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

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 :- A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution.

The normal distribution is also known as a Gaussian distribution or probability bell curve. It is symmetric about the mean and indicates that values near the mean occur more frequently than the values that are farther away from the mean

Basic examples of normal distribution: Height and weight

Height is one simple example of values that follow a normal distribution pattern. Most people are of average height -- whatever that may be for a given population. If the heights of these people are represented in graphical format along with the heights of people who are taller and shorter than the average, the distribution will always be a normal distribution. This is because the people of average height will be clustered near the middle, while those who are taller and shorter will be farther away

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

Ans :- Missing data can skew anything for data scientists, from economic analysis to clinical trials .Missing data reduces the statistical power of the analysis, which can distort the validity of the results.

When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data.

The imputation method develops reasonable guesses for missing data. The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias

Instead of deletion, data scientists have multiple solutions to impute the value of missing data. Depending why the data are missing, imputation methods can deliver reasonably reliable results. These are examples of single imputation methods for replacing missing data.

The imputation technique which I would recommend is 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.

12. **What is A/B testing?**

Ans :- A/B testing is one of the most popular controlled experiments used to optimize web marketing strategies. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B.

13. **Is mean imputation of missing data acceptable practice?**

Ans:- No ,mean imputation of missing data is not acceptable practice
Mean imputation does not preserve the relationships among variables. True, imputing the mean preserves the mean of the observed data. So, if the data are missing completely at random, the estimate of the mean remains unbiased.
If all you are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estimate.
Mean Imputation Leads to An Underestimate of Standard Errors
A second reason is applying to any type of single imputation. Any statistic that uses the imputed data will have a standard error that’s too low.

14. **What is linear regression in statistics?**

Ans :- 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). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression
Linear regression has two primary purposes—understanding the relationships between variables and forecasting.

15. **What are the various branches of statistics?**

Ans :- The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data 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 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.
Inferential Statistics -
Inferential statistics, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics.
Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics.