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

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?

a) Central Limit Theorem b) Central Mean Theorem

c) Centroid Limit Theorem d) All of the mentioned

Ans. A) Central Limit Theorem

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

Ans. 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

Ans. D)

5. _____ random variables are used to model rates.

a) Empirical b) Binomial c) Poisson d) All of the mentioned

Ans. C)

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

a) True b) False

Ans. b) False

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

a) Probability b) Hypothesis c) Causal d) None of the mentioned

Ans. b) Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data. a) 0 b) 5 c) 1 d) 10

Ans. A) 0

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

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?

In a normal distribution, data is symmetrically distributed with no skew. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the centre. Normal distributions are also called Gaussian distributions or bell curves because of their shape.

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

Missing data can be handled in a variety of ways. isnull() and dropna() will help to find the columns/rows with missing data and drop them. fillna() will replace the wrong values with a placeholder value. Fill the missing values with mean, median or mode depending on the variable.

Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values. Imputation techniques like knn imputer and iterative imputers recommended.

12. What is A/B testing?

A/B testing is a form of statistical hypothesis testing and two-sample hypothesis testing. Statistical hypothesis testing is a method in which a sample dataset is compared against the population data. Two-sample hypothesis testing is a method in determining whether the differences between the two samples are statistically significant or not.

13. Is mean imputation of missing data acceptable practice?

Although imputing missing values by using the mean is popular imputation technique, there are serious problems with mean imputation. The variance of a mean-imputed variable is always biased downward from the variance of the un-imputed variable. Mean imputation does not preserve the relationships among variables. Because the imputations are themselves estimates, there is some error associated with them. Ultimately, because standard errors are too low, so are p-values which may rise Type I errors without realizing it. That's not good.

14. What is linear regression in statistics?

Linear regression shows the linear relationship between two variables by applying a linear equation to observed data. One variable is supposed to be an independent variable, and the other is to be a dependent variable. For example, the weight of the person is linearly related to his height. Hence this shows a linear relationship between the height and weight of the person. As the height is increased, the weight of the person also gets increased.

15. What are the various branches of statistics?

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 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 are used to get a brief summary of data. You can have the summary of data in numerical or graphical form.