Interview Preparation and Doubt Resolution

-Rahul Nandanwar

Statistics

- Types of Statistics
- Hypothesis Testing
- Use of Hypothesis Testing
- Important terms in Hypothesis testing
- Tests in Hypothesis Testing
- Errors in Hypothesis Testing

Question - **How can we assess the statistical significance of an insight?**

Types of Statistics

- Descriptive Statistics - Descriptive statistics is a term given to the analysis of data that helps to describe, show and summarize data in a meaningful way. It is a simple way to describe our data. Descriptive statistics is very important to present our raw data ineffective/meaningful way using numerical calculations or graphs or tables. This type of statistics is applied on already known data.

- Inferential Statistics - In inferential statistics predictions are made by taking any group of data in which you are interested. It can be defined as a random sample of data taken from a population to describe and make inference about the population. Any group of data which includes all the data you are interested is known as population. It basically allows you to make predictions by taking a small sample instead of working on whole population.

Question – **What are the types of statistics?**

Hypothesis Testing

- Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter.
- In statistical analysis, we have to make decisions about the hypothesis. These decisions include deciding if we should accept the null hypothesis or if we should reject the null hypothesis. Every test in hypothesis testing produces the significance value for that particular test. In Hypothesis testing, if the significance value of the test is greater than the predetermined significance level, then we accept the null hypothesis. If the significance value is less than the predetermined value, then we should reject the null hypothesis.

Question – **What is Hypothesis Testing?**

Important terms in Hypothesis testing

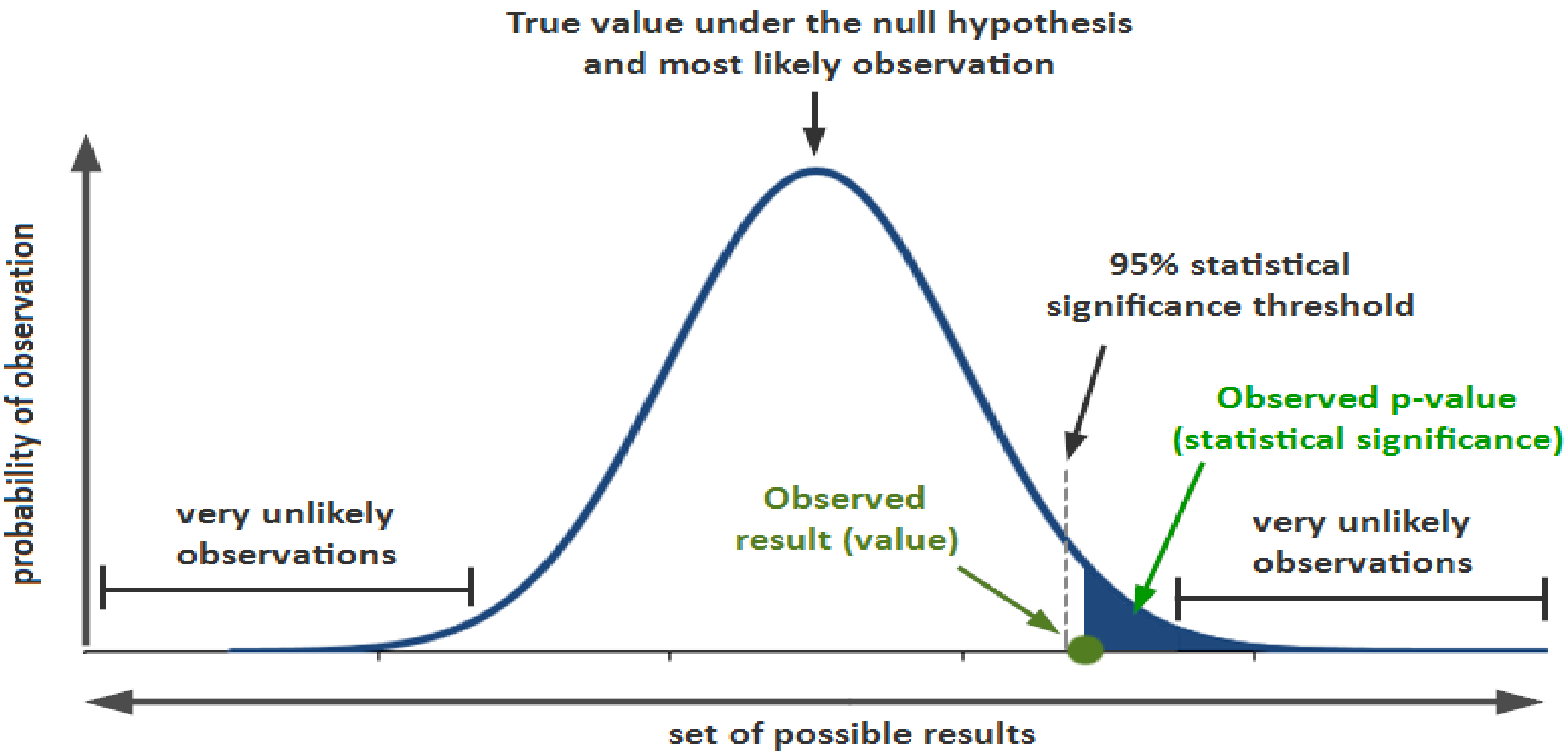
- Null hypothesis(H0): Null hypothesis is a statistical hypothesis that assumes that the observation is due to a chance factor. Null hypothesis is denoted by; H0: μ1 = μ2, which shows that there is no difference between the two population means.

