upGrad

#LifeKoKaroLift

Post-Graduate Diploma in ML/AI

upGrad

Course : Machine Learning

Lecture On : Pre exam session

Instructor : Dr Reena Duggal

Modules Included

Statistics Essentials

- Inferential Statistics
- Hypothesis Testing
- Exploratory Data Analysis
- Python

Modules Included

ML - 1

- Linear Regression
- Logistic Regression
- Naïve Bayes
- Model Selection

15 MCQs 15 X1=15

5 Multi-select options 5x3 = 15

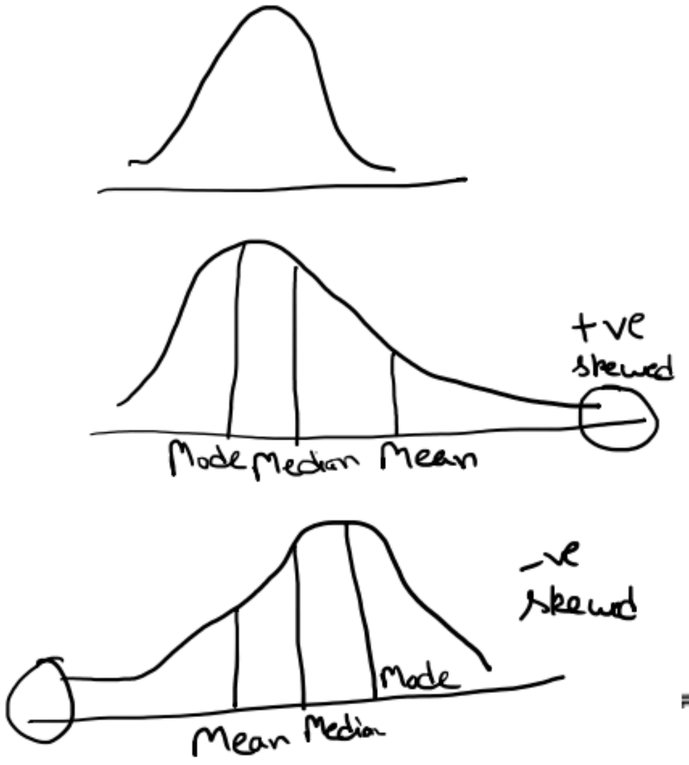
2 Python Coding Q 2x5 = 10

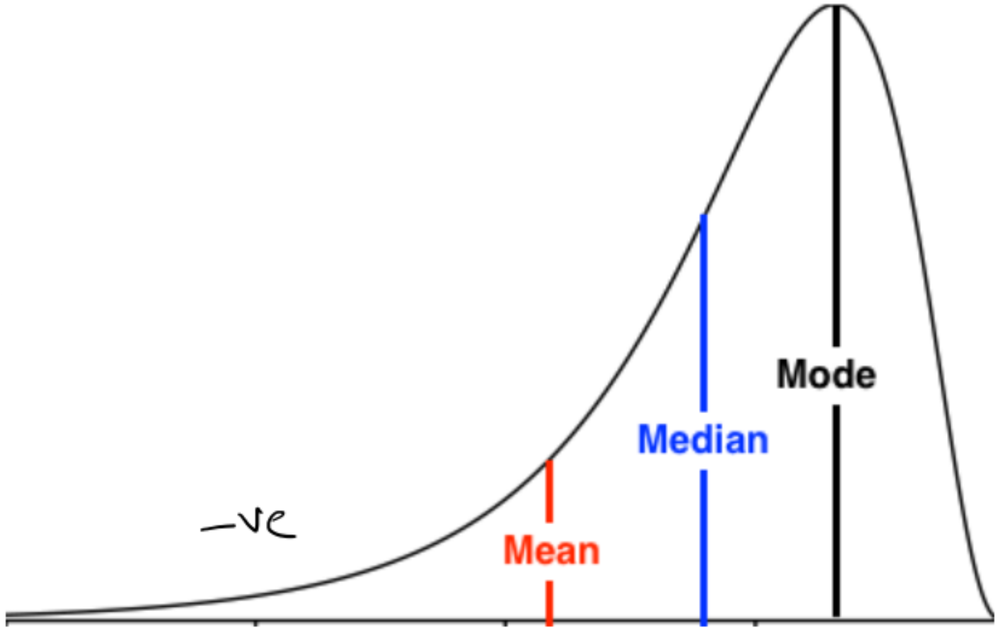
40

Measures of Central Tendency

upGrad

" The measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data."





Poll

upGrad

If a positively skewed distribution has a median of 50, which of the following statement is true?

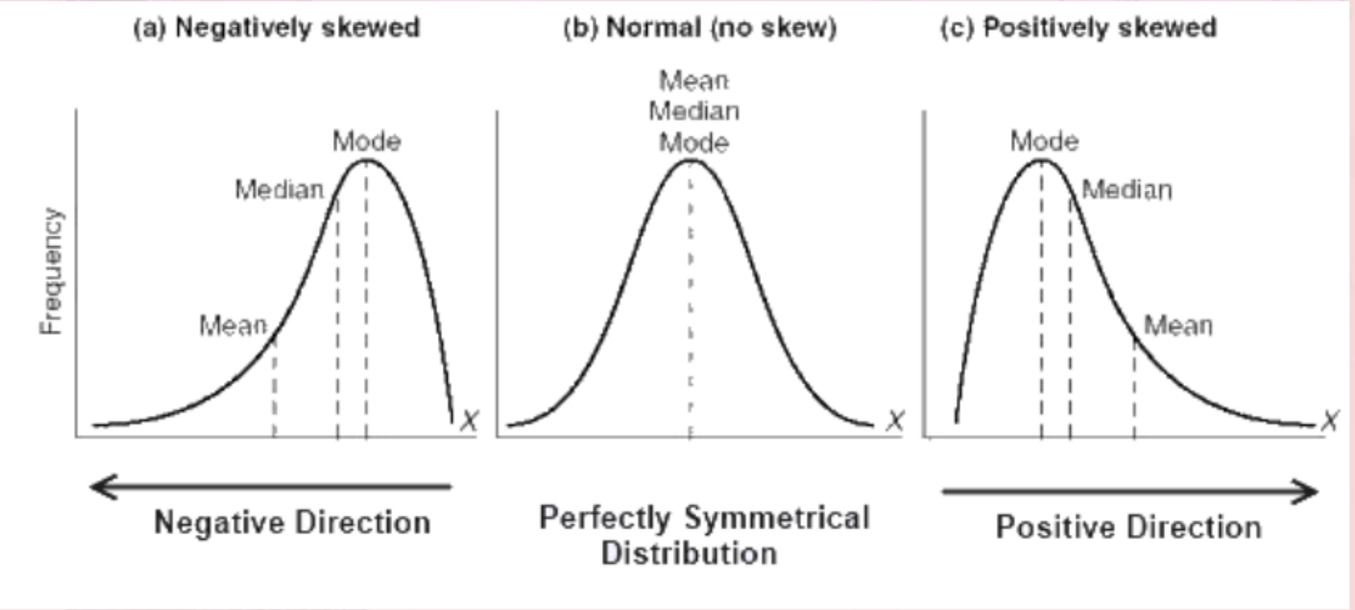
A) Mean is greater than 50
B) Mean is less than 50
C) Mode is less than 50
D) Mode is greater than 50
E) Both A and C
F) Both B and D

Poll

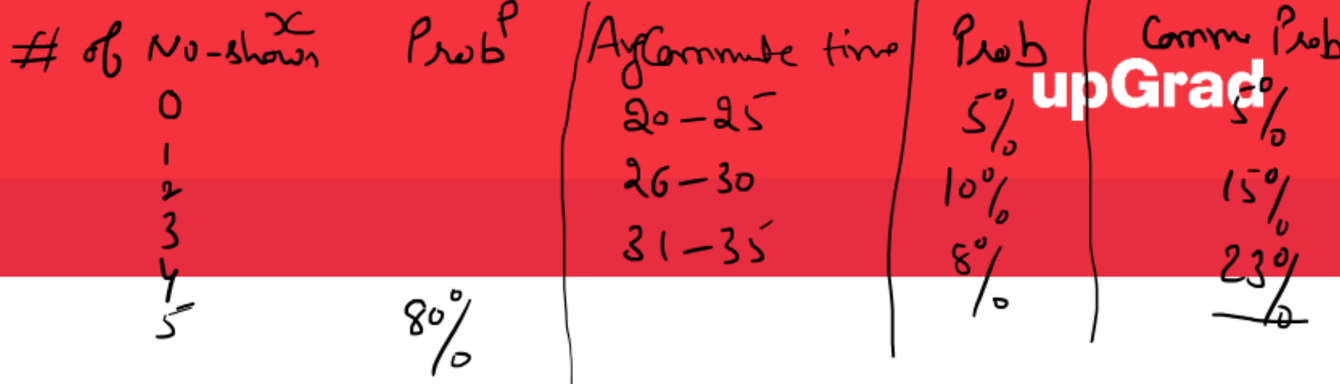
upGrad

If a positively skewed distribution has a median of 50, which of the following statement is true?

A) Mean is greater than 50
B) Mean is less than 50
C) Mode is less than 50
D) Mode is greater than 50
E) Both A and C
F) Both B and D



INFERENTIAL STATISTICS



Expected Value = Avg.

The **expected value** for a variable X is the value of X we would "expect" to get after performing the experiment once. It is also called the **expectation**, **average**, and **mean value**

$$EV(X) = x_1 * P(X = x_1) + x_2 * P(X = x_2) + \ldots\ldots\ldots\ldots\ldots + x_n * P(X = x_n)$$

Discrete Probability Distributions

A discrete distribution describes the probability of occurrence of each value of a discrete random variable. A discrete random variable is a random variable that has countable values, such as a list of non-negative integers. Ex. A coin flip

Continuous Probability Distributions

A continuous distribution describes the probabilities of the possible values of a continuous random variable. A continuous random variable is a random variable with a set of possible values (known as the range) that is infinite and uncountable Ex. Commute time of an upgrad employee

Poll

upGrad

One game that is popular at some carnivals and amusement parks involves selecting a floating plastic duck at random from a pond full of ducks.

In most cases, the letter S, M, or L appears on the bottom of the duck, signifying that the winner receives a small, medium, or large prize, respectively.

The duck is then returned to the pond for the next game. Although the prizes are typically toys, crafts, etc., suppose that the monetary values of the prizes are as follows:

Small is $ 0.50, Medium is$1.50, and Large is $5.00

The probabilities of winning an item on duck selection are as follows:

Small 60%, Medium 30%, and Large is 10%.

x	$P(x)$	$x \cdot P(x)$
0.50	0.60	
1.50	0.30	
5	0.10	
	$\sum x P(x)$	+1.25

What is the expected value of money that the person will lose or win playing this game?

- +1
- +1.25
- -1.25

Poll

upGrad

One game that is popular at some carnivals and amusement parks involves selecting a floating plastic duck at random from a pond full of ducks.

In most cases, the letter S, M, or L appears on the bottom of the duck, signifying that the winner receives a small, medium, or large prize, respectively.

The duck is then returned to the pond for the next game. Although the prizes are typically toys, crafts, etc., suppose that the monetary values of the prizes are as follows:

Small is $ 0.50, Medium is$1.50, and Large is $5.00

The probabilities of winning an item on duck selection are as follows:

Small 60%, Medium 30%, and Large is 10%.

What is the expected value of money that the person will lose or win playing this game?

- +1
- +1.25
- -1.25

Central Limit Theorem

upGrad

HT

μ = avg salary = 7l σ = 0.9l

$\bar{X}$ = 6.7l

$P(\bar{x} = 6.7 \text{ when } \mu = 7)$

P_value High 80% ⇒ Assumption was still true

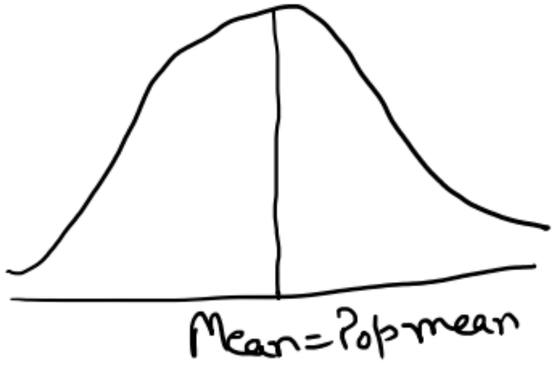
Low 2% ⇒ Assumption is false

Obj : To find Prob. of a sample mean

>30

$SD = \frac{0.9}{\sqrt{100}}$ $\frac{\sigma}{\sqrt{n}}$

CLT helps us to find prob of a sample mean (P-value)



$$SD = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

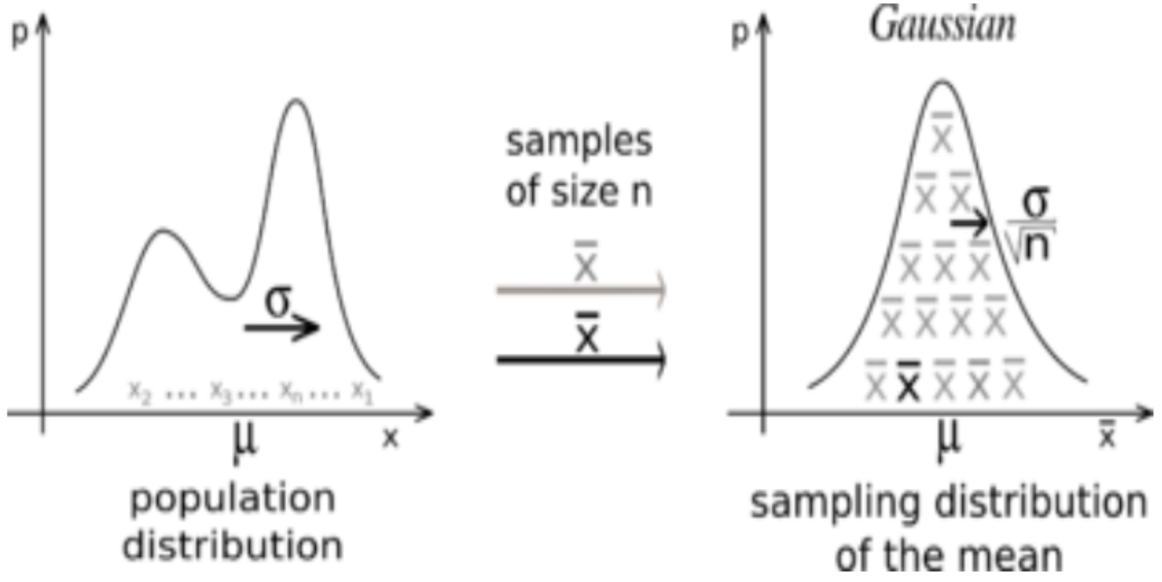
upGrad

The central limit theorem says that, for any kind of data, provided a high number of samples has been taken, the following properties hold true:

