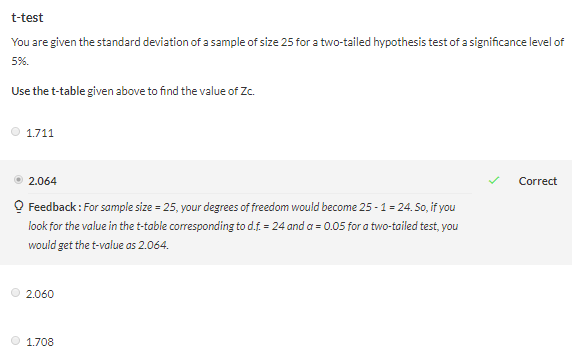


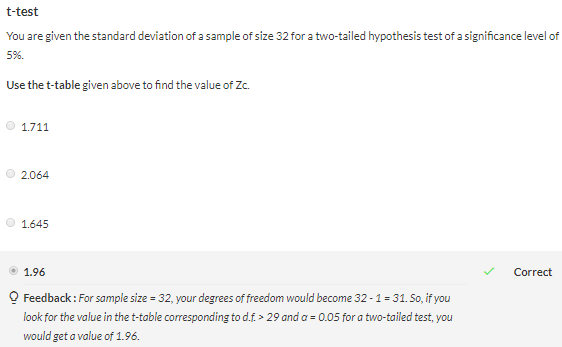
## Industry Demonstration Hypothesis Testing

