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

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

3. Which of the following is incorrect with respect to use of Poisson distribution?

b) Modeling bounded count data

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

Answer) d) all of the mentioned

5. _____ random variables are used to model rates.

c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.

b) False

7. Which of the following testing is concerned with making decisions using data?

b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

a) 0

9. Which of the following statement is incorrect with respect to outliers?

c) Outliers cannot conform to the regression relationship Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Answer 10). The normal distribution, also known as the Gaussian or standard normal distribution, is the probability distribution that plots all of its values in a symmetrical fashion, and most of the results are situated around the probability's mean. Values are equally likely to plot either above or below the mean.
The normal distribution is a probability distribution that (roughly) describes many common datasets in the real world. It is the most common type of distribution, and it arises naturally in statistics through random sampling techniques.
Nowadays, it is more common to show up as a model for the "lifespan" of a product, like a light bulb, or the outcome of standardized tests, like IQ. Biological measurements, like height or weight, are often estimated with normal distributions.

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

Answer 11. The real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values.

6 ways to correct the data are as follows:

Deleting Rows with missing values

Impute missing values for continuous variable

Impute missing values for categorical variable

Other Imputation Methods

Using Algorithms that support missing values

Prediction of missing values

The following are common methods for Imputations:

Mean imputation. Simply calculate the mean of the observed values for that variable for all individuals who are non-missing.
Substitution
Hot deck imputation
Cold deck imputation
Regression imputation
Stochastic regression imputation
Interpolation and extrapolation

12. What is A/B testing?

A/B testing is a basic randomized control environment. It is a way to compare the two versions of a variable of a variable to find out which performs better in a controlled environment.

For instance, let's say you own a company and want to increaae the sales of your product . Here,

Either you can use random experiments, or you can apply scientific and statistical methods.

A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts- A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now , on the basis of the response from customers groups who used A and B respectively you try to decide which is performing better .

13. Is mean imputation of missing data acceptable practice?

Answer 13. NO mean imputation sometimes reduced the overall accuracy of the model. while using a mean or mode can significantly reduce the model's accuracy and bias the results.

Let's assume the fitness score of 30 people ranging from 20-60.

Fitness score of 20-30 age groups is 8.

Fitness score of 30-40 age groups is 6.

Fitness score of 40-50 age groups is 4.

Fitness score of 50-60 age groups is 2.

The mean is 5.

So if there is any value missing from age group 50-60 and we replace it with mean it will definitely alter the results and overall accuracy of the model is reduced.

So, it's better to delete the rows of missing data rather replacing it with mean, Or we use different algorithms to increase the accuracy.

14. What is linear regression in statistics?

Answer 14). Linear regression is a basic and commonly used type of predictive analysis. This Linear regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. 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.

15. What are the various branches of statistics?

Answer 15). The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics

Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).

Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

Inferential Statistics

Inferential statistics is a statistical method that concludes from a small but representative sample the characteristics of a bigger population. In other words, it allows the researcher to make assumptions about a wider group, using a smaller portion of that group as a guideline.