**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.
   1. True
   2. False

Answer :- a) True

1. 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?
   1. Central Limit Theorem
   2. Central Mean Theorem
   3. Centroid Limit Theorem
   4. All of the mentioned

**Answer :- a) Central Limit Theorem**

1. Which of the following is incorrect with respect to use of Poisson distribution?
   1. Modeling event/time data
   2. Modeling bounded count data
   3. Modeling contingency tables
   4. All of the mentioned

**Answer :- b) Modeling bounded count data is incorrect.**

1. Point out the correct statement.
   1. The exponent of a normally distributed random variables follows what is called the log- normal distribution
   2. Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   3. The square of a standard normal random variable follows what is called chi-squared distribution
   4. All of the mentioned

**Answer :-** c) The square of a standard normal random variable follows what is called chi-squared distribution

1. random variables are used to model rates.
   1. Empirical
   2. Binomial
   3. Poisson
   4. All of the mentioned

**Answer :- c)Poisson**

1. 10. Usually replacing the standard error by its estimated value does change the CLT.
   1. True
   2. False

**Answer :- b)False**

1. 1. Which of the following testing is concerned with making decisions using data?
   1. Probability
   2. Hypothesis
   3. Causal
   4. None of the mentioned

**Answer:- b) Hypothesis**

1. 4. Normalized data are centered at and have units equal to standard deviations of the original data.
   1. 0
   2. 5
   3. 1
   4. 10

**Answer:- a) 0**

1. Which of the following statement is incorrect with respect to outliers?
   1. Outliers can have varying degrees of influence
   2. Outliers can be the result of spurious or real processes
   3. Outliers cannot conform to the regression relationship
   4. None of the mentioned

**Answer:- 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?

**Answer :-**

Normal distribution, also known as the Gaussian distribution or bell curve, is a continuous probability distribution that is characterized by a symmetric bell-shaped curve. The normal distribution is widely used in statistics and data analysis due to its simplicity, mathematical tractability, and broad applicability to a wide range of natural and social phenomena.

The normal distribution is completely defined by two parameters: the mean (μ) and the standard deviation (σ). The mean represents the center of the distribution and the standard deviation represents the spread or dispersion of the distribution. The shape of the distribution is such that approximately 68% of the observations fall within one standard deviation of the mean, 95% fall within two standard deviations, and 99.7% fall within three standard deviations. The normal distribution is symmetric around the mean, and its tails extend indefinitely in both directions, but the probability of observing extreme values decreases rapidly as we move away from the mean.

The normal distribution arises naturally in many situations, such as the distribution of errors in measurements, the distribution of IQ scores, the distribution of heights and weights in a population, and the distribution of stock returns. It is also a key assumption in many statistical techniques, such as linear regression, hypothesis testing, and confidence intervals.

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

**Answer :-**

Handling missing data is a common problem in data analysis and there are several ways to deal with it. The choice of method depends on the amount and pattern of missing data, the type of analysis to be performed, and the assumptions about the missingness mechanism. Here are some common strategies for handling missing data:

Complete case analysis: In this approach, we simply delete all the observations that contain missing data. This method is easy to implement, but it can lead to biased results if the missingness is related to the outcome or other variables of interest.

Imputation methods: Imputation methods involve filling in the missing values with estimated values based on the available data. There are several imputation techniques, including:

Mean imputation: Replacing missing values with the mean of the available data for that variable. This method assumes that the missing values are missing completely at random (MCAR) and that the variable is normally distributed.

Median imputation: Replacing missing values with the median of the available data for that variable. This method is similar to mean imputation but is more robust to outliers and non-normal data.

Regression imputation: Using a regression model to predict the missing values based on the available data for that variable and other relevant variables. This method assumes that the missingness is missing at random (MAR) and that there is a linear relationship between the missing variable and the other variables.

Multiple imputation: Creating multiple copies of the dataset with different imputed values for the missing data and combining the results to produce an overall estimate. This method accounts for the uncertainty in the imputed values and produces more accurate results than single imputation methods.

Other methods: Other methods for handling missing data include weighting methods, such as inverse probability weighting or propensity score weighting, and model-based methods, such as maximum likelihood or Bayesian methods.

