STATISTICS WORKSHEET-1

ANSWERS

1.**Bernoulli random variables take (only) the values of 1 and 0.**

Ans:-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?

Ans:-a)Central Limit Theorem

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

Ans:-b)Modelling 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.

Ans:-d)All of the mentioned.

5. ______ random variables are used to model rates.

a) Empirical b) Binomial c) Poisson d) All of the mentioned .

Ans:-c)Poisson

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

Ans:-b)False

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

a) Probability b) Hypothesis c) Causal d) None of the mentioned

Ans:-b)Hypothesis

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

a) 0 b) 5 c) 1 d) 10

Ans:-a)0

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

a) Outliers can have varying degrees of influence.

b) Outliers can be the result of spurious or real processes.

c) Outliers cannot conform to the regression relationship.

d) None of the mentioned.

Ans:- c) Outliers cannot conform to the regression relationship.

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

Ans:-Normal Distribution, also known as Gaussian Curve is one of the most important probability distribution in statistics and describes how the values of variable are distributed.

Normal Distribution means it is fully characterized by its mean and standard deviation. This make the distribution symmetrical and is depicted as a bell-shaped Curve.this means that the distribution has more data around the mean and the distribution decreases as you move away from the centre. A Normal distribution is defined by a mean of 0 and standard deviation of ±1.

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

Ans:- There are few common ways to handle missing data:-

1.Mean or median Imputation- when the data is missing at random, we can use the list wise or pair wise deletion of the missing observations.

2.Multivariate imputation by chained equation(MICE)-It assumes that the missing data are missing at random(MAR).It imputes data on a variable-by-variable basis by specifying an imputation model.

3.Random Forest-This is a non- parametric imputation method applicable to various variable types that works well with both missing at random and not missing at random.

We can find missing data in a dataset and either drop those rows and columns, or decide to replace them with another value. In Panda, there are 2 very useful methods:isnull() and dropna(),that will help to find the missing value and drop those values.

12. What is A/B testing?

Ans:-A/B Testing (also known as bucket testing /split run test)is a basic randomized control experiment with two variants, A and B.It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

1**3. Is mean imputation of missing data acceptable practice?**

Ans:-MEAN IMPUTATION – is a process of replacing the null values in a data collection with the mean data. Firstly, Mean imputation is also considered unacceptable practice since it ignores feature correlation. Secondly , mean imputation decreases the variance of the data while increasing the bias. As a result, reduced variance, the model is less accurate.

14. What is linear regression in statistics?

Ans:-In Statistics, Linear regression is used to predict the value of a variable based on the value of another variable.It attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be a dependent variable .A linear regression line has an equation of the form y=a+bx , where ‘x’ is the explanatory variable and ‘y’ is the dependent variable.

15. What are the various branches of statistics?

Ans:-There are two branches of statistics:-Descriptive statistics and inferential statistics.

1)**DESCRIPTIVE STATISTICS**- These are used to summarize, organize and make sense of a set of scores or observations. It has further two sub categories :-Central Tendency and dispersion of data.

2)**INFERENTIAL STATISTICS-** These are used that allow researchers to infer or generalize, observations made with samples to the larger population from which they are selected.