WORKSHEET 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) Cental 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) All of the mentioned

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Ans. c) Poisson

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

a) True

b) False

Ans. a) True

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?

Ans. Normal distribution aka Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

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

Ans. Ways to handle missing data :

a) Deleting rows with missing values
b) Impute missing values for continuous variable
c) Impute missing values for categorical variable
d) Other Imputation Methods
e) Using algorithms that support missing values
f) Prediction of missing values
g) Imputation using Deep Learning – Datawig

Common imputation techniques recommended are: Mean or median imputation, Multivariate Imputation by Chained Equations (MICE) and Random Forest.

12. What is A/B testing?

Ans. A/B testing aka Split testing or Bucket testing is a method of comparing two versions of a webpage or app against each other to determine which performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

13. Is mean imputation of missing data acceptable practice?

Ans. 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. But Mean reduces a variance of the data. Since most research studies are interested in relationship among variables, mean imputation is not a good solution.

14. What is linear regression in statistics?

Ans. In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. 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?

Ans. There are three real branches of statistics

a) Data Collection
b) Descriptive Statistics
    - Measures of central tendency: mean, median and mode
    - Measures of frequency: Count, percent and frequency
    - Measures of variation or dispersion: Range, variance, standard deviation
    - Measures of position: Percentile ranks, Quartile ranks
c) Inferential Statistics
    - Parameter Estimates: Point estimate and Interval Estimate
    - Hypothesis Testing: Null Hypothesis and Research or Alternative Hypothesis

❖ The following types of inferential statistics are extensively used and relatively easy to interpret:
    - ➢ One sample test of difference/ One sample hypothesis test
    - ➢ Confidence Interval
    - ➢ Contingency Tables and Chi Squared Statistic
    - ➢ T-test or Anova
    - ➢ Pearson Correlation
    - ➢ Bi-variate regression
    - ➢ Multi-variate regression