STATISTICS WORKSHEET-1 (Answers)

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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

Answer: 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?

Answer: a) Central Limit Theorem

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

4. Point out the correct statement.
a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
c) The square of a standard normal random variable follows what is called chi-squared distribution
d) All of the mentioned

Answer: d) All of the mentioned

5. _____ random variables are used to model rates.

Answer: c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.
Answer: b) False

7. Which of the following testing is concerned with making decisions using data?
Answer: b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
Answer: a) 0

9. Which of the following statement is incorrect with respect to outliers?
Answer:

(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?
Answer:

A normal distribution curve is symmetrical, bell shaped curve defined by the mean and standard deviation of a data set. The normal curve is a probability distribution with a total area under the curve of 1.

11. How do you handle missing data? What imputation techniques do you recommend?
Answer:

To handle missing data I suggest 3 techniques,

1. Delete the record missing value.

   Only if it is a huge dataset, delete the record.

2. Create a separate modes to handle missing value.

   By creating a separate model we can predict the output for the missing value. This method takes huge time to complete. When this method we can use is if the dataset is small otherwise no.

3. Then Using statistical methods Mean, Median or Mode.
   - By using the Average of the mean data (Then replacing the NaN / missing value).
   - By using the Median method (change the data to Sorting format- then replacing the missing value).
   - By using Mode (which ever the value is having more frequency. That value will be replacing to the NaN or missing data).

12. What is A/B testing?
Answer:

Its a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. A/B testing is one of the most prominent and widely used statistical tools.

13. Is mean imputation of missing data acceptable practice?
Answer:

It’s a bad practice.

Mean imputation preserves the mean of the observed data. Leads to an underestimate of the standard deviation. Distorts relationships between variables by “pulling” estimates of the correlation toward zero.

14. What is linear regression in statistics?
Answer:

Linear Regression is a linear model that assumes a linear relationship between input variables (independent variables 'x') and output variable (dependent variable 'y') such that 'y' can be calculated from a linear combination of input variables (x).

15. What are the various branches of statistics?
Answer:

There are two main branches of statistics:

1. Descriptive statistics.
2. Inferential statistics.

Statistics are used to describe or summarize the characteristics of a sample or data set, such as a variable's mean, standard deviation, or frequency. Inferential statistics, in contrast, employs any number of techniques to relate variables in a data set to one another, for example using correlation or regression analysis. These can then be used to estimate forecasts or infer causality.