Parth Makwana (DS0722)

ANSWERS_STATISTICS WORKSHEET-1

Q1 TO Q9 ANSWERS

1. Bernoulli random variables take (only) the values 1 and 0.
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?
b) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
b) Modeling bounded count data

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

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

6. Usually replacing the standard error by its estimated value does change the CLT.
a) False

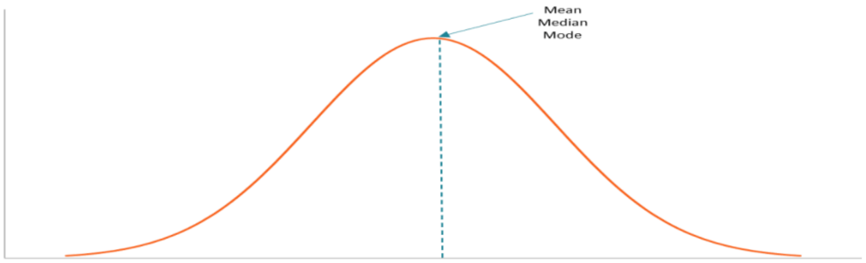
7. Which of the following testing is concerned with making decisions using data?
b) Hypothesis

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

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

Q10 AND Q15 ANSWERS

10. What do you understand by the term Normal Distribution?
The Normal Distribution is a core concept of statistics & it's application in Data Science. The Normal Distribution when visualized it's more of a Bell Curved dipiction of Data. Normal Distribution is symmetric about the mean (center) the data near the mean are more frequent in occurence than data far from the mean.



11. How do you handle missing data? What imputation techniques do you recommend?
Missing data (NAN Values).This missing data will impact on the prediction of model. In order to Handle missing value's we should use mean for continues data and mode for Discrete data. But not all the time mean or mode method is recommended. If data set is small we can always use FillNa method, But while working on BigData set's we should use Imputers
Recommended Imputers:

- ✓ Knn Imputer
- ✓ Iterative Imputer
- ✓ Mean Imputation

12. What is A/B testing?
A/B testing is also known as split testing, we split data in two groups, and show two different versions of data sets with the goal of comparing the results to find the more successful version.

The process of replacing the NAN (NULL) values in data set with the data's mean is known as mean imputation. Typically Mean Imputation is considered a terrible pracitice as it ignores the features correlation. This method also decreases the variance of out data while increasing bias.

13. Is mean imputation of missing data acceptable practice?
As a result of reduced variance, the model is less accurate and the confidence interval is narrower

14. What is linear regression in statistics?
Linear regression is part of Supervised ML. It consist of Input Data & Output Data this terms are known as Features(Independent variables/predictors) & Labels(target/class).
Regression in statistics is the process of predicting a Label(or Dependent Variable) based on the features(Independent Variables) at hand. Regression is used for time series modelling and finding the causal effect relationship between the variables and forecasting. For example, the relationship between the stock prices of the company and various factors like customer reputation and company annual performance etc. can be studied using regression.
Regression analysis is an important tool for analysing and modelling data. Here, we fit a curve/line to the data points, in such a manner that the differences between the distance of the actual data points from the plotted curve/line is minimum. The topic will be explained in detail in the coming sections.

15. What are the various branches of statistics?

- Types of Statistics
  - Descriptive
  - Inferential
- Population and Sample
  - Parameter and Statistics (Mean, Median and Mode)