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?

This is a well known distribution of values symmetrically around it's mean. Most of the observations spread out around the peak of the dataset equally in both directions. It looks like a bell shaped curve.

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

Good data = good analysis. If there are missing values in the data, the conclusion or the analysis could be biased and incomplete. In order to deal with missing data, two primary methods are used. Imputation or removal of missing data. First it is important to understand why the data is missing .

Imputation is method to develop reasonable guesses for the missing data. This technique works well when the amount of missing data is relatively lower than the total amount of data.

The techniques based on factors like the amount of missing data, used to impute data are:

1) mean, median
2) Time-Series specific methods
3) LOCF & NOCB (Observation carried forward or backward)
4) Linear interpolation
5) Multiple imputation
6) k nearest neighbors

12. What is A/B testing?

It is an experiment run in several cases to compare and find out a better option. A very common scenario is creating 2 different versions of a web page, version A and version B and let different website visitors view these versions randomly and then statistically comparing which version brings better conversion and engagement on the website.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is a method in which the missing values are filled in by the mean of all the values in a variable. Doing so could lead to a severely biased estimates while we use this dataset. Hence there are other better alternatives to this practice.

14. What is linear regression in statistics?

Linear regression is a statistical method used to be able to predict a value based on another value. Let us say there are 2 variables x and y where, x is the independent variable and y is the dependant variable. By plotting the available x and y variables on an xy- plane and using a technique like RSME, we could come up with a function like $y = f(x)$ or formula for a straight line like $y = mx + b$.

Using this formula, we could predict what the y value would be like, if we know an x value.

15. What are the various branches of statistics?

The 3 branches of statistics are data collection, descriptive statistics and inferential statistics.