### T Distribution

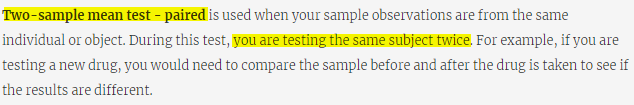


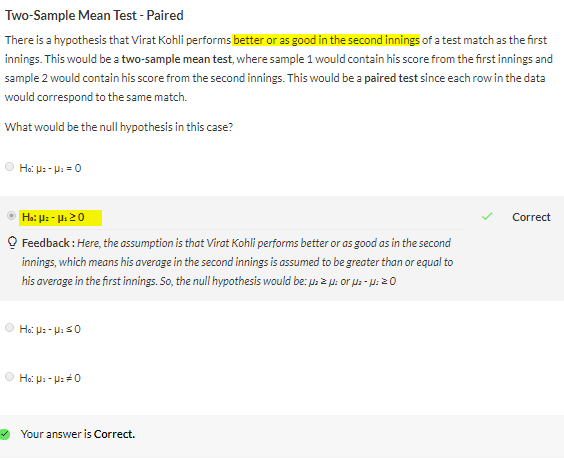




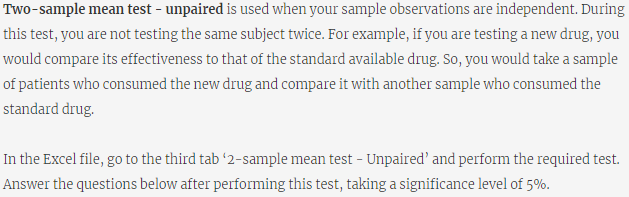


### Two-Sample Mean Test (Paired)





### Two-Sample Mean Test (Unpaired)

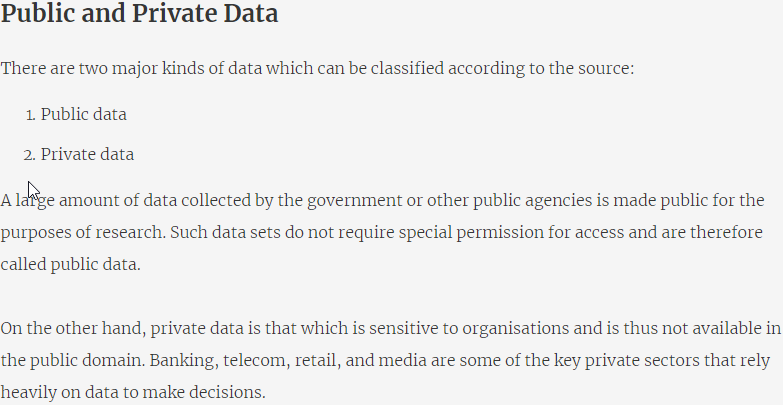


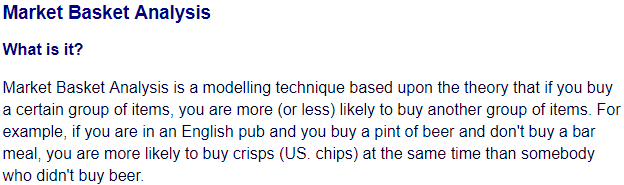
### Two-Sample Proportion Test

# Module: 6 (Exploratory Data Analysis)



## Data Souring





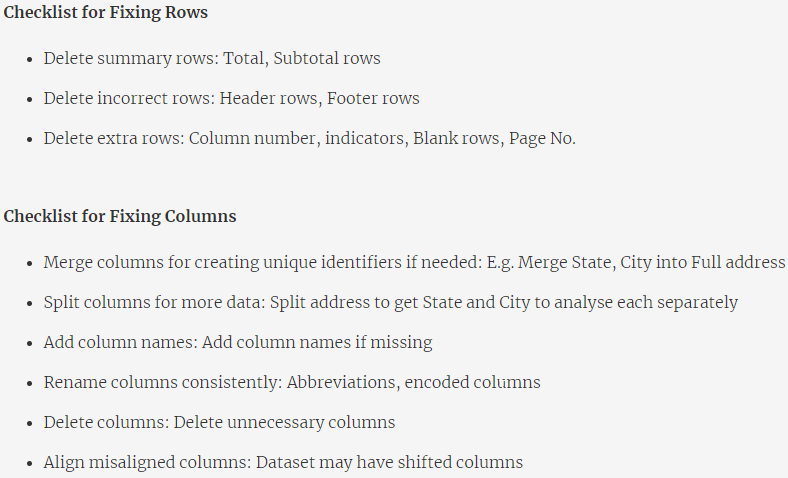
<https://data.gov.in/>

<https://github.com/datameet>

<https://github.com/awesomedata/awesome-public-datasets>

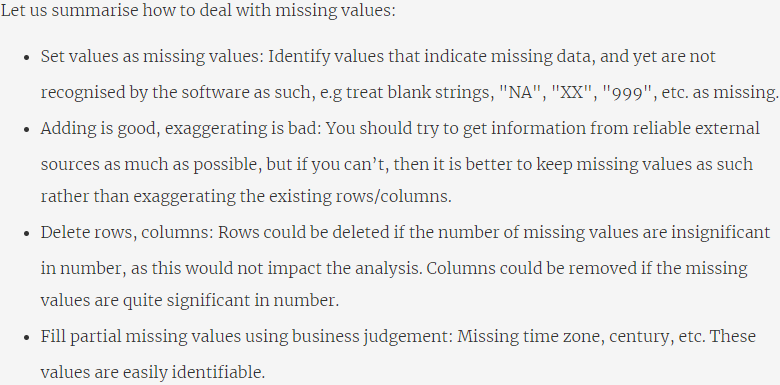
## Data Cleaning

### Fix Rows and Columns

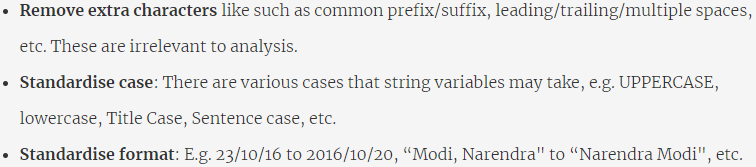




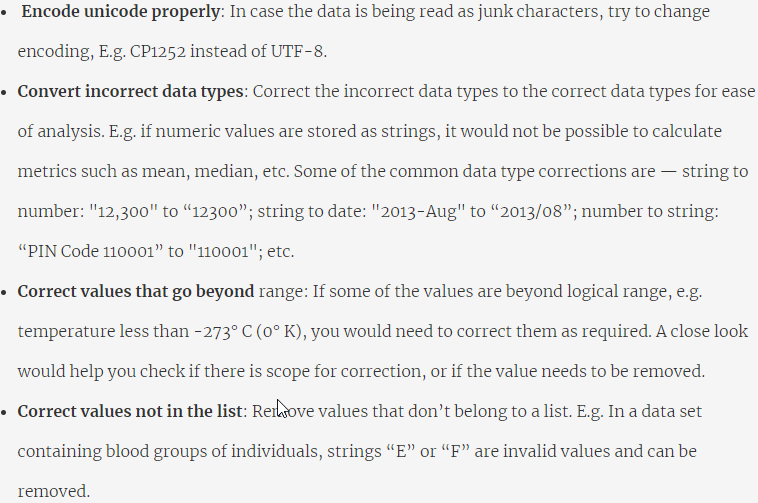
### Fix Missing Values

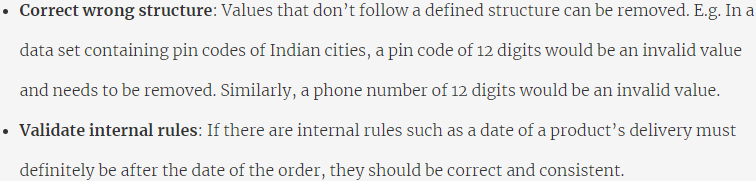


### Standardise Values

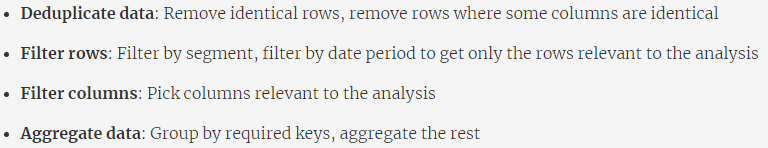


### Invalid Values

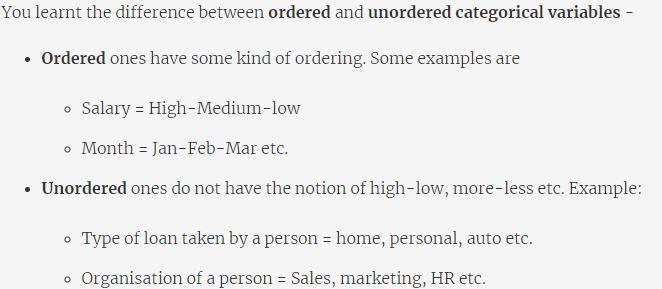


