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

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

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

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

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

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

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

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

- **Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.**

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

Answer :-

It refers to a probability distribution which Normal distribution, also known as the Gaussian distribution, is a probability distribution that shows the data which are near to the mean showing that data near the mean are more frequently occurring than data which away from the mean value. In graph form, normal distribution will appear as a bell curve.

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

Answer :-

There are three ways to handle missing data which include:-

1. **Deletion Methods**

This method is applicable with certain types of datasets in case of which participants have missing values.

So in that case we can use this method in two ways

a. **Listwise Deletion-** It means deleting any participants or data entries with missing values. This method can be used in case when we have large amount of data as deletion will not affect the quality of the dataset.

b. **Pairwise Deletion.**
It is the process of deleting/eliminating information which is vital for testing but missing. It is more suitable than Listwise deletion as it deletes entire entries if any data is missing, regardless of its importance.

2. **Regression Analysis**

Regression can be applied for handling missing data as it can help in predicting the null value with the help of using other information from the dataset but this largely depends on how well connected the remaining data is.

3. **Imputation Techniques**

Imputation is replacing missing values with substitute values.

There are two ways to apply this techniques in order to handle missing data:

a. **Average Imputation** uses the average value of the responses from other data entries in order to fill out missing values due to which the variability of the dataset get reduces.

b. **Common-Point Imputation Common-Point Imputation** uses the middle point or most commonly chosen value. For example, on a five-point scale, substitute a 3, the midpoint, or a 4, the most common value (in many cases). This is a bit more structured than guessing, but it's still among the more risky options. Use caution unless you have good reason and data to support using the substitute value.

c. **Average Imputation**: Use the average value of the responses from the other participants to fill in the missing value. If the average of the 30 responses on the question is a 4.1, use a 4.1 as the imputed value. This choice is not always recommended because it can artificially reduce the variability of your data but in some cases makes sense.

d. **Multiple Imputation:** The most sophisticated and, currently, most popular approach is to take the regression idea further and take advantage of correlations between responses. In multiple imputation [pdf], software creates plausible values based on the correlations for the missing data and then averages the simulated datasets by incorporating random errors in your predictions. It is one of a number of examples where computers continue to change the statistical landscape. Most statistical packages like SPSS come with a multiple-imputation feature. More on multiple imputation.

I will recommend **Multiple imputation** because it uses several complete data sets and provides both the within-imputation and between-imputation variability.

12. What is A/B testing?

Answer :-

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

- For instance, let's take an example of two product A & B. Suppose want to increase the sales of your product which can be done either by using random experiments, or we can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, let's suppose 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.

- Basically, It is a hypothetical testing method in order to make decisions that estimate population parameters based on sample statistics. The **population** refers to all the customers buying A & B product, while the **sample** refers to the number of customers that participated in the test.

- A/B testing works best when testing incremental changes, such as UX changes, new features, ranking, and page load times. Here you may compare pre and post-modification results to decide whether the changes are working as desired or not.

A/B testing doesn't work well when testing major changes, like new products, new branding, or completely new user experiences. In these cases, there may be effects that drive higher than normal engagement or emotional responses that may cause users to behave in a different manner.

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.

Mean imputation is typically considered as bad practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level 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 :-

Linear regression technique which 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 :-

Descriptive Statistics and Inferential Statistics

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.

While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.