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?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

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

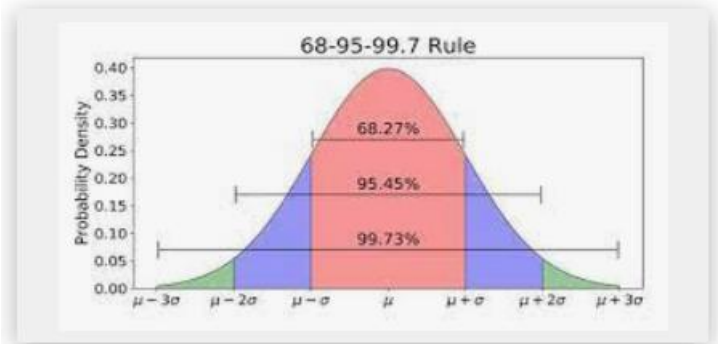
The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. Most people recognize its familiar bell-shaped curve in statistical reports.

The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely. While the normal distribution is symmetrical, not all symmetrical distributions are normal. For example, the Student's t, Cauchy, and logistic distributions are symmetric.

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena. Characteristics that are the sum of many independent processes frequently follow normal distributions. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

Types of normal distribution:

- Skewed Right
- Symmetric Distribution
- Skewed right



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

Missing data appear when no value is available in one or more variables of an individual. Following are the way we can handle the missing data:

- Deletions. Pairwise Deletion. Listwise Deletion/ Dropping rows. Dropping complete columns.
- Basic Imputation Techniques. Imputation with a constant value. Imputation using the statistics (mean, median, mode)
- K-Nearest Neighbor Imputation.
- Iterative imputer works like regression, NaN column value considered as label. It predicts the values for NaN

13. What is A/B testing?

A/B testing is a type of split testing and is commonly used to drive improvements to specific variables or elements by measuring user or audience engagement. The approach is commonly used to optimize marketing campaigns or digital assets like websites. In A/B testing a specific variable is altered such as a title, image, or element layout. A sample of the audience is shown the control version and the altered version in a 50/50 split. Half traffic will interact with the original version, the other half will interact with the newer version. Engagement or the completion of a defined goal is the metric that is compared between the versions after a set period of time.

A/B testing can be used to:

- Refine marketing campaign messaging and design.
- Improve conversion rates through enhancements to user experience.
- Continuously optimise assets like web pages by considering user engagement.

14. Is mean imputation of missing data acceptable practice?

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: 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.

15. What are the various branches of statistics?

Two branches, descriptive statistics and inferential statistics, comprise the field of statistics.

Descriptive Statistics:

CONCEPT The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

EXAMPLES The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory's production line.

CONCEPT The branch of statistics that analyzes sample data to draw conclusions about a population.

Inferential Statistics:

CONCEPT The branch of statistics that analyzes sample data to draw conclusions about a population.

EXAMPLE A survey that sampled 2,001 full-or part-time workers ages 50 to 70, conducted by the American

Association of Retired Persons (AARP), discovered that 70% of those polled planned to work past the traditional mid-60s retirement age. By using methods discussed in Section 6.4, this statistic could be used to draw conclusions about the population of all workers ages 50 to 70.