1.Sampling distribution's mean ($\mu_{\bar{X}}$) = Population mean (μ)

2.Sampling distribution's standard deviation (Standard error) = σ/√n

3.For n > 30, the sampling distribution becomes a normal distribution



Hypothesis Testing

upGrad

Hypothesis : a claim or an assumption that we make about one or more population parameters

H_0	H_1
$\mu = 7\ell$	$\mu \neq 7\ell$
$\mu \geq 7\ell$	$\mu < 7\ell$
$\mu \leq 7\ell$	$\mu > 7\ell$

Types of hypothesis:

Null hypothesis (H_0)

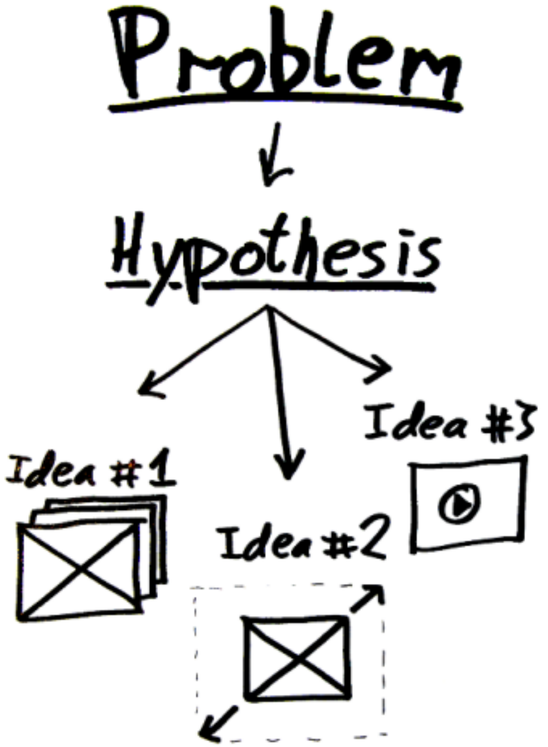
- Makes an assumption about the status quo
- Always contains the symbols '=', '≤' or '≥'

$\mu = 500 m\ell$

Alternate hypothesis (H_1)

- Challenges and complements the null hypothesis
- Always contains the symbols '≠', '<' or '>'

$\mu \neq 500 m\ell$



Hypothesis Testing

upGrad

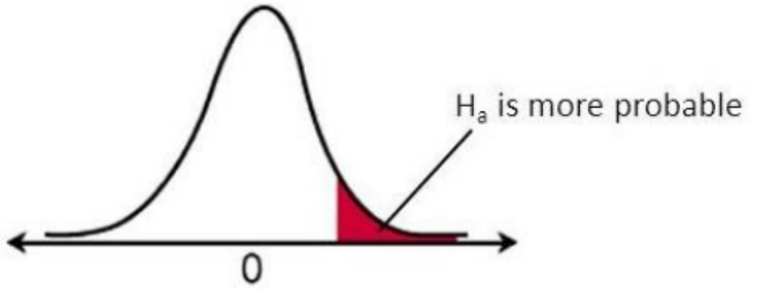
Types of tests:

Upper-tailed test

$H_0: \mu \leq 7\ell$

$H_1: \mu > 7\ell$

- The critical region lies on the right side of the distribution
- The alternate hypothesis contains the > sign

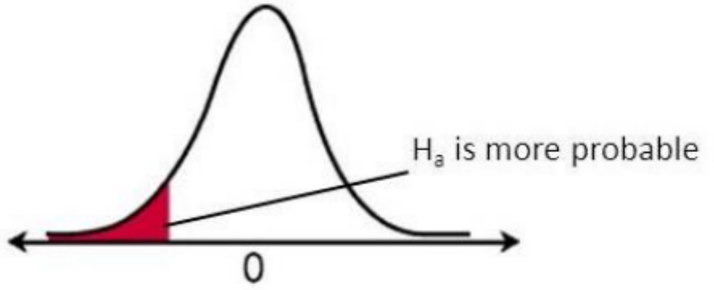


Lower-tailed test

$H_0: \mu \geq 7\ell$

$H_1: \mu < 7\ell$

- The critical region lies on the left side of the distribution
- The alternate hypothesis contains the < sign

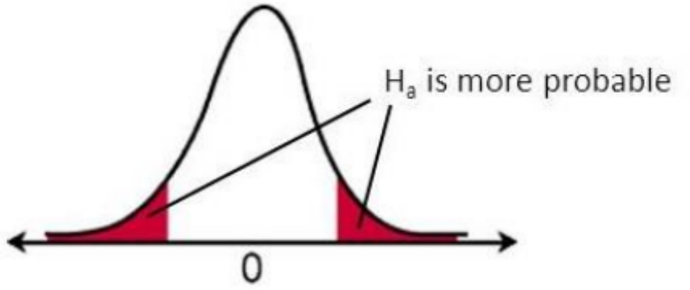


Two-tailed test

$H_0: \mu = 7\ell$

$H_1: \mu \neq 7\ell$

- The critical region lies on both sides of the distribution
- The alternate hypothesis contains the ≠ sign



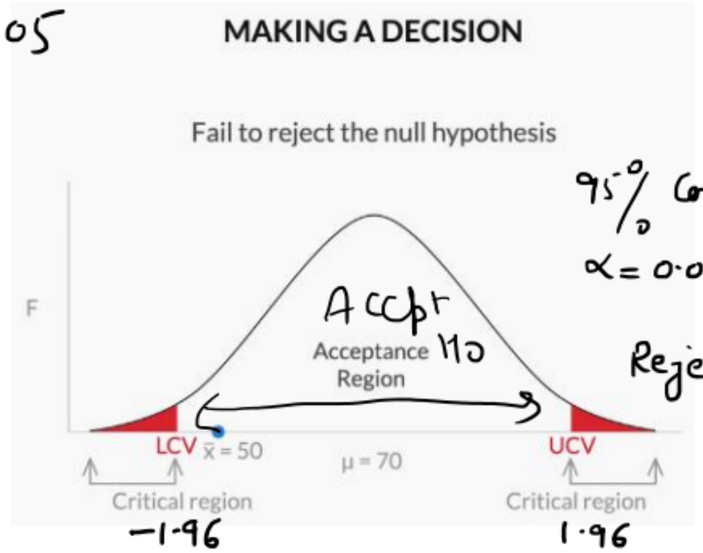
Hypothesis Testing

upGrad

Making a decision

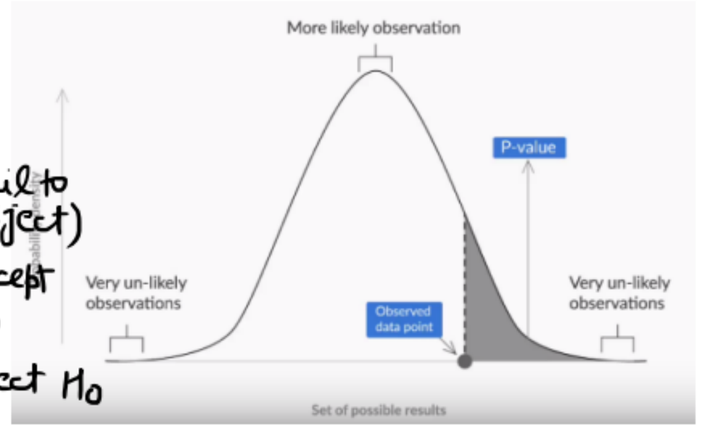
Critical value method:

- Calculate the value of Z_c from the given value of α (significance level)
- Calculate the critical values (UCV and LCV) from the value of Z_c
- Make the decision on the basis of the value of the sample mean(x-bar) with respect to the critical values (UCV AND LCV)



P-value method:

- Calculate the value of z-score for the sample mean point on the distribution
- Calculate the p-value from the cumulative probability for the given z-score using the z-table
- Make a decision on the basis of the p-value (multiply it by 2 for a two-tailed test) with respect to the given value of α (significance value).



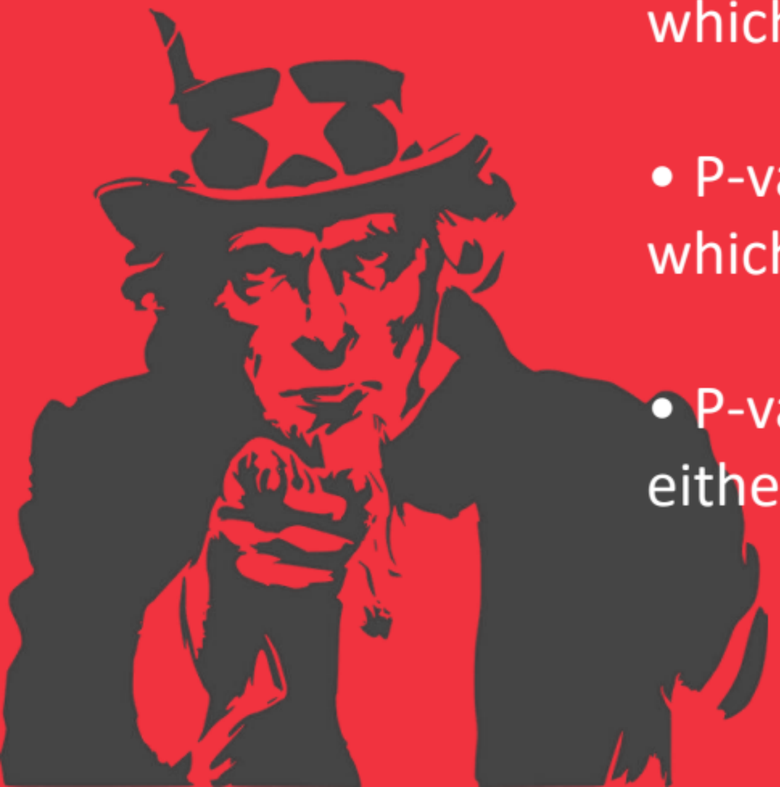
What does P-value signify about the statistical data?

P-value is used to determine the significance of results after a hypothesis test in statistics. P-value helps the readers to draw conclusions and is always between 0 and 1.taking alpha value as 0.05

- P- Value > 0.05 denotes weak evidence against the null hypothesis which means the null hypothesis cannot be rejected.

- P-value < 0.05 denotes strong evidence against the null hypothesis which means the null hypothesis can be rejected.

- P-value = 0.05 is the marginal value indicating it is possible to go either way.



Example – Waiting Time

upGrad

You are the manager of a fast food restaurant. You want to determine if the population mean waiting time has changed from the 4.5 minutes. You can assume that the population standard deviation is 1.2 minutes. You select a sample of 36 orders in an hour. Sample mean is 4.1 minutes. Use the relevant hypothesis test to determine if the population mean has reduced from the past value of 4.5.

$H_0 : \mu \geq 4.5$

$H_1 : \mu < 4.5$

$\sigma = 1.2 \qquad \bar{x} = 4.1$

$n = 36$

Conf. level = 95%

$\alpha = 1 - 0.95 = 0.05$

$$Z_{Stat} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{4.1 - 4.5}{\frac{1.2}{\sqrt{36}}} = -$$

$P(Z) =$

Test it with Conf. level 99%. $\alpha = 0.01$

$< 5\%$ Reject H_0

$> 5\%$ Accept H_0

Example – Waiting Time

- Step 1: Given: n = 36, $\bar{x}$ =4.1 σ = 1.2

- Step 2: Formulate Hypothesis

  H_0 : μ >= 4.5

  H_1 : μ < 4.5

- Step 3: Define test statistic

  $Z_{stat} = (\bar{x} - \mu) / (\sigma / \sqrt{n})$

  $= (4.1 - 4.5) / (1.2/\sqrt{36}) = -2$

Example – Waiting Time

- Step 4: Draw diagram

- Step 5: (critical value method)

  Determine critical values

  $\alpha = 5\% = 0.05$.

  $-Z_\alpha = -1.64$

- Step 6: (critical value method)

  Compare whether Zstat value is in reject region and make decision

  $-2 < -1.64$

  $\mu < 4.5$

  Since Zstat is in Reject region, H0 is rejected. i.e. H1 is accepted.

