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

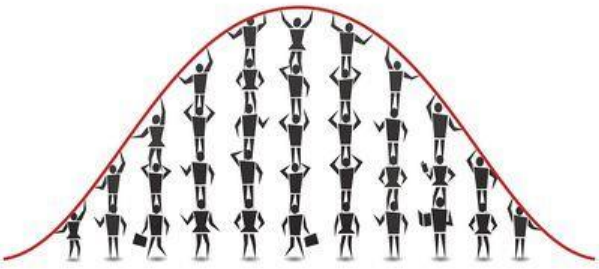
10. What do you understand by the term Normal Distribution?

Ans. The Normal Distribution is defined by the probability density function for a continuous random variable in a system. If f(x) is the probability density function and X is the random variable. Then, it defines a function which is integrated between the range or interval (x to x + dx), giving the probability of random variable X, by considering the values between x and x + dx.

$$f(x) \geq 0, \forall x \in (-\infty, +\infty)$$

$$\int_{-\infty}^{+\infty} f(x)\, dx = 1$$

The height of people is an example of normal distribution. Most of the people in a specific population are of average height. The number of people taller and shorter than the average height people is almost equal, and a very small number of people are either extremely tall or extremely short. Several genetic and environmental factors influence height. Therefore, it follows the normal distribution.



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

Ans. I will analyze each column with missing values carefully to understand the reasons behind the missing of those values, as this information is crucial to choose the strategy for handling the missing values.
There are 2 primary ways of handling missing values:

1)Deleting the Missing values
2)Imputing the Missing Values

Imputing the Missing Values
There are many imputation methods for replacing the missing values. Different python libraries such as Pandas, and Sci-kit Learn are useful to do this.
1) making an educated guess about the missing value to replace it with some arbitrary value.
2) Replacing with the mean, mode or median.
3) Missing values can also be imputed using interpolation. Pandas' interpolate method can be used to replace the missing values with different interpolation methods like 'polynomial,' 'linear,' and 'quadratic.' The default method is 'linear.'
4) Missing values are imputed using the k-Nearest Neighbors approach, where a Euclidean distance is used to find the nearest neighbors.

12. What is A/B testing?

Ans. An A/B test is usually run to test the success of any change in an existing feature or to test the impact of a new feature. A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions.

.

In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable

13.Is mean imputation of missing data acceptable practice?

Ans. Mean imputation allows for the replacement of missing data with a plausible value, which can improve the accuracy of the analysis. Additionally, mean imputation can help to reduce the bias in the results of a study by limiting the effects of extreme outliers. Mean imputation is not always applicable, however. It is only reasonable if the distribution of the variable is known. This means that it cannot be used in situations where values are missing due to measurement error, as is the case with some psychological tests.

14.What is linear regression in statistics?

Ans. Linear regression is used to predict the relationship between two variables by applying a linear equation to observed data. There are two types of variable, one variable is called an independent variable, and the other is a dependent variable. Linear regression is commonly used for predictive analysis.

Agricultural scientists often use linear regression to measure the effect of fertilizer and water on crop yields. For example, scientists might use different amounts of fertilizer and water on different fields and see how it affects crop yield. They might fit a multiple linear regression model using fertilizer and water as the .. predictor variables and crop yield as the response variable

15.What are the various branches of statistics?

Ans. The two main branches of statistics are descriptive statistics and inferential statistics

Descriptive statistics deals with the presentation and collection of data.

Inferential statistics, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics**.**