STATISTICS WORKSHEET-1

****

Name:-Poonam Gawade Batch:-1838

****

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
   b) False
   Answer: True

2. Which of the following theorem states that the distribution of averages of iid variables, properlynormalized, becomes that of a standard normal as the sample size increases?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned
   Answer: Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned
   Answer: 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- normaldistribution
   b) Sums of normally distributed random variables are again normally distributed even if the variablesare dependent
   c) The square of a standard normal random variable follows what is called chi-squareddistribution
   d) All of the mentioned
   Answer: All of the mentioned

5. ______random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
   Answer: Poission

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Answer: False

7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned

   Answer: Hypothesis

8. 4. Normalized data are centered at____and have units equal to standard deviations of theoriginal data.
   a) 0
   b) 5
   c) 1
   d) 10

   Answer: 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

   Answer: 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

The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions.

Extreme values in both tails of the distribution are similarly unlikely. While the normal distribution is symmetrical, not all symmetrical distributions are normal. For example, the Student’s t, Cauchy, and logistic distributions are symmetric.
Parameters of the Normal Distribution

1)Mean:-The mean is the central tendency of the normal distribution. It defines the location of the peak for the bell curve. Most values cluster around the mean.

2)**Standard deviation:-**The standard deviation is a measure of variability. It defines the width of the normal distribution. The standard deviation determines how far away from the mean the values tend to fall. It represents the typical distance between the observations and the average.

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

Answer

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
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
Imputation using Deep Learning Library

Imputaion:- Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset.

Frequent Category Imputation:-

This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the Mode of that column. This technique is also referred to as Mode Imputation.

- **Assumptions:-**
  - Data is missing at random.
  - There is a high probability that the missing data looks like the majority of the data.

- **Advantages:-**
  - Implementation is easy.
  - We can obtain a complete dataset in very little time.
  - We can use this technique in the production model.
- **Disadvantages:-**
  - The higher the percentage of missing values, the higher will be the distortion.
  - May lead to over-representation of a particular category.
  - Can distort original variable distribution.
- **When to Use:-**
  - Data is Missing at Random(MAR)
  - Missing data is not more than 5% – 6% of the dataset.

12. What is A/B testing?

Answer

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.

A/B Testing is a widely used concept in most industries nowadays, and data scientists are at the forefront of implementing it.

For instance, let's say you own a company and want to increase 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.

Example- 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 customer 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

Yes, you get the same mean from mean-imputed data that you would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small. Because the imputations are themselves estimates, there is some error associated with them. But your statistical software doesn't know that. It treats it as real data.

14. What is linear regression in statistics?

Answer

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

15. What are the various branches of statistics?

Answer

There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.

1)Data collection :- It is all about how the actual data is collected. For the most part, this needn't concern us too much in terms of the mathematics (we just work with what we are given), but there are significant issues to consider when actually collecting data. For data such as marks in a class test, this is fairly straightforward. Each student has a defined mark associated with them, so the marks are simply collected together to make the data set.

2)Descriptive statistics:- It is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically.
The basic aim of descriptive statistics is to 'present the data' in an understandable way. If you simply write down every piece of data, it means little to someone who sees it; it needs to be summarised.

3)Inferential statistics:- It is the aspect that deals with making conclusions about the data. This is quite a wide area; essentially you are asking 'What is this data telling us, and what should we do?' For example, a council might be considering altering the speed limit on a main road, after a number of accidents. They might do this by surveying the speeds of cars (data collection) and then arrive at a conclusion as to whether the speed limit needs to be lowered