Statistics worksheet-1

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

Answer- 1. A). True

Question-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.

Answer-2. A). Central Limit Theorem

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

Answer-3.B) Modeling bounded count data

Question-4. Point out the correct statement.

Answer-4. D) All of the mentioned

Question-5. _____ random variables are used to model rates.

Answer-5. C) Poisson

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

Answer-6. B) False

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

Answer-7.B)) Hypothesis

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

Answer-8. A) 0

Question-9. Which of the following statement is incorrect with respect to outliers.

Answer-9.C) Outliers cannot conform to the regression relationship

Question-10. What do you understand by the term Normal Distribution.

Answer-10. The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions.

The normal distribution model is motivated by the central limit theoram. This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

Question-11. How do you handle missing data? What imputation techniques do you recommend.

Answer-11. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations.

Imputation Techniques

- Complete Case Analysis(CCA):- This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data i.e we consider only those rows where we have complete data i.e data is not missing. ...
- Arbitrary Value Imputation. ...
- Frequent Category Imputation.

Question-12. What is A/B testing

Answer-12. A/B testing also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics.

A/B testing is one of the components of the overarching process of conversion rate optimization, using which you can gather both qualitative and quantitative user insights. You can further use this collected data to understand user behavior, engagement rate, pain points, and even satisfaction with website features, including new features, revamped page sections, etc. If you're not A/B testing your website, you're surely losing out on a lot of potential business revenue.

Question-13. Is mean imputation of missing data acceptable practice.

Answer-13. The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-

year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Question-14. What is linear regression in statistics

Answer-14. . In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

Linear Regression is one of the regression technique and can be defined as the following: "Linear Regression is a field of study which emphasizes on the statistical relationship between two continuous variables known as Predictor and Response variables".

Question-15. What are the various branches of statistics.

Answer-15. Statistics is a study of presentation, analysis, collection, interpretation and organization of data

There are two main branches of statistics

- Inferential Statistic.
- Descriptive Statistic.

Inferential Statistics:

Inferential statistics used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population.

Descriptive Statistics:

Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphycal form.