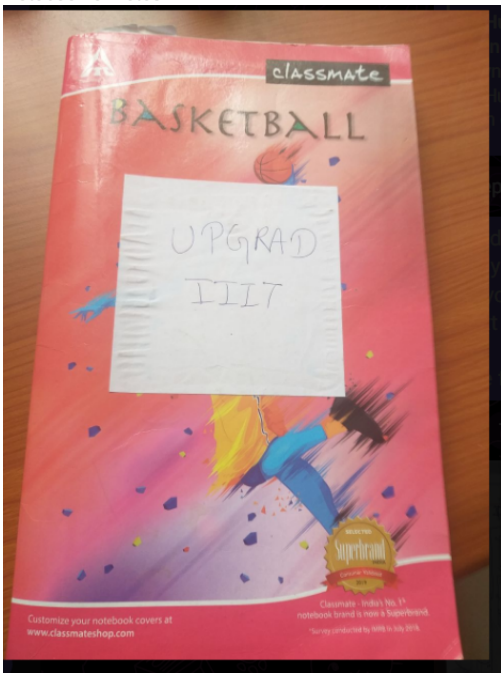
Exam - 1 Agenda

05 September 2022 08:45

Notebook of notes:

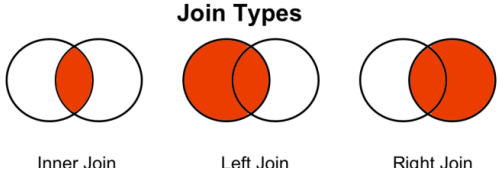


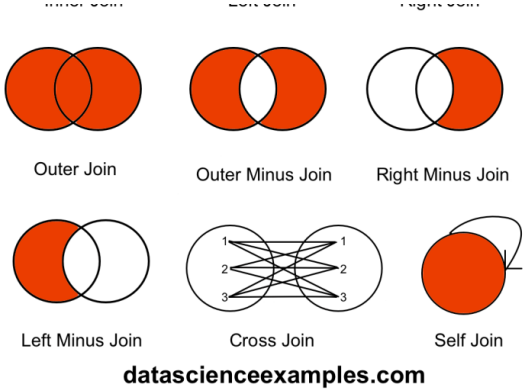
Syllabus:

Statistics essentials

1. **EDA - It is about finding patterns in data, and drawing some inferences from data**
    a. Data sourcing
        i. Private and Public data
            1. Private:
                a. Banking data
                b. Telecom data
                c. Human Resource data
                d. Retail data
                e. Media data
            2. Public:
                a. awesomedata/awesome-public-datasets: A topic-centric list of HQ open datasets. (github.com)
                b. Data.gov
                c. Home | Open Government Data (OGD) Platform India
                d. Home | Government of India (censusindia.gov.in)
                e. DataMeet: https://github.com/datameet
    b. Data cleaning
        i. Fixing rows and columns
            1. Delete incorrect rows, summary rows and extra rows
            2. Add , rename , split (for more data) , merge, align misaligned, or delete columns / column names
        ii. Fixing Missing values
            1. Treat blanks, "N/A", "-" etc. as missing
            2. Fill missing values with some value
            3. Remove missing values
            4. Fill partial missing values
        iii. Standardising values
            1. Remove outliers
            2. Converting to the same unit (Standardising units)
            3. Scale values
            4. Standardise case
            5. Remove extra characters (prefix/suffix etc)
        iv. Fixing Invalid values
            1. Encode unicode properly (most common: CP1252, UTF-8)
            2. Convert incorrect data types (string -> number etc)
            3. Correct wrong structure (phone number > 10 digits)
            4. Correct values beyond range (temp)
        v. Filtering data
            1. De-duplicate data
            2. Filter rows/ columns
            3. Aggregate data

Imp: Joins in Pandas





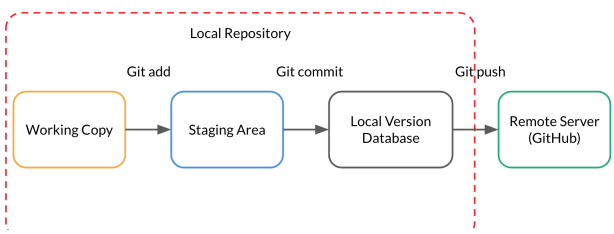
In the above diagram,
Left join == left outer join
Right join == right outer join
Outer join == Full outer join

c.
  i. Univariate analysis (UVA)
    1. Data Description
    2. Categorical Variables
      a. Unordered
        (i) not possible to say that a certain category is 'more or less' or 'higher or lower' than others. For example, color is such a variable (red is not greater or more than green etc.)
      b. Ordered
        (i) ordered categories have a notion of 'higher-lower', 'before-after', 'more-less' etc. For e.g. the age-group variable having three values - child, adult and old is ordered categorical because an old person is 'more aged' than an adult etc.

