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
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
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
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
5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   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
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
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

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 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 far from the mean. In graph form, normal distribution will appear as a bell curve.

The standard normal distribution has two parameters-the mean and the standard deviation.

The mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.

Normal distributions are symmetrical, but not all symmetrical distributions are normal.

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

Ans: There are 3 Common Methods to Handle Missing Data.

1. Mean or Median Imputation.When data is missing at random,we can use list-wise or pair-wise deletion of the missing observations.
2. Multivariate Imputation by Chained Equations MICE assumes that the missing data are Missing at Random (MAR).
3. Random Forest.

12.What is A/B testing?

Ans: A/B testing is basically statistical hypothesis testing or statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics.

Start the A/B testing process by making a claim (hypothesis). Launch your test to gather statistical evidence to accept or reject a claim (hypothesis) about your website visitors. The final data shows you whether your hypothesis was correct, incorrect or inconclusive.

13. Is mean imputation of missing data acceptable practice?

Ans: No, It's a Bad practice in general. It is a non-standard, but a fairly flexible imputation algorithm. It uses RandomForest at its core to predict the missing data. It can be applied to both continuous and categorical variables which makes it advantageous over other imputation algorithms.

14. What is linear regression in statistics?

Ans: Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula

$$y = c + b*x$$

where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Types of Linear Regression:

1.Simple Linear Regression
2.Multiple Linear Regression
3.Logistic Linear Regression
4.Ordinal Linear Regression
5.Multinomial Linear Regression
6.Discriminant Linear Regression

15. What are the various branches of statistics?
Ans: The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.