Example – Waiting Time

- Step 5: (p-value method) $Z_{stat} = -2$
  - Find p-value.
  - P-value = *stats.norm.cdf(-2)* $H_1 : <$ (left-tail, lower)
    = 0.023 < 0.05 (2.3%) 0.01 (99% Conf. level)
- Step 6: (p-value method)
  - Since p-value < α, therefore Reject Null Hypothesis, i.e. Accept Alternate Hypothesis

$\mu < 4.5$

Type I and Type II Error

upGrad

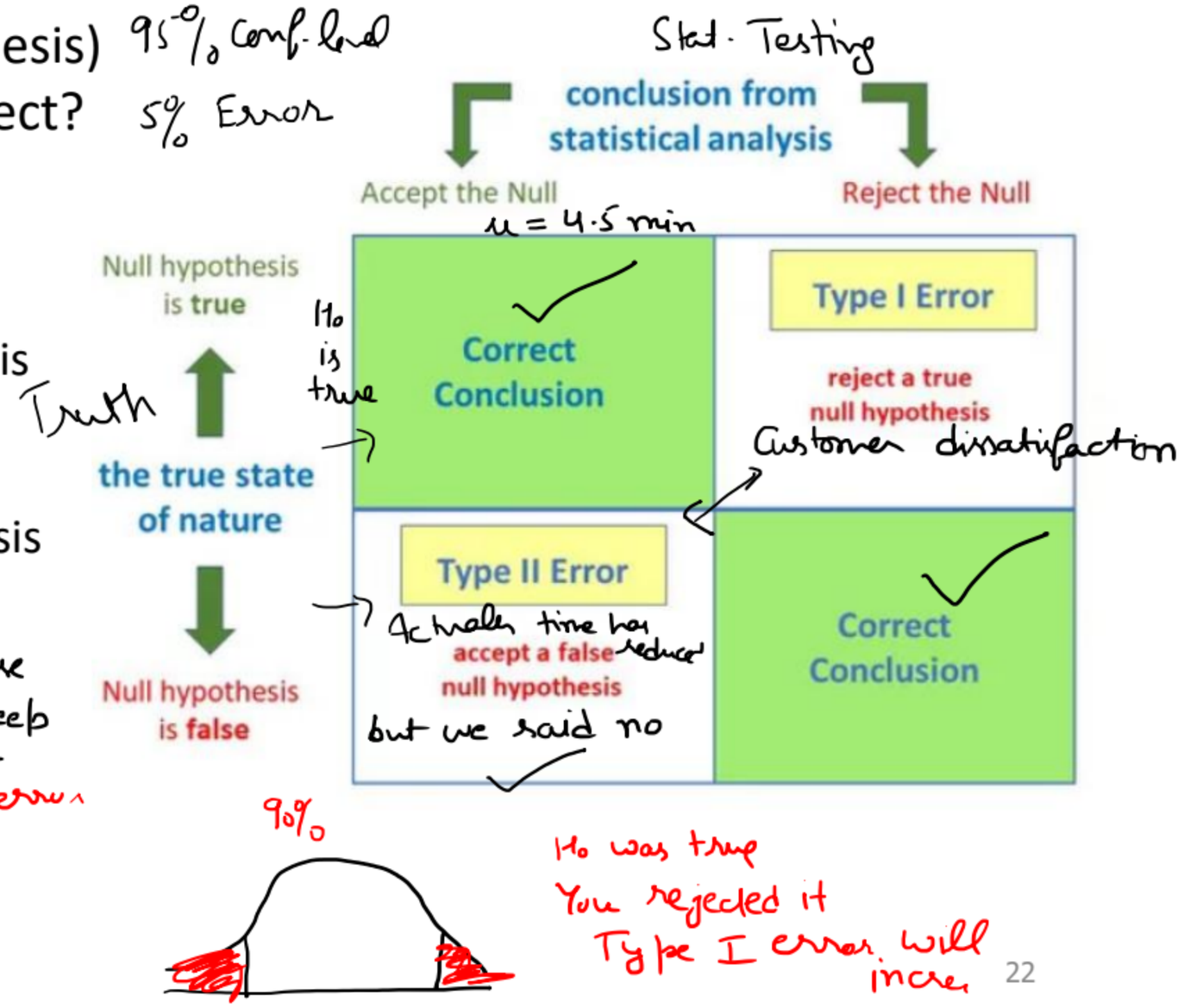
Are the decision (e.g. Reject Null Hypothesis) made using Hypothesis test, always correct?

95% Conf. level
5% Error

A type I error occurs when the null hypothesis is true but is rejected.

A type II error occurs when the null hypothesis is false, but erroneously fails to be rejected.

H0 was false
You will keep accepting it
Type II error



H0 was true
You rejected it
Type I error will incre

Suppose we want to test whether boys, on average, score higher than 600 on the SAT verbal section or not.

Suppose we choose our α to be .05. Since this is a one-tailed test, we find our critical value in the upper tail of the sampling distribution, which is z = 1.645. Suppose we also happen to know that the standard deviation for boys SAT verbal section scores is 50.

Now we collect the data using a random sample of 20 Boys and their verbal section scores:

589 633 597 631 639
589 608 638 592 641
645 631 647 607 628
638 603 626 586 608

Determine the null and alternative hypotheses.

This, gives us the hypotheses:

$$H0 : \mu \leq 600$$
$$H1 : \mu > 600$$

Determine the Z – statistic.

$$z = (618.8 - 600) / (50/\sqrt{20}) = 1.6815$$

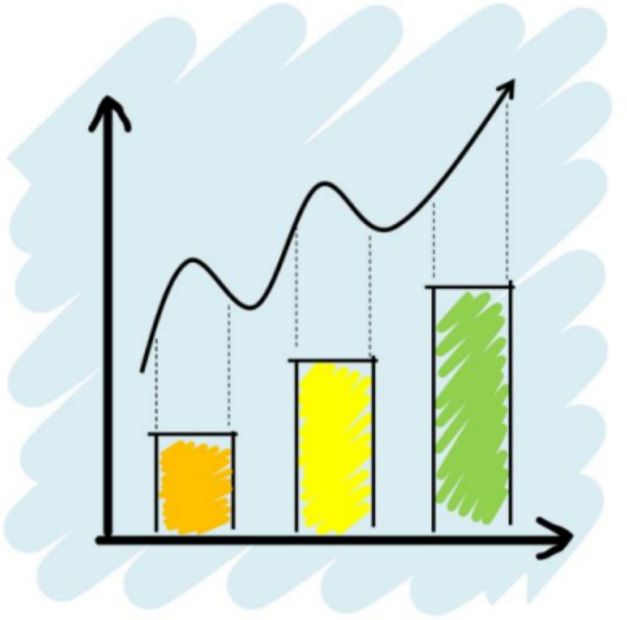
Determine inference based on z-critical

Since 1.6815 > 1.645, we reject the null hypothesis in favor of the alternative explanation that boys score, on average, better than 600 on the verbal section of the SATs

What is Exploratory Data Analysis?

upGrad

" Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations."



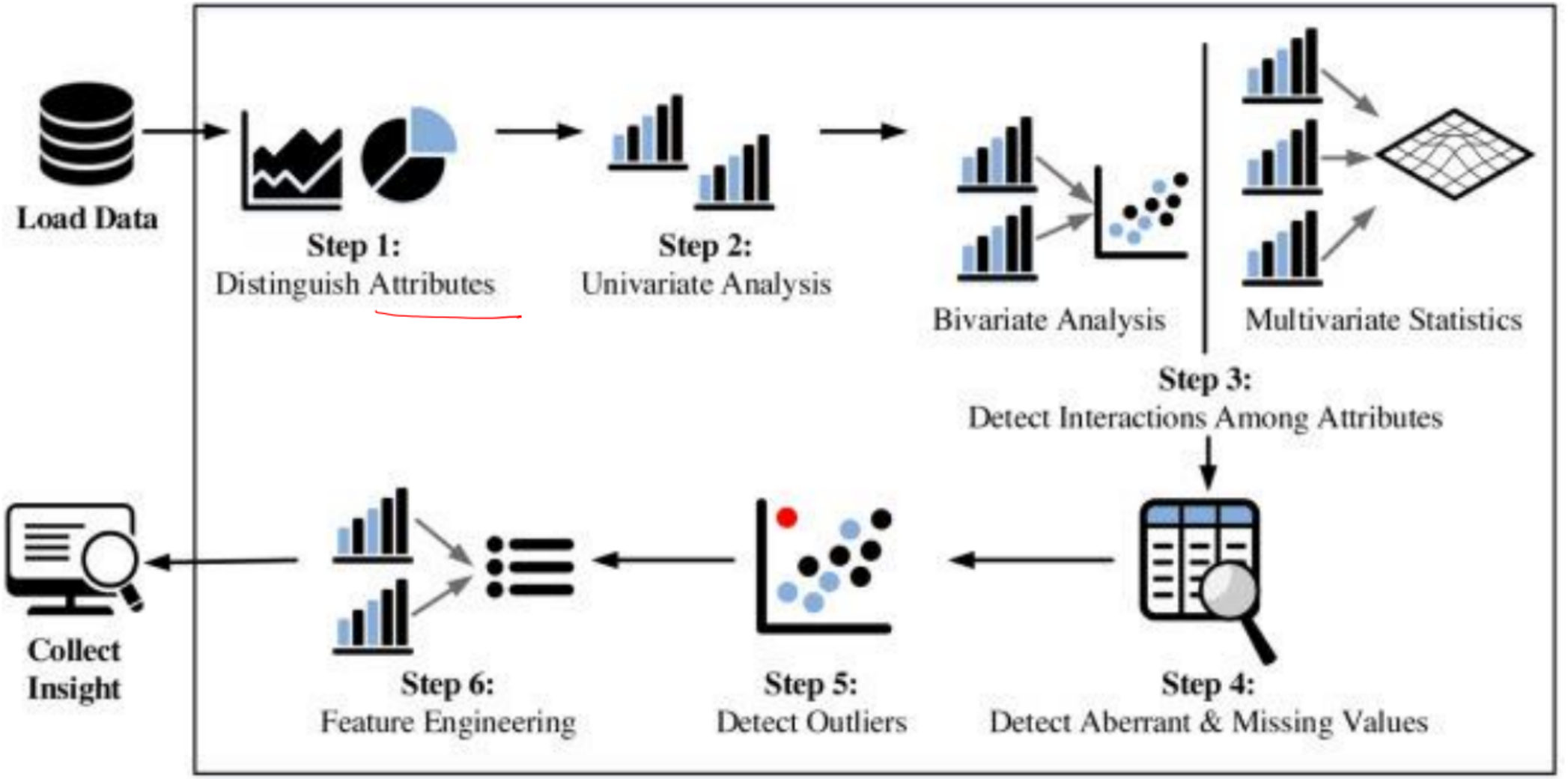
kinds of data source

Public data: Data collected by the government or other public agencies that are made public for the purposes of research are known as Public data.

Private data: Data generated by Banking, telecom, retail, and media are some of the key private sectors that rely heavily on data to make decisions.

Fundamental steps of EDA process

upGrad

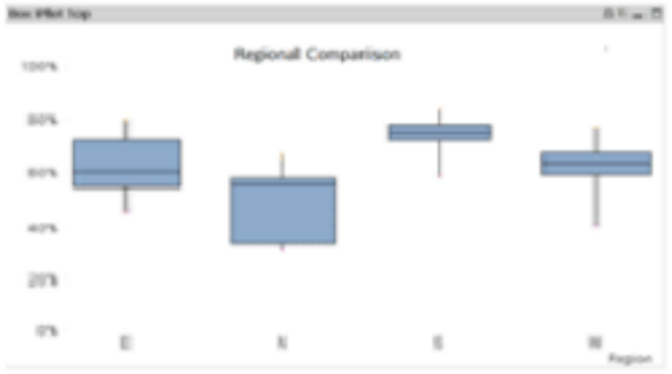


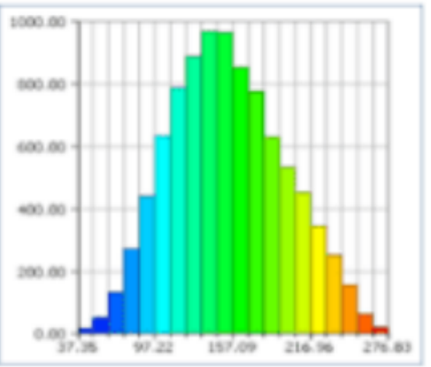
Univariate Analysis

upGrad

At this stage, we explore variables one by one. Method to perform univariate analysis will depend on whether the variable type is categorical or continuous.

Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	

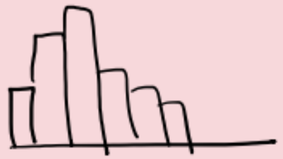




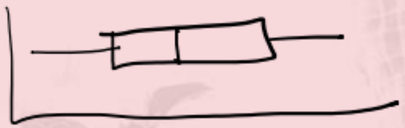
Poll

upGrad

Please select the plots which could be used for univariate analysis

A. Histogram 

B. Correlation matrix

C. Box-plot 

D. Scatter plot

Poll - solution

upGrad

Please select the plots which could be used for univariate analysis

A. **Histogram**

B. Correlation matrix

C. **Box-plot**

D. Scatter plot

Bi-variate Analysis

upGrad

Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at a pre-defined significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables.