      (Imp:
      A power law distribution has the form $Y = k X^{\alpha}$, where: X and Y are variables of interest, α is the law's exponent, k is a constant.)
    3. Quantitative variables
      a. Summary metrics
        i. Mean, median, mode, variance, standard deviation, InterQuartile-distance (better if outliers)
  ii. Segmented UVA
    1. Basis of Segmentation
      a. Group data and compare
    2. Quick way of segmentation
    3. Comparison of averages
    4. Comparison of other metrics
d. Bivariate analysis (BVA) - influence of one variable over another variable
  i. BVA on Continuous variables
  ii. Correlation
  iii. BVA on Categorical variables
e. Derived metrics
  i. Types:
    1. Type-driven
      a. Steven's typology
        i. Cardinal, ordinal, interval, ratio variables ('x' is twice as long as 'y') What is the difference between ordinal, interval and ratio variables? Why should I care? - FAQ 1089 - GraphPad
      b. Wrapping around numbers:
        i. Latitude and longitude
        ii. Time (hours...)
      c. Location
        i. Demographics (urban, rural)
        ii. Ruling party
        iii. Time zone
        iv. Directions
    2. Business-driven
    3. Data-driven

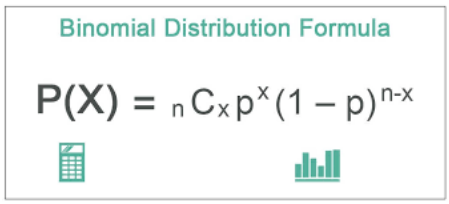
  Note:
  Type driven - Extracting info based on relations between types / categories of data.
  Business domain-driven - Impact on cost with various factors.
  Data-driven - Using Past data for reference.

2. **Intro to Git and GitHub**
  a. Version Control System (VCS)
    i. Background
    ii. Git vs GitHub
    iii. Installation
    iv. Basic Git commands
    v. Resetting vs Reverting
    vi. Branches
  b. Contributing to open-source libraries
  c. Portfolio website creation



3. **Lending club case study (Assignment)** (1_(Assignment)_Lending club case study (completed))

4. **Inferential statistics**
   a. Basics
      i. Random variables
      ii. Probability distributions
      iii. Expected value
   b. Discrete Probability Distribution
      i. Probability without experiment (theoretical probability)
      ii. Binomial distribution



x	P(X=x)
0	$(1-p)^4$
1	$4p(1-p)^3$
2	$6p^2(1-p)^2$
3	$4p^3(1-p)$
4	p^4

Figure 1 - Probability Distribution for General Probability p

However, as Prof. Tricha said, there are some **conditions** that need to be followed in order for us to be able to apply the formula.

1. **Total number** of trials is **fixed** at n
2. Each trial is **binary**, i.e., has **only two possible outcomes** - success or failure
3. **Probability of success** is **same** in all trials, denoted by p

      iii. Cumulative Probability
         1. P(X<=2) = P(X=0) + P(X=1) + P(X=2)

   c. Continuous Probability Distribution
      i. Probability Density Functions (PDF) vs Cumulative Distribution Function (CDF)
         1. In PDF graph, Height of the rectangle is PDF, area of rectangle (upto a point 'x') is cumulative probability of 'x'.
      ii. Normal distribution
         1. Bell-curve
         2. All data that is normally distributed follows the **1-2-3 rule**. This rule states that there is a -
            a. **68%** probability of the variable lying **within 1 standard deviation** of the mean
            b. **95%** probability of the variable lying **within 2 standard deviations** of the mean
            c. **99.7%** probability of the variable lying **within 3 standard deviations** of the mean
         3.

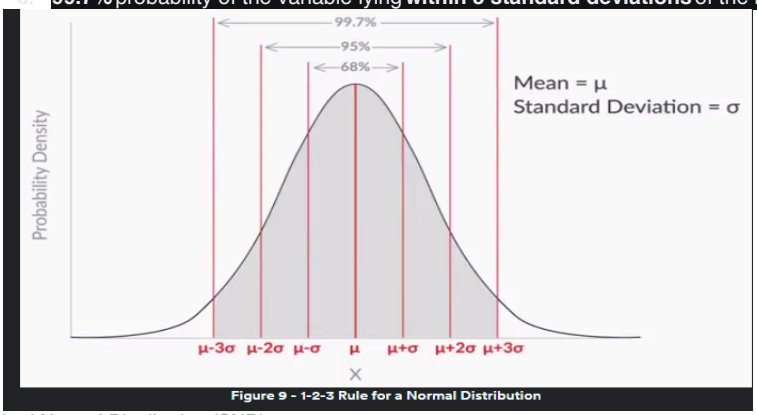


Figure 9 - 1-2-3 Rule for a Normal Distribution

      iii. Standard Normal Distribution (SND)