The choice of imputation method depends on the assumptions about the missingness mechanism and the specific research question. Multiple imputation is generally recommended as the preferred method, as it produces more accurate estimates and allows for uncertainty quantification. However, the method should be chosen based on the nature of the data and the research question, and sensitivity analysis should be conducted to assess the impact of missingness assumptions on the results.

窗体顶端

窗体底端

1. What is A/B testing?

Answer:-

A/B testing is a statistical technique used in marketing and product development to compare two versions of a webpage, advertisement, or other marketing material to determine which one performs better in terms of user engagement or conversion rate.

In A/B testing, two different versions of a marketing asset are randomly presented to users in equal proportions, and the results are compared to see which version performs better. The version that generates a higher engagement or conversion rate is typically chosen as the winner, and further optimization can be performed on that version to improve its performance even further.

A/B testing can be used to test different variables such as headlines, images, colors, calls-to-action, or even entire landing pages. It is an important tool for businesses to optimize their marketing efforts and improve their return on investment.

1. Is mean imputation of missing data acceptable practice?

**Answer :-**

Mean imputation is a commonly used method for handling missing data, where missing values in a dataset are replaced with the mean value of the non-missing values in the same variable. While mean imputation is a simple and convenient way to handle missing data, it has several limitations and may not always be an acceptable practice.

One of the main limitations of mean imputation is that it can lead to biased and inaccurate results, especially if the missing data are not missing completely at random. Mean imputation assumes that the missing values are missing at random, which means that the probability of a value being missing is independent of its true value and any other variables in the dataset. If this assumption is not met, then mean imputation can introduce bias and inaccuracies in the data analysis.

Another limitation of mean imputation is that it can reduce the variability of the data and underestimate the standard error of the mean. This can result in inflated statistical significance and confidence intervals that are too narrow, leading to false conclusions.

Therefore, while mean imputation is a quick and easy method for handling missing data, it is not always an acceptable practice. Alternative methods such as multiple imputation or maximum likelihood estimation should be considered in situations where missing data are not missing completely at random.

1. What is linear regression in statistics?

**Answer :**

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes that there is a linear relationship between the dependent variable and the independent variables.

In simple linear regression, there is only one independent variable, and the relationship between the dependent variable and the independent variable is described by a straight line. The goal of linear regression is to find the equation of this line, which can be used to predict the value of the dependent variable for any given value of the independent variable.

In multiple linear regression, there are multiple independent variables, and the relationship between the dependent variable and the independent variables is described by a linear equation with multiple coefficients. The goal of multiple linear regression is to estimate the values of these coefficients, which can be used to predict the value of the dependent variable for any combination of values of the independent variables.

Linear regression is commonly used in many fields, including economics, finance, social sciences, and engineering. It can be used for prediction, estimation, and hypothesis testing, and is a powerful tool for understanding the relationships between variables. However, it assumes that the relationship between the variables is linear, and is sensitive to outliers and influential observations. Careful diagnostic checking is important to ensure that the assumptions of linear regression are met before using the results for decision-making.

1. What are the various branches of statistics?

**Answer :-**

Statistics is a broad field that has several branches, each focusing on different aspects of data analysis and modeling. The main branches of statistics are:

Descriptive statistics: This branch deals with the summarization and visualization of data using measures such as means, medians, standard deviations, and histograms.

Inferential statistics: This branch deals with making inferences and predictions about a population based on a sample of data, using methods such as hypothesis testing and confidence intervals.

Probability theory: This branch deals with the study of random events and their likelihood of occurrence, and forms the basis of statistical inference.

Biostatistics: This branch applies statistical methods to biological and medical research, including clinical trials and epidemiology.

Econometrics: This branch applies statistical methods to economic data, including forecasting, modeling, and causal inference.

Statistical machine learning: This branch focuses on developing algorithms and models that can learn from data and make predictions or decisions.

Bayesian statistics: This branch deals with the use of probability theory to model uncertainty and update beliefs based on new data.

Spatial statistics: This branch deals with the analysis of data that are spatially correlated, such as data from geographic information systems (GIS).

Time series analysis: This branch deals with the analysis of data that are collected over time, such as stock prices or weather data.

Each branch of statistics has its own set of techniques, methods, and applications, and together they form the foundation of modern data analysis and modeling.

