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
   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?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned
   Ans)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
   Ans) 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
   Ans)d) All of the mentioned.

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

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Ans)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
   e) Ans)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

Ans)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

   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) 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.

Graphically, a normal distribution is a bell curve because of its flared shape. The precise shape can vary according to the distribution of the values within the population. The population is the entire set of data points that are part of the distribution.

Regardless of its exact shape, a normal distribution bell curve is always symmetrical about the mean. A symmetrical distribution means that a vertical dividing line drawn through the maximum/mean value will produce two mirror images on either side of the line, in which half the population is less than the mean and half is greater. However, the reverse is not always true; that is, not all symmetrical distributions are normal.

In the bell curve, the peak is always in the middle, and the mean, mode and median are all the same.

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

Ans) Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset

There are three main types of missing data:

1-Missing completely at random (MCAR)
2-Missing at random (MAR)
3-Not missing at random (NMAR)

Below are the most common method to handle the missing data:

1-Mean imputation
2-Substitution
3-Hot deck imputation
4-Cold deck imputation
5-Regression imputation
6-Stochastic regression imputation

12. What is A/B testing?

Ans) A/B testing (also known as split testing or bucket testing) is a methodology for 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.

A/B testing allows individuals, teams and companies to make careful changes to their user experiences while collecting data on the impact it makes. This allows them to construct hypotheses and to learn what elements and optimizations of their experiences impact user behavior the most. In another way, they can be proven wrong—their opinion about the best experience for a given goal can be proven wrong through an A/B test.

13. Is mean imputation of missing data acceptable practice?

Ans)Mean imputation (MI) is one such method in which the mean of the observed values for each variable is computed and the missing values for that variable are imputed by this mean. This is the most common method of data imputation, where you just replace all the missing data with the mean, median, or mode of the column. If you're in a rush this is useful because it's easy and fast, it changes the statistical nature of the data. The drawback is it skews the histograms and also underestimates the variance in the data. No doubt, this practice is very common, but it should be avoided. It should never be used if your data is MNAR.

14.What is linear regression in statistics?

Ans) Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation.

15.What are the various branches of statistics?

There are two types of statistics are: Descriptive and inferential.

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.

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.