Q1) Explain and discuss various PMF/PDF ?

Ans:- Probability Density/mass function. depicts the probability of observing measurement along with a particular value. Therefore the integral over density is 1 always. Probability density function (PDF) of random variable X, that is denoted as f can ~~be~~ defined as follows:-

$$f(x) = \frac{df(x)}{dx}$$

Here, F is cumulative Distribution Function (cdf) of x. The f(x) is derivative of cdf F with respect to x. But for discrete distributions, the density function is $f(x) = Pr(X=x)$

Here, in such cases f is sometimes called Probability function or Probability mass function.

The probability that a random variable X adopts on value in interval [a, b] would be integral of pdf evaluated between a and b. This is depicted as follows:-

$$Pr(a \le x \le b) = \int_a^b f(x)dx$$

In case of discrete distributions, equation ③ converts to summing up probabilities of all the values in the interval.

$$Pr(a \le x \le b) = \sum_{x \in [a,b]} f(x) = \sum_{x \in [a,b]} Pr(X=x)$$

The probability Density Function (PDF) might plot the values of Pdf against qualities of some particular distributions. The theoretical pdf plots are plotted with empirical pdf plots in certain cases, so that the histograms or bar graphs can assess visually to know whether data has a specific distribution. The pdf plot will return a set of coordinates of points that are already plotted or to be plotted in future. Qualities are used for plots & probability. Densities are values of pdf that are related to quantities.

Eg:-

```
dev.new()
pdfplot (param. list = list (mean = 2, sd = 2), curve.fill = FALSE,
        ylim = c(0, dnorm (0)), main = "")

pdfplot (add = TRUE , pdf .col = "red" )
legend ("topright", legend = c ("N(2,2)" , "N(0,1), col = c("black",
        "red", ), lwd = 5* par ("cex"))

title ("PDF plots for two Normal Distibutions").
```

(Q2) How do you solve problems with Hypothesis testing?

Ans:- Hypothesis Testing:-

The statistical hypothesis can be defined as an assumption with respect to a population that may or may not be true. It is a set of formal procedures that is used by statisticians for accepting or rejecting statistical hypothesis. In fact it is a process of validating the hypothesis that is made by researchers. To validate the hypothesis. The complete population is considered.

In this process it makes use of random sample from the population. The selection or rejection of hypothesis depends on result of testing over the sample data.

→ Process of hypothesis testing:-

Hypothesis Testing consists of following steps:-

1) Static the Hypothesis:-

In this step the type of hypothesis must be stated whether it is null or alternative hypothesis. If one is true then other must be false.

2) Formulate an Analysis plan:-

In this step the process of using sample data in evaluating null hypothesis is determined. This process must focus only one single test stastic.

3) Analyse sample data:-

In this step, value of test statistic is computed by using various properties such as mean score, proportion stastic, z-score etc.

4) Interpret Results:-

In this step apply the ~~design~~ desicions defined in analysis plan. If the value of test statistic depends on hypothesis then the null hypothesis must be rejected.

escample of hypothesis testing:-

Consider an example, to check whether a coin was fair and balanced. According to null hypothesis the half flips would be of heads and other of tails. And according to alternative hypothesis the flips of head and tail must be different.

$$H_0 : P = 0.5$$
$$H_a : P \neq 0.5$$

Flipping of coin for 50 times might result 40 heads & 10 tails. Based on the desult the null hypothesis must be rejected & concluded according to the evidence that coin was not fair & balanced probably.

(Q3) Explain Linear regression & Logistic Regression Model building in R?

Ans:- A simple linear regression model can be built by using lm() function or glm() function. Both perform same operation and produce similar output. Steps to build simple linear regression model are:-

Consider the example of fitting simple linear regression model for dataset "faithful" in R.

1. collecting data and understanding it.
2. fitting the linear regression model.
3. Predincting the dependant variable based on independent variable through regression model.

e.eruption = intercept + (slope * waiting)

For the independant variable value of waiting = 80, the predicted value of variable is 4.14.

4. Test of significance:-

It p value is less than 0.005 then null hypothesis can be rejected because there is significant relationship b/w dependant & independant variables.

5. coefficient of determining R-squared value :-

R value lies between 0 and 1. If R is 1.00 then there is relationship b/w dependant & independent variables. But if R is 0.00 then there isn't any relationship.

6. Finding the confidence interval:-

It is the confidence interval (95%) for mean value of dependant variable, given value of independent variable for waiting time is 80 mins then the 95% prediction interval for eruption is 4.17.

7. Finding the prediction interval:-

It is the interval estimate of independant variable for given value of independant variable. The lower & upper limits will be 3.19 and 5.15 respectively.

* Logistic Regression:-

It is a powerful model which is commonly used in the field of marketing and medicine. The formulae for LR is:-

$$P(y_i = 1) = \text{logit}^{-1}(x_i\beta)$$

In eq① y_i denotes response of i^{th} element ,$x_i\beta$ is linear predictor & value of $logit^{-1}$ function can be given as :-

$$logit^{-1}(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^x}$$

The inverse converts the continuous output generated from linear predictor and makes it lie b/w range of 0 & 1. This can be considered as inverse of link function.

