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: A probability Bell Curve is more appropriately described as a Normal Distribution. The mean of a normal distribution is zero. Any positive standard deviation belongs to the normal distribution. We are aware that the mean aids in establishing the symmetry line of a graph and that the standard deviation aids in determining how widely distributed the data are. The graph gets narrower and the data are closer together if the standard deviation is less. The data are more evenly distributed and the graph becomes wider as the standard deviation increases. The area under the normal curve is divided using the standard deviations. The percentage of data that falls within each split section's respective graph region is indicated.

11. How do you handle missing data? What imputation techniques do you recommend?
Answer: The Things to handle missing data are given as:
1. We Can Use each column's "mean" value. The mean along each column is used to fill in the NaN values.
2. We Can Use the value marked "most frequent" in each column.
3. Let's now have a look at a fresh DataFrame containing category features.
4. In each column, use "interpolation."
5. Alternatively, we might use K-Nearest Neighbour.

The Imputations Techniques I Would Recommend Would Be As Follows:

1. Zero Replacement: In Zero Replacement case, we can substitute zero for the missing value in all circumstances.
2. Maximum or Minimum Replacement: Replace the missing value with the minimum or maximum value of the feature using the min or max replacement method.
3. Mean , Median or Mode Replacement: Use the mean, median, or mode value to fill in any missing data.
4. Additionally, the value of the preceding cell may be used to replace the value of the missing cell. This kind of method is frequently used for entering time series data. For instance, it makes sense to substitute the price from the (i-1)-th day for an instrument whose price is missing on the i-th day.

12. What is A/B testing?
Answer: Statistical hypothesis testing, or statistical inference, is essentially what A/B testing entails. It is a decision-making analytical technique that uses sample statistics to estimate population parameters.
The population of your website (or particular group of pages) refers to all visitors, whereas the sample refers to the number of visitors that took part in the test.
Imagine that we decide to make a change to one of your product pages based on the results of an A/B test that evaluated a "sample" of your website's users. In the end, just a portion of the visitors viewed the challenger, so naturally, not all of the visitors did. With A/B testing, you assume that if the challenger (i.e., variation) in the test increased conversions for a group of visitors on product pages, it will subsequently have the same result for all visitors to your product pages (we will explore the accuracy of a variation's validity in more detail later).

13. Is mean imputation of missing data acceptable practice?
Answer: Mean imputation of missing data is generally bad practise.
If only estimating, then Mean imputation keeps the observed data's mean intact. causes the standard deviation to be understated has the effect of "pushing" estimates of the correlation toward zero, which distorts correlations between variables.

14. What is linear regression in statistics?

Answer: A fundamental and widely used form of predictive analysis is linear regression. Regression analysis' main goal is to look at two things: (1) Is it possible to accurately forecast an outcome (dependent) variable using a set of predictor variables? (2) Which individual variables—as shown by the size and sign of the beta estimates—are highly important predictors of the outcome variable, and how do they affect the outcome variable? The link between one dependent variable and one or more independent variables is explained using these regression estimations.

The formula y = c + b*x, where y is the estimated score of the dependent variable, c is a constant, b is the regression coefficient, and x is the score on the independent variable, defines the simplest form of the regression equation with one dependent and one independent variable.The variables' identities, The dependant variable in a regression has many different names. It can be referred to as a regressand, endogenous variable, criteria variable, or outcome variable. The independent variables may also be referred to as predictor variables, regressors, or exogenous variables.

15. What are the various branches of statistics?

Answer: There are Two Branches of Statistics Mainly Known as:

1. Descriptive Statistics
2. Inferential Statistics

1. Descriptive Statistics : The first area of statistics that deals with data collecting is descriptive statistics. People believe it to be too simple, but it is not. The design and experiments must be known to the statisticians. Additionally, they must choose the appropriate focus group and avoid biases. Descriptive statistics, on the other hand, are utilised to do numerous types of analysis on diverse research.

   Descriptive statistics have two parts :

   - Central tendency measures
   - Variability measures

2. Inferential Statistics: Through using inference statistics, statisticians can draw conclusions, make decisions, or forecast the activity of a certain population using data from a sample. By applying descriptive statistics, inference statistics frequently use probabilistic language. Additionally, a statistician uses these methods to generate, evaluate, and draw conclusions based on scant data. That is discovered by collecting samples and evaluating their dependability. The majority of future projections and generalization based on a smaller sample population research fall within the purview of inference statistics. Additionally, the majority of social science experiments concentrate on analysing a small sample population that sheds light on how the community behaves.

   Inferential statistics have five parts:

   - Regression analysis
   - Analysis of variance (ANOVA)
   - Analysis of covariance (ANCOVA)
   - Statistical significance (t-test)
   - Correlation analysis