- Alternative hypothesis(H1): Contrary to the null hypothesis, the alternative hypothesis shows that observations are the result of a real effect.

- Level of significance: Refers to the degree of significance in which we accept or reject the null-hypothesis. 100% accuracy is not possible for accepting or rejecting a hypothesis, so we therefore select a level of significance that is usually 5%.

Question – **Terms in Hypothesis testing?**

Probability & Statistical Significance Explained



- Type I error: When we reject the null hypothesis, although that hypothesis was true. Type I error is denoted by alpha. In hypothesis testing, the normal curve that shows the critical region is called the alpha region.

- Type II errors: When we accept the null hypothesis but it is false. Type II errors are denoted by beta. In Hypothesis testing, the normal curve that shows the acceptance region is called the beta region.

- Power: Usually known as the probability of correctly accepting the null hypothesis.

Question – **What is Type – I & Type – II error in hypothesis testing?**

P-value

- Statistical hypothesis test may return a value called p or the p-value. This is a quantity that we can use to interpret or quantify the result of the test and either reject or fail to reject the null hypothesis. This is done by comparing the p-value to a threshold value chosen beforehand called the significance level.
- The significance level is often referred to by the Greek lower case letter alpha.
- A common value used for alpha is 5% or 0.05. A smaller alpha value suggests a more robust interpretation of the null hypothesis, such as 1% or 0.1%.
- The p-value is compared to the pre-chosen alpha value. A result is statistically significant when the p-value is less than alpha. This signifies a change was detected: that the default hypothesis can be rejected.
- **If** p-value > alpha: Fail to reject the null hypothesis (i.e. not significant result).
- **If** p-value <= alpha: Reject the null hypothesis (i.e. significant result).

Different Tests

- Z-test - In a z-test, the sample is assumed to be normally distributed. A z-score is calculated with population parameters such as **“population mean”** and **“population standard deviation”** and **is used to validate a hypothesis that the sample drawn belongs to the same population.**

  Null: Sample mean is same as the population mean

  Alternate: Sample mean is not same as the population mean

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

  Null: True difference between these group means is zero.

  Alternate: True difference is different from zero

- Chi Square - The Chi Square statistic is commonly used for testing relationships between categorical variables.

  Null - No relationship exists on the categorical variables in the population

  Alternate - Assumes that there is an association between the two variable

- ANOVA Test - ANOVA is a statistical technique that assesses potential differences in a scale-level dependent variable by a nominal-level variable having 2 or more categories.