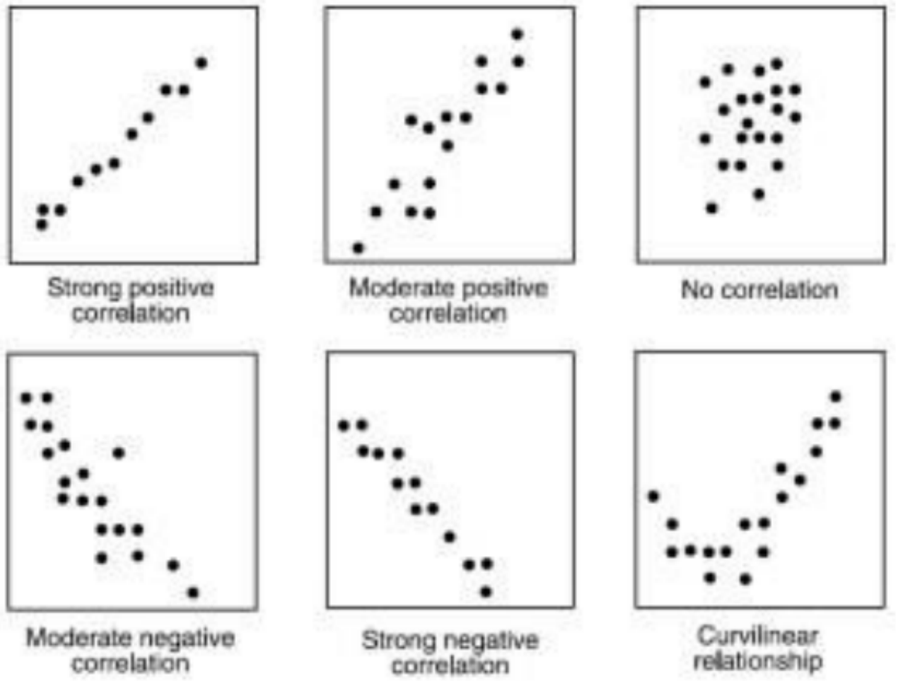
The combination can be:

1. Continuous & Continuous
2. Categorical & Categorical
3. Categorical & Continuous

Continuous vs continuous

upGrad

While doing bi-variate analysis between two continuous variables, we should look at scatter plots. It is a nifty way to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.



Categorical & Categorical

upGrad

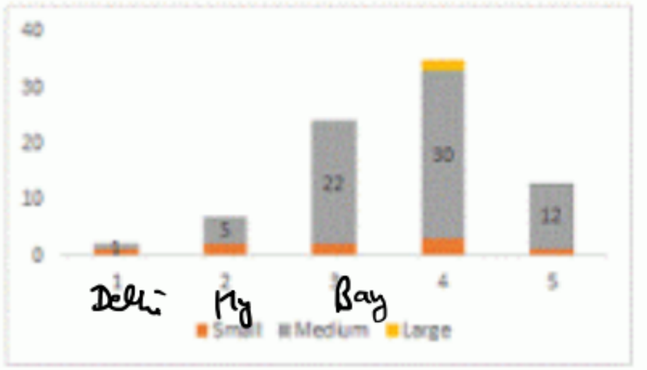
Pivot table

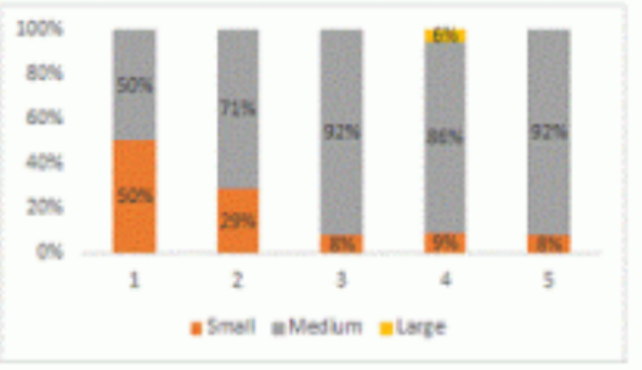
Product Category

Frequency Row Pct	1	2	3	4	5	Total
Small	1 11.11	2 22.22	2 22.22	3 33.33	1 11.11	9
Medium	1 1.43	5 7.14	22 31.43	30 42.86	12 17.14	70
Large	0 0.00	0 0.00	0 0.00	2 100.00	0 0.00	2
Total	2	7	24	35	13	81

Frequency Missing = 77

Bar Plot



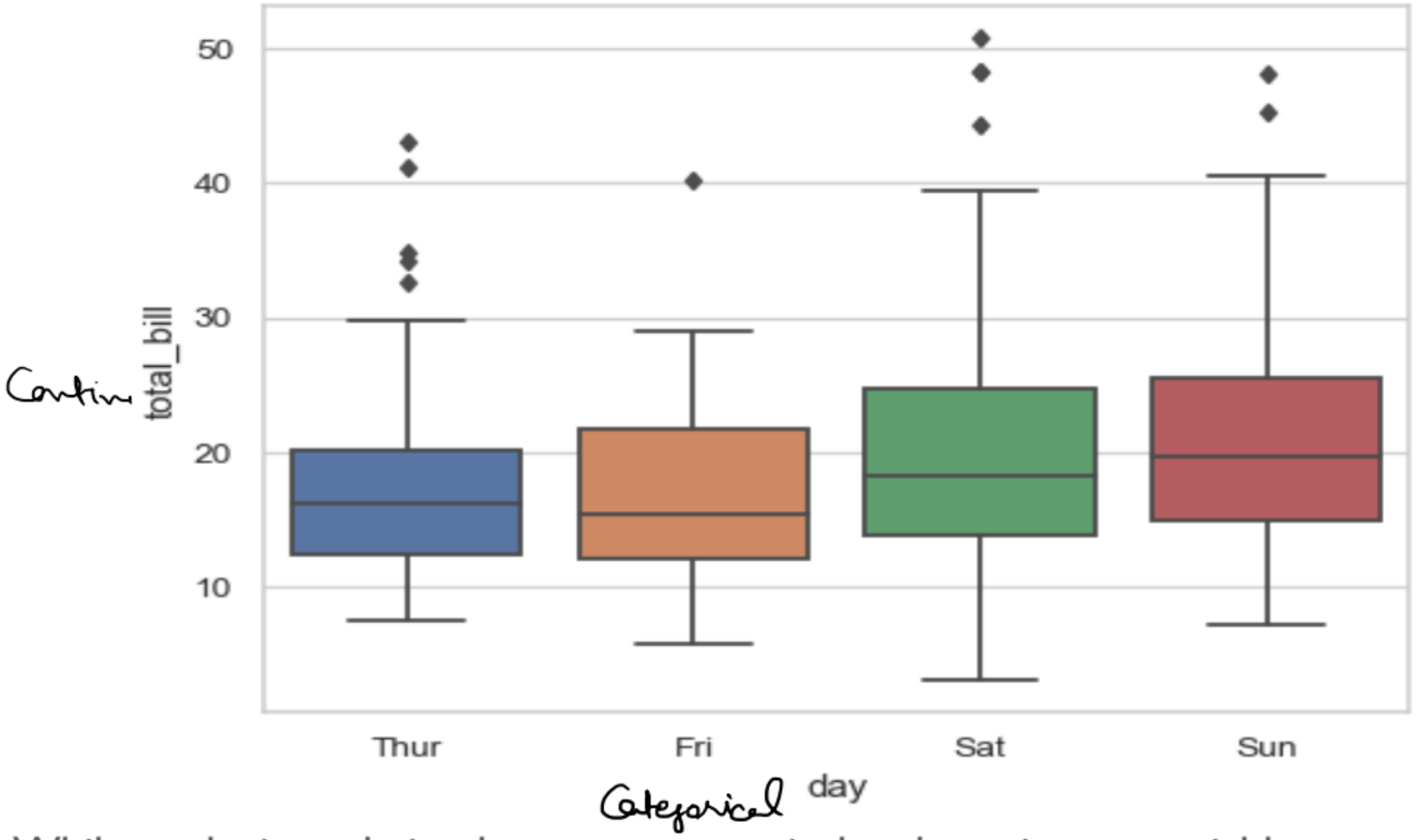


Two-way table: We can start analyzing the relationship by creating a two-way table of count and count%. The rows represents the category of one variable and the columns represent the categories of the other variable. We show count or count% of observations available in each combination of row and column categories.

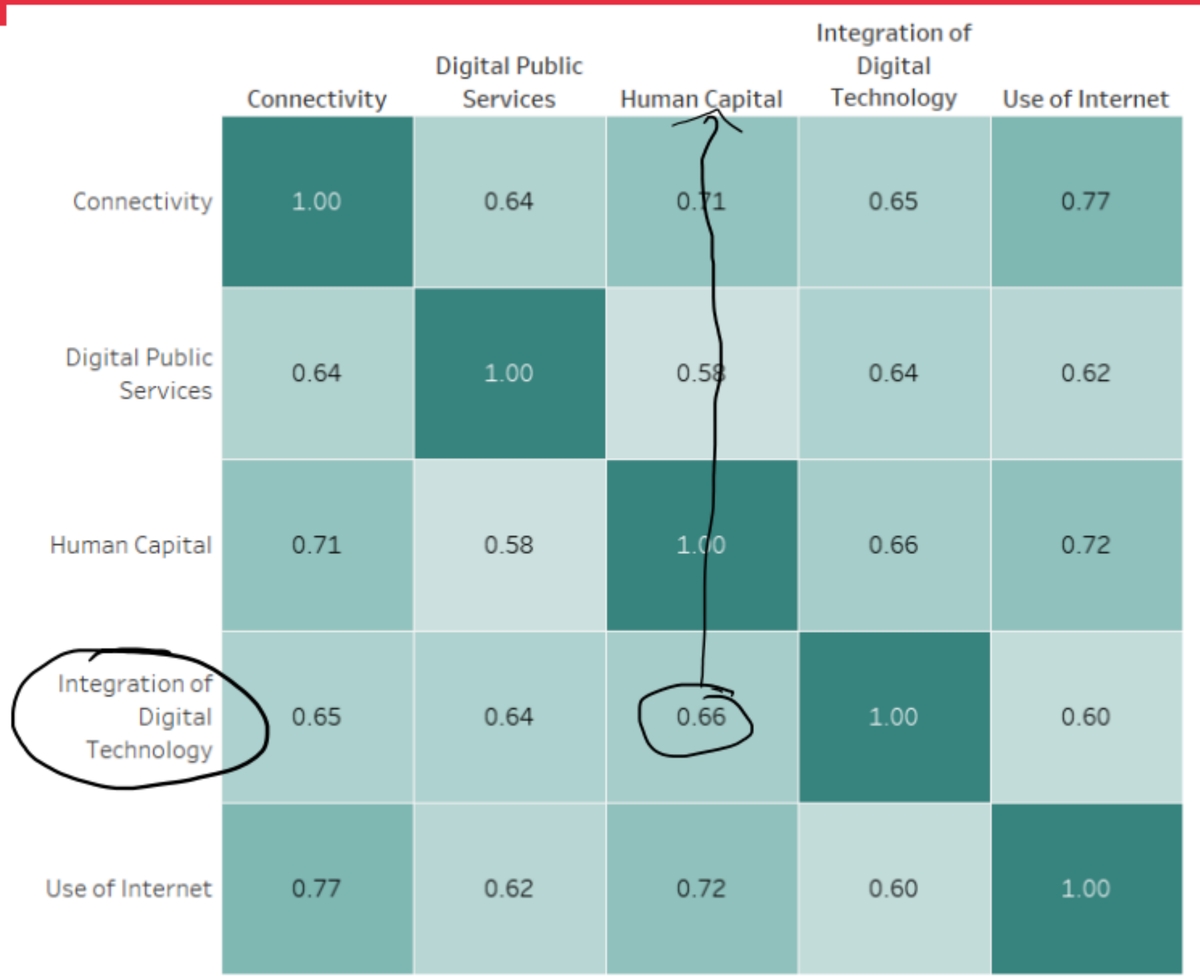
Stacked Column Chart: This method is more of a visual form of Two-way table.

Categorical & continuous

upGrad



While exploring relation between categorical and continuous variables, we can draw box plots for each level of categorical variables



	Connectivity	Digital Public Services	Human Capital	Integration of Digital Technology	Use of Internet
Connectivity	1.00	0.64	0.71	0.65	0.77
Digital Public Services	0.64	1.00	0.58	0.64	0.62
Human Capital	0.71	0.58	1.00	0.66	0.72
Integration of Digital Technology	0.65	0.64	0.66	1.00	0.60
Use of Internet	0.77	0.62	0.72	0.60	1.00

Poll 1

Please select a category which would lead to highest improvement in "**Integration of Digital Technology**" category?

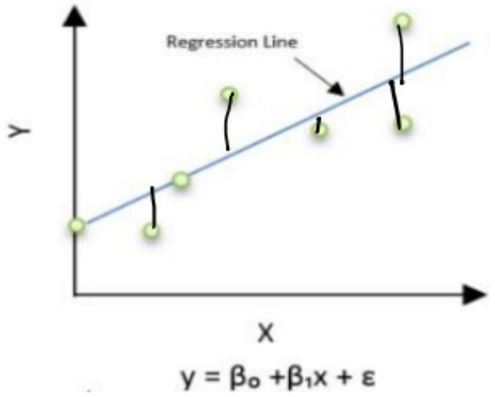
a) Use of internet
b) Human Capital
c) Connectivity
d) Digital public services

	Connectivity	Digital Public Services	Human Capital	Integration of Digital Technology	Use of Internet
Connectivity	1.00	0.64	0.71	0.65	0.77
Digital Public Services	0.64	1.00	0.58	0.64	0.62
Human Capital	0.71	0.58	1.00	0.66	0.72
Integration of Digital Technology	0.65	0.64	0.66	1.00	0.60
Use of Internet	0.77	0.62	0.72	0.60	1.00

Poll 1 (Answer)

Please select a category which would lead to highest improvement in "**Integration of Digital Technology**" category?

a) Use of internet
b) Human Capital
c) Connectivity
d) Digital public services



$$\sum \left(y - \hat{y}\right)^2$$

Minimize

Loss function

What Is Linear Regression?

It is the simplest form of regression. It is a technique in which the **dependent variable is continuous** in nature. The relationship between the dependent variable and independent variables is assumed to be linear in nature.

Simple Linear Regression

$Y = a + bx_1$

- Dependent (response) variable
- One independent (predictor) variable

Multiple Linear Regression

An extension of simple linear regression

$Y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \ldots\ldots b_nx_n$

- Dependent (response) variable
- Multiple independent (predictor) variables

***Simple linear regression**, gets its adjective 'simple', because it concerns the study of only one predictor variable, similarly **multiple linear regression** concerns the study of two or more predictor variables.*

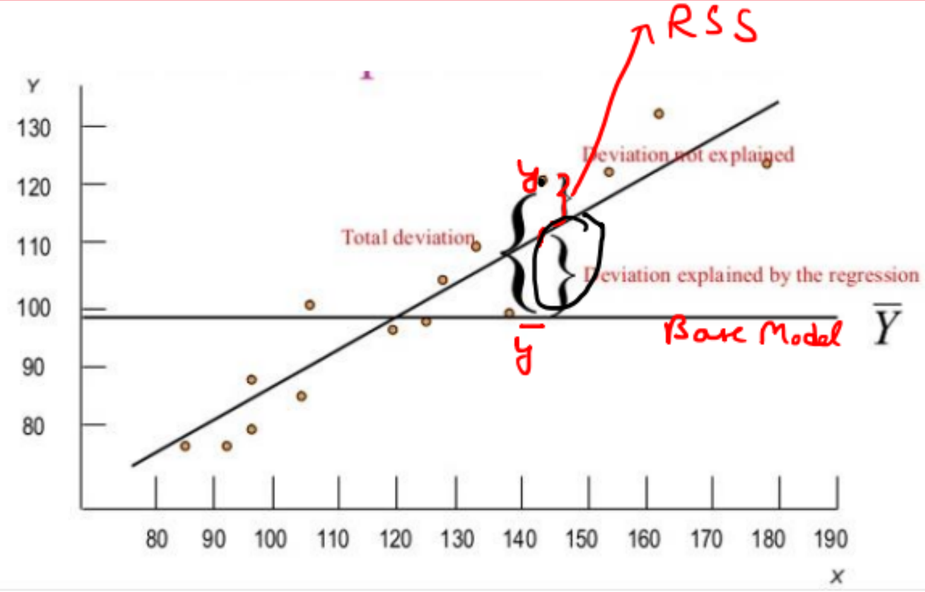
Residual sum of squares (RSS)

Unexplained variation

This gives information about how much the target value varies around the regression line (predicted value)

$$\text{Error} = \sum_{i=1}^{n} (Actual\_output - predicted\_output)^2$$

The best-fit line is obtained by minimizing a quantity called Residual Sum of Squares (RSS) which could be optimized using Gradient Descent to get parameters of the best fit line. It gives more emphasis on larger loss.



Total sum of squares (TSS)

This tells how much the data point move around the mean.

$$\text{Error} = \sum_{i=1}^{n} (\text{Actual\_output} - \text{average\_of\_actual\_output})^2$$

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{Explained\ Variation}{Total\ Variation} = 1 - \frac{\text{Unexp. var (RSS)}}{\text{Total var}}$$

total variation $= \sum(y - \bar{y})^2$

explained variation $= \sum(\hat{y} - \bar{y})^2$

unexplained variation $= \sum(y - \hat{y})^2$

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

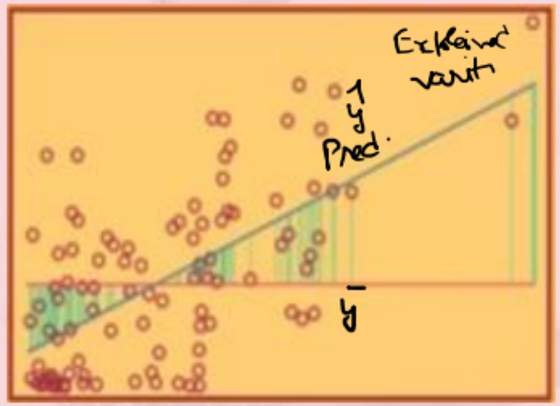
p = Number of predictors

N = Total sample size.

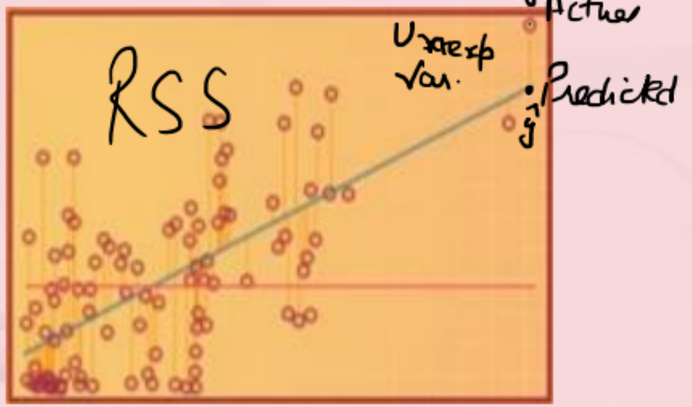
Poll 1

Which of the figures best represent RSS & TSS respectively?

(1)



(2)

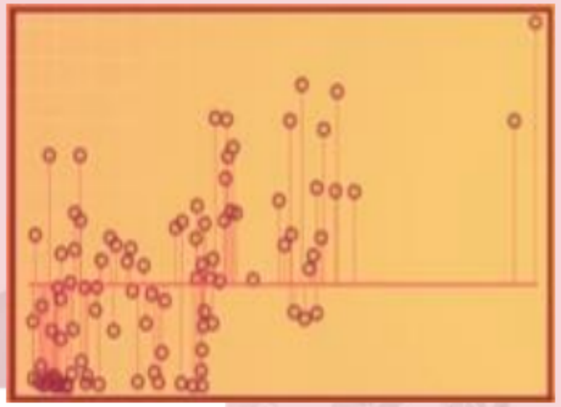


(3)

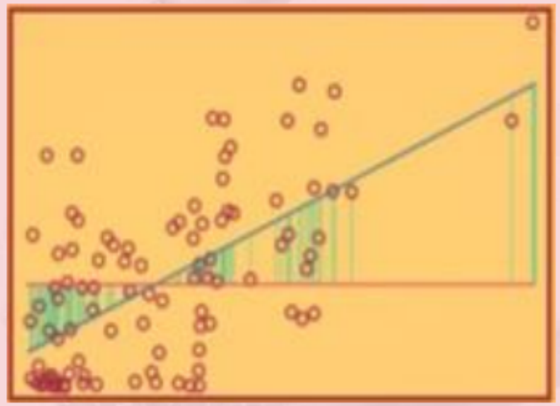
A) 1 & 2
B) 2 & 3
C) 1 & 3
D) 3 & 1

Poll 1(Answer)

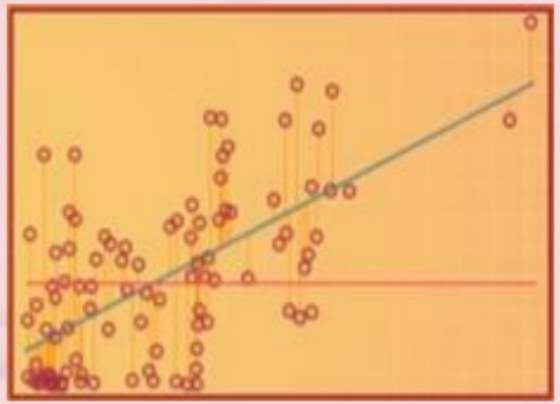
Which of the figures best represent RSS & TSS respectively?



(1)



(2)



(3)

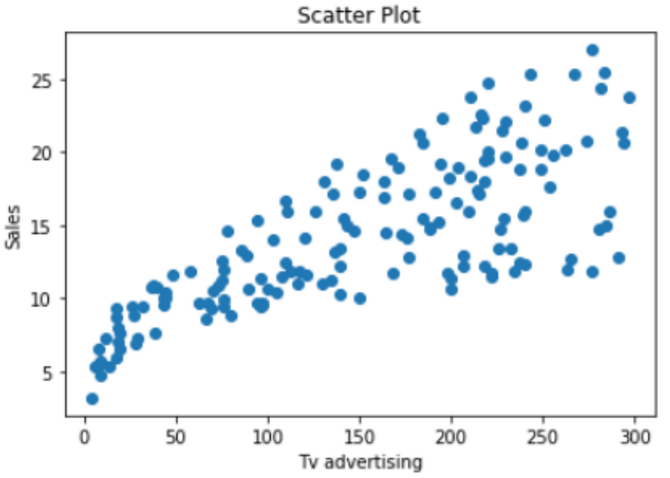
A) 1 & 2
B) 2 & 3
C) 1 & 3
D) 3 & 1

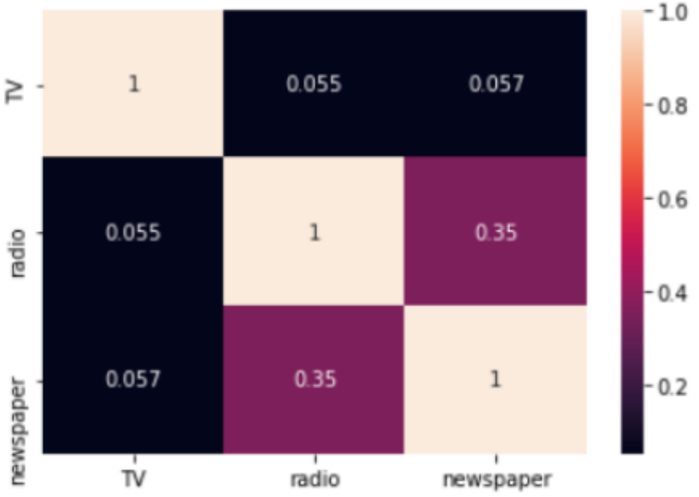
Assumptions of Linear Regression

upGrad

Linear Relationship between the features and target

According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target.





VIF
RFE

Little or no Multicollinearity between the features

Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables. It is therefore a type of disturbance in the data if present weakens the statistical power of the regression model. Pair plots and heat maps(correlation matrix) can be used for identifying highly correlated features.

http://people.duke.edu/~rnau/testing.htm

https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/

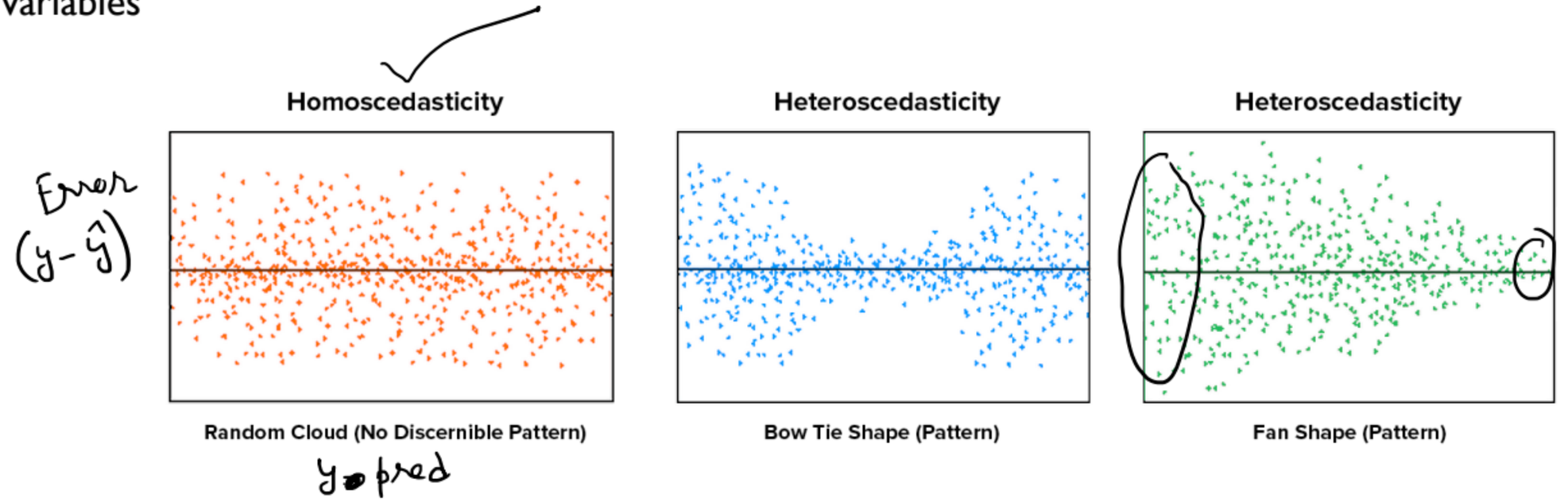
Assumptions of Linear Regression

upGrad

Residual Analysis
(Actual - Predicted)

Homoscedasticity Assumption

Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables



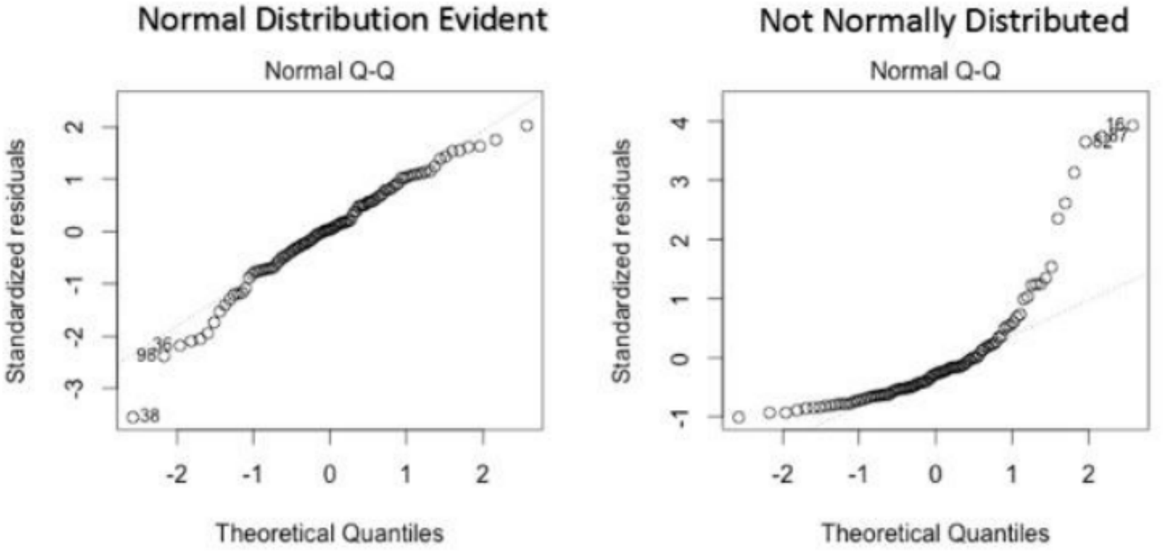
Assumptions of Linear Regression

upGrad

Error terms are normally distributed with mean zero

The fourth assumption is that the error(residuals) follow a normal distribution. One particular repercussion of the error terms not being normally distributed is that the p-values obtained during the hypothesis test to determine the significance of the coefficients become unreliable.

