WORKSHEET-1

STATISTICS

Answer 1:- A) True

Answer 2:- B) Central limit theorem

Answer 3:- C) Modelling contingency tables

Answer 4:- C) The square of a standard normal random variables follows what is called chi – squared distribution

Answer 5:- A) empirical

Answer 6:- B) False

Answer 7:- B) Hypothesis

Answer 8:- A) 0

Answer 9:- C) outliers cannot conform to the regression relationship

Answer 10:-

Normal distribution is a probability distribution that describes a continuous variable such as height or weight, that tends to cluster around a central or average value , with a symmetrical bell – shaped curve. The curve is characterized by two parameters , the mean and standard deviation. The mean represents the central tendency of the distribution , whiloe standard deviation measures the spread or variability of the data.

Answer 11:_

Handling missing data is an important aspects of data analysis , as missing data can affect the validity and reliability of statistical analysis. Some of common technique for handling missing data are as follow :-

1. **Complete case analysis :-this technique involves removing any observ ation with missing values from the dataset , this method van be used when the amount data is small .**
2. **Mean/mode/median imputation :- this involves replacing missing values with mean, median, mode of the available data fro that**

variable. this can be used missing data is not too extensive when missing values are missing at random

3. **Regression imputation:- it uses regression model to predict missing values based on available data .used data is correlated with other variable.**
4. **Multiple imputation:- this involve creating multiple dataset and then combining them to estimate the final results.**

In all the above I will recommend 2 & 3 because both of vcan be used on extensive data set and check data is correlated.

Answer 12:-

A/B testing is also known as split testing , the goal of this testing is to determine whether there is statistically significant difference in performance between the two version , and to identify which version performance better. A/B testing can be used to test various element of a product or campaign such as design content ,pricing or messaging.

To conduct this test it is important to ensure that two groups are comparable in terms of demographic and other relevant factors.

Answer 13:-

Mean imputation is commonly used method to handle missing data but when the data is small and if data is big or large it lead to biased estimate and incorrect estimates or if data is not missing at random. Here are some potential issue with mean imputation:-

1. **Bias :- mean imputation can introduce bias into the data , as it assumes that missing values are similar to the observed values. If the missing data is not missing at random or if there is a systematic difference between the missing values and observed values .**
2. **Loss of information:- ,ean imputation does not take into the account the uncertainty associated with missing values , and it can lead to loss of information in the data**
3. **Variance inflation:- mean imputation can inflate the variance of the data , as it increases the similarity between the observed values and reduces the variability of the data , as it increases the similarity:-**

between the observed values and reduces the variability of missing values .

Answer 14:-

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variable . the goal of linear regression is to find a linear relationship that best describe the data.

In simple linear regression ,there is only one independent variable and the relationship between the independent variable and dependent variable is described by a straight line. The equation of line is :-

Y=b0 + b1x 6+ e

Where y is dependent variable , x is the independent vaiable , b0 and b1 are the intercept and slope coefficient and e is the error term.

Answer 15:-

Statistics is a broad field that encompasses many different branches , each of them focus on different aspects of data analysis and modelling.

1. **Descriptive statistics: it deal with summary and presentation of data. It includes measure of central tendency , standard deviation and grapichal methods (histograms and plots)**
2. **Inferential statistics :- it is branch that deals with makinf inferences and prediction about population based on data from a sample. Uses regression analysis, hypothesis testing etc.**
3. **Bayesian statistics:- it uses probability theory to model and analyze uncertain events**
4. **Probability theory :- deal with study of random variables evenst and their outcome**
5. **Data science:- data science is a field that combine statistics , computer science and domain specific knowledge to extract insights and knowledge from data,. Technique such as machine learning , data mining and big data analytics.**