  Null - There is no difference in means

  Alternate - That the means are not all equal

- Shapiro-Wilk test
- Durban-Watson test

Question – **What are the tests in Hypothesis testing?**

Statistical Power

- Statistical power, or the power of a hypothesis test is the probability that the test correctly rejects the null hypothesis.
- That is, the probability of a true positive result. It is only useful when the null hypothesis is rejected.
- The higher the statistical power for a given experiment, the lower the probability of making a Type II (false negative) error. That is the higher the probability of detecting an effect when there is an effect. In fact, the power is precisely the inverse of the probability of a Type II error.
- Power = 1 - Type II Error
- Pr(True Positive) = 1 - Pr(False Negative)
- **Low Statistical Power**: Large risk of committing Type II errors, e.g. a false negative.
- **High Statistical Power**: Small risk of committing Type II errors.

Question - **What is Statistical power?**

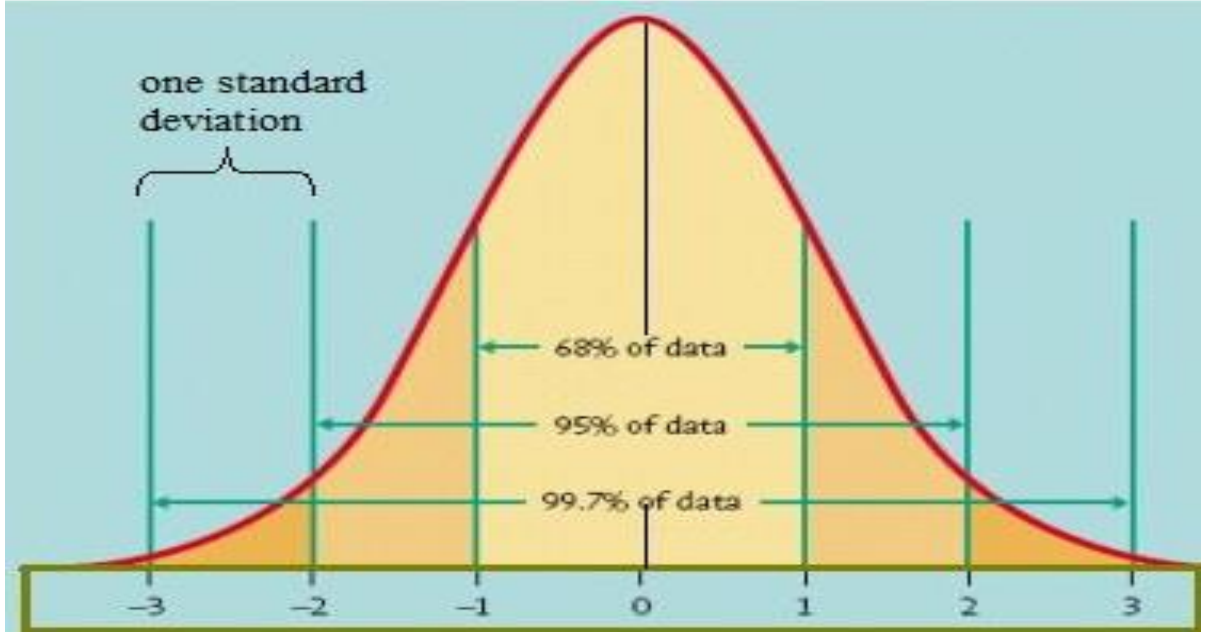
Selection Bias

- Selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved
- Selection bias is the term used to describe the situation where an analysis has been conducted among a subset of the data (a sample) with the goal of drawing conclusions about the population, but the resulting conclusions will likely be wrong (biased), because the subgroup differs from the population in some important way. Selection bias is usually introduced as an error with the sampling and having a selection for analysis that is not properly randomized.
- Mechanisms for avoiding selection biases include:
- Using random methods when selecting subgroups from populations.
- Ensuring that the subgroups selected are equivalent to the population at large in terms of their key characteristics (this method is less of a protection than the first, since typically the key characteristics are not known).