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

Z-score => Z - standardised normal variable

Equation to find cumulative probability:

Alternatively, you can use the following equation to find the cumulative probability:

$$F(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{\frac{-t^2}{2}} dt$$

d. Central Limit Theorem (CLT)
   i. Samples
      1. To reiterate, these are the notations and formulae related to populations and their samples:

Population/Sample	Term	Notation	Formula
Population $(X_1, X_2, X_3, \ldots\ldots, X_N)$	Population Size	N	Number of items/elements in the population
	Population Mean	μ	$\frac{\sum_{i=1}^{i=N} X_i}{N}$
	Population Variance	σ^2	$\frac{\sum_{i=1}^{i=N}(X_i-\mu)^2}{N}$
Sample $(X_1, X_2, X_3, \ldots\ldots, X_n)$ (Sample of Population)	Sample Size	n	Number of items/elements in the sample
	Sample Mean	$\bar{X}$	$\frac{\sum_{i=1}^{i=n} X_i}{n}$
	Sample Variance	S^2	$\frac{\sum_{i=1}^{i=n}(X_i-\bar{X})^2}{n-1}$

Figure 1 - Notations and Formulae Related to Populations and Their Samples

   ii. Sampling Distributions
   iii. CLT

So, 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**) = $\sigma/\sqrt{n}$
3. **For n > 30**, the sampling distribution becomes a **normal distribution**

Sampling Distributions (onlinestatbook.com)

   iv. So, the central limit theorem says that, for any kind of data, provided a high number of samples h 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**) = $\frac{\sigma}{\sqrt{n}}$
3. **For n > 30**, the sampling distribution becomes a **normal distribution**

   v. Estimating Mean using CLT
   vi. Confidence Interval
      1. Margin of Error:
         a. Z*S/(n**0.5)
      2. Confidence Interval:
         a. Feedback:
            As you know, sample mean $\bar{X}$ = 530 and $\frac{Z^*S}{\sqrt{n}}$ = 19.6. Now, you know the confidence interval is $(\bar{X} - \frac{Z^*S}{\sqrt{n}}, \bar{X} + \frac{Z^*S}{\sqrt{n}})$. Putting in the values, you can calculate the confidence interval as (510.4, 549.6).

5. **Hypothesis testing**
   a. Concepts of Hypothesis testing
      i. Null and Alternate Hypotheses

Hypothesis Testing starts with the formulation of these two hypotheses:

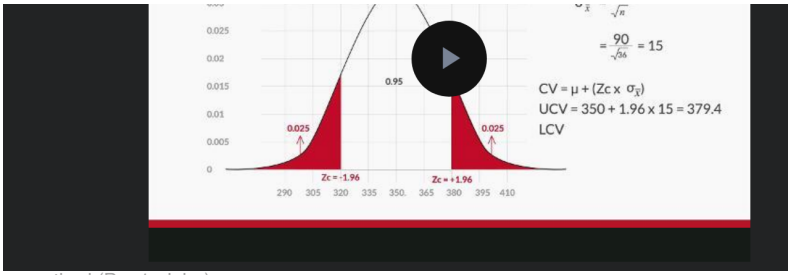
- **Null hypothesis** (H_0): The status quo
- **Alternate hypothesis** (H_1): The challenge to the status quo

$\neq$ in H_1 → Two-tailed test → Rejection region on **both sides** of distribution

$<$ in H_1 → Lower-tailed test → Rejection region on **left side** of distribution

$>$ in H_1 → Upper-tailed test → Rejection region on **right side** of distribution

      ii. Making a decision
      iii. Critical Value Method

1. 

iv. P-value method (P = 1-alpha)

P-value Table

The P-value table shows the hypothesis interpretations:

P-value	Decision
P-value > 0.05	The result is not statistically significant and hence don't reject the null hypothesis.
P-value < 0.05	The result is statistically significant. Generally, reject the null hypothesis in favour of the alternative hypothesis.
P-value < 0.01	The result is highly statistically significant, and thus rejects the null hypothesis in favour of the alternative hypothesis.

v. Types of errors

1. Hypothesis — a claim or an assumption that you make about one or more population parameters

2. Types of hypothesis:

- **Null hypothesis** (H_0) - Makes an assumption about the status quo
  - Always contains the symbols '=', '≤' or '≥'
- **Alternate hypothesis** (H_1) - Challenges and complements the null hypothesis
  - Always contains the symbols '≠', '<' or '>'