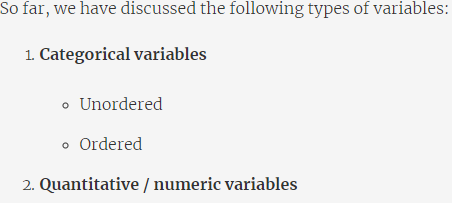


### Filtering Data

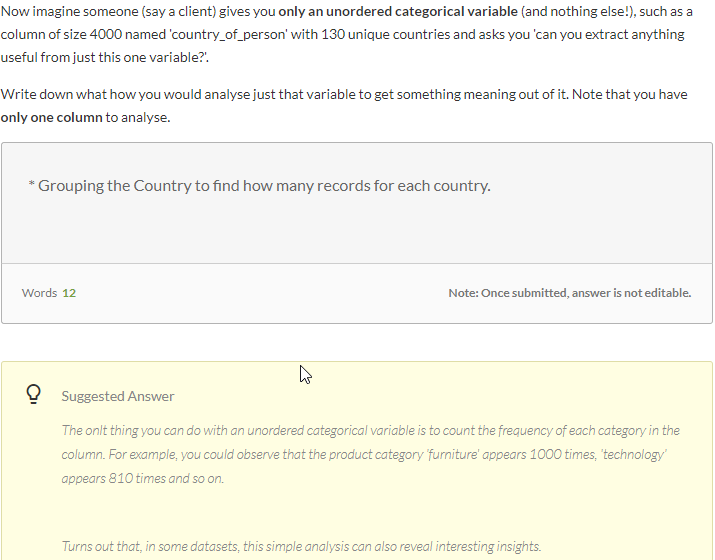


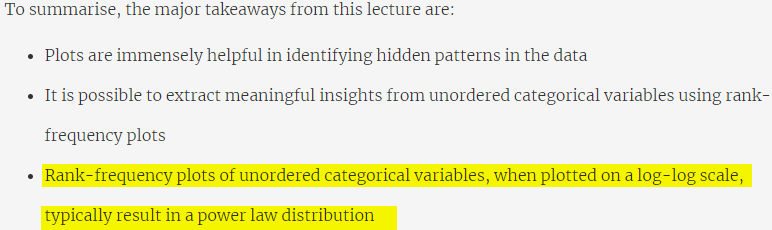
## Univariate Analysis





### Unordered Categorical Variables (Univariate Analysis)

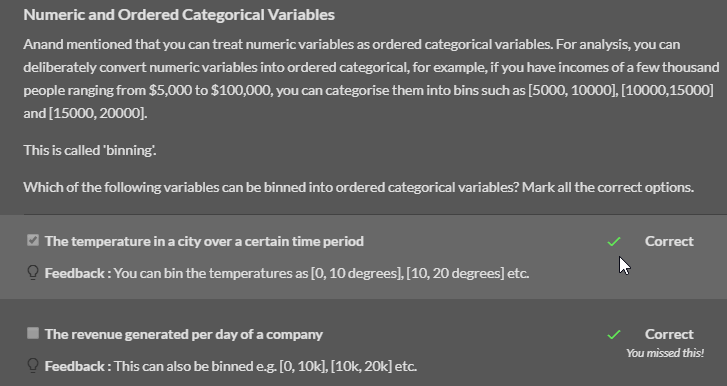




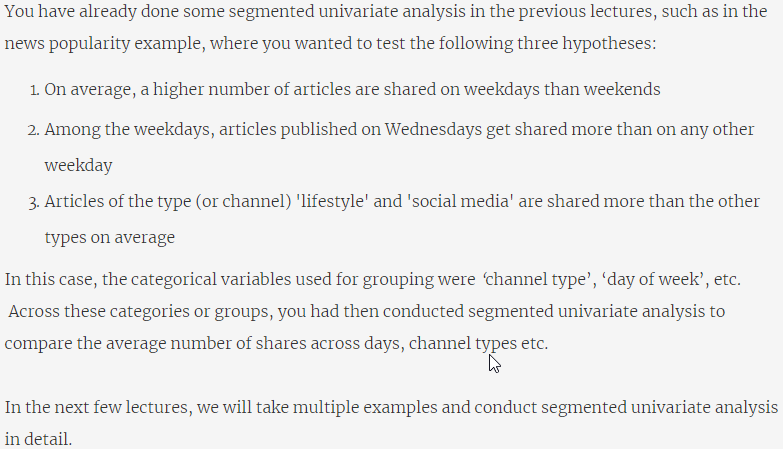
### Order Categorical Variables (Univariate Analysis)

* Meter Reading is Fixed in 100, 200 Limit in one survey. Bar chart shows the spikes in these limits.
* In the score sheet of the student, with bar chart it can be shown in 35 (passing mark) the spikes are high.
* Tendulkar batting analysis can be refer further.

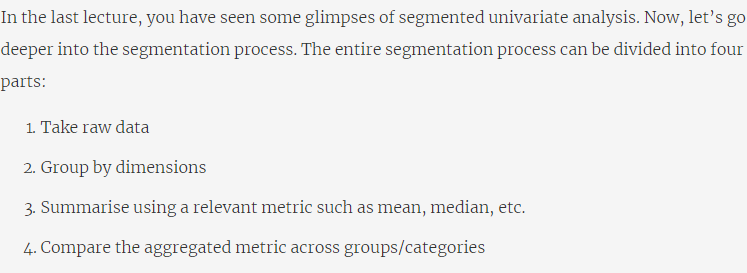
### Quantitative Variables (Univariate Analysis)



## Segmented Univariate

****

### Basics of Segmentation



The example which was given in this part was:

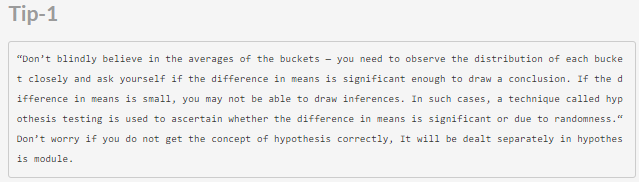
