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

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

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 : b

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 : c

5. ______ random variables are used to model rates. a) Empirical b) Binomial c) Poisson d) All of the mentioned

Ans : c

6. Usually replacing the standard error by its estimated value does change the CLT. a) True b) False

Ans : b

7. 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

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

Ans : a

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 : d

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

Ans : he term "Normal Distribution" refers to a symmetric, bell-shaped probability distribution that is characterized by two parameters: the mean (μ) and the standard deviation (σ). It is also known as the Gaussian distribution after Carl Friedrich Gauss, who contributed significantly to its understanding.

Key characteristics of the Normal Distribution include:

1. **Symmetry**: The distribution is symmetric around its mean μ.
2. **Bell-shaped curve**: It has a single peak at the mean μ, with the probability density function reaching its maximum at μ and decreasing symmetrically in both directions.
3. **Parameterized by mean and standard deviation**: The mean μ determines the center of the distribution, while the standard deviation σ measures the spread or dispersion of the distribution. The variance, σ^2, is also commonly used.
4. **Empirical rule**: A large proportion (about 68%) of the observations fall within one standard deviation of the mean, an even larger proportion (about 95%) within two standard deviations, and nearly all (about 99.7%) within three standard deviations.
5. **Foundation in theory and application**: The Normal Distribution arises naturally in many natural and social phenomena due to the Central Limit Theorem, which states that the distribution of the sum (or average) of a large number of independent, identically distributed random variables tends toward a normal distribution, regardless of the original distribution of the variables.

The Normal Distribution is widely used in statistics, probability theory, and many fields of science and engineering due to its mathematical tractability and its applicability to describe real-world data.

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

Ans : Handling missing data is a crucial step in data analysis and can significantly impact the validity and outcomes of statistical analyses. Here are some common approaches to handle missing data:

1. **Identify the nature and mechanism of missingness**: Before choosing a method, it's important to understand why data are missing. Missing data can occur randomly (Missing Completely at Random - MCAR), systematically (Missing at Random - MAR), or non-randomly (Missing Not at Random - MNAR). The mechanism of missingness can guide the choice of imputation method.
2. **Delete observations**: This is the simplest approach but may lead to loss of valuable information, especially if data are missing non-randomly.
3. **Mean or median imputation**: Replace missing values with the mean or median of the observed data for that variable. This is straightforward but can distort the distribution and variability of the data.
4. **Mode imputation**: For categorical variables, replace missing values with the mode (most frequent category).
5. **Regression imputation**: Use other variables that are correlated with the variable with missing data to predict missing values using regression models.
6. **Multiple imputation**: Generate several plausible values for each missing data point to account for uncertainty in the imputation process. This involves creating multiple datasets with imputed values and analyzing each dataset separately before combining results.
7. **K-nearest neighbors (KNN) imputation**: Replace missing values with the average of values from similar observations, based on other variables.
8. **Expectation-Maximization (EM) algorithm**: An iterative algorithm that estimates parameters of a statistical model with missing data.
9. **Domain-specific knowledge**: Use knowledge about the data and the domain to inform imputation decisions, such as using temporal patterns or expert judgment.

The choice of imputation technique depends on the specific context of the data, the nature of missingness, the distribution of the data, and the goals of the analysis. There is no one-size-fits-all solution, and it's often recommended to compare results from different methods or to perform sensitivity analyses to understand the impact of imputation on the results.

In practice, multiple imputation is often preferred when feasible because it accounts for uncertainty in imputation and provides more robust estimates compared to single imputation methods like mean or median imputation. However, the appropriateness of each method should be evaluated based on the specific dataset and research objectives.

12.What is A/B testing?

Ans : A/B testing, also known as split testing, is a controlled experiment used to determine the effectiveness of a change (such as a new feature, design element, or marketing strategy) on user behavior or outcomes. It involves comparing two versions of a web page, app, email, advertisement, or other digital content to see which one performs better according to a predefined metric.

13.Is mean imputation of missing data acceptable practice?

Ans : Mean imputation of missing data is a straightforward method to handle missing values by replacing them with the mean of the observed data for that variable. While it is widely used due to its simplicity, there are important considerations and potential drawbacks to be aware of:

Pros:

1. **Simple and easy to implement**: Mean imputation is easy to understand and apply, making it accessible even for non-statisticians.
2. **Preserves sample size**: It avoids the reduction in sample size that would occur if observations with missing data were removed.
3. **Maintains statistical power**: By preserving sample size, mean imputation can help maintain statistical power in analyses.

Cons:

1. **Distorts variable distribution**: Mean imputation can distort the distribution of the variable, particularly if missing data are not missing at random (MAR). This can lead to biased estimates of variability and relationships between variables.

2. **Underestimates variability**: Mean imputation tends to underestimate the variability of the data, as it reduces the variance of the imputed variable.
3. **Assumption of normality**: Mean imputation assumes that the variable follows a normal distribution. If the variable is highly skewed or has outliers, mean imputation may not be appropriate.
4. **Does not account for uncertainty**: It does not account for uncertainty in the imputed values, which can lead to underestimation of standard errors and incorrect inference.

14.What is linear regression in statistics?

Ans : In statistics, linear regression is a statistical method used to model the relationship between a dependent variable (often denoted as YYY) and one or more independent variables (often denoted as X1,X2,…,XpX_1, X_2, \ldots, X_pX1,X2,…,Xp). It assumes that this relationship can be approximated by a linear equation of the form:

Y=β0+β1X1+β2X2+…+βpXp+ϵY = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilonY=β0+β1X1+β2X2+…+βpXp+ϵ

where:

- YYY is the dependent variable (also called the response or outcome variable).
- X1,X2,…,XpX_1, X_2, \ldots, X_pX1,X2,…,Xp are the independent variables (also called predictors, features, or explanatory variables).
- β0,β1,…,βp\beta_0, \beta_1, \ldots, \beta_pβ0,β1,…,βp are the coefficients, representing the parameters of the model that quantify the effect of each independent variable on the dependent variable.
- ϵ\epsilonϵ is the error term, representing the variability in YYY that is not explained by the linear relationship with the XXX variables.

The goal of linear regression is to estimate the coefficients β0,β1,…,βp\beta_0, \beta_1, \ldots, \beta_pβ0,β1,…,βp that best fit the observed data, minimizing the sum of squared differences between the observed values of YYY and the values predicted by the linear model.

Key Concepts:

- **Simple vs. Multiple Regression**: Simple linear regression involves one independent variable XXX, whereas multiple linear regression involves two or more independent variables X1,X2,…,XpX_1, X_2, \ldots, X_pX1 ,X2,…,Xp.

- **Assumptions**: Linear regression assumes that the relationship between the dependent and independent variables is linear, that the errors ϵ\epsilonϵ are normally distributed with mean zero, constant variance (homoscedasticity), and are independent of each other.
- **Interpretation of Coefficients**: The coefficients β0,β1,…,βp\beta_0, \beta_1, \ldots, \beta_pβ0,β1,…,βp quantify the change in the dependent variable YYY associated with a one-unit change in each independent variable XXX, holding other variables