STATISTICS WORKSHEET

Answers

Q1 A) True

Q2 A) Central Limit Theorem

Q3 B) Modeling bounded count data

Q4 D) All of the mentioned

. Q5 C) Poisson

Q6 B) False

Q7 B) Hypothesis

Q8 A) 0

Q9 C) Outliers cannot conform to the regression relationship

Q10

- Normal distribution, also known as the Gaussian distribution, 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. In a normal distribution the mean is O and the standard deviation is 1.
- It has zero skew and the normaldistribution is the most important probability distribution in statistics because it fits many natural phenomenal

Q11 Data can be missed due to the following ways that are shown below:

- Missing Completely At Random (MCAR) : When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random.

- Missing At Random (MAR} : The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but ismissing randomly only within sub-samples of data.

- No Missing At Random (NMAR) : When the missing data has a structure toit, we cannot treat it as missing at random.

- Missing data is a huge problem for data analysis because it distorts findings.It's difficult to be fully confident in the insights when you know that some entries are missing values. Hence, why they must be addressed. According todata scientists, there are three types of missing data that I have explained above.

Imputation Techniques:

1. Mean or Median Imputation
2. Multivariate Imputation by Chained Equations (MICE)
3. Random Forest

You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value. In Pandas, there are two very useful methods: isnull () and dropna() that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill theinvalid values with a placeholder value (for example, 0), you could use the fillna() method.

Q12

- A/B testing, 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 drives business metrics
- Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to ‘control’ or the original testing variable. Whereas B refers to ‘variation’ or a new version of the original testing variable.

Q13

- True, 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 and that's a good thing. Further it is non-standard and it uses Random Forest. It is use to predict the missing data. It also can be used for both i.e. continuous as well as categorical data and so it makes advantageous over otherimputations.

There are some limitations too :-

1. Mean imputation does not preserve the relationship among variables. It preserves the mean of observed data. If data is missing completely at random,the estimate of the mean remains unbiased.
2. Mean Imputation leads to an underestimate of standard errors.

Q14

- Linear regression analysis is used to predict the vale of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. Linearregression attempts to model the relationship between two variables by fittinga linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model. A linear regression line has an equationof the form Y= mx + c, where X is the explanatory variable and Y is the dependent variable. The slope of the line is m, and c is the intercept (the value of y when x=0).

Types of linear regression:

Simple linear regression, Multiple linear regression Logistic regression, Ordinal regression, multinominal regression

Q15

- Various branches of statistics are explained as follows :

1. Descriptive Methods :- This type of method consists of all the preliminary steps to final analysis and interpretation. As such this method includes the method of collection, methods of tabulation, measures of central tendency, measures of dispersion, measures of skewness, and analysis of time series. These methods bring out the various characteristics of data and help in summarizing and interpreting the salient features of the data. This method isalso otherwise called Descriptive Statistics.

2. Analytical Methods :- This type of method consists of all those methods which help in the matter of analysis and comparison between any two or more variables. This includes the methods of correlation, regression analysis, association of attributes and the like. This method is also otherwise called Analytical Statistics.

3. Inductive Methods :- This type of method consists of all those procedures thathelp in the generalization or estimation over a phenomenon on the basis of random observation or partial data. This includes the procedure of interpolation,extrapolation, theory of probability and the like. This method is also otherwise called Inductive Statistics.

4. Inferential Methods :- This type of method consists of those procedures which help in drawing inferences about the characteristics of the population on the basis of samples. As such, this method includes the theory of sampling, differenttests of significance, statistical control etc. This method is also otherwise called Inferential Statistics.

5. Applied Methods :- This type of method consists of those procedures whichare applied to the problems of real life. This includes the method of statistical quality control, sample survey, linear programming and inventory control.