3. Types of tests:

- **Two-tailed test** - The critical region lies on both sides of the distribution
  - The alternate hypothesis contains the ≠ sign
- **Lower-tailed test** - The critical region lies on the left side of the distribution
  - The alternate hypothesis contains the < sign
- **Upper-tailed test** - The critical region lies on the right side of the distribution
  - The alternate hypothesis contains the > sign

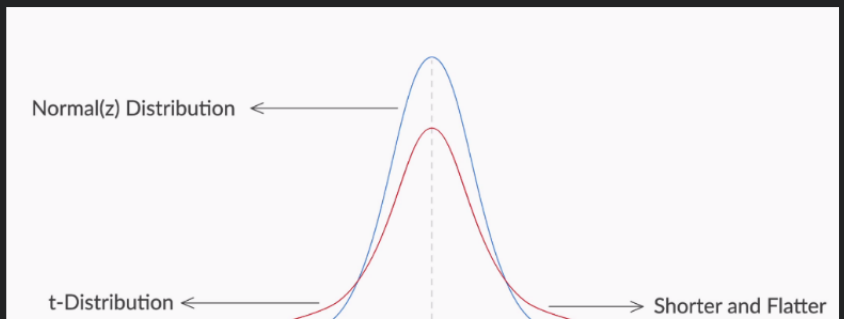
4. Making a decision - Critical value method:

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

b. Industry demonstration of Hypothesis testing

i. T Distribution

A t-distribution is also referred to as **Student's T distribution**. A t-distribution is similar to the normal distribution in many cases; for example, it is symmetrical about its central tendency. However, it is shorter than the normal distribution and has a flatter tail, which would eventually mean that it has a larger standard deviation.



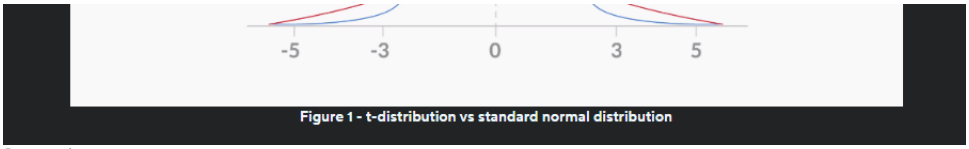


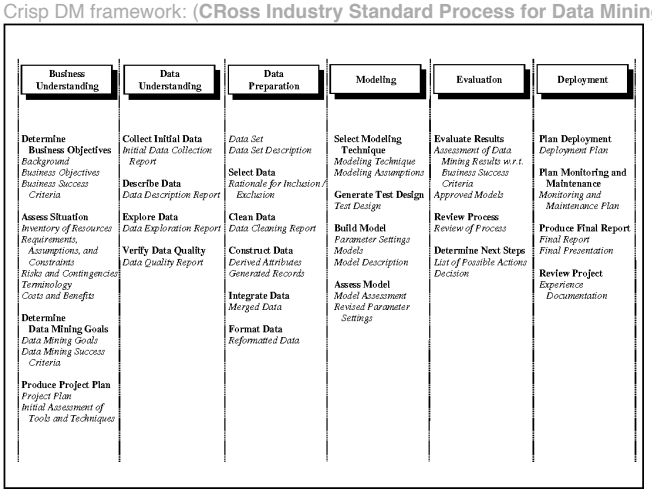
Figure 1 - t-distribution vs standard normal distribution

ii. 2-sample mean test
iii. 2-sample proportion test
iv. A/B testing demonstration
v. Hypothesis testing in Python
vi. Industry relevance

