STATISTICS WORKSHEET-1

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.

   a) True

   b) False

Solution: a

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?

   a) Central Limit Theorem

   b) Central Mean Theorem

   c) Centroid Limit Theorem

   d) All of the mentioned

Solution: a

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

   a) Modeling event/time data

   b) Modeling bounded count data

   c) Modeling contingency tables

   d) All of the mentioned

Solution: b

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

Solution: d

5. _____ random variables are used to model rates.

   a) Empirical

   b) Binomial

c) Poisson

d) All of the mentioned

Solution: c

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

a) True

b) False

Solution: b

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Solution: b

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

a) 0

b) 5

c) 1

d) 10

Solution: a

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

a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes

c) Outliers cannot conform to the regression relationship

d) None of the mentioned

Solution: c

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?

Solution: A normal distribution, sometimes called the bell curve where a constant proportion of data points lies under the curve between the mean and a specific number of standard deviation.

In normal distribution, the mean and median are equal, and 68% of the data lies within 1 standard deviation. It is also known as Gaussian distribution.

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

Solution: As we know that the missing data can reduce the statistical power of the analysis, which can distort the validity of the results. It is very important to understand why the data is missing before deciding which approach to employ.

I strongly suggest the following imputational techniques to handle this better. They are

1. **Mean, median or mode imputation:** Here we use mean or median or mode value to use of the non-missing observations. This can be useful in cases where the number of missing data is low. However, for large number of missing values, using mean or median can result in loss of variation in data.
2. **Regression analysis:** Regression is useful for handling missing data because it can be used to predict the null value using other information from the dataset. There are several methods of regression analysis, like stochastic regression.
3. **Random forest:** It works well with both data missing at random and not missing at random. It is applicable to various variable types since it uses multiple decision tree to estimate missing values.

12. What is A/B testing?

Solution: A/B testing (also known as split testing or bucket testing) which consists of a randomized experiment with two variants. 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?

Solution: True, if the data are missing completely at random, the estimation of mean remains unbiased. But outliers data points will have a significant impact on the mean and hence, in such cases, it is not recommended to use mean for replacing the missing values.

14. What is linear regression in statistics?

Solution: The linear approach for modelling the relationship between a scalar response and one or more explanatory variables is called linear regression. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

15. What are the various branches of statistics?

Solution: The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data.