Question - **What is Selection bias with regards to a dataset, not variable selection?**

Normal Distribution

The normal distribution is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper off equally in both directions.



Question – Different type of Distributions, What is Normal/Gaussian Distribution?

EDA

- 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.
- Scatter Plot
- Bar Graph
- Histogram
- Box plot
- Violin plot
- Q-Q plot
- Heat map
- Pair plot

Question – Explain on any plot?

- Difference between a box plot and a histogram

Bivariate Analysis

When we talk about bivariate analysis, it means analyzing 2 variables.

- **Numerical vs. Numerical**

  1. Scatterplot
  2. Line plot
  3. Heat map for correlation
  4. Joint plot

- **Categorical vs. Numerical**

  1. Bar chart
  2. Violin plot
  3. Categorical box plot
  4. Swarm plot

- **Two Categorical Variables**

  1. Bar chart
  2. Grouped bar chart
  3. Point plot

Data Preprocessing Techniques

- Handling Missing Values
- Removing duplicates
- Outlier Treatment
- Normalizing and Scaling(Numerical Variables)
- Encoding Categorical variables(Dummy Variables)
- Feature Transformations

Question – What is an outlier? Explain how you might screen for outliers?
What are all techniques to handle missing values and categorical encoding techniques?

Normalization vs. Standardization Feature Scaling

- **Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.**
- Here’s the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- **Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.**

$$X' = \frac{X - \mu}{\sigma}$$

Question – What is Standardization & Normalization?
Why & When Feature scaling is required in Machine Learning?

Feature Engineering Techniques

- Indicator Variables
  - Indicator variable from thresholds
  - Indicator variable from multiple features
  - Indicator variable for special events
  - Indicator variable for groups of classes
- Interaction Features
  - Sum of two features
  - Difference between two features
  - Product of two features
  - Quotient of two features
- Feature Representation
  - Date and time features
  - Numeric to categorical mappings
  - Grouping sparse classes
  - Creating dummy variables

https://elitedatascience.com/feature-engineering-best-practices

Machine Learning Model Development Steps

- Data Selection
- Data Description
- EDA
- Data Transformation
- Selection of ML algorithms
- Data Standardization & Normalization
- Train-Test split
- Model training
- Model Evaluation
- Hyper parameter Training
- Deployment

Question – What are the basic steps to build any Machine learning model?

Machine Learning

- Supervised Machine learning

  Regression Analysis, Classification(Predictive Analytics)

- Unsupervised Machine learning

  Clustering algorithms

- Reinforcement Machine learning

  Reinforcement learning is the training of machine learning models to make a sequence of decisions

Question – What are all the types of Machine Learning?

Linear Regression

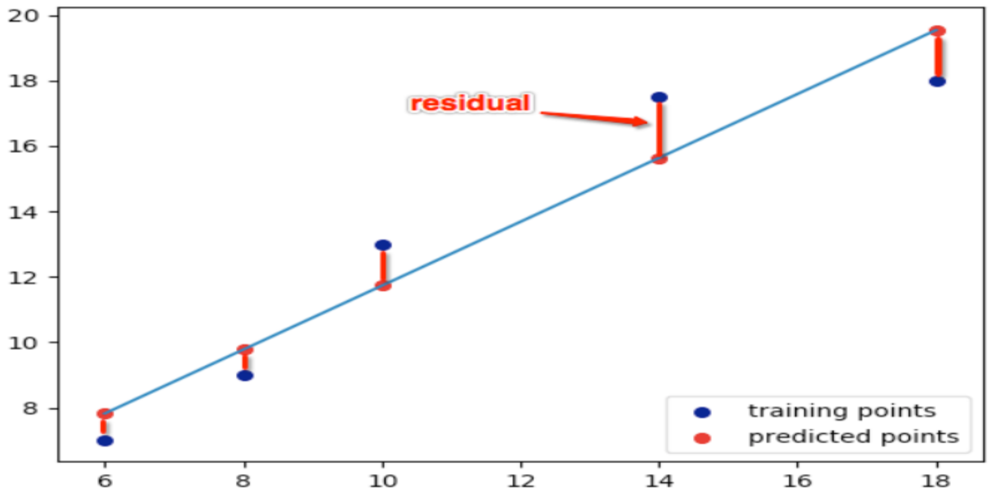
- The term "linearity" in algebra refers to a linear relationship between two or more variables. If we draw this relationship in a two dimensional space (between two variables, in this case), we get a straight line.
- Linear Regression Equation - y = mx + b
- Multiple Linear Regression Equation - $y=\beta 0+\beta 1x1+...+\beta nxn$
- The best fit line is obtained by minimizing the *residual*.
- In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