In R programming. glm() function is used to handle logistic regression. This Fn is also capable of handling linear regression. The formular interface of glm() is sames as lm() but with some additional options required to be set.

Q4) Explain KNN Algorithm and how it is implemented in R?

Ans:- KNN can be defined as supervised machine learning Algorithm that can classify a new data point into ~~tags~~ target class based on neighbouring datapoints features. For example consider a KNN algorithm for a machine that differentiates b/w apples & mangoes.

Algorithm.

1) Assume k as the number of nearest neighbours, T as the set of training examples and z = ('x', y') as test example (x = attribute set y = class label).

2) For each test example perform step 3 & step 4.

3) Calculate the distance b/w z . i.e (x' . x) & each example

i.e (x , y) ∈ D.

&

4) select the set of k-nearest training examples to z.

5) $y'' = \text{argmax} \sum_{(x_i, y_i) \in D_z} I \, (V = y_i)$

6) End of loop.

* KNN implementation in R.

Step1 :- Importing Data

Step2 :- Data cleaning

Step3 :- Data Normalization

Step4 :- Data splitting

Step5 :- Developing Machine Learning Model.

Step6 :- Evaluating the model

Step 7 : Optimisation

(Q5) Explain K-means algorithm and how it is implemented in R?

Ans :- k Means :-

The K-means is an iterative clustering algorithm in which objects are moved among set of clusters until the desired set is achieved. It is most popular & commonly used method. This algorithm is built on the concept of user specified input parameter (K).

A set of 'n' objects are divided into 'K' clusters by the algorithm.

A high degree of similarity among elements in clusters is obtained, while a high degree of dissimilarity among the elements in different clusters is achieved simultaneously. The clusters centroid gives measure of cluster's similarity.

→ Algorithm:-

Algorithms input

The no. of desided clusters are denoted by 'K'.

A dataset containing 'n' objects is denoted by 'D'

Algorithms output

K: A set consisting of k clusters.

Procedure:-

Step1:- initially select 'k' ~~outputs~~ objects randomly from initial cluster centers.

Step 2:- Depending on distance b/w the object and cluster mean, each remaining object is arranged to the cluster which is most similar to mean.

Step 3:- Calculate new mean value of the object for each cluster.

Step 4:- Step-3 is repeated and process iterates until criterion fn changes. The resulting 'k' clusters are compact and separate. K-means method, typically uses the square error criterion fn. which is:-

$$S = \sum_{a=1}^{k} \sum_{e \in C_a} |e - mean_a|^2 \text{ and}$$

$$mean_a = \frac{1}{Mean} \sum_{j=1}^{mean} n_{aj}$$

K-means clusteing algoorithm can be implemented by predefining. Here, k represents the number of clusters to be defined.

<u>Steps to implement k-means are as follows</u>:-

(1) In the 1st step k-centers are defined and every cluster is assigned to the cluster that has closest center to it.

(2) In the 2nd step, the centers are redefined using the observation of each cluster. The column means are used for defining the centroid.

The above steps are repeated until the centers converge. Therefore k-means alg. is said to be iterative.

The k-means fn. that is in R-base will not handle NAs. For eg:- Fill the NAs with 0's. Selection of method to fill the missing data and moreover implementing that particular method must be done correctly.

```
x-0 ← x
x.0 [is.na(x-0)] ← 0

k ← kmeans (x-0, centers = 10)
```

The clusters assigments are available in the cluster components.

```
groups ← k_clusters
split (names(groups), groups).
```

Here the final cluster is observed to be random because the first centre is selected at random. So same stability needs to be imposed by repeating the complete thing for 'n' no. of times and then averaging the outputs.

```
k ← K-means (x-0 , centers =10 , nstart=25)
    groups ← k $ cluster
    split (names(groups), groups)
```

Q6) Write Rcode for reading data from relational databases - MySQL, Reading data from noSQL databases - MangoDB?

Ans:- Using package RMySQL we can easily query MySQL as well as Maria DB databases and store result in R-dataframe.

→ My SQL :-

1 code :-

```
library (RMYSQL)

mydb ← dbconnect(MySQL(), user='user' , password ='password',
        dbname ='dbname', host ='127.0.0.1')

Query string ← "select* from table1 t1 JOIN table2 t2 on
                t1.id = t2.id "
```

~~Query string ← "select * from table1 t1~~ -

```
Query ← dbsendQuery (mydb, query string)
data ← fetch (query ,n=-1). # n=-1 to return all strings
        using limits.
```

⟶

→ **MangoDB** :-

2 code :-

```
library (json lite)
Library (mangolite)
db ← mango (collection= "tweets" , db= " Tweet collector ",
            url = " mango db:// username : password @
                                                hostname")

documents ← db $ find ( limit = 100000 , scip = 0, fields =
            '{" id": false , "Text" : true }' )
```