STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. a) True
2. a) Central Limit Theorem
3. b) Modeling bounded count data
4. d) All of the mentioned
5. c) Poisson
6. b) False
7. b) Hypothesis
8. a) 0
9. c) Outliers cannot conform to the regression relationship

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

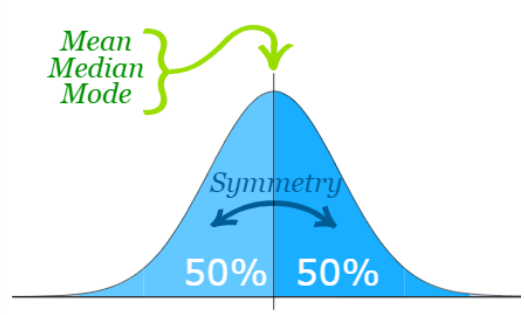
Q10. What do you understand by the term Normal Distribution?

Answer:

Normal distribution, also known as the Gaussian distribution

Properties of a normal distribution

- The mean, mode and median are all equal.
- The curve is symmetric at the center (around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- In a normal distribution the mean is zero and the standard deviation is 1.



Normal Distribution is often called a "Bell Curve" because it looks like a bell. We can take any Normal Distribution and convert it to The Standard Normal Distribution, that is called "Standardizing". It can help us to make decisions about the data.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Answer: Handling missing data

a. Dropping rows with null values or Dropping features with high nullity

If dataset is having many rows we either delete a particular row if it has a null value for a particular feature. Also If particular feature is having many null values then drop that feature. Complete removal of data with missing values results in robust and highly accurate model. It Loss of information and data if the percentage of missing values is high compared to the whole dataset

b. Replacing With Mean

In a mean substitution, the mean value of a variable is used in place of the missing data value for that same variable. This is a better approach when the data size is small.It can prevent data loss which results in removal of the rows and columns.

c. Replacing With Median could be better choice than median if outliers are present.

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 the invalid values with a placeholder value (for example, 0), you could use the fillna() method.

Imputation techniques to handle missing data.

The idea behind the imputation approach is to replace missing values with other sensible values. As you always lose information with the deletion approach when dropping either samples (rows) or entire features (columns), imputation is often the preferred approach.
The many imputation techniques can be divided into two subgroups: single imputation or multiple imputation.

In **single imputation**, a single imputation value for each of the missing observations is generated. The imputed value is treated as the true value, ignoring the fact that no imputation method can provide the exact value. Therefore, single imputation does not reflect the uncertainty of the missing values.

Replacement by	**Numerical Features Only**	**Numerical and Nominal Features**
Existing values	Minimum / Maximum	Previous / Next / Fixed
Statistical values	(Rounded) Mean / Median / Moving Average, Linear / Average Interpolation	Most Frequent
Predicted values	Regression Algorithms	Regression & Classification Algorithms, k-Nearest Neighbours

Table Ref: https://www.kdnuggets.com/2020/09/missing-value-imputation-review.html

In **multiple imputation**, many imputed values for each of the missing observations are generated. This means many complete datasets with different imputed values are created. The analysis is performed on each of these datasets and the results are polled. Creating multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations. A number of algorithms have been developed for multiple imputation. One well known algorithm is Multiple Imputation by Chained Equation (MICE).

Q12. What is A/B testing?

Answer: - A/B testing also known as split testing. 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 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.

Q13. Is mean imputation of missing data acceptable practice?

Answer: - 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. If all you are doing is estimating means , and if the data are missing completely at random, mean imputation will not bias your parameter estimate.

Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That's good too.. 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 other imputations.

There are some limitations too: -

1. Mean substitution leads to bias in multivariate estimates such as correlation or regression coefficients. Values that are imputed by a variable's mean have, in general, a correlation of zero with other variables. Relationships between variables are therefore biased toward zero.
2.Mean Imputation leads to an underestimate of standard errors and variance.

Q14. What is linear regression in statistics?

Answer:- In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula **y = c + b*x ,** where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

1. Simple Linear Regression
   With simple linear regression when we have a single input, we can use statistics to estimate the coefficients.
   This requires that you calculate statistical properties from the data such as means, standard deviations, correlations and covariance. All of the data must be available to traverse and calculate statistics.
2. Ordinary Least Squares
   When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients.
3. Gradient Descent

When there are one or more inputs you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.

4. Regularization

There are extensions of the training of the linear model called regularization methods.
Two popular examples of regularization procedures for linear regression are Lasso Regression and Ridge Regression:

Q15. What are the various branches of statistics?

Answer: - Various branches of statistics are given below: -

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 is also 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 that help 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 which in drawing inferences about the characteristics of the population on the basis of samples. As such, this method includes the theory of sampling, different tests of significance, statistical control etc. This method is also otherwise called inferential statistics.

5. Applied Methods: -

• This type of method consists of those procedures which are applied to the problems of real life. This includes the method of statistical quality control, sample survey, linear programming, inventory control and the like.