STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.

Ans - a(True)

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

Ans – a(Central Limit Theorem)

3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans- b(Modeling bounded count data)

4. Point out the correct statement.

Ans - d) All of the mentioned

5. _____ random variables are used to model rates

Ans- (c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

Ans - (b)False

7. 1. Which of the following testing is concerned with making decisions using data?

Ans – (b) Hypothesis

8 . 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.

Ans – (a) 0

8. Which of the following statement is incorrect with respect to outliers?

Ans – (c) Outliers cannot conform to the regression relationship

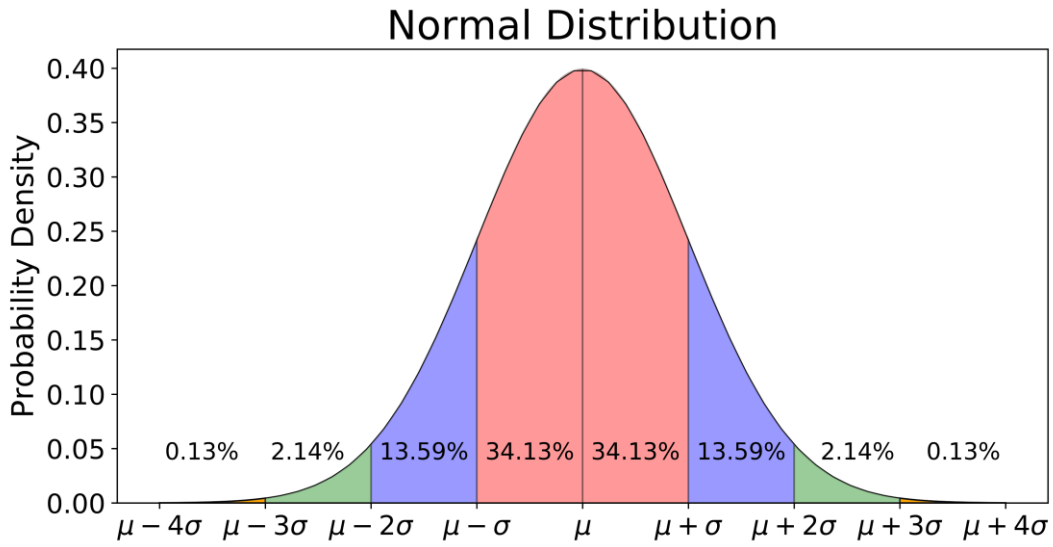
Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

9. What do you understand by the term Normal Distribution?

Ans – Normal Distribution is also called. Gaussion Distribution because Great Germen Mathematician Karl Friedrich Gauss, has invented the Normal Distribution Graph.

What Actually Do Normal Distribution.

Normal Distribution also called bell Curve. Because when we plot a normal distribution in graph . it follow some shape . which is look like bell. (See figure Below)



In normal Distribution we think that mean median mode are exactly same. Because maximum data lie on the mean.
And half of the data lie in left and right side of the distribution.according to frequency. I think you can see in the figure how data distributed symmetrically.

<u>What is the intuition behind the scene ?</u>

So, In normal distribution he use 2 parameters.

1- Mean

2- Standard Deviation.

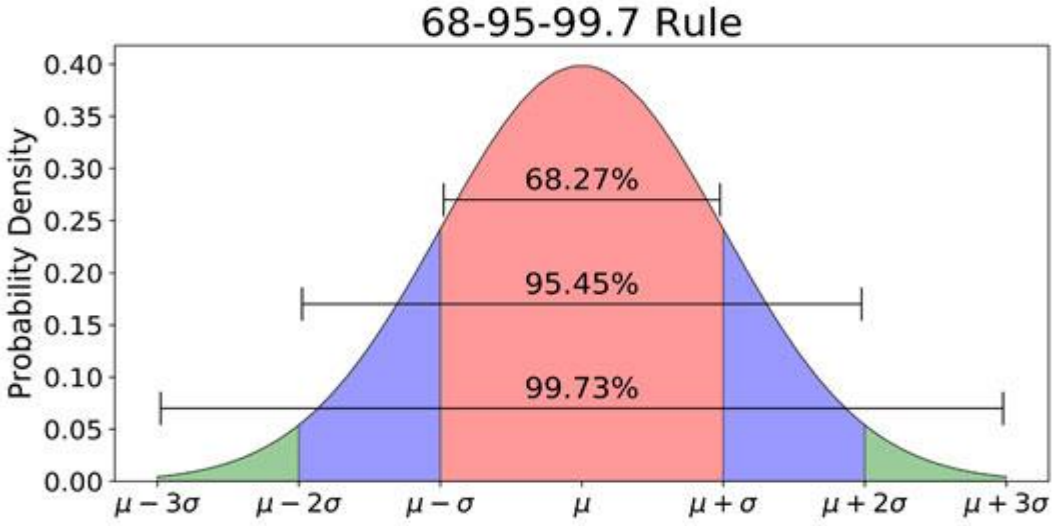
So , the centre of the graph is mean and right and left side +- 1 is standard deviation.