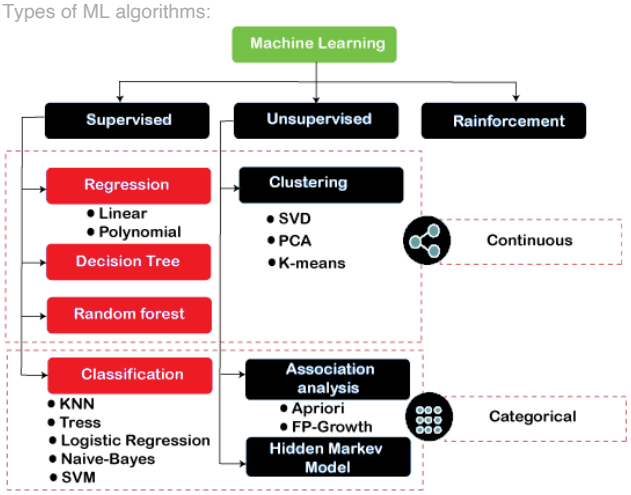
==========

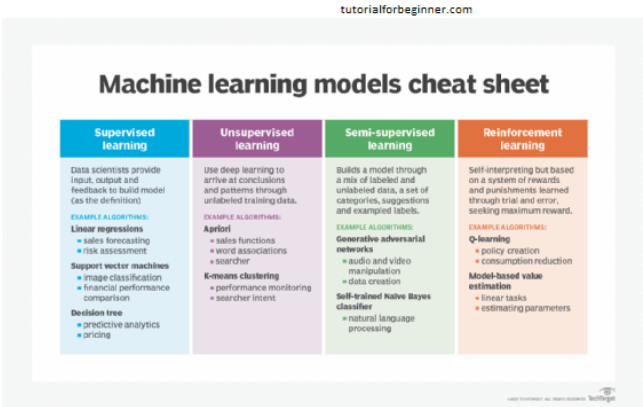
Machine learning

Crisp DM framework: (**CRoss Industry Standard Process for Data Mining (CRISP-DM)**)



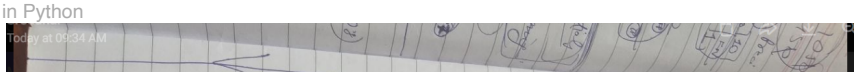
Types of ML algorithms:

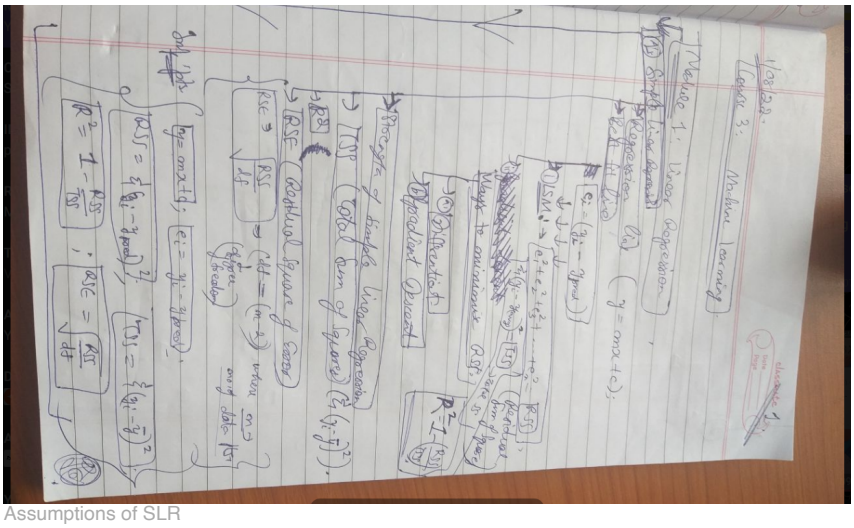




1. **Linear Regression**
    a. Intro to Simple Linear Regression (SLR)
        i. Intro to ML
        ii. Regression line
            1. y = mx + c
        iii. Best-Fit line
            1. Ordinary Least Squares Method (see below)
        iv. Strength of SLR

    b. SLR in Python





i. Assumptions of SLR
ii. Reading and Understanding the data
iii. Hypothesis testing in Linear Regression
iv. Building a Linear model
v. Residual analysis and Predictions
vi. Linear Regression using SKLearn

c. Multiple Linear Regression (MLR)
   i. Moving from SLR to MLR: New considerations
      1. Formula and conditions:

Most of the concepts in multiple linear regression are quite similar to those in simple linear regression.

The formulation for predicting the response variable now becomes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

Apart from the formulation, there are some other aspects that still remain the same:

1. The model now fits a hyperplane instead of a line
2. Coefficients are still obtained by minimising the sum of squared errors, the least squares criteria
3. For inference, the assumptions from simple linear regression still hold - zero-mean, independent and normally distributed error terms with constant variance

The new aspects to consider when moving from simple to multiple linear regression are:

1. **Overfitting**
   - As you keep adding the variables, the model may become far too complex
   - It may end up memorising the training data and will fail to generalise
   - A model is generally said to overfit when the training accuracy is high while the test accuracy is very low
2. **Multicollinearity**
   - Associations between predictor variables, which you will study later
3. **Feature selection**
   - Selecting the optimal set from a pool of given features, many of which might be redundant becomes an important task

   ii. Multicollinearity
      1. VIF: (Variance Inflation Factor)

The VIF is given by:

$$VIF_i = \frac{1}{1-R_i^2}$$

         a. How well a predictor variable is correlated with all the other variables, excluding the target variable

✓ Correct

Feedback:
VIF measures how well a predictor variable can be predicted using all other predictor variables

   iii. Dealing with Categorical variables
      a. Feature scaling:

i. It is important to note that **scaling just affects the coefficients** and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

There are two major methods to scale the variables, i.e. standardisation and MinMax scaling.

Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1

The formulae in the background used for each of these methods are as given below:

- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

b. T-statistic:

i.

Formula

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{\mathrm{SE}(\hat{\beta})}$$

$t_{\hat{\beta}}$ = t-statistic

$\hat{\beta}$ = estimator of parameter

β_0 = non-random known constant

$\mathrm{SE}(\hat{\beta})$ = standard error of the estimator

c. F-statistic:

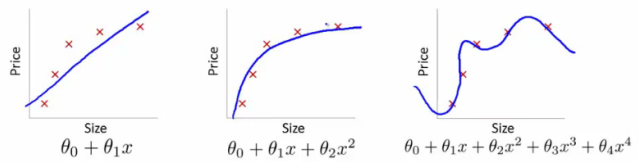
Table A4: 5% Critical Values of the *F* Distribution

i.

		Numerator Degrees of Freedom									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Denominator Degrees of Freedom	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
	120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
	∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

d. Bias vs variance:

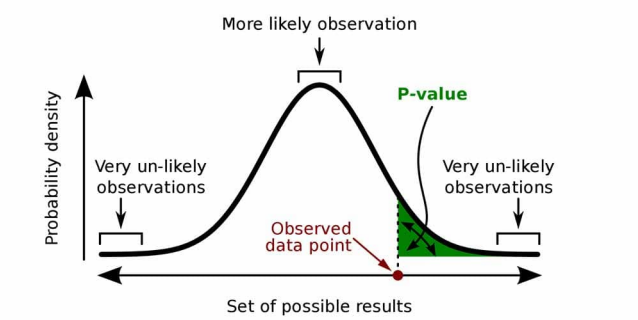
i.



High bias (underfit) | "Just right" | High variance (overfit)

e. P-value:

i.



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

f. Dummy variable:

i. In statistics and econometrics, particularly in regression analysis, a dummy variable is **one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to**

shift the outcome

ii. Each of the dummy variables has 'm' levels. So to represent one categorical variable, you would require (m-1) levels. Hence, to represent 'n' categorical variables, you would need (m-1)*n dummy variables.

iv. Model Assessment and Comparison
1. Adjusted R**2: (and AIC: Akaike information criterion)
(R**2 is called Coefficient of Determination)

i.

Hence, there are two new parameters that come into picture:

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

$$AIC = n \times log(\frac{RSS}{n}) + 2p$$

Here, n is the sample size meaning the number of rows you'd have in the dataset and p is the number of predictor variables.

v. Feature selection
1. RFE:
Recursive feature elimination is based on the idea of repeatedly constructing a model (for example, an SVM or a regression model) and choosing either the best or worst performing feature (for example, based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated. As such, it is a greedy optimisation for finding the best performing subset of features.

- Manual feature selection - A very tedious task in order to select the correct set of features.
- Automated feature selection - The three step process is involved.
- Select top 'n' features
- Forward/backward/Stepwise selection based on AIC
- Regularization
- Finding a balance between the two - A balance of both manual and automatic feature selection is required to attain the features.

d. MLR in Python
i. Reading and understanding the data
ii. Data preparation
iii. Building the model
iv. Residual analysis and predictions
v. Variable selection using RFE (Recursive Feature Elimination)

2. **Linear Regression assignment**
a. **Bike sharing assignment** (2_(Assignment)_Bike_sharing_assignment (completed))

3. **Logistic Regression**
a. Univariate Logistic regression (ULR)
i. Binary classification
ii. Sigmoid curve
1. Formula:

i. $y = P(Diabetes) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$

2. Likelihood function: Best fit sigmoid curve would be the one that maximises the below product.

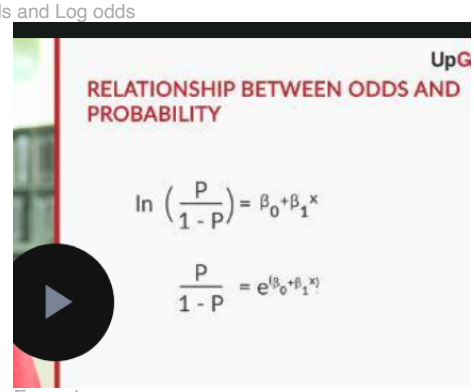
i.