Question – What is Linear Regression? How do you know it is Best Fit Line?

https://www.upgrad.com/blog/machine-learning-interview-questions-answers-ii/

Loss Function in Linear Regression

- MSE - **MSE** is the average of the squared error that is used as the loss function for least squares regression: It is the sum, over all the data points, of the square of the difference between the predicted and actual target variables, divided by the number of data points.
- RMSE - RMSE is the square root of **MSE**.
- MAE - Mean Absolute Error (**MAE**) refers to a the results of measuring the difference between two continuous variables.



Question – What are evaluation metric in regression models?

Assumptions in Linear Regression

- It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the ‘linearity assumption’.
- Assumptions about the residuals:
  - Normality assumption: It is assumed that the error terms, $\varepsilon^{(i)}$, are normally distributed.
  - Zero mean assumption: It is assumed that the residuals have a mean value of zero.
  - Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ^2 This assumption is also known as the assumption of homogeneity or homoscedasticity.
  - Independent error assumption: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.
- Assumptions about the estimators:
  - The independent variables are measured without error.
  - The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.

Question – What are the Assumptions of Linear Regression?

R square and adjusted R square

- R-square statistics - R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model.

- Adjusted R-squared statistic - The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable.

- Problems with R-squared statistic - The R-squared statistic isn’t perfect. In fact, it suffers from a major flaw. Its value never decreases no matter the number of variables we add to our regression model. That is, even if we are adding redundant variables to the data, the value of R-squared does not decrease. It either remains the same or increases with the addition of new independent variables. This clearly does not make sense because some of the independent variables might not be useful in determining the target variable. Adjusted R-squared deals with this issue.

Question – What is difference between R-square and Adjusted R-square?

Differences between correlation and regression

- Regression establishes how *x* causes *y* to change, and the results will change if *x* and *y* are swapped. With correlation, *x* and *y* are variables that can be interchanged and get the same result.
- Correlation is a single statistic, or data point, whereas regression is the entire equation with all of the data points that are represented with a line.
- Correlation shows the relationship between the two variables, while regression allows us to see how one affects the other.
- The data shown with regression establishes a cause and effect, when one changes, so does the other, and not always in the same direction. With correlation, the variables move together.

Question – What is difference between Correlation and Regression?
https://learn.g2.com/correlation-vs-regression#:~:text=Correlation%20is%20a%20single%20statistic,how%20one%20aff

Regularization Techniques in Linear Regression

- Ridge Regression - **Ridge regression** adds “*squared magnitude*” of coefficient as penalty term to the loss function. Here the *highlighted* part represents L2 regularization element.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

- Lasso Regression - LASSO full form is *Least Absolute Shrinkage Selector Operator* . It is quite similar to ridge regression. LASSO adds “*absolute value of magnitude*” of coefficient as penalty term to the loss function. This way,

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

How Can I Deal With Multicollinearity?

- If multicollinearity is a problem in your model -- if the VIF for a factor is near or above 5 -- the solution may be relatively simple.
- **Remove highly correlated predictors from the model.** If you have two or more factors with a high VIF, remove one from the model. Because they supply redundant information, removing one of the correlated factors usually doesn't drastically reduce the R-squared. Consider using stepwise regression, best subsets regression, or specialized knowledge of the data set to remove these variables. Select the model that has the highest R-squared value.
- **Use Partial Least Squares Regression (PLS) or Principal Components Analysis**, regression methods that cut the number of predictors to a smaller set of uncorrelated components.

