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?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

Q10.

Normal Distribution:

The normal distribution is a probability function that describes how the values of variables in a sample are distributed.

The graphical representation of normal distribution appears as a bell curve.

The normal distribution has two parameters, mean and standard deviation.

The normal curve is symmetrical about the mean. The data near mean is more frequent in occurrence than data far from the mean.

The mean is at the middle of the curve and divides the area into halves

The total area under the curve is equal to 1.

Q11:
There are different ways of dealing with missing values.
1)Deleting Rows with missing values:

Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of rows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped.

2)Impute missing values for continuous variable:

Columns in the dataset which are having numeric continuous values can be replaced with the mean, median, or mode of remaining values in the column. Replacing the above two approximations (mean, median) is a statistical approach to handle the missing values.

3)Impute missing values for categorical variable:

When missing values is from categorical columns (string or numerical) then the missing values can be replaced with the most frequent category(mode). If the number of missing values is very large then it can be replaced with a new category.

4)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. This strategy will add more information into the dataset which will result in the change of variance.

Q12

An A/B test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

A/B testing is a 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 example, If a company want to modify its product and check sales whether the old product has higher sales or the modified product. Here, either you can use random experiments, or you can apply scientific and statistical methods.

We will use A/B testing and collect data to analyze which product has better sales.

A/B testing works best when testing incremental changes, such as UX changes, new features, ranking, and page load times. Here you may compare pre and post-modification results to decide whether the changes are working as desired or not.

A/B testing doesn't work well when testing major changes, like new products, new branding, or completely new user experiences.

In A/B testing we have to make two hypotheses i.e Null hypothesis and alternative hypothesis.

the null hypothesis states that there is no difference between the control and variant groups.

The alternative hypothesis challenges the null hypothesis and is basically a hypothesis that the researcher believes to be true.

we have to collect enough evidence through our tests to reject the null hypothesis.

Q13:

By imputing missing data with mean, we will be replacing the missing values with the same value (i.e average of all values)

1. Imputing missing values with mean decreases the variance. the more data goes missing, the mean will be added to more no.of items the data.
2. Because it substitutes missing data with the mean of data points, mean imputation may considerably change the values of correlations.

Q14:

Linear Regression is used to predict out come of a variable based on other variables.

The variable which will be predicted is called as dependent variable ,where as rest of the variables are independent variables.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable.

Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables

Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates after certain time.

Linear Regression is classified into two types

1) **Simple linear regression:**

In Simple Linear Regression, we try to find the relationship between a single independent variable (input) and a corresponding dependent variable (output). This can be expressed in the form of a straight line.
The equation of straight line is: y=ax+b
Y is the output (dependent variable) which will be predicted.
A is intercept and b is slope.

2) **Multiple linear regression**

In Multiple Linear Regression, we try to find the relationship between 2 or more independent variables (inputs) and the corresponding dependent variable (output). The independent variables can be continuous or categorical.
The equation for multiple linear regression is:
Y=ax1+bx2+cx3+…….+mxn+c

Here a,b,c,…m are coefficients.
X1,x2,x3…xn are predictor variables.

Q15:

There are two branches in statistics

1) Descriptive statistics

2) Inferential Statistics

Descriptive statistics:

- Descriptive statistics summarizes or describes the characteristics of collected data.
- Descriptive statistics consists of two basic categories of measures: measures of central tendency (generally termed as mean) and measures of variability or spread (termed as standard deviation).
- Measures of central tendency includes mean, median and mode.
- Mean is the average of the data, median is the middle value and mode is the most frequently occurring value.
- Measures of variability or spread describe the dispersion of data within the set.

INFERENTIAL STATISTICS:

Inferential statistics is used to analyze the results and draw conclusions.Inferential statistics uses statistical models to help you compare your sample data to other samples or to previous research. Mostly we use statistical models called the Generalized Linear model and include Student's t-tests, ANOVA (Analysis of Variance), regression analysis.

OBJECTIVE ANSWERS:

Q1:A

Q2:A

Q3:B

Q4:D

Q5:C

Q6:A

Q7:A,B

Q8:A

Q9:D