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 Theorm

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

Anwer: 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

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: Normal Distribution is probability distribution that is symmetric about the mean. There is no skewness. Most data cluster at the center. We can say mean, median and mode are same.

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

Answer: Best ways to deal with missing data is to substitute using mean, median and mode.

12. What is A/B testing?

Answer: It's randomized experimentation process where two or more versions of a variable are exposed to different segment of customer at the same time to determine which version leaves maximum impact and drives business.

Let's consider two alternative design A & B. Visitors of a website are randomly served with one of any two. Then data about their activity is collected by web analytics. Now using A/B testing it can be determined which design has better business metrics.

13. Is mean imputation of missing data acceptable practice?

Answer: In some cases it doesn't make sense. For example suppose we have a table having value age and fitness score and an 8 year old is missing fitness score. If we substitute it by mean of people having age b/w 15-30 year old then an 8 year will have fitness score more than what he actually has.

It means mean imputation ignores feature correlation.

14. What is linear regression in statistics?

Answer: Linear Regression is process of predicting the value of dependent variable from independent variable.

$Y = mx + C$

Y = Dependent variable

x = Independent variable

C = Intercept

M = slope

Change in value of x leads to change in value of y.

15. What are the various branches of statistics?

Answer:

Variable: Variable can be numerical and categorical.

Descriptive Statistics: It deals with collection of data. Descriptive statistics is measure of central tendency. We are able to measure central tendency through mean, mode and median.

Variance and std. deviation helps in finding measure of dispersion.

Inferential Statistics: Drawing the right conclusion from statistical analysis that has been performed by descriptive statistics.

Hypothesis helps us in drawing an inference and coming to right conclusion.