Logistic Regression

- Logistic regression is famous because it can convert the values of logits (logodds), which can range from -infinity to +infinity to a range between 0 and 1.
- Logistic Regression Equation - $f(z) = 1/(1+e^{-z})$
- Multiple Logistic Regression Equation - $f(z) = 1/(1+e^{-(\alpha+1X1+2X2+....+kXk)})$
- Logistic regression uses functions called the logit functions,that helps derive a relationship between the dependent variable and independent variables by predicting the probabilities or chances of occurrence.

Question – What is Logistic Regression?
https://www.upgrad.com/blog/machine-learning-interview-questions-answers-logistic-regression/

Why can’t linear regression be used in place of logistic regression for binary classification?

- **Distribution of error terms**: The distribution of data in case of linear and logistic regression is different. Linear regression assumes that error terms are normally distributed. In case of binary classification, this assumption does not hold true.
- **Model output**
- **Variance of Residual errors**: Linear regression assumes that the variance of random errors is constant. This assumption is also violated in case of logistic regression.

Cost Function in Logistic Regression

- Log loss

  For logistic regression, the Cost function is defined as

$$Cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & \text{if } y = 1 \\ -log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Question – Why can’t we use Mean Square Error (MSE) as a cost function for logistic regression?

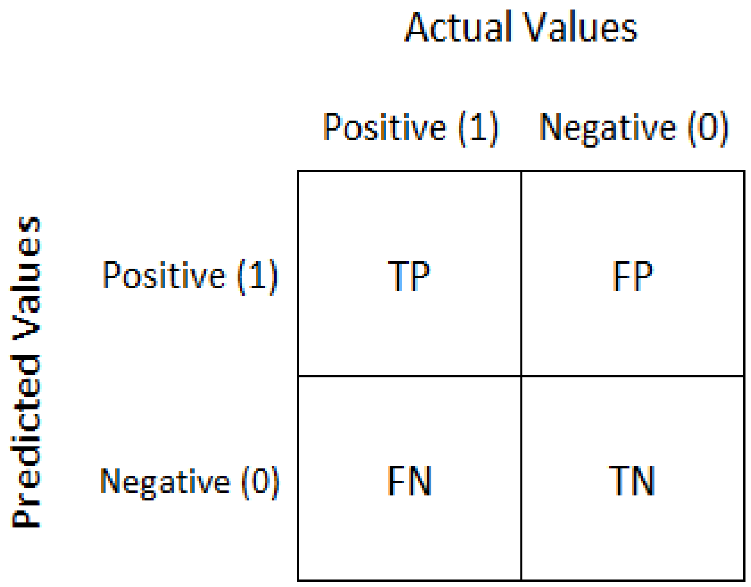
Decision Boundary

- Since logistic regression prediction function returns a probability between 0 and 1, in order to predict which class this data belongs we need to set a threshold. For each data point, if the estimated probability is above this threshold, we classify the point into class 1, and if it's below the threshold, we classify the point into class 2.

Question - Is the decision boundary linear or nonlinear in the case of a logistic regression model?

