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

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

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

A) A probability function that specifies how the values of a variable are distributed is called Normal distribution. It is symmetric since most of the observations assemble around the central peak of the curve. The probabilities for values of the distribution are distant from the mean narrow off evenly on both directions. It is also called the Gaussian Distribution, is the most important continuous probability distribution. Sometimes it is also called a bell curve.
Some of the important properties of the normal distribution are -

• In a normal distribution, the mean, mean and mode are equal. (i.e., Mean = Median= Mode).
• The total area under the curve should be equal to 1.
• The normally distributed curve should be symmetric at the centre.
• There should be exactly half of the values are to the right of the centre and exactly half of the values are to the left of the centre.
• The normal distribution should be defined by the mean and standard deviation.
• The normal distribution curve must have only one peak. (i.e., Unimodal)
• The curve approaches the x-axis, but it never touches, and it extends farther away from the mean.

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

A) Missing data can be handled in following ways –
Deleting Rows This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values.
Replacing With Mean/Median/Mode -This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values.
Assigning An Unique Category - A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values.
Predicting The Missing Values - Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm. This method may result in better accuracy, unless a missing value is expected to have a very high variance.
It is not necessary to handle a particular dataset in one single manner.
One can use various methods on different features depending on how and what the data is about. Having a small domain knowledge about the data is important, which can give you an insight about how to approach the problem.

12. What is A/B testing?

A) A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment, for instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

13. Is mean imputation of missing data acceptable practice?

A) Mean imputation is acceptable practice; however, this method should be avoided as far as possible because -

1. Mean imputation does not preserve the relationships among variables.
2. Mean Imputation Leads to An Underestimate of Standard Errors

14. What is linear regression in statistics?

A) Linear regression strives to show the relationship between two variables by applying a linear equation to observed data. One variable is supposed to be an independent variable, and the other is to be a dependent variable. For example, the weight of the person is linearly related to his height. Hence this shows a linear relationship between the height and weight of the person. As the height is increased, the weight of the person also gets increased.

15. What are the various branches of statistics?

A) The study of statistics can be categorized into two main branches. These branches are
descriptive statistics and inferential statistics.
Descriptive statistics –
Descriptive statistics are used to summarise and describe data (information that has been collected).
1. Data are usually organised and presented in tables or graphs that summarise information, such as histograms, pie charts, bars or scatterplots.
2. Descriptive statistics are only descriptive and, thus, do not involve generalising beyond the data that has been collected.
Inferential statistics -With Inferential statistics, data are usually collected from a sample; that is, a smaller representative subset of the larger population we wish to investigate.
1. Inferential statistics use the theory of probability to investigate whether patterns found in the sample of study can be generalised to the wider population where the sample comes from.
2. Inferential statistics aim to test hypotheses and explore relationships between variables, and can be used to make predictions about the population.
3. Inferential statistics are used to draw conclusions and inferences; that is, to make valid generalisations from samples.