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
   Ans.(a) True
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
   Ans.(a)Central limit theorem
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
   Ans.(c) Modeling bounded count data
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
   Ans.(D) All of the mentioned
5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
   Ans.(C)Poisson
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Ans.(10) 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
   Ans.(B).Hypothesis.
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
   Ans.(a) 0

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

Ans.(C)Outliers cannot conform to the regression relationship

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

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

Ans. Normal distribution is a type of continuous probability distribution. which means most of data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution.
In graph form, normal distribution will appear as a bell curve.

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

Ans.

- We often encounter missing values (due to some collection error or because of the corrupted
- data) while we are trying to analyze or understand our data.
- Missing values can cause bias and can reduce the efficiency & statistical power of the analysis of
- how the model performs & can distort the validity of the results. There are many ways in which
- we can handle missing data.

There are two primary methods to solve the missing data error:

A) Imputation method: Imputation is the process of replacing missing values with reasonable guesses for missing data. It is done as a pre-processing step. It is most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation.

B) Deletion or Removal of data method: While dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

- Before deciding which approach to employ, data scientists must understand why the data is missing.

- There are various imputation techniques used to cater missing data, some of which are given below:

1. **Mean, Median and Mode:** This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can calculate the mean or median of the existing observations. However, when there are many missing variables, mean or median results can result in a loss of variation in the data. This method does not depend on the relationship between the variables.

2. **Time-Series Specific Methods:** The time series methods of imputation assume the adjacent observations will be like the missing data. These methods work well when that assumption is valid. However, these methods won't always produce reasonable results, particularly in the case of strong seasonality.

(1) **K Nearest Neighbors:** In this method, data scientists choose a distance measure for k-neighbors, and the average is used to impute an estimate. The data scientist must select the number of nearest-neighbors and the distance metric. KNN can identify the most frequent value among the neighbors and the mean among the nearest-neighbors.

(2) **Multiple Imputation:** Multiple imputation is considered a good approach for data sets with a large amount of missing data. Instead of substituting a single value for each missing data point, the missing values are exchanged for values that encompass the natural variability and uncertainty of the right values. Using the imputed data, the process is repeated to make multiple imputed data sets. Each set is then analyzed using the standard analytical procedures, and the multiple analysis results are combined to produce an overall result. The various imputations incorporate natural variability into the missing values, which creates a valid statistical inference. Multiple imputations can produce statistically valid results even when there is a small sample size or a large amount of missing data.

3. What is A/B testing?

Ans.A/B testing is a method of comparing two versions of a webpage or app against each other to determine which one performs better.Its like hypothesis testing and two-sample hypothesis testing to compare two versions. Its commonly use for improve user experience and Advertising.

4. Is mean imputation of missing data acceptable practice?

- Mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.
- It's a popular solution to missing data, despite its drawbacks. Mainly because it's easy.
- But that doesn't make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population.
- There are many alternatives to mean imputation that provide much more accurate estimates and standard errors, so there really is no excuse to use it.

There are two big problems associated with mean imputation:

- Mean imputation does not preserve the relationships among variables.
- Mean Imputation Leads to An Underestimate of Standard Errors

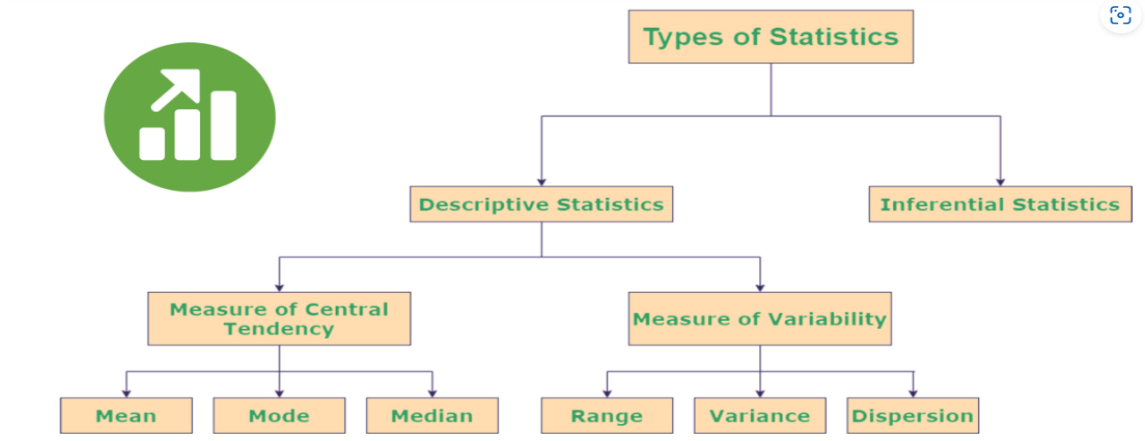
5. What is linear regression in statistics?

Ans:

- Linear regression is a basic and commonly used type of predictive analysis.
- this technique will identify the strength of the impact that the independent variables show on dependend variables.
- The overall idea of regression is to examine two things:
- Which variables in particular are significant predictors of the outcome variable, and in what way do they impact the outcome variable?
- These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

Three major uses for regression analysis are
(1) determining the strength of predictors
(2) forecasting an effect
(3) trend forecasting.

6. What are the various branches of statistics?



Statistics is a set of mathematical methods and tools that enable us to answer important questions about data.

1). Descriptive Statistics - this offers methods to summarise data by transforming raw observations into meaningful information that is easy to interpret and share.

A). Measures of Central Tendency- Central tendency measures specifically help statisticians evaluate the distribution center of values.

- **Mean** is a conventional method used to describe the central tendency. Typically, calculate the average of values, count all values, and then divide them with the number of available values.
- **Median** is the result that is in the middle of a set of values. An easy way to calculate the median is to edit the results in numerical journals and locate the result that is in the center of the distributed sample.
- **Mode** is the frequently occurring value in the given data set.

B). Measures of Variability: The variability measure helps statisticians to analyze the distribution that is spreading from a specific data set. Some of the variables of variability include quartiles, ranges, variances, and standard deviation.

- **Range** can be measured by subtracting the lowest value from the highest value. The wide range indicates high variability, and the small range specifies low variability in the distribution.
- **Variance** measures how far each number in the dataset from the mean. To compute variance first, calculate the mean and squared deviations from a mean.

2). Inferential Statistics - this offers methods to study experiments done on small samples of data and to conclude, bring decisions, or predict a defined population.

- Different types of inferential statistics include:

- **Regression analysis:** It is a set of statistical methods used to estimate relationships between a dependent variable and one or more independent variables. It includes several variations, like linear, multiple linear, and nonlinear. The most well-known models are simple linear and multiple linear.
- **Analysis of variance (ANOVA):** ANOVA is a statistical method that distributes observed variance data into various components. A one-way ANOVA is applied for three or more data groups to gain information about the relationship between the dependent and independent variables.
- **Analysis of covariance (ANCOVA):** It is used to test categorical variables' main and interaction effects on constant dependent variables and keep control for the impact of selected other constant variables. The control variables are known as covariates.
- **Statistical significance (t-test):** It is used to determine a significant difference between the means of two groups related to particular features. A t-test studies the t-statistic, the t-distribution values, and the degree of freedom to learn the statistical significance.
- **Correlation analysis:** It is a statistical method that is used to find the relationship between two variables or datasets and discover how strong the relationship may be.

FLIP ROBO