Evaluation Matrices for Classification Model

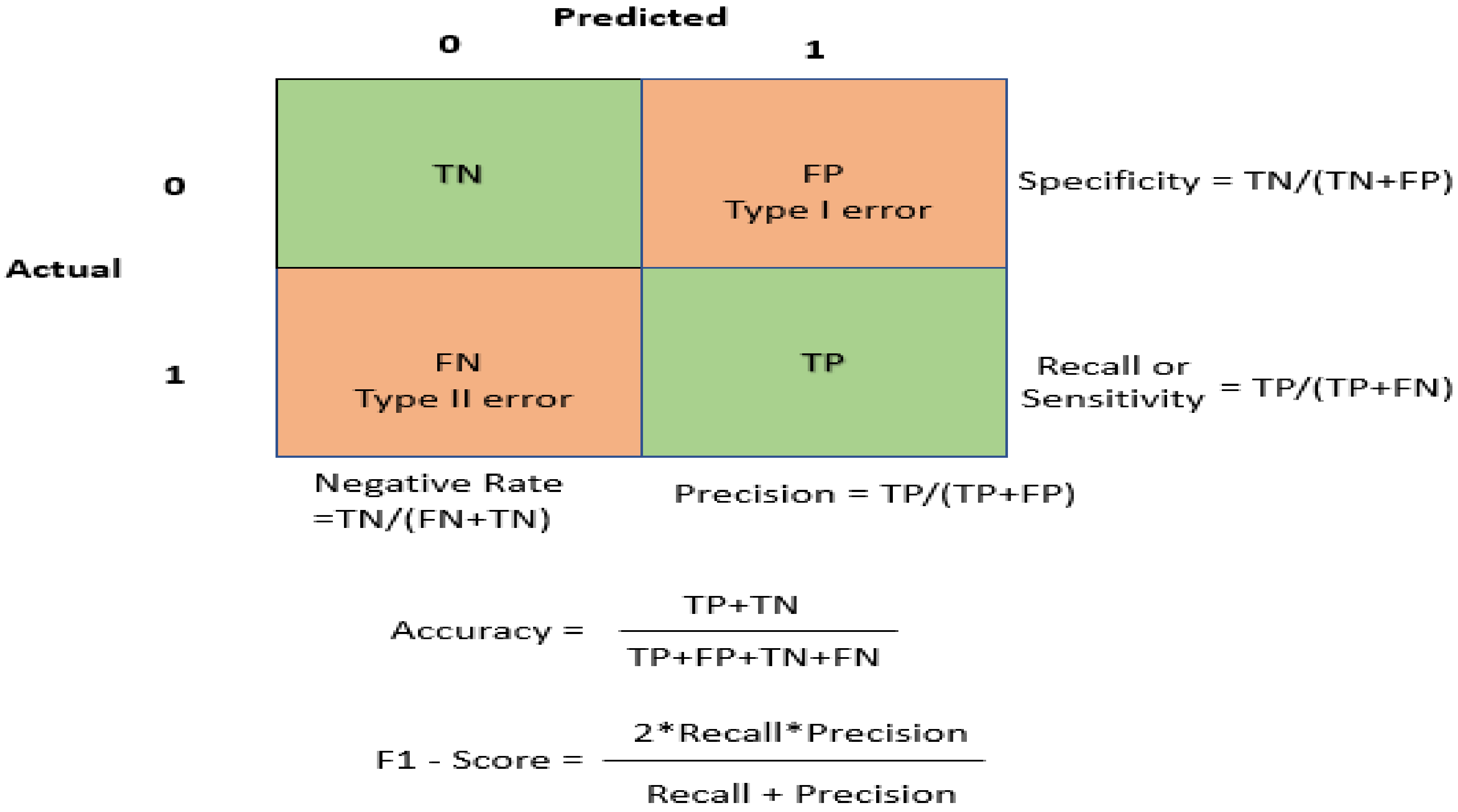
Confusion Matrix



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Question – What is confusion matrix ?

Evaluation Matrices



		Predicted	
		0	1
Actual	0	TN	FP Type I error
	1	FN Type II error	TP

Specificity = TN/(TN+FP)

Recall or Sensitivity = TP/(TP+FN)

Negative Rate =TN/(FN+TN)

Precision = TP/(TP+FP)

