WORKSHEET 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

Answer: B

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

Answer: A

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

Answer: B

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

Answer: B

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

Answer: C

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

Answer: A

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

Answer: B

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

Answer: A

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 WORKSHEET

Answer: C

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

A normal distribution is the continuous probability distribution with a probability density function that gives you a symmetrical bell curve. Simply put, it is a plot of the probability function of a variable that has maximum data concentrated around one point and a few points taper off symmetrically towards two opposite ends.

Main concepts in normal distribution are:

1. Continuous Probability Distribution: A probability distribution where the random variable, X, can take any given value, e.g., amount of rainfall. You can record the rainfall received at a certain time as 9 inches. But this is not an exact value. The actual value can be 9.001234 inches or an infinite amount of other numbers. There is no definitive way to plot a point in this case, and instead, you use a continuous value.
2. Probability Density Function: An expression that is used to define the range of values that a continuous random variable can take.

A normal distribution has a probability distribution that is centered around the mean. This means that the distribution has more data around the mean. The data distribution decreases as you move away from the center. The resulting curve is symmetrical about the mean and forms a bell-shaped distribution

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

Deleting the Missing Values

Imputing the Missing Values

Imputing the Missing Values for Categorical Features

Imputing the Missing Values using Sci-kit Learn Library

Using “Missingness” as a Feature

12. What is A/B testing?

A/B testing is a experimentation process where two or more variants (A and B) are compared, in order to determine which variable is more effective.

To conduct an A/B test, you must produce two variants of the same content, each with a single variable altered. Afterward, you’ll present these two versions to two groups of people with identical sizes and compare their performance over a certain amount of time long enough to draw precise judgments regarding your findings.

A/B testing enables marketers to compare the performance of two different versions of marketing content. The following is the examples of A/B test types. Imagine you run an online store that sells clothing, and your goal is to boost the lead generation on your product pages. You choose to do A/B testing on two variants of your product page.

You keep the initial product page layout in control group A, which places the image of the product, synopsis, pricing, and “**Add to Cart**” button above the fold. You make a minor adjustment in experimental group B by moving the product review area under the “Add to Cart” button.
In the A/B test, **50% of the people** who visit your website are randomly assigned to **Group A**, and the remaining 50% are randomly assigned to **Group B**. You keep track of the number of visitors from each category that add the product to their cart and then check out.
You evaluate the data from the A/B test after a week of operation and discover that Group B had a 10% greater conversion rate than Group A. This indicates that the inclusion of the product review area had a favorable effect on website users’ purchasing decisions.

Based on these findings, **you decide to add a product review** area to every one of your product pages to boost sales and conversion rates. This is but one illustration of how A/B testing in marketing may be utilized to make data-driven decisions and enhance the efficacy of your marketing initiatives.

13. Is mean imputation of missing data acceptable practice?

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 terrible practice since it ignores feature correlation.

e.g. 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?

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?

There are **two main branches** of statistics
- Inferential Statistic.
- Descriptive Statistic.

Inferential Statistics:
Inferential statistics used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population.

Descriptive Statistics:
Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphycal