Statistics-worksheet-1

Q1-Bernoulli random variables take (only) the values 1 and 0.
A-True

Q2-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?

A-Central Limit Theorem

Q3-Which of the following is incorrect with respect to use of Poisson distribution?

A- Modeling bounded count data

Q4- Point out the correct statement.

A-The square of a standard normal random variable follows what is called chi-squared distribution

Q5-_____ random variables are used to model rates.

A- Poisson

Q6-Usually replacing the standard error by its estimated value does change the CLT

A- False

Q7-Which of the following testing is concerned with making decisions using data?

A- Hypothesis

Q8-Normalized data are centered at_____and have units equal to standard deviations of the original data

A- 0

Q9- Which of the following statement is incorrect with respect to outliers?

A- Outliers cannot conform to the regression relationship

Q10-What do you understand by the term Normal Distribution?

A- The normal distribution is explained by two parameters Mean & Standard deviation.The normal distribution with Mean=0 and Standard deviation=1 is called as Standard normal distribution and this curve usually defines that how the collection of observations(data) is behaving is it normal? ,Skewed right?,Skewed left? This curve is also helpful to identify the percentage of data lying within 1 S.D.(68%)
2 S.D.(95%) and 3 S.D.(99.7%). When the curve falls from the peak on either sides,at first the slope is convex (bulging outwards) as it falls faster and faster but at a certain point this point is knowns

as point of inflection('bending back). The distance between point of inflection on either sides are equal to the Standard Deviation.

Q11-How do you handle missing data? What imputation techniques do you recommend?

A- Usually missing data is found in form of 'NAN', there are different kind of imputers to solve this situation at the ground depending upon the situation these imputers are-

a) *Simple imputer* – It works on the basic mean method for imputing the values.

b) *KNN imputer* - Alike KNN algorithm this imputer finds the two closest neighbors but the important part to remember is you have to pass continuous data alongwith the column where you want to impute the values,here it will find the number of neighbors(as mentioned in code example n_neighbors=2) and it will take the average of two close neighbors to fill the value.

c)*Iterative imputer* – This imputer works on prediction basis here you need to pass the continuous data and the outcome will be predicted by this imputer itself, this imputer internally predicts the outcome to fill NAN values

Q12- What is A/B testing?

A- A/B testing is example of statistical hypothesis testing ,a process where a hypothesis is made about the relationship of two data sets and those data sets are compared against each other to determine if there is a statistically significant relationship present or not present.

Q13- Is mean imputation of missing data acceptable practice?

A- No, it can somehow be used in symmetrical data distribution which is very scarce hereby using the mean imputation it distorts the relationship between variables.Their is a possibility of biasness which tends to affect confidence intervals and other inferential statistics.

Q14-What is linear regression in statistics?

A- Linear regression is simpler words can be defined as line of maximum likelihood / Best fit line where we have data points close to a line ,internally it follows a method of OLS(Ordinary least square technique) which tries to reduce the squared error between each data point and the line,where the error is distance between each point and the line you have.
It uses the formula y = mx+c(y=Prediction,x=Input value). The angle wgich line makes is your 'm' and the point from where we are drawing the line is your intercept i.e. 'c'

Q15-What are the various branches of statistics?

A- There are two Branches of statistics-
a) *Descriptive Statistics* – Descriptive Statistics defines the data where we are able to describe by using

- **Measure of Central Tendency – Mean(Average),Median(Mid-Value),Mode(highest occurring frequency)**

- **<u>Measure of Dispersion</u>- Variance and Standard deviation,here you can find how much data is varying from the mean or in simpler words how much dispersion is their in data.**

b) *<u>Infrential statistics</u>* – This statistics is purely based on sampling here the essence of central limit theorem plays a vital role and says that if you take any sample from your normal distribution and you plot the same it will also form normal distribution.
It also says that, Mean of sample mean = Mean of population mean