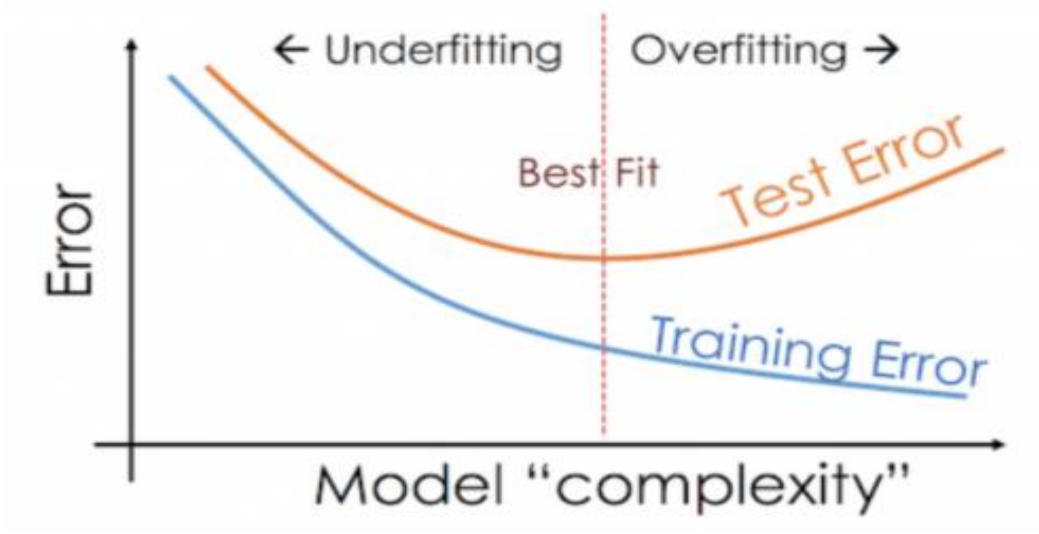
$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{F1 - Score} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

Question - Explain evaluation metrics with the help of confusion matrix ?

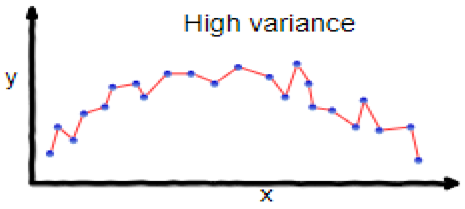
Overfitting & Underfitting

- *This situation where any given model is performing too well on the training data but the performance drops significantly over the test set is called an overfitting model.*
- *On the other hand, if the model is performing poorly over the test and the train set, then we call that an underfitting model.*

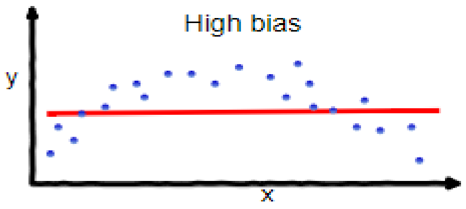


Question – What is Overfitting and Underfitting ?

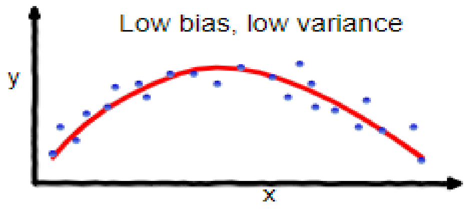
Bias – Variance Trade off



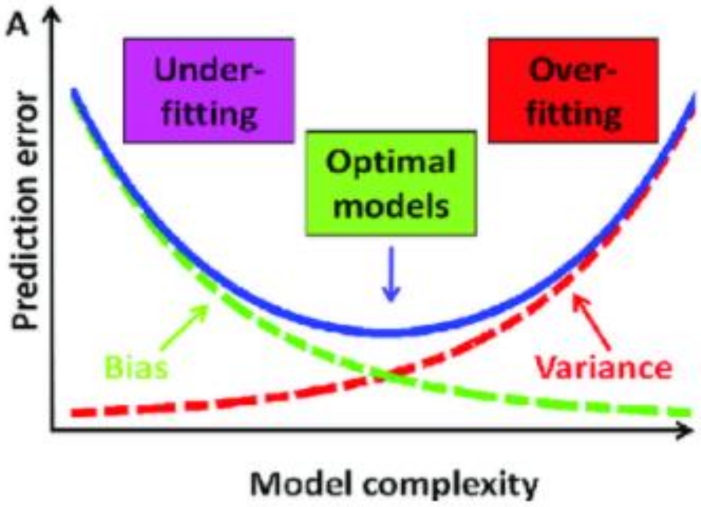
overfitting



underfitting



Good balance



Question – What is bias-variance Tradeoff ?

Thanks...!!!