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)

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)

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)

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)

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)

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)

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)

Note: Q10 to Q15 are on next page.

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

10. What do you understand by the term Normal Distribution?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

Answers:

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

Ans: Normal Distribution is also known as the 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 than data far from the mean. The normal distribution appears as a 'bell curve' when graphed.

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

Ans: Replacing missing entries with the average (mean), median (middle value) or the most frequent Value of the corresponding column. This is a quick and easy approach but it can introduce bias if the Missing data isn't randomly distributed.
Imputation techniques are: -
-Next or Previous Value
-K Nearest Neighbors
-Missing Value Prediction
-Most Frequent Value
-Average or Linear Interpolation
- Mean, Median and Mode (above discussed)
-Fixed Value

12. What is A/B testing?

Ans: A/B testing (also called split testing or bucket testing) is methodology for comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random and statical analysis is used to determine which variation performs better for a given conversation goal.

13. Is mean imputation of missing data acceptable practice?

Ans: No, it depends on situation.
Mean imputation is typically considered terrible practice since it ignores feature correlation.
Consider a scenario: we have a table with age and fitness scores and an eight-year-old has a missing Fitness score. If we average the fitness scores of people between the ages of 15 and 80, the wight year old boy will appear to have a significantly greater fitness level than he actually does.
Second, mean imputation decreases the variance of our data while increasing bias. As a result of reduced variance, the model is less accurate and confidence level is narrower.

14. What is linear regression in statistics?

Ans: Linear regression in statistics predicts the relationship between two variables by assuming they have a straight
line connection. It finds the best line that minimizes the difference between predicted and actual values. Ex. The wight of the person is linearly related to their height. So, this shows a linear relationship between the height and weight of the person. According to this, as we increase the height, the weight of the person will also increase.

15. What are the various branches of statistics?

Ans: There are two main branches of statistics- descriptive statistics and Inferential statistics.

Descriptive Statistics- Descriptive Statistics is considered as the first part of statistical analysis which deals with Collection and presentation of data. Scientifically, descriptive statistics can be defined as brief explanatory coefficients that are used by stasticians to summarize a given data set.

Inferential Statistics: Inferential statistics are techniques that enable statisticians to use the gathered information from a sample to make inferences, decisions or predictions about a given population. Inferential statistics often talks in probability terms by using descriptive statistics. These techniques are majorly used by statisticians to analyze data, make estimates and draw conclusions from the limited information.