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 b) False

Answer :- 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

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

Answer :- a) 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: - B) Modeling bounded count data

----

4. Point out the correct statement.

a) The exponent of 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.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer :- C) Passion

--

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

a) True

b) False

Answer :- B) 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 :-B) Hypothesis

--

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

a) 0

b) 5

c) 1

d) 10

Answer :- 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

Answer :- 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 :- Normal distribution is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. . It is also known as the Gaussian distribution. In graph form, normal distribution will appear as a bell curve. Some important points about normal distribution are -

- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- In reality, most pricing distributions are not perfectly normal.
- A normal distribution is the proper term for a probability bell curve.

----

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

Answer :-Imputation is use to handle missing data. Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset. These techniques are used because removing the data from the dataset every time is not feasible and can lead to a reduction in the size of the dataset to a large extend, which not only raises concerns for biasing the dataset but also leads to incorrect analysis. Techniques used in Imputation are –

1. Complete Case Analysis(CCA)- This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data i.e. we consider only those rows where we have complete data i.e. data is not missing. This method is also popularly known as “Listwise deletion”.

Assumptions:-Data is Missing At Random (MAR).

Missing data is completely removed from the table.

Advantages:- Easy to implement.

No Data manipulation required.

Limitations:-Deleted data can be informative.

Can lead to the deletion of a large part of the data.

Can create a bias in the dataset, if a large amount of a particular type of variable is deleted from it.

The production model will not know what to do with Missing data.

When to Use:- Data is MAR(Missing At Random).

Good for Mixed, Numerical, and Categorical data.

Missing data is not more than 5% – 6% of the dataset.

Data doesn't contain much information and will not bias the dataset.

2. Arbitrary Value Imputation- This is an important technique used in Imputation as it can handle both the Numerical and Categorical variables. This technique states that we group the missing values in a column and assign them to a new value that is far away from the range of that column. Mostly we use values like 99999999 or -9999999 or "Missing" or "Not defined" for numerical & categorical variables.

Assumptions:-Data is not Missing At Random.

The missing data is imputed with an arbitrary value that is not part of the dataset or Mean/Median/Mode of data.

Advantages:- Easy to implement.

We can use it in production.

It retains the importance of "missing values" if it exists.

Disadvantages:- Can distort original variable distribution.

Arbitrary values can create outliers.

Extra caution required in selecting the Arbitrary value.

When to Use:- When data is not MAR(Missing At Random).

Suitable for All.

3. 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 allows to compare two versions (A and B) of app or webpages to determine which is more effective and perform better. It is also known as split testing. It is the best method in quantifying changes in a product or changes in a marketing strategy. And this is becoming increasingly important in a data-driven world where business decisions need to be back by facts and numbers.

----

13. Is mean imputation of missing data acceptable practice?

Answer :- The process of replacing null values in a data collection with the data's mean is known as mean imputation. It ignores feature correlation so mean imputation is considered terrible practice.

e.g. : we have a table with height, weight and age, and 10 year-old has a missing weight . If we average the weight of people between the ages of 10 and 80, the 10 year-old will appear to have more weight than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

----

14. What is linear regression in statistics?

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). This linear regression analysis is very helpful in several ways like it helps in foreseeing trends, future values, and moreover predict the impacts of changes. Linear regression divided into two types:-

1) **Simple Linear Regression -** In simple linear regression, we aim to reveal the relationship between a single independent variable and a corresponding dependent variable.
2) **Multiple Linear Regression -**In this type of linear regression, we always attempt to discover the relationship between two or more independent variables and the corresponding dependent variable and the independent variables can be either continuous or categorical.

----

15. What are the various branches of statistics?

Answer :- Statistics is mainly divided into the following two branches.

1) Descriptive Statistics
2) Inferential Statistics

1) Descriptive Statistics-In the descriptive Statistics, the Data is described in a summarized way. The summarization is done from the sample of the population using different parameters like Mean or standard deviation. Descriptive Statistics are a way of using charts, graphs, and summary measures to organize, represent, and explain a set of Data.

- Data is typically arranged and displayed in tables or graphs summarizing details such as histograms, pie charts, bars or scatter plots.
- Descriptive Statistics are just descriptive and thus do not require normalization beyond the Data collected.

2) Inferential Statistics- In the Inferential Statistics, we try to interpret the Meaning of descriptive Statistics. After the Data has been collected, analyzed, and summarised we use Inferential Statistics to describe the Meaning of the collected Data.

- Inferential Statistics use the probability principle to assess whether trends contained in the research sample can be generalized to the larger population from which the sample originally comes.
- Inferential Statistics are intended to test hypotheses and investigate relationships between variables and can be used to make population predictions.
- Inferential Statistics are used to draw conclusions and inferences, i.e., to make valid generalizations from samples

---****--**