According to Normal Distribution .

67% data contain by Std 1

95% data contain by Std 2

97% data contain by Std 3



So. This trick we can also use for removing outliers . because when we use std 3 to remoze z score . you will cover 99% of data .

What is the difference between a normal distribution and a standard normal distribution ?

A normal distribution is determined by two parameters the mean and the variance. A normal distribution with a mean of 0 and a standard deviation of 1 is called a standard normal distribution.

10. How do you handle missing data? What imputation techniques do you recommend?

Ans- So , when I get a dataset with missing values like Nan. And sometimes we also get in missing values in blank space .

So , when we get this type of issue .

First things I have to do is . Counting the columns . to know about how many columns have missing values.

If I get only 2 -3 columns have nan value .and that columns is important for target column. so I use .**FillNa method.**

Syntax – data.fillna(data['Column_Name'].mean()) (When we have continuous data)

If we have Categorical column . we have to use **mode()** in the place of mean.

So. That’s how I am handle missing values.

What imputation techniques do you recommend ?

Imputation is a technique used for replacing the missing data with some substitute value to get most of the data and information of the dataset. These techniques are used because removing the data from the dataset every time is not good and can lead to a reduction in the size of the dataset to a large extend.

Which not only raises concerns for biasing the dataset but also leads to incorrect analysis.

Suppose – You have a dataset with 1000 features and there is also a nan values . in more then 600 columns and this columns which have nan. Are also important for label. And you have to keep this feature . and predict the label.

Now , you think that how can replace nan values in some substitute value the imputation technique come into picture.

When you have lots of columns with nan. You will use imputation technique to replace all nan with mean and mode .

(I am not saying all the time when you get nan . just use impution technique and replace with substitute value its all depend upon the dataset,how data is big and what dataset is want)

There are few Imputation Technique and How we can import:-

1- Simple Imputer (from sklearn.impute import SimpleImputer)
2- KNN Imputer (from sklearn.impute import KNNImputer)
3- Iterative Imputer (from sklearn.experimental import enable_iterative_imputer)

So , this all are imputation technique and all using different different method to impute the null values.

If You Ask Me. imputation techniques do you recommend?

So , In my point of view its depend on how dataset is big .how many feature have missing values . if there 2-3 features have missing values then definatily I will not use imputation technique . but if we have 500 columns to impute the missing values . so I use the imputation technique .Because it is very easy to use and it fill missing values with different different technique. Like .

KNN Imputer - It will select neighbors values and then take a decision . and fill the missing values .

Iterative Imputer - This method treat other columns (which does not have nulls as feature and train on them and treat null column as label .finally it will predict the Nan data and impute . Its just like regression problem .Here null columns is label..

Now . You can see that how we can impute the missing values with uses of different different methods.

11. What is A/B testing?

Ans - A/B testing is also known as **split testing**. In AB testing, we create and analyze two variants of an application in order to find which variant performs better in terms of user experience, leads, conversions or any other XYZ goal and then eventually keeping the better performing variant.

Lets Understand with example.

Suppose . you are working in amazon. And your boss will come to and say to you . if you change your interface to different style . then may be our sales will be boosted .

So , for that I just using some coding technique I just change my user interface. But not deploy in real website .

So , before deploy in different server you have to upload both design and distributed the customer to the both same web sites . for 1 week . and after one week now you can check which interface the customer likes . if more customer like 1st interface then we did not change our interface . and if more customer like 2nd interface then we will change to tha all customer . According to our profit.

So this is the intuition behind A/B testing.There is something to take care when you apply A/B testing :-

- Sampling Bias – Means only random samples come to the websites
- Not sample to Under Biased
- Enough Sample size to select which one is best.
- And This is powerfull technique..

Uses of A/B testing :-

1- Decide the User Interface in Website
2- A/B Testing also use in Machine Learning

Why we Only use A/B testing ? Why we do not use ABC testing ?

Ans- We can also use ABC testing .but when we try more then 2 testing . then there is a chances to get confuse in which one we have to select for our decision . This reason for we are only using A/B testing . in A and B which one can get majority we can go with the.

We have lot to talk about that . in A/B testing we can also use Hypothesis .

So this all are about the A/B testing.

13. Is mean imputation of missing data acceptable practice?

Ans - See 1st if missing value is smaller we directly eliminate the values .but when we consider that the missing portions larger we use these imputation methods here mean imputation is considered bad practice because it just implement the mean values Mean imputation is typically considered terrible practice since it ignores

feature correlation. But we can always check this by checking the change in model's performance.

14. What is linear regression in statistics?

Ans - > So What is Linear Regression ? Before that we have to know about Regression.

Regression is nothing . But this regression method is used to estimate the relationship between different entities .

Y=(F)X

Here X -> Indipendent Variable

Y -> Dependent Variable

F is nothing but Function.

Regression Analysis -> Regression analysis help us to understand the relationship between dependent and independent variables. By creating model.

