Statistics Worksheet

Q1. Bernoulli random variables take (only) the values 1 and 0.
Ans – 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?
Ans - a) Central Limit Theorem

Q3. Which of the following is incorrect with respect to use of Poisson distribution?
Ans. b) Modeling bounded count data

Q4. Point out the correct statement.
Ans. b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

Q5. _____ random variables are used to model rates.
Ans. c) Poisson

Q6. Usually replacing the standard error by its estimated value does change the CLT.
Ans. b) False

Q7. Which of the following testing is concerned with making decisions using data?
Ans. b) Hypothesis

Q8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
Ans. a) 0

Q9. Which of the following statement is incorrect with respect to outliers?
Ans. c) Outliers cannot conform to the regression relationship

Q10. What do you understand by the term Normal Distribution?
Ans.
In normal distribution, most of data occurs more around 'mean' of the data set, while the mean is set at ZERO, and standard deviation is 1.
It appears in bell curve shape.
As per central limit theorem, all datasets will appear normal as number of samples goes higher and higher.
It is symmetrical distribution.

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

a) Handling of missing data depends on the type of parameter that has data missing.

For some parameters for which adding data by calculated and compared estimation is not going to raise errors e.g., Rank, Score data, etc for these we can add data using Imputation techniques.
For other parameters that can raise certain errors for real world e.g., Year of manufacturer, etc we cannot impute as in this case Calculation and Estimation wont work. For such parameters we have to drop those rows with missing values entirely.

b) There are several imputation techniques which can be used. From them I present to you two methods :-
    i) Iterative Imputer – This Imputer works in the same manner as Regression Mod2el

ii) KNN Imputer – This Imputer works as the K-Nearest Neighbor Model

Q12. What is A/B testing?
Ans.
This testing method applies Two samples testing method from Hypothesis testing, and studies a parameter's relation with 2 different variables i.e A and B to determine which is causing more reaction/variation on the values of that parameter.
e.g A youtuber uploads same video with 2 different titles.
1st is titled "I'm going to consume heavy amount of calorie in 1 sitting"
2nd is titled "10,000 calories consumed in 1 sitting"
Purpose of this experiment was to find out which title brings in more viewers and using it to boost channel's views.
Result was that the video with B title i.e., 2nd title was viewed more than A i.e., 1st
In professional statistical analysis, this method would be used in more calculated, sophisticated manner.

Q13. Is mean imputation of missing data acceptable practice?
Ans.
Mean Imputation generates values and fills empty spaces based on the data of that particular column, instead of studying values and relations with other parameters/features.
This could end up causing more harm than good.
e.g if we have to impute missing salary of an employee of high-end Management position, but by mean imputation he would get assigned salary of mid-level associate or at best a senior associate, causing our data to faulter and statistical analysis to give wrong results.
For this reason, mean imputation is not acceptable practice.

Q14. What is linear regression in statistics?
Ans.
Linear Regression is basically studying the relation between one dependant variable and one independent variable, studying the result of variation in dependant variable with respect to the other independent variable.
$Y = mX + C$ is the formula that is used to calculate the dependency.
Where,
Y – Dependant variable
X = Independent variable

Q15. What are the various branches of statistics?
Ans.
Below are the parts in which Statistics can be divided:-

1) Theoretical Statistics – This is the branch that deals with consolidating sets of data to form statistical distribution to be analysed.
2) Statistical functions – This one deals with analysing the data into groups and makes it available to present data further to be interpreted. This includes tabulation, calculations, etc.
3) Descriptive Statistics – This branch helps plot diagrams, graphs and charts using the provided data.
4) Inferential Statistics – This branch deals with making interpretations based on presented data and analyses relations between the variables.