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 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)

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

   Answer (C) Poisson

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) The null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research or alternative 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 can 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, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables, 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.

Properties of a normal distribution

1. The mean, mode and median are all equal.
2. The curve is symmetric at the center (i.e. around the mean, μ).
3. Exactly half of the values are to the left of center and exactly half the values are to the right.
4. The total area under the curve is 1.

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

Answer: We will use the imputation technique as this 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. Below were some of the commonly used methods.

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 "List wise deletion".

Assumptions:-
1. Data is missing At Random (MAR).
2. Missing data is completely removed from the table.

Advantages:-
1. Easy to implement.
2. No Data manipulation required.

Limitations:-
1. Deleted data can be informative.
2. Can lead to the deletion of a large part of the data.
3. Can create a bias in the dataset, if a large amount of a particular type of variable is deleted from it.

When to Use:-
1. Data is MAR(Missing At Random).
2. Good for Mixed, Numerical, and Categorical data.
3. Missing data is not more than 5% – 6% of the dataset.
4. 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.

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:-
1. Easy to implement.
2. We can use it in production.
3. It retains the importance of "missing values" if it exists.

Disadvantages:-
1. Can distort original variable distribution.
2. Arbitrary values can create outliers.
3. Extra caution required in selecting the arbitrary value.

When to Use:-
1. When data is not MAR (Missing At Random).
2. 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:-
1. Data is missing at random.
2. There is a high probability that the missing data looks like the majority of the data.

Advantages:-
1. Implementation is easy.
2. We can obtain a complete dataset in very little time.
3. We can use this technique in the production model.

Disadvantages:-
1. The higher the percentage of missing values, the higher will be the distortion.
2. May lead to over-representation of a particular category.
3. Can distort original variable distribution.

When to Use:-
1. Data is Missing at Random(MAR)
2. Missing data is not more than 5% – 6% of the dataset.

12. What is A/B testing?

Answer: A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

How it works

In A/B test, we take a webpage or app screen and modify it to create a second version of the same page. This change can be as simple as a single headline, button or be a complete redesign of the page. Then, half of the traffic is shown the original version of the page (known as the control) and half are shown the modified version of the page (the variation**).** As visitors are served either the control or variation, their engagement with each experience is measured and collected in a dashboard and analyzed through a statistical engine.

13. Is mean imputation of missing data acceptable practice?

Answer – The Mean imputation of missing data was not commonly practiced as this method does not preserve the relationships among variables. Below were some of the disadvantages of this method.
1. Mean imputation reduces the variance of the imputed variables.
2. Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
3. Mean imputation does not preserve relationships between variables such as correlations.

14. What is linear regression in statistics?

Answer: Linear regression is a kind of statistical analysis that attempts to show a relationship between two variables. It looks at various data points and plots a trend line. It can create a predictive model on apparently random data, showing trends in data.

15. What are the various branches of statistics?

Answer: Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data. There are two major areas of statistics are known as **descriptive statistics**, which describes the properties of sample and population data, and **inferential statistics**, which uses those properties to test hypotheses and draw conclusions .