WORKSHEET
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.
Ans: 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?
Ans: a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
Ans: a) Modeling event/time data

4. Point out the correct statement.
Ans: d) All of the mentioned

5. ______ random variables are used to model rates.
Ans: c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
Ans: a) True

7. 1. Which of the following testing is concerned with making decisions using data?
Ans: b) Hypothesis

8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.
Ans: a) 0

9. Which of the following statement is incorrect with respect to outliers?
Ans: c) Outliers cannot conform to the regression relationship

WORKSHEET

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

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

11. How do you handle missing data? What imputation techniques do you recommend?
Ans: Missing value can bias the results of the machine learning models and/or reduce the accuracy of the model.

We have a few algorithms like K-nearest and Naive Bayes support data which helps to handle missing values.

1. Mean or Median Imputation

When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

2. Multivariate Imputation by Chained Equations (MICE)

MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

3. Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

12. What is A/B testing?

Ans: It is also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metri

13. Is mean imputation of missing data acceptable practice?

Ans: Yes, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased.

14. What is linear regression in statistics?

Ans: It is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable

15. What are the various branches of statistics?

Ans: There are three branches of statistics:

1. **Data Collection**: It is all about how the actual data is collected. For the most part, this needn't concern us too much in terms of the mathematics (we just work with what we are given), but there are significant issues to consider when actually collecting data.
2. **Descriptive Statistics** :It is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on)
3. **Inferential Statistics**
   Inferential statistics is the aspect that deals with making conclusions about the data.