How the marks % get impacted with various factors of the student. (Parents Education, Number of Siblings and so on.

Note: But we need to be careful to derive the relation. As there is high possibility of, “No Relation”

e.g: See below marks of Student

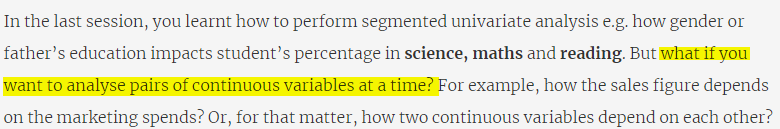
|  |  |
| --- | --- |
|  |  |

In Both the dataset, average difference is 1. But in 1st data set, in all the data point Boys are having less marked than Girls but in 2nd data set all records are random and scattered.

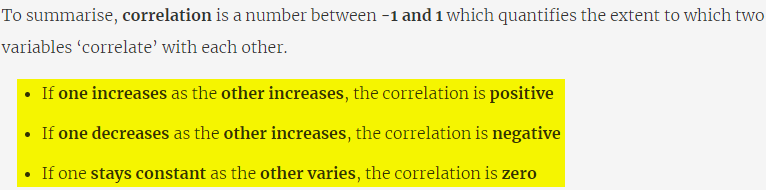


## Bivariate Analysis

### Bivariate Analysis on Continuous Variables:



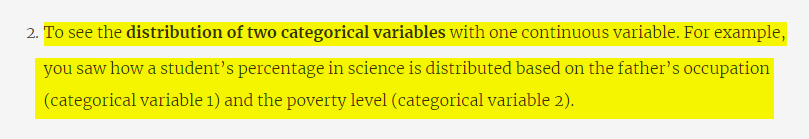
**This method is basically to check the correlation between 2 variables.**



### Bivariate Analysis on Categorical Variables:

\* In Continuous Bivariate analysis, we try to find the correlation of 2 entities.

