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

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. Usually replacing the standard error by its estimated value does change the CLT.

a) True

b) False

7. Which of the following testing is concerned with making decisions using data?

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

8. 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

WORKSHEET

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. For example, The bulk of students will score the average (C), while smaller numbers of students will score a B or D. An even smaller percentage of students score an F or an A. This creates a distribution that resembles a bell (hence the nickname). The bell curve is symmetrical. Half of the data will fall to the left of the mean; half will fall to the right.

The empirical rule tells you what percentage of your data falls within a certain number of standard deviations from the mean:
• 68% of the data falls within one standard deviation of the mean.
• 95% of the data falls within two standard deviations of the mean.
• 99.7% of the data falls within three standard deviations of the mean.

Properties of a normal distribution

- The mean, mode and median are all equal.
- The curve is symmetric at the center (i.e. around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.

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

Ans - Below are some techniques to handle missing data

1. Use deletion methods to eliminate missing data. The deletion methods only work for certain datasets where participants have missing fields.
2. Use regression analysis to systematically eliminate data.
3. Data imputation techniques.

The simplest imputation method is replacing missing values with the mean or median values of the dataset at large, or some similar summary statistic. This has the advantage of being the simplest possible approach, and one that doesn't introduce any undue bias into the dataset.

12. What is A/B testing?

Ans - A/B testing (also known as split testing) is the process of comparing two versions of a web page, email, or other marketing asset and measuring the difference in performance.

13. Is mean imputation of missing data acceptable practice?

Ans – Yes, mean imputation of missing data acceptable practice because it doesn't introduce any undue bias into the dataset.

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.

15. What are the various branches of statistics?

Ans - Statistics has two main branches namely:

Descriptive statistics:

This is a branch of statistics which deals with methods of collection of data, its presentation and organization in various forms, such as distribution tables, graphs (e.g., ogive, Lorenz curves, etc.), diagrams (e.g., pie charts) and finding measures of central tendency and measures of dispersion or spread which are used in the description of data.

Descriptive statistics is used to present the data in an understandable way, so that a meaningful description can be made.

Inferential or predictive statistics:

This is a branch of statistics which deals with techniques used for analysis of data, making estimates that lead to predictions and drawing conclusions or inferences from limited information taken on sample basis and testing the reliability of the estimates or predictions.
Inferential statistics is used to make comparisons or predictions about a larger group, known as population, using information gathered about a small part of that population called a sample.

Inferential statistics answers questions, such as "what is this data telling us about?" and "what should we do?" Techniques used are forecasting trends, hypothesis testing, kurtosis, skewness, etc.