This product is called the **likelihood function**. It is the product of:

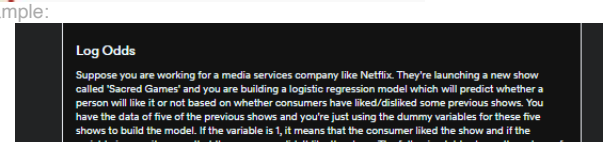
$$[(1-P_i)(1-P_i)\text{------ for all non-diabetics --------}] * [(P_i)(P_i)\text{ -------- for all diabetics -------}]$$

iii. Finding the best-fit sigmoid curve
iv. Odds and Log odds

1.



2. Example:



i.

variable is zero, it means that the consumer didn't like the show. The following table shows the values of the coefficients for these five shows that you got after building the logistic regression model.

Variable Name	Coefficient Value
TrueDetective_Liked	0.47
ModernFamily_Liked	-0.45
Mindhunter_Liked	0.39
Friends_Liked	-0.23
Narcos_Liked	0.55

Now, you have the data of three consumers Reetesh, Kshitij, and Shruti for these 5 shows indicating whether or not they liked these shows. This is shown in the table below:

Consumer	TrueDetective_Liked	ModernFamily_Liked	Mindhunter_Liked	Friends_Liked	Narcos_Liked
Reetesh	1	0	0	0	1
Kshitij	1	1	1	0	1
Shruti	0	1	0	1	1

Based on this data, which one of these three consumers is most likely to like to new show 'Sacred Games'?

Reetesh ✓ Correct

Feedback:
Correct!

To find the person who is most likely to like the show, you can use log odds. Recall the log odds is given by:

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n$$

Here, there are five variables for which the coefficients are given. Hence, the log odds become:

$$ln(\frac{p}{1-p}) = 0.47X_1 - 0.45X_2 + 0.39X_3 - 0.23X_4 + 0.55X_5$$

As you can see, we have ignored the β_0 since it will be the same for all the three consumers. Now, using the values of the 5 variables given, you get -

$$(Log\ Odds)_{Reetesh} = (0.47 \times 1) - (0.45 \times 0) + (0.39 \times 0) - (0.23 \times 0) + (0.55 \times 1) = 1.02$$

$$(Log\ Odds)_{Kshitij} = (0.47 \times 1) - (0.45 \times 1) + (0.39 \times 1) - (0.23 \times 0) + (0.55 \times 1) = 0.96$$

$$(Log\ Odds)_{Shruti} = (0.47 \times 0) - (0.45 \times 1) + (0.39 \times 0) - (0.23 \times 1) + (0.55 \times 1) = -0.13$$

As you can clearly see, the log odds of Reetesh is the highest, hence, the odds of Reetesh liking the show is the highest and hence, he is most likely to like the new show, Sacred Games.

b. Multivariate Logistic Regression (MVLR)

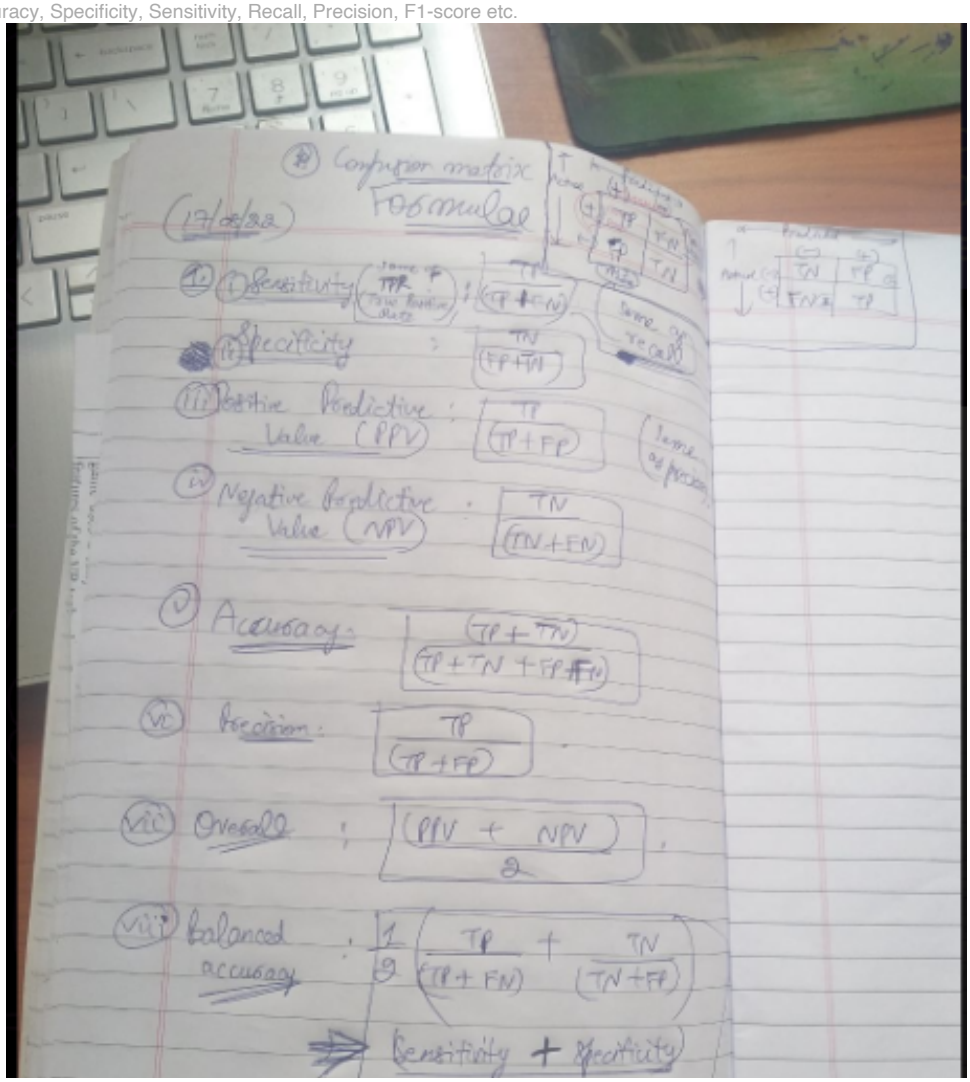
such that the equation now becomes:

