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
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. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Answer: - 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
   Answer: - 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

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

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 also known as bell curve can be defined as the probability distribution in which all the values are plotted in symmetrical fashion and most of the data is situated around the probability mean. The normal distribution has two main parameters, the mean and the standard deviation. The mean, median and the mode and equal.

Below are some of the properties of normal distribution

1). The mean, median and mode are all equal.
2). The curve is symmetric at the center.
3). The half of the values are to the left side of the center and half of the values are to the right side of the center
4). The total are under the curve is 1.

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

Answer: - Data can be missing in the following ways:-

- **Missing Completely At Random (MCAR):** When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random.
- **Missing At Random (MAR):** The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data.
- **Not Missing At Random (NMAR):** When the missing data has a structure to it, we cannot treat it as missing at random.

Imputation Techniques: -

1. Mean or Median Imputation
2. Multivariate Imputation by Chained Equations (MICE)
3. Random Forest

You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value.

In Pandas, there are two very useful methods: isnull () and dropna() that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill the invalid values with a placeholder value (for example, 0), you could use the fillna() method.

12. What is A/B testing?

Answer: A/B testing is a basic randomized control experiment. It is way to compare two versions of a variable to find out which performs better in a controlled environment. A/B testing is one of the most prominent and widely used statistical tools.

13. Is mean imputation of missing data acceptable practice?

Answer: Mean imputation is not an acceptable practice as there lot of repercussion one can face after implementing this.

1).Mean imputation can lead to bias in multivariate estimates such as correlation and regression coefficient. Values that are imputed by a variable's mean have, in general, a correlation of zero with other variables. Relationships between variables are therefore biased toward zero.

2). Standard errors and variance of imputed variables are biased. For instance, let's assume that we would like to calculate the standard error of a mean estimation of an imputed variable. Since all imputed values are exactly the mean of our variable, we would be too sure about the correctness of our mean estimate. In other words, the confidence interval around the point estimation of our mean would be too narrow.

3).If the response mechanism is MAR or MNAR, even the sample mean of your variable is biased (compare that with point 3 above). Assume that you want to estimate the mean of a population's income and people with high income are less likely to respond; your estimate of the mean income would be biased downwards.

14. What is linear regression in statistics?

Answer: In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

15. What are the various branches of statistics?

Answer: There are two branches of statistics

Descriptive Statistics:- Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Inferential Statistics:-involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.