STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

Q1. Bernoulli random variables take (only) the values 1 and 0.

a) True
b) False

Answer: a) True

Q 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) Central Limit Theorem

Q3 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) Modeling bounded count data

Q4 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: d) All of the mentioned

Q5 ______ random variables are used to model rates

a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

Answer: c) Poisson Distribution

Q6 10. Usually replacing the standard error by its estimated value does change the CLT.

a) True
b) False

Answer: a) True

Q7 I . 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) Hypothesis

Q8 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) 0

Q9 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

Answer: c) Outliers cannot conform to the regression relationship

FLIP ROBO

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?**

Answer: The most significant probability distribution in statistics for independent, random variables is the normal distribution, sometimes referred to as the Gaussian distribution. In statistical reports, its well-known bell-shaped curve is generally recognized. The majority of the observations are centered around the middle peak of the normal distribution, which is a continuous probability distribution that is symmetrical around its mean. The probabilities for values that are farther from the mean taper off equally in both directions.

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

Answer: The definition of missing data is the values or data that are not available (or not stored) for one or more variables in the dataset. Missing values are typically represented by NaN. This phrase means "Not a Number." There are two ways of handling missing data

1. Deleting the Missing Values
2. Imputing the Missing Values

- **Deleting the Missing values**:

  1. **Deleting the entire row:-** If a row has many missing values, you can drop the entire row

  2. **Deleting the entire column:** If a column has many missing values, you can drop the entire column.

- **Imputing the Missing Values**:

  1. **Replacing with an arbitrary value :-** You can substitute an arbitrary value for the missing value if you can make a well reasonable assumption about it.

  2. **Replacing with the mean:-** The most used technique for estimating missing values in numerical columns. The 'fillna' method can be used to impute values using the means of the corresponding column values.

  3. **Replacing with the mode:-** The most frequent value is the mode. When there are categorical traits, it is applied.

  4. **Replacing with the median:-** The center value is called median. For imputation in the case of outliers, it is preferable to utilize the median value.

  5. **Forward fill & Backward fill:-** Sometimes it makes more sense to imput values using the preceding value rather than the mean, mode, or median. This is referred to as forward and backfill. Most often, it is applied to time series data.

  6. **Impute Missing Values Using Sci-kit Learn Library:-**

     - **Univariate approach:** Use the class SimpleImputer to replace any missing data with the mean, mode, median, or some other constant value when using a univariate approach. Consideration is given to a single feature.

     - **Multivariate approach**: In a multivariate approach, multiple features are taken into account. Taking the multivariate method into consideration, there are two techniques to impute missing data. utilizing the IterativeImputer or KNNImputer classes.

12. **What is A/B testing?**

Answer: A/B testing, also referred to as split testing, is a randomized process in which two or more variations of a variable (web page, page element, etc.) are displayed to various groups of website visitors at the same time to see which version has the greatest impact and influences business metrics. A/B tests are useful for companies in many sectors.

The use of A/B testing removes any uncertainty from website optimization and empowers experienced optimizers to make informed choices. In A/B testing, "control" or the original testing variable is referred to as "A." B, on the other hand, denotes "variation" or a new iteration of the initial testing variable.

In a B2C e-commerce website, run an experiment in which customers are offered two different calls-to-action to determine whether it affects clickthrough rates. An A/B test on the subject line of a B2B company's nurturing emails might be conducted to determine which subject line increases open rates.

13. **Is mean imputation of missing data acceptable practice?**

Answer: When data is combined across lengthy time periods from various sources to address real-world issues, missing values are frequently present, and accurate machine learning modeling necessitates careful treatment of missing data.

One of the methods is mean imputation, where the mean value of the whole feature column is used to replace any missing data. The data may be distorted in variables like salary, as seen in the previous section. It might not be a good idea to replace the missing values with mean imputation in these circumstances. Keep in mind that only numerical data can be used to impute missing data using mean values.

14. **What is linear regression in statistics?**

Answer: Since it is the simplest sort of model to fit and analyze, linear regression was the first statistical form statisticians investigated. The associations between at least one explanatory variable and an outcome variable are modeled using linear regression. The independent and dependent variables are recognized as just that—variables. The process is referred to as simple linear regression when there is only one independent variable (IV). Multiple regression is the statistical term for a situation when there are more IVs.

15. **What are the various branches of Statistics?**

Answer: **Branches of Statistics are:**

1. Descriptive Statistics: The first stage of statistical analysis, descriptive statistics, is concerned with gathering and presenting data. Brief explanatory coefficients that statisticians employ to sum up a specific data set are the scientific definition of descriptive statistics. A data set typically represents either a sample of a population or the full population. There are major categories for descriptive statistics.

   - Measures of central tendency: Statistics experts can estimate the center of the distribution of values with the aid of measures of central tendency. These inclination measurements are: Mean, Median, and Mode

   - Measures of variability:- The variability metric aids statisticians in examining how a particular collection of data is distributed. Measures of variability include quartiles, range, variance, and standard deviation

2. Inferential Statistics:- It is possible for statisticians to draw conclusions, make choices, or make predictions about a specific population using inferential statistics techniques. By utilizing descriptive statistics, inferential statistics frequently uses probabilistic language. Statisticians primarily use these methods to evaluate data, develop estimates, and draw conclusions from the scant information they have access to through sampling and verifying the accuracy of their estimations.

   Inferential statistics can be calculated in a variety of ways, including:

   - Analysis of regression
   - ANOVA, or analysis of variance
   - Covariance analysis (ANCOVA)
   - (T-test) Statistical significance
   - correlational research

15. What are the various branches of statistics?