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

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

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. 10. Usually replacing the standard error by its estimated value does change the CLT.
   b) False

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

8. 4. 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

FLIP ROBO

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

d) What do you understand by the term Normal Distribution?
e) How do you handle missing data? What imputation techniques do you recommend?
f) What is A/B testing?
g) Is mean imputation of missing data acceptable practice?
h) What is linear regression in statistics?
i) What are the various branches of statistics?

FLIP ROBO

1. What do you understand by the term Normal Distribution?
   Normal Distribution is the distribution of sample data on a graph that forms the symmetric form such that the mean, median and mode will be similar or close to each other. The frequency of the data points will be more in the middle as compare to sides of the graph. There will be no skewness in the data in any side. The standard deviation will be smaller in Normal Distribution.

2. How do you handle missing data? What imputation techniques do you recommend?
   Most of the time there will be missing data in the entire data which will cause distortion in ML model and distribution of the data. So, it is important to handle missing data so there are some techniques to handle missing data
   1. Mean – taking a mean of the entire data and filling those numbers in NaNs will help the data to form normal distribution which leads to building a tuned ML model.
   2. Median – in some cases we cannot take mean of the data to fill NaNs so we can use the median of the data to fill NaNs.
   3. Mode – this can be mostly use in categorical data (True or False, Yes or No, 1 or 0) where we can not use mean so there, we can fill NaNs with Mode method.
   4. Random – Depending on the data type we can use the random numbers or data points to fill NaNs.
   5. Keep the NaNs as it is – in case if there are missing names, address, any personal info, phone numbers or any other info we can ignore the missing data depending on the type of data.

3. What is A/B testing?
   A/B testing is the way to compare two outcomes. It helps to choose the best fitted outcome or result as per requirements. In making any hypothesis it is important to build Null hypothesis and alternative hypothesis. A/B testing allows researcher to make one decision that is backed by some statistics like P-value.

FLIP ROBO

4. Is mean imputation of missing data acceptable practice?
   Yes, mean imputation of missing data is acceptable practice, however it depends on the data type. For categorical data such as Yes or No, True or False, 1 or 0 we cannot use mean imputation. We can use mean in case
   E.g: There is a data of Nation-wide sales of the company representing sales number in different stats on different day. So, in this case suppose there is a missing data of Goa stats for the one “Sunday” so we can take mean of the other Sunday’s sales data of that month.
   E.g: There is a dataset of diabetes patients and there are some patients having some missing data in blood pressure however they share same symptoms with other patients. So we can take the mean of other patients’ blood pressure numbers which are having same symptoms and can fill NaNs.

5. What is linear regression in statistics?
   It is the relationship between two or more datapoints those are having some linear relationship. We can use linear regression to predict the other value based on the values we have. There are two elements in the linear regression, independent and dependent variable.
   E.g: with the assumption of “Other things being equal”
   There is a linear relationship between number of years of experience and the salary amt. The greater number of experiences one has will get more salary. In this case salary is the dependent variable and experience is the independent variable.

6. What are the various branches of statistics?
   There are two type of branches 1 Descriptive statistics 2 Inferential statistics
   1. Descriptive statistics – it is the statistics that uses different ways to describe and present the data such as Graphs, tables, standard deviation and formals such as mean, median and mode.
   2. Inferential statistics - Inferential statistics makes the use of various analytical tools to draw findings about the data. It can be used to make decisions.