Linear Regression -> Linear regression is a predictive model used for finding the linear relationship between dependent and independent variables.

There are two types of Regression .

- Simple linear regression
- Multiple linear regression

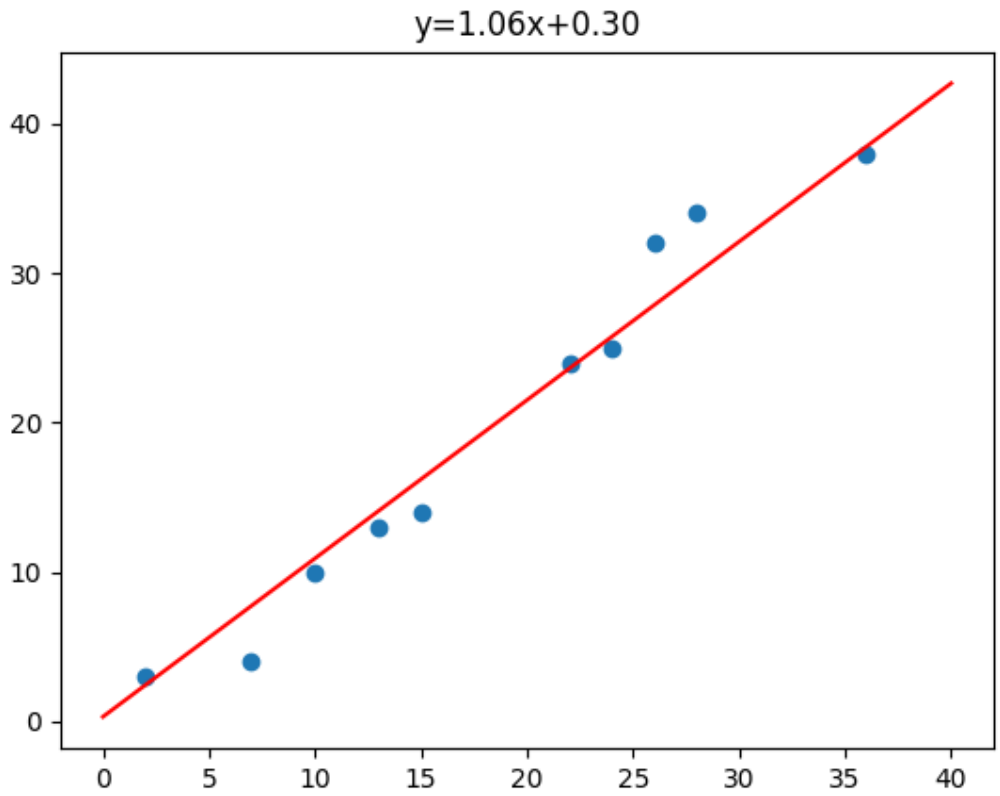
And both are using same formula .

Y=MX+C

So ,Here M= Slope
C= intercept.

Formula extended in the basis of which Regression you use.

So . if you know in pure laymon term . You can say that . if we have to find the relationship between dependent and independent varibles for predicting the model. And uses some formula to find the relationship. . So this is know as Linear Regression.



So, in this picture red line which can show the positive relationship . this is a linear line and all the data points are close to the the line . so now this is showing a some relationship between dependent and independent variables .

This is is intuition behind the Linear Regression in Statistics.

15. What are the various branches of statistics?

Ans – **Branches of Statistics : -**

There are two Major Branches of Statistics :

1- Descriptive Statistics
2- Inferential Statistics

Descriptive Statistics - > Involves the organization, summarization, and display of data ..

Lets understand in pure laymon term :

Suppose . you have to find the average marks of class 10^{th} student in maths .
So what you do .
You just conduct the exam for whole class . and calcalute the average marks of all student .

So . what it is mean .

Its mean you just apply one method to get average marks of whole population ..

It means if you apply any things to whole population so this is knows as Discriptive Statistics .

Inferential Statistics - Inferential Statistics is a statistical method that deduces from a small but representative sample the characterstics of a bigger population .

Lets Understand in pure laymon term.

Suppose – In our country we organised the elections for PM . So , what is does . before the election result we all heard that one word is EXIT POLE . so what is does mean . it mean is .?

Its mean . news channel anchors are going all the places in india and ask about the next PM in india .So what majority people says . they think as he is the next PM of our country.

So what it does actually ?

He select the sample peoples from different different places and ask them . and who will be the our next pm .so , who one can get majority vote . they tell us. In figure of exit pole.

So , same concept apply on apply on . when we have to find the average height of indian peoples ..

- We select 100 people
- Then take a height all 100 peoples
- Then take average height of all 100 peoples .
- Then we call as indian people average height is that ..

Lets Understand what is sample and population

Sample - From 1000 People we select only 100 as sample. Is knows as sample

Poplulation – From 1000 people we select whole 1000 peoples. This is known as population.

ALL the questions in covered in this PDF.

Internsip Batch - 33

Student of DataScience

Student of DataTrained – Saurav

Date – 21 : 10 : 2022

Time – 11:21