Normal distribution of the residuals can be validated by plotting a q-q plot.



Q1. In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?

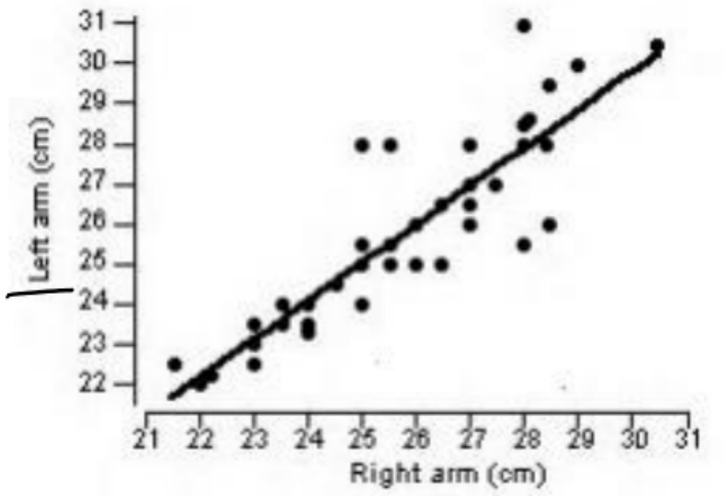
$\hat{y} = b_0 + b_1 x$

a) by 1
b) no change
c) by intercept
d) by its slope

Q2 . The following scatterplot shows the relationship between the left and right forearm lengths (cm) for 55 college students along with the regression line, where y = left forearm length x = right forearm length.

Which of the following linear equation is correct?

a) $\hat{Y} = 1.22 + 0.95x$
b) $\hat{Y} = 1.22 - 0.95x$
c) $\ddot{X} = 1.22 + 0.95y$
d) $\ddot{X} = 1.22 - 0.95y$

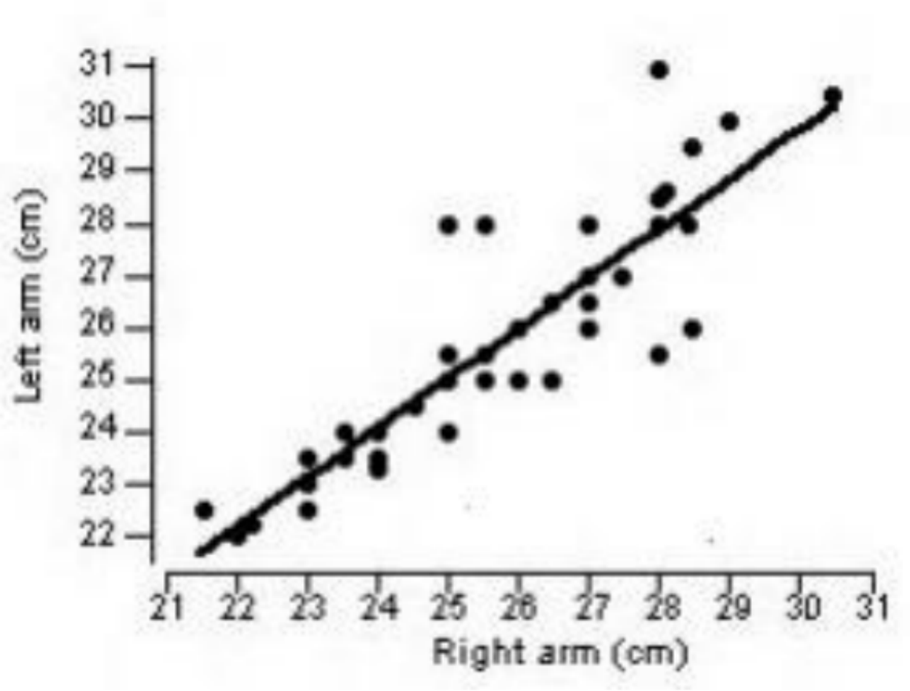


Polls

upGrad

Q3. One of the four choices is the value of the correlation for this situation. The correlation is

a) –0.88
b) 0.00
c) 0.88
d) 1.00



Q4. The proportion of total variation explained by x, R2 , is closest to

a) –78.3%
b) 0.0%
c) 78.3%
d) 100.0%

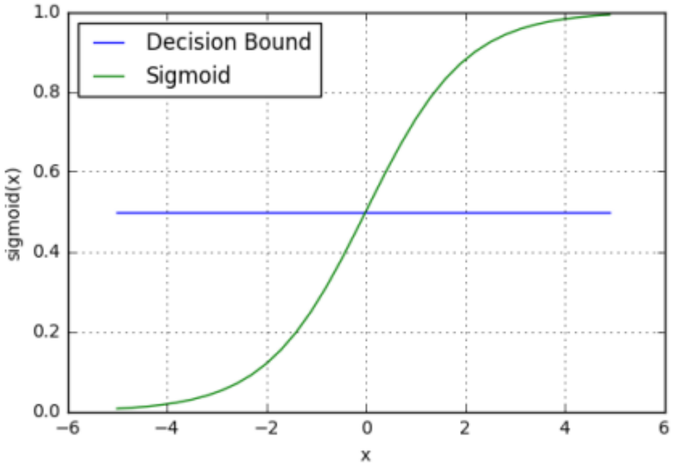
upGrad

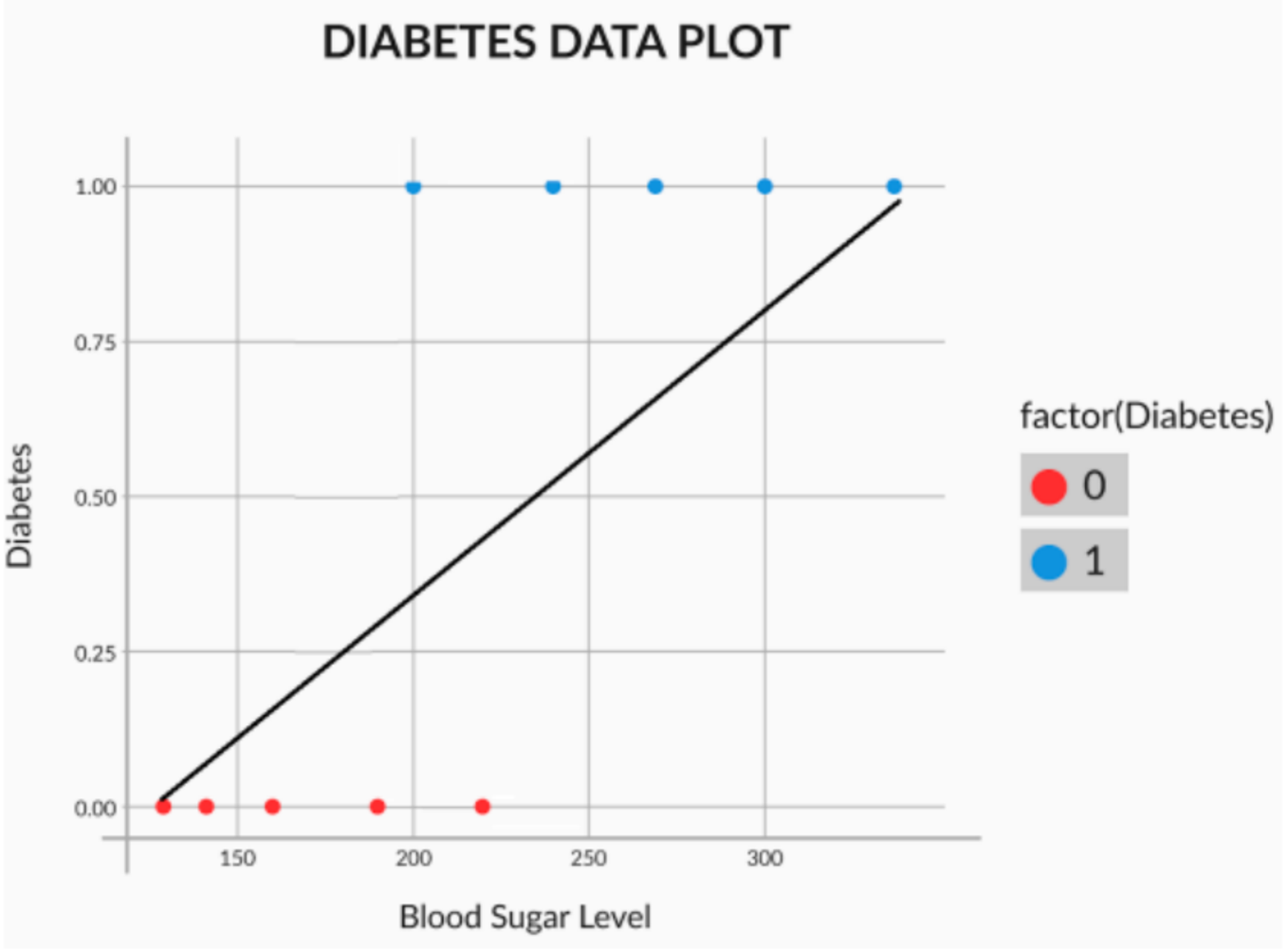
Difference between Linear & Logistic regression

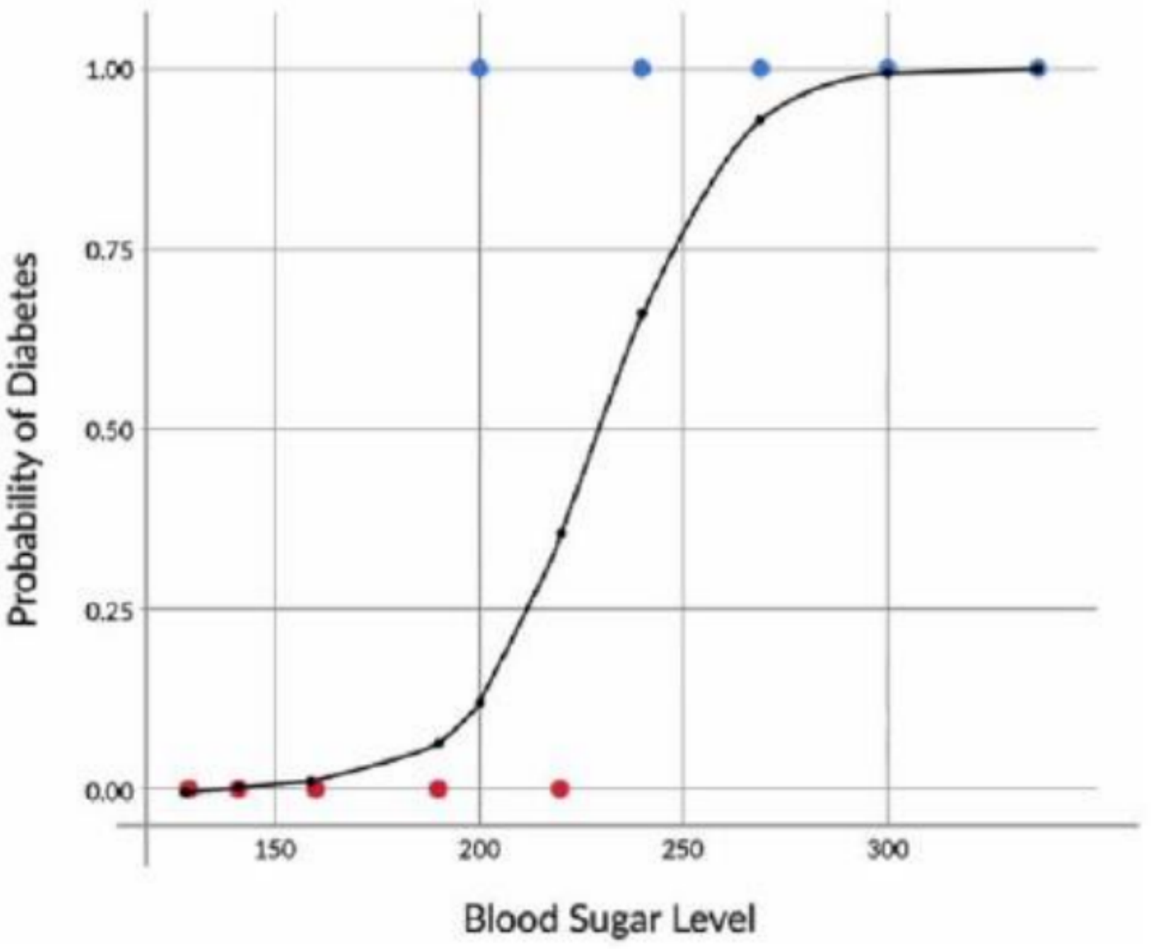
upGrad

What is logistic regression?