\* In Bivariate analysis in categorical level, we narrow down each entity taken Continuous Bivariate analysis and try to find out if the correlation is correct in all the category based on other features.



Summary:

Bivariate analysis is basically to do the analysis of the data by taking two columns at a time.

There are 2 kind of bivariate analysis.

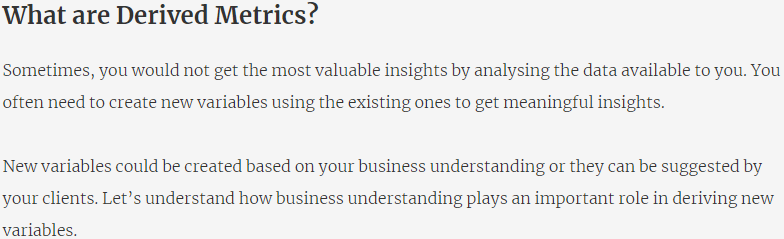
1. Analysis on Continuous

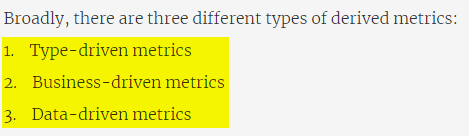
2. Analysis on Categorical Variables

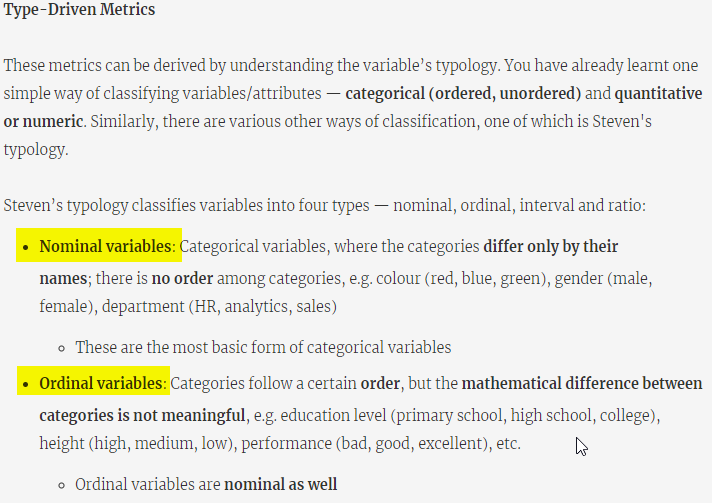
In continuous variable analysis, we can check the direct relation of one column to another. We can check if one column value is increasing, if other is increasing. By doing the correlation analysis it can be found it.

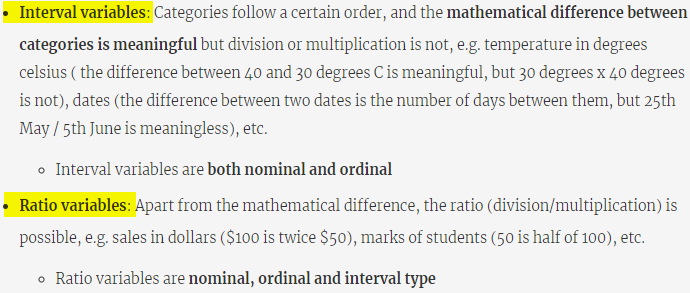
In categorical variable analysis, we can split the column data in to multiple category and see if the correlation is correct for each category.

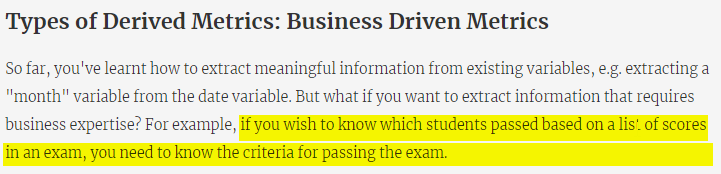
## Derived Metrics











\* From passing cutoff (data need to get), we can derive columns for each marks (Math, Reading, Science, Social) if the student passed or failed.

\* If there is a overall cuttoff for total marks, then we can derive if student pass or failed by summing up all the category marks.

\* If there is any special category where student can go to next level with passing max 3 subject,then same can be used to derive some special column.

