Statistics_worksheet_1 :

1. A.) True
2. A.) Central limit theorem
3. B.) Modelling bounded count data
4. D.)All of the mentioned
5. C.) Poisson
6. B.) False
7. B.)Hypothesis
8. A.) 0
9. C.) outliers cannot conform to the regression relationship

10. Normal distribution is also known as gaussian distribution . In statistics it is the continuous probality distribution it can be defined as probability density function for a continuous random variable . It forms a bell curve when plotted and it can be used in a real world ex – like : in the competitive exams here , the normal distribution will define the max students will score the average marks , while smaller amount of students will score the grades (b,d) and a smaller amt of students score will be (a,f) and it can be derived from the empirical rule or formula of the normal distribution curve. Normal distribution posses some very ossom properties like It is symmetrical , this means we can divide the curve in 2 equal parts .
    If a data is skewed data it can be made normal by presenting it inside the normal distribution's bell curve as normal distribution have no skewness in it .

11.) If we talk about missing data lets first define what are these missing datas or values. In a huge/small dataset there some cells or blocks which have some null values ,so we call them missing values and for interpreting the whole dataset we need to handle them so that analysis on the dataset should be accurate and correct . while we do this process of handling the missing values we call it just a part of Data cleaning.

So now the question comes that how we can handle them

a)using the DROPNA FUNCTION :-

in python we have a famous library which is pandas there 1st we have to load the dataset and then pass a python command named as dataframe.Drop.na= (axis =0 , inplace =True), and by this all the Na value will be removed from the dataset.

Now let me explain the command I used here 1st dataframe.dropna will select the loaded dataset and make it ready to drop with dropna func

2nd axis =0 is used for dropping values from rows as ex: lets take we have random rows where we have the Null values so it will remove it .

And 3rd is , inplace =True this is used to make the change permanent in the Dataset further loaded)

B)using the FILLNA function:-

Inside python pandas we have another technique that helps in handling the missing values which is fiilna

The concept of (fillna) function is that we will be replacing the Na values with some other defined value by us .

So it can be used as

Dataframe['xyz'].fillna('U').head()

So it will replace all the missing data with U value or any value which will be given by us .

C .) Replacing the missing values with mean/median/mode

This technique is almost same as the above fillna technique only the difference here is that it used the mean , median or mode to replace the null values of that Dataset.With that it also takes help of Numpy one more very famous liabary of python for data analysis .

Dataframe['abc'].replace(np.Nan; dataframe['abc'] .mean()).head()

In all the above 3 technique there have some pros and some cons in all of them, according to me though the 1st technique has risk of data loss but still it is efficient to work on the huge datasets and it may not create any problem while we do the data encoding for creating a model .so I will prefer the 1st one .

12.)

A/B testing is a scientific experiment which helps us to compare the performance of two versions of contents to see which works better and which one will be appealing to the customer /viewers. It test variant (A) against the a variant (B) Version to measure which one is more successful in the s given metrics parameters , & it is very commonly used by any organization by their marketing team to launch a better product in the market . A/B Testing will be very clear , with help of a example here –

A new I.T startup firm want to create a website with a help of a Web -devloper so they hired a developer who can create well managed website for their company , and now when the website is created he has made 2 variants of the same webpage with 2 versions (v1,v2) now the team is confused that they should go, with which version of the website as both of them are great according to them -even both were appealing , so atlast they decided to perform a A/B testing by launching the website with two different versions and added a Q&A feedback at last of each website versions . when they checked the feedback the team got to know that the V2 is working better on the metrics which was given by the senior management and thus they were able to decide that the V2 version of the website will be better to launch than launching other one . Thus the team was able to launch their 1st website and they were very successful in market and even able to give a better customer experience .

So this example clearly states the true definition of A/b Testing and also how it works , to solve the real life problems .

13.)

If we talk about filling the Null values then , we can easily ans that we have many methods to do the same , from their we have one technique where we fill the , NA values with the help of , calculating their mean values . so now the questions If we are using this technique to handle null values is it a good practice or not . As we know all the techniques has some more or less pros and cons so in this case we have many pros thus we can surely use this method . if we are using this technique we will safe in case

of data -loss from the dataset which might be possible in dropping function but the only cons is that it not useful for all type of datasets we have choose this replacing technique according to type of dataset otherwise , this may further can creating a problem in making ML models.

14.)

Linear regression is the degree of relationship between two variables . Regressions analsysis helps in determining the cause and the effect relationship between the variables. It shows us the change which has taken place in the , dependent variable when there was a change in the independent variable. It can be very easily understood using a graph method.

The equation for the Regression analysis

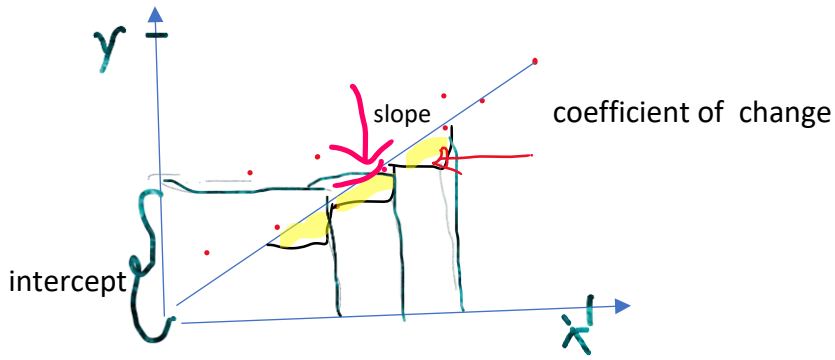
Y = a + bX

Here the Y = Dependent variable

X = Independent variable

a = intercept

b = coefficient showing the slope line



one real world example of where we can calculate regression is - the car sales-purchase and the petrol price hike

here petrol is the independent variable or 'X' while change in the petrol price will surely effect the sales and purchases of the of the cars as cars sales&purchases is the dependent variable 'Y' which will bear the change done on petrol price . Thus if the the petrol prices rises with a given percentage then surely it will effect the sales and purchases of the cars too.Thus checking this cause and effect relationship of two commodities or variables is know as linear regression in statistics.

15.)

In statistics we have two branches in total which gives us the definition of how stats work and how important it is for us :- statistics is a part of mathematics used to perform different operations – Data collection , analysis, and so on. Statistics examine the methodology of collecting , reviewing , and making Data conclusions.

Thus the two branches of statistics are :-

a.) Descriptive statistics:=In this type of statistics data is summarized through the given observations. The summarization is in form of sample of population using parameters such as mean or standard deviation . descriptive statistics is the way to collect , organize and display the data using the table , graphs and summary measures.

It has further divided

Measures of frequency

Measures of central tendency

Measures of dispersions

b.)Inferential statistics :- The type of statistics is used to interpret the conclusions of the Descriptive stats. That means as the descriptive stats is used to collect, organize and display data on the other hand inferential stats is used to , analyze that data and get some insights.It allows us to use information collected from the samples to make decisions, predictions or a inferences from a population.