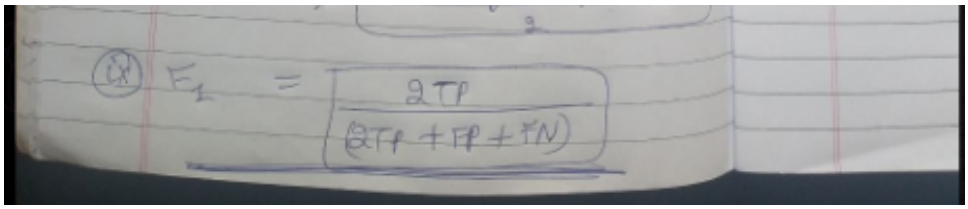
$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_n X_n)}}$$

i. Model Evaluation Metrics

1. Accuracy, Specificity, Sensitivity, Recall, Precision, F1-score etc.

a.





2. ROC Curve (Receiver Operating Characteristic curve)
    a. In Python
3. Finding the Optimal threshold (generally, cutoff is 0.5)
4. Making predictions

4. **Naïve Bayes**
    a. Bayes theorem
        i. Conditional Probability and its intuition
            1. Prior vs posterior probability:
               A posterior probability is the probability of assigning observations to groups given the data.
               A prior probability is the probability that an observation will fall into a group before you collect the data.
        ii. Joint probability vs Conditional probability -
            Feature is the variable based on which classification is done.
    b. Naïve Bayes for categorical data
        i. With one feature
        ii. Comprehension
        iii. Conditional Independence in Naïve Bayes
        iv. Deciphering Naïve Bayes
            1. Maximum Aposteriori Classification Rule (MAP)
    c. Naïve Bayes for Text Classification
        i. Document Classifier (Bag of Words Array)
            1. Pre-processing steps , and Worked out example
            2. Stopwords (ranks.nl)
        ii. Laplace smoothing (when probability in text-based classification is 0, this method helps)

1.

Word	P(word | Sports)	P(word | Not Sports)
a	$\frac{2+1}{11+14}$	$\frac{1+1}{9+14}$
very	$\frac{1+1}{11+14}$	$\frac{0+1}{9+14}$
close	$\frac{0+1}{11+14}$	$\frac{1+1}{9+14}$
game	$\frac{2+1}{11+14}$	$\frac{0+1}{9+14}$

        iii. Bernoulli Naïve Bayes
        iv. Python lab (Naive-Bayes/Naive Bayes for text classification at main · ContentUpgrad/Naive-Bayes (github.com))
            1. SMS Span Ham Classifier
                a. Multinomial
                b. Bernoulli
        v. Practice questions

5. **Model selection**
    a. Principles of Model selection
        i. Model and Learning algorithm
            The learning algorithm is instructed what needs to be done; it figures out how it needs to be done and returns a model.
        ii. Simplicity, Complexity and Overfitting
        iii. Bias-variance tradeoff
        iv. Regularization
    b. Model evaluation
        i. Regularization and hyperparameters
        ii. Model evaluation and Cross validation
            1. Model evaluation
                a. Python demo
            2. Cross-validation
                a. Motivation
                b. Python demo
                c. Hyperparameter tuning

        iii. Practice questions

==========