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.




DIABETES DATA PLOT
Diabetes
1.00
0.75
0.50
0.25
0.00
150
200
250
300
Blood Sugar Level
factor(Diabetes)
0
1

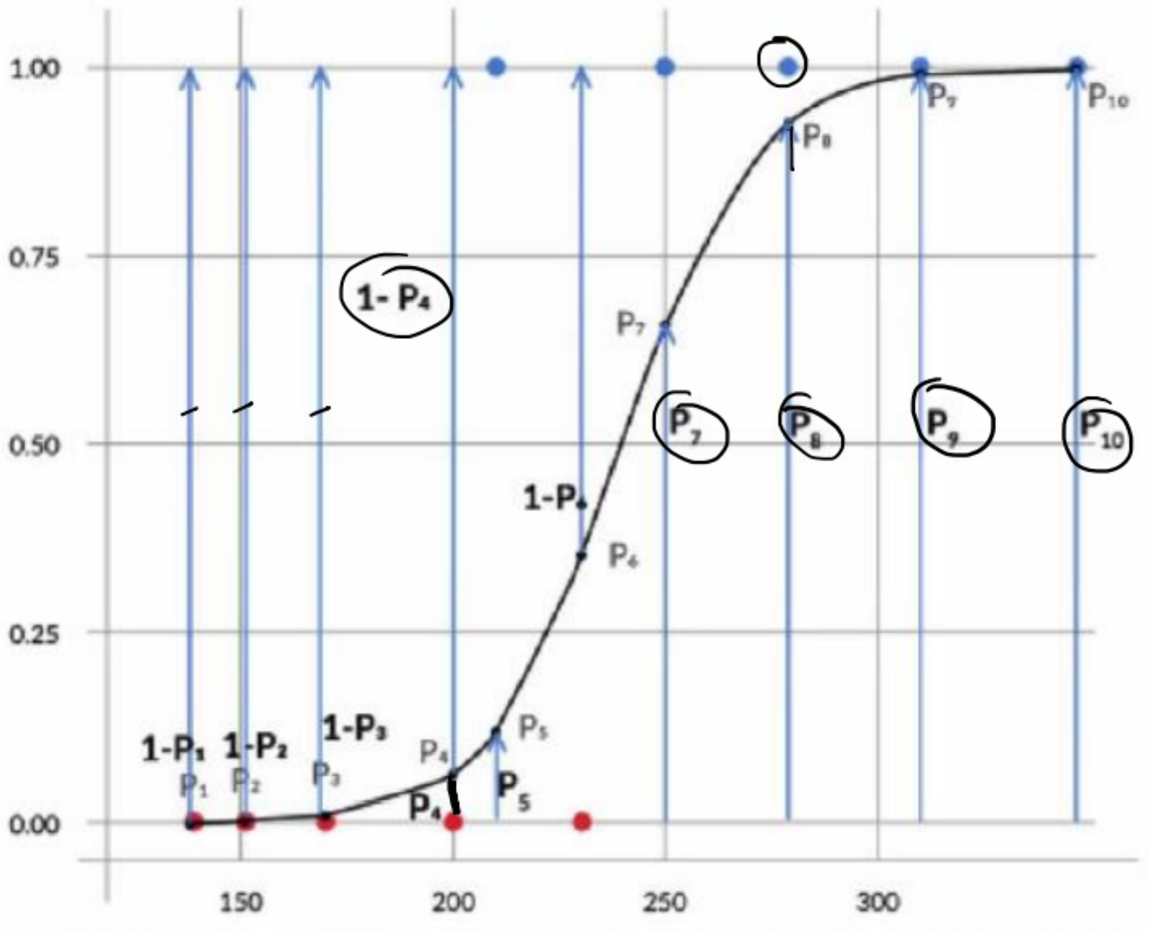

Probability of Diabetes
1.00
0.75
0.50
0.25
0.00
150
200
250
300
Blood Sugar Level

Max. Likelihood

odd ratio $= \frac{P(\text{win})}{1-P(\text{win})}$

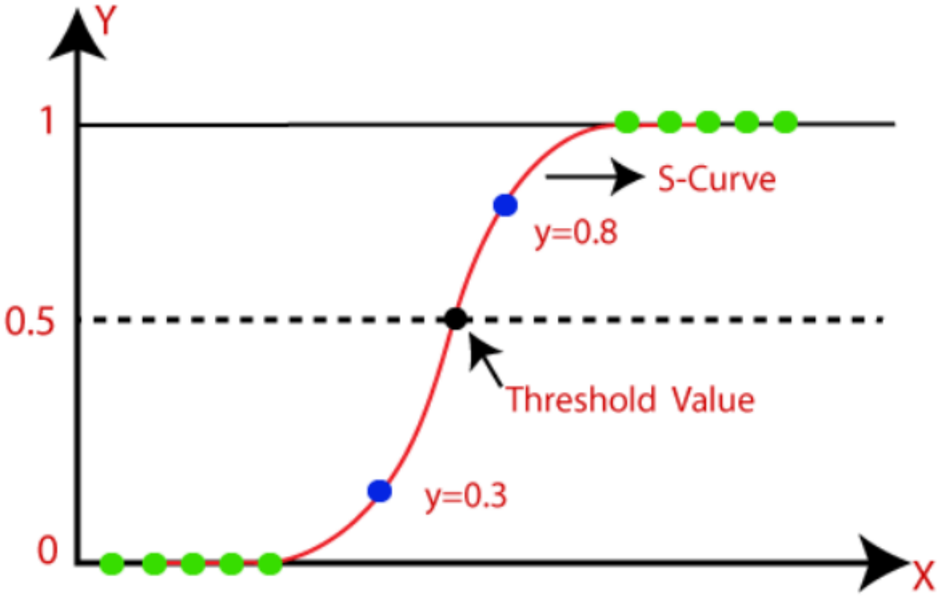
upGrad

Betas obtained by maximizing likelihood



$$P(\text{Diabetes}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \qquad \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x$$

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots)}}$$



Likelihood = $(1-P_1)(1-P_2)(1-P_3)(1-P_4)(P_5)(1-P_6)(P_7)(P_8)(P_9)(P_{10})$

Generally, it is the product of -

[(1-Pᵢ)(1-Pᵢ) ------ for all non-diabetics --------] X [(Pᵢ)(Pᵢ) -------- for all diabetics -------]

Polls

upGrad

Suppose that seven out of 10 males are admitted to an engineering school while three of 10 females are admitted.

Q1. The odds for admitting a male are:

a) 2.5

b) 2.3

c) 0.48

If you are male, the probability of being admitted is 0.7 and the probability of not being admitted is 0.3.

Q2. The odds for admitting a female are:

a) 2.5

b) 2.3

c) 0.42

If you are female it is just the opposite, the probability of being admitted is 0.3 and the probability of not being admitted is 0.7.

Q3. What is the odds ratio for admission for both males and females?

a) 2.5

b) 5.4

c) 4.8

Thus, for a male, the odds of being admitted are 5.44 times as large as the odds for a female being admitted.

Naive Bayes

$P(Play) = 50\%$

$P(Rain) = 30\%$

$$P(Play \mid Rain) = \frac{P(Rain \mid Play) \cdot P(Play)}{P(Rain)}$$

upGrad

What is Naive Bayes?

A Naive Bayes classifier is a probabilistic machine learning model that is based on the Bayes theorem

Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Using Bayes theorem, we can find the probability of **A** happening, given that **B** has occurred. Here, **B** is the evidence and **A** is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$ ✓

Naive Bayes algorithms are mostly used in sentiment analysis, spam filtering, recommendation systems etc. They are fast and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent. In most of the real life cases, the predictors are dependent, this hinders the performance of the classifier.

$P(SPAM) = 2\%$

$P(SPAM \mid lottery \text{ and } Bank\ acount) = \frac{P(lottery \mid SPAM) \cdot P(SPAM)}{P(lottery)}$

$=$

upGrad

Why is “Naive” Bayes naive?

Despite its practical applications, especially in text mining, Naive Bayes is considered "Naive" because it makes an assumption that is virtually impossible to see in real-life data: the conditional probability is calculated as the pure product of the individual probabilities of components. This implies the absolute independence of features – a condition probably never met in real life.

Example : a Naive Bayes classifier that figured out that you liked pickles and ice cream would probably naively recommend you a pickle ice cream.

Model Evaluation

$\frac{P(win)}{1-P(win)}$ $\frac{P(Loss)}{1-P(Loss)}$

upGrad

A confusion matrix shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data. The matrix is *N*x*N*, where *N* is the number of target values (classes). Performance of such models is commonly evaluated using the data in the matrix. The following table displays a 2x2 confusion matrix for two classes (Positive and Negative).

Confusion Matrix	Target: Positive	Target: Negative		
Positive	a	b	*Positive Predictive Value*	a/(a+b)
Negative	c	d	*Negative Predictive Value*	d/(c+d)
	Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	

- **Accuracy** : the proportion of the total number of predictions that were correct.
- **Positive Predictive Value** or **Precision** : the proportion of positive cases that were correctly identified.
- **Negative Predictive Value** : the proportion of negative cases that were correctly identified.
- **Sensitivity** or **Recall** : the proportion of actual positive cases which are correctly identified.
- **Specificity** : the proportion of actual negative cases which are correctly identified.

Validating Model

upGrad

1. Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

$$\frac{100 + 900}{100 + 900 + 5 + 100} = \frac{1000}{1105} = 95\%$$

2. Confusion Matrix

$$TPR = \text{Sensitivity} = \text{Recall} = \text{Detection Rate} = \frac{TP}{TP + FN} = \frac{100}{200} = 50\%$$

> "The confusion matrix shows the ways in which your classification model is confused when it makes predictions."

Actual Class \ Predicted class	P	N
P	True Positives (TP) ✓	False Negatives (FN) ✗
N	False Positives (FP) ✗	True Negatives (TN) ✓

Predicted: Guilty / Not guilty

Actual: Guilty / Not guilty

	Guilty	Not guilty
Guilty	100	100
Not guilty	500	900

$$\text{Recall} = \frac{100}{110} = 99\%$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{100}{100 + 500} = \frac{1}{6} = 16\%$$

HIV Test		
Actual/Predicted		
	Infected	Not Infected
Infected	100	100
Not Infected	5	900

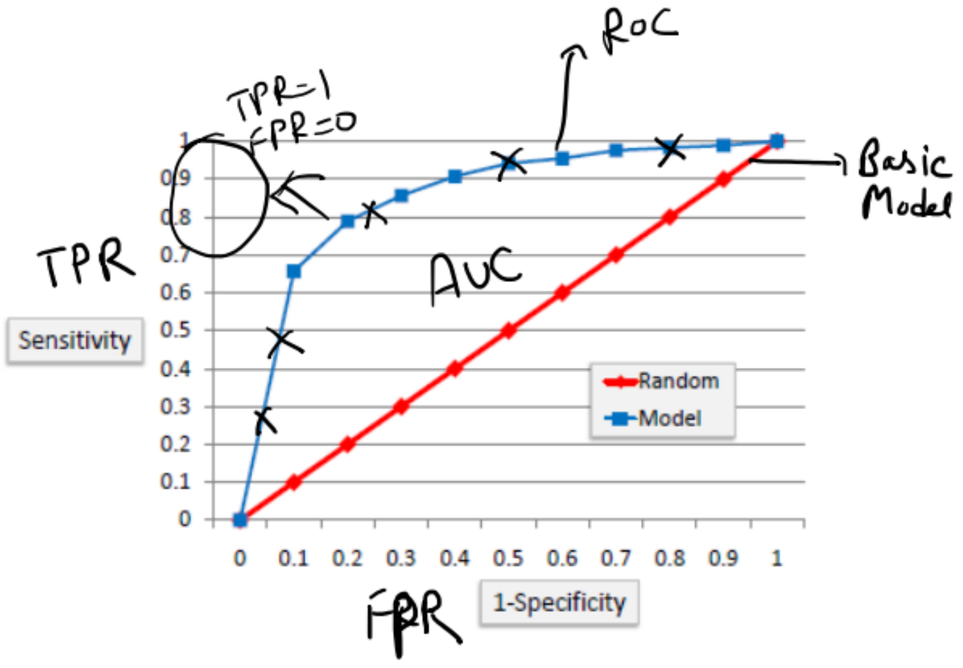
Model Evaluation

upGrad

Threshold = 0.50

ROC Curves summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds

It's often used as a proxy for the trade-off between the sensitivity of the model (true positives) vs the fall-out or the probability it will trigger a false alarm (false positives).



$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

FPR = 1 − Specificity

TPR → Highest

FPR = lowest

upGrad

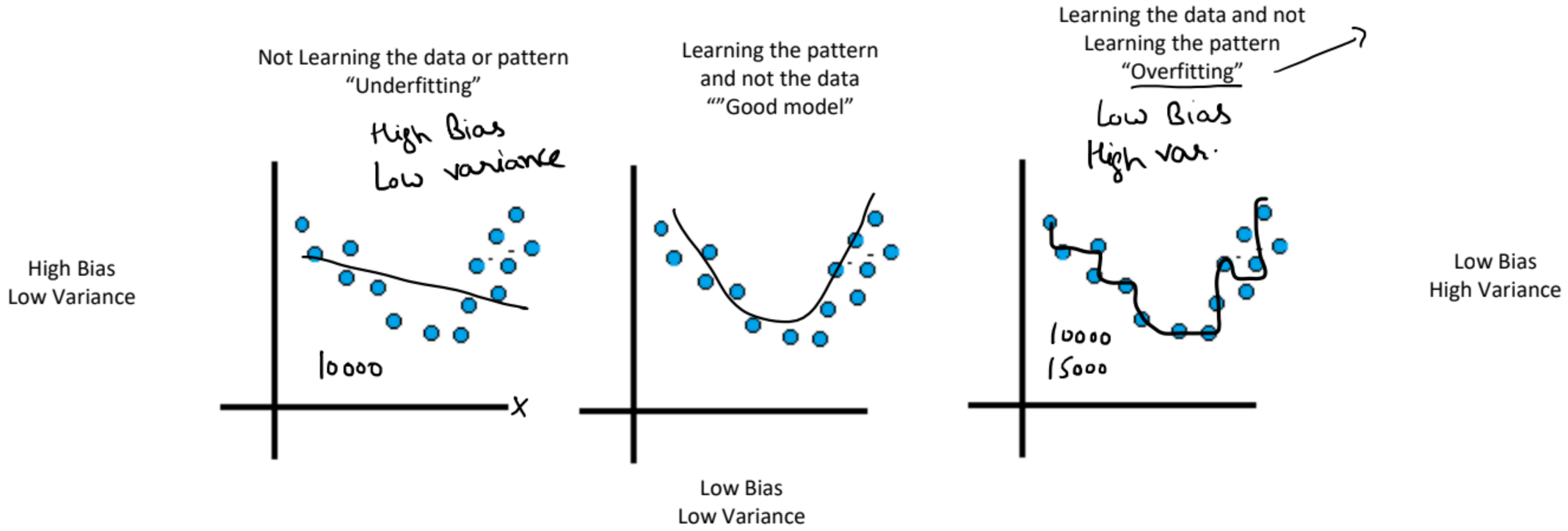
What is Bias Variance Tradeoff ?

Overfitting and Underfitting

Error = Bias + Variance

upGrad

What is Bias and Variance

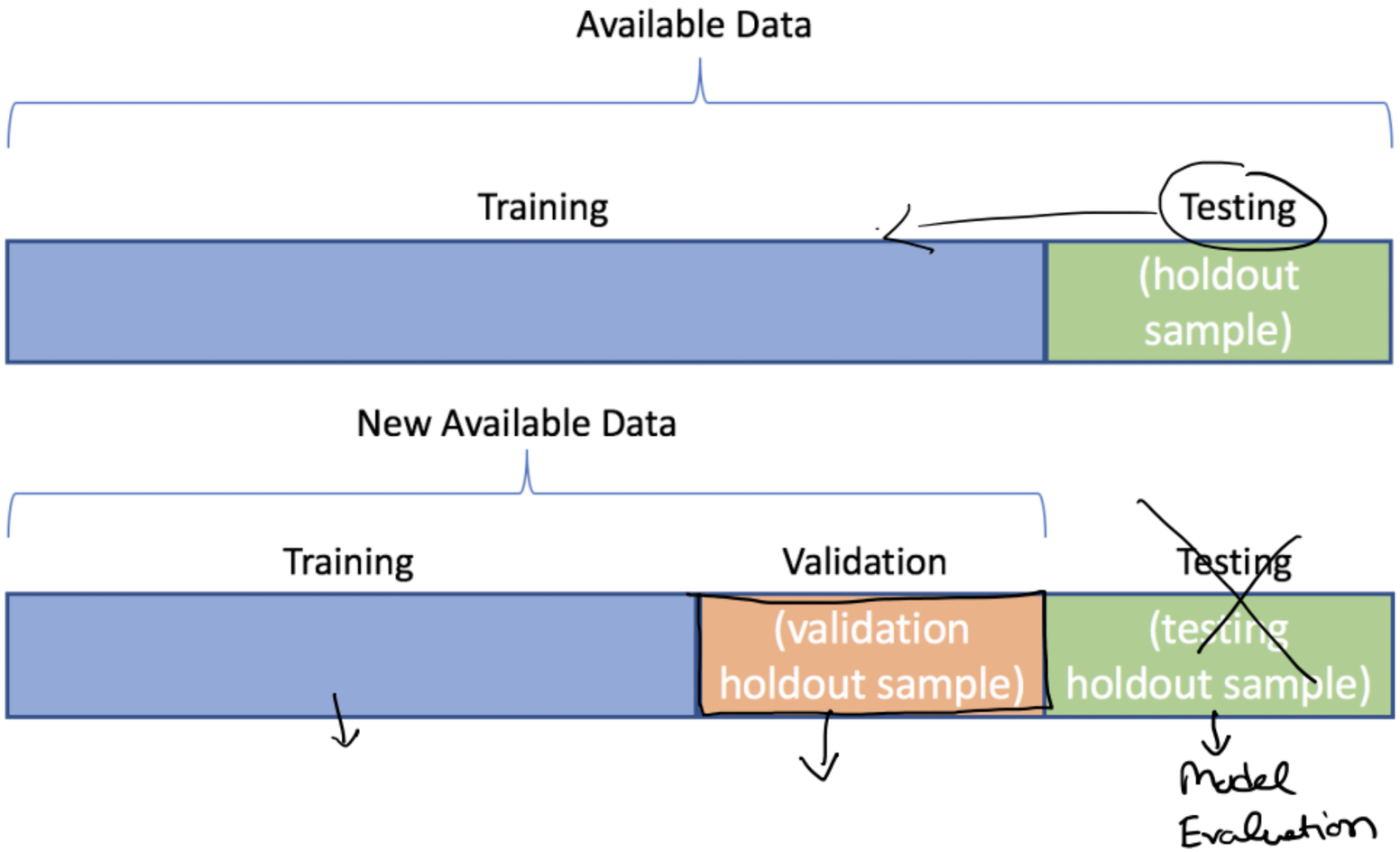


Bias - Amount of Error that the model is making on train data

Variance - Amount that the model(target variable) will change given different training data

Problems with manual hyperparameter tuning

upGrad



The problems with manual hyperparameter tuning are:

- Split into train and test sets: Tuning a hyperparameter makes the model 'see' the test data.
- Split into train, validation, test sets: The validation data would eat into the training set.

However, in cross-validation, you split the data into train and test sets and train multiple models by sampling the train set. Finally, you just use the test set to test the hyperparameter once.

CROSS VALIDATION – THE IDEAL PROCEDURE

upGrad

1. Divide data into three sets, training, validation and test sets

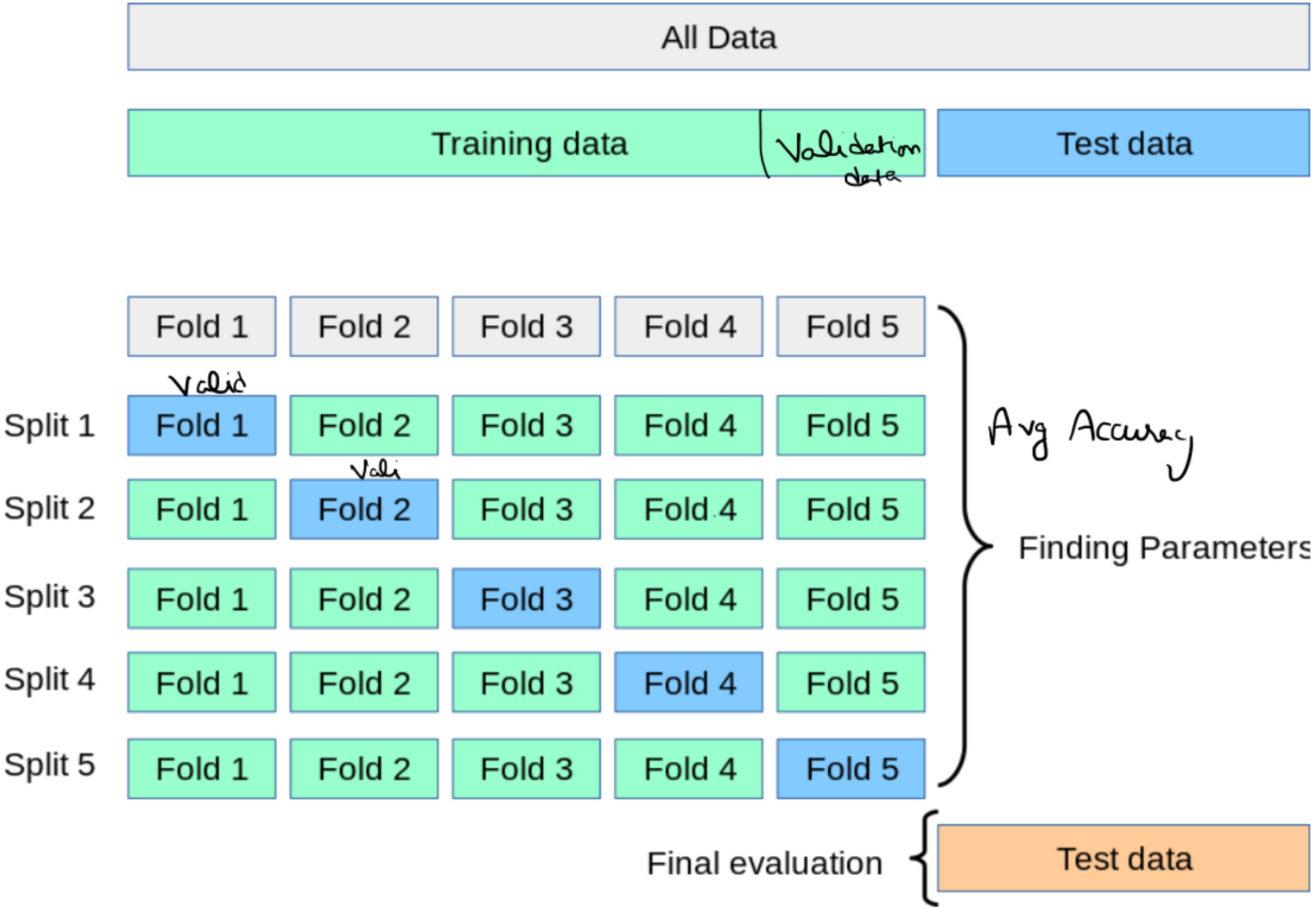
2. Find the optimal model on the training set, and use the test set to check its predictive capability

3. See how well the model can predict the test set

4. The validation error gives an unbiased estimate of the predictive power of a model

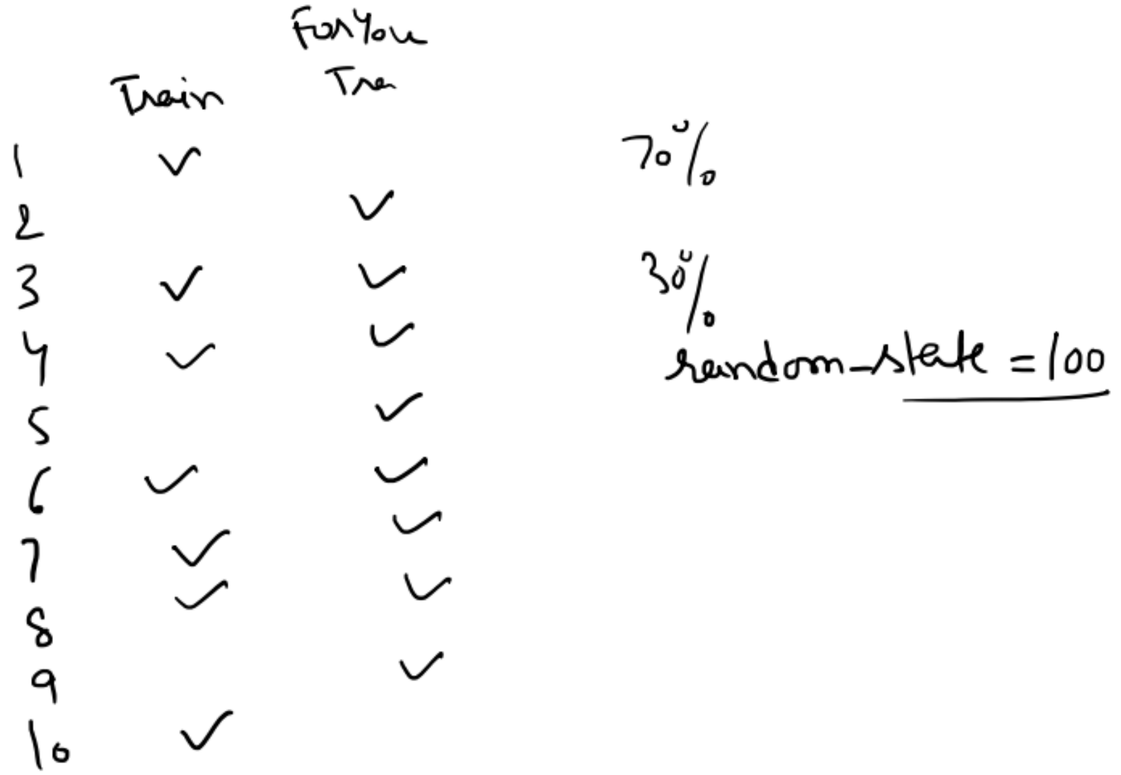
K-fold Cross Validation Example

upGrad



upGrad

#LifeKoKaroLift



Thank You!

References :
https://www.superprof.co.uk/resources/academic/maths
https://courses.lumenlearning.com/introstats1/chapter/null-and-alternative-hypotheses/
https://www.ck12.org/book/CK-12-Advanced-Probability-and-Statistics-Concepts/
https://www.indiabix.com/data-interpretation