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.

   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) Modeling bounded count data

4. Point out the correct statement.

   Ans-**d) All of the mentioned**

5. _____ random variables are used to model rates

   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?

   Ans-**b) Hypothesis**

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

   Ans-**a) 0**

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

   Ans-**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?

    Ans-

    Normal distribution, also known as the Gaussian 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.

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

- Use deletion methods to eliminate missing data

The deletion methods only work for certain datasets where participants have missing fields. There are several deleting methods – two common ones include Listwise Deletion and Pairwise Deletion. It means deleting any participants or data entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings

- **Use regression analysis to systematically eliminate data**

Regression is useful for handling missing data because it can be used to predict the null value using other information from the dataset. There are several methods of regression analysis, like Stochastic regression. Regression methods can be successful in finding the missing data, but this largely depends on how well connected the remaining data is.

- **imputation techniques**

  Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. Common-point imputation, on the other hand, is when the data scientists utilise the middle point or the most commonly chosen value

12. What is A/B testing?

A/B testing—also called split testing or bucket testing—compares the performance of two versions of content to see which one appeals more to visitors/viewers. It tests a control (A) version against a variant (B) version to measure which one is most successful based on your key metrics.

A/B testing provides the most benefits when it operates continuously. A regular flow of tests can deliver a stream of recommendations on how to fine-tune performance. And continuous testing is possible because the available options for testing are nearly unlimited.

As noted above, A/B testing can be used to evaluate just about any digital marketing asset including:

- emails
- newsletters
- advertisements
- text messages

- website pages
- components on web pages
- mobile apps

13.Is mean imputation of missing data acceptable practice?

Ans

- Bad practice in general
- If just estimating means: mean imputation preserves the mean of the observed data
- Leads to an underestimate of the standard deviation
- Distorts relationships between variables by “pulling” estimates of the correlation toward zero

14. What is linear regression in statistics?

Ans-

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.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

15. What are the various branches of statistics?

- Data Collection
- Descriptive Statistics
- Inferential Statistics

1. Data Collection- First and foremost method of any process which requires some kind of analysis using data. Data collection can be done in many ways

   Including

   - Surveys
   - Transactional Tracking
   - Interviews and focus group
   - Observation
   - Online Tracking
   - Social Media Monitoring

2. 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.

3. 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.

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.