Statistcs Worksheet 1

QUESTION 1

Answer- a)

QUESTION 2

Answer-a)

QUESTION 3

Answer-b)

QUESTION 4

Answer- d)

QUESTION 5

Answer-c)

QUESTION 6

Answer-b)

QUESTION 7

Answer- b)

QUESTION 8

Answer- a)

QUESTION 9

Answer – c)

QUESTION 10

Answer-

A normal distribution is a statistical phenomenon representing a symmetric bell-shaped curve. Most values are located near the mean; also, only a few appear at the left and right tails.

It follows the empirical rule or the 68-95-99.7 rule.

Here, the mean, median, and mode are equal; the mean and standard deviation of the function are 0 and 1, respectively.

This mathematical function has two key parameters:
The mean (μ) and the standard deviation (σ).

QUESTION 11

Answer-- Imputation is a method in which the missing values in any variable or data frame (in Machine learning) are filled with numeric values for performing the task. By using this method, the sample size remains the same. Only the blanks which were missing are now filled with some deals.

Sklearn.impute package provides 2 types of imputations algorithms to fill in missing values**:**

1)simpleImputer

SimpleImputer is used for imputations on univariate datasets; univariate datasets have only a single variable*.* SimpleImputer allows us *to* attribute values in any feature column using only missing values in that feature space*.*

For numeric data, it uses the strategy of mean, meadian and for string data it uses strategy most frequent which is like mode of string values.

2) iterativeImputer

IterativeImputer is used for imputations on multivariate datasets, and multivariate datasets are datasets have more than two variables or feature columns per observation. IterativeImputer allows us to use the entire dataset of available features columns to impute the missing values.

In IterativeImpute**,** each feature with a missing value is used as a function of other features with known output and models the function for imputations. The same process is then iterated in a loop for some iterations, and *at* each step, a feature column is selected *as* output y**.** Other feature columns are treated as inputs X, then a regressor is fit on (X, y) for known y and is used to predict the missing values of y.

QUESTION 12

Answer--A/B testing is a type of split testing and is commonly used to drive improvements to specific variables or elements by measuring user or audience engagement. The technique of A/B testing can be used to test and improve machine learning models. The technique can be used to decide whether a new model is an improvement over a current model. The organisation should choose a metric to compare the control model with the new model. This metric will be used to measure success, and outline the difference between the two deployments. The two models will need to be simultaneously deployed on a sample of data for a defined period of time. Half of the users will be interacting with the control model, and the other half will be interacting with the new model.

QUESTION 13

Answer—The mean imputation of missing data is not acceptable practice. It leads to an underestimate of standard deviation and distorts relationship between variables by pulling estimates of the correlation towards zero.

QUESTION 14

Answer--

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called *simple linear regression* for more than one, the process is called multiple linear regression. Linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y.

Y= A+BX +E,

Where Y= Dependent variable that we are trying to predict

X= independent variable we are using to predict.

A= the intercept

B= Slope

E= regession residual errors

QUESTION 15

Answer—In mathematics, there are two branches of statistics:

1) Descriptive
2) Inferential

DESCRIPTIVE:

Descriptive statistics have two parts:

a) Central tendency measure
b) Variability measure

Central tendency measure:

1. **Mean:** Mean is a conventional method used to describe the central tendency. Typically, calculate the average of values, count all values, and then divide them with the number of available values.
2. **Median:** It is the result that is in the middle of a set of values. An easy way to calculate the median is to edit the results in numerical journals and locate the result that is in the center of the distributed sample.
3. **Mode**: The mode is the frequently occurring value in the given data set.

Variability measure:

The variability measure helps statisticians to analyze the distribution that is spreading from a specific data set. Some of the variables of variability include quartiles, ranges, variances, and standard deviation.

INFERENTIAL:

Inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population. Different types of inferential statistics include:

- **Regression analysis:** It is a set of statistical methods used to estimate relationships between a dependent variable and one or more independent variables. It includes several variations, like linear, multiple linear, and nonlinear. The most well-known models are simple linear and multiple linear.
- **Analysis of variance (ANOVA):** ANOVA is a statistical method that distributes observed variance data into various components. A one-way ANOVA is applied for three or more data groups to gain information about the relationship between the dependent and independent variables.
- **Analysis of covariance (ANCOVA):** It is used to test categorical variables' main and interaction effects on constant dependent variables and keep control for the impact of selected other constant variables. The control variables are known as covariates.
- **Statistical significance (t-test):** It is used to determine a significant difference between the means of two groups related to particular features. A t-test studies the t-statistic, the t-distribution values, and the degree of freedom to learn the statistical significance.
- **Correlation analysis:** It is a statistical method that is used to find the relationship between two variables or datasets and discover how strong the relationship may be.