Statistics Assignment

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? a) Modeling event/time data

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

Ans: b) False

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

Ans: b) Hypothesis

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

Ans: a) 0

9. 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

10. What do you understand by the term Normal Distribution?
Ans: The Normal Distribution is called the Gaussian Distribution it is the most significant continuous probability distribution. A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics. It can be used to approximate other probability distributions.

A normal distribution is the proper term for a probability bell curve.
In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
Normal distributions are symmetrical, but not all symmetrical distributions are normal

11. How do you handle missing data? What imputation techniques do you recommend?
Ans: When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data. The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.
The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

One of the most common methods of imputing values when dealing with missing data is using Mean, Median and Mode.

12.What is A/B testing?
Ans: A/B testing also known as bucket testing or split-run testing is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective

13. Is mean imputation of missing data acceptable practice?
Ans: Yes, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased.
And by imputing the mean, you are able to keep your sample size up to the full sample size. This is the original logic involved in mean imputation.

Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

14. What is linear regression in statistics?
Ans: linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables which is also known as dependent and independent variables.

The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable

15. What are the various branches of statistics?
Ans: There are two main branches of statistics – A) Inferential Statistic. B) Descriptive Statistic.

Inferential Statistics: Inferential statistics used to make inference and describe about the population. These stats are more useful when it's not easy or possible to examine each member of the population. Descriptive Statistics: Descriptive statistics are used to get a brief summary of data. You can have the summary of data in numerical or graphical form.