

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

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?

- a) Central Limit Theorem
- b) Central Mean Theorem
- c) Centroid Limit Theorem
- d) All of the mentioned

ANSWER: a) Central Limit Theorem(CLT)

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

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.

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

ANSWER: c) Poisson

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

- a) True
- b) False

ANSWER: b) False

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

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

ANSWER: b) Hypothesis

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

ANSWER: 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

ANSWER: c) Outliers cannot conform to the regression relationship

WORKSHEET

Q10 and 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 is the proper term for a probability bell curve . In a normal distribution the mean is 0 and the standard deviation is 1. It is further divided into three types: symmetric, left skewed, right skewed.

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

Answer : Missing values in any data creates problem for model building/analysis. Treatment of such values is priority.

If the missing values in the data are very small as in less than 1% then Deleting the particular Row will be my first choice

If the missing values can also be eliminated in pairs or groups

If the missing values in the data are not relevant for analysis/modelling then Deleting the Entire Column will be my first choice

If above mentioned doesn't apply then Imputing the Missing Value is the only option some the methods are Replacing With Mean, Replacing With Mode (categorical variable), Replacing With Median, Replacing with Previous Value – Forward Fill, Replacing with Next Value – Backward Fill, Interpolation Nearest Neighbors Imputations (KNNImputer), use of mice package.

12. What is A/B testing?

Answer : A/B testing is the experimental methodology used in order to pick the better version of the same application/webpage/task . Two variant of the same type A and B are taken and then based on performance/user experience one of them is selected to be put forward in the production phase. It is an application of sample hypothesis testing that is done to avoid unnecessary risks by allowing us to target our resources for maximum effect and efficiency.

13. Is mean imputation of missing data acceptable practice?

Answer : 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 if there are outliers in the dataset, Outliers data points will have a significant impact on the mean and hence, in such cases, it is not recommended to use the mean for replacing the missing values. Using mean values for replacing missing values in such cases will not create a great model.

14. What is linear regression in statistics?

Answer : Linear regression is used to deduce a the relationship between two continuous variables with the objective to predict the value of an output variable (or response) based on the value of an input (or predictor) variable. It shows the value with help of a trend line.

15. What are the various branches of statistics?

Answer: There are 3 branches of statistics:

- 1) Data collection
- 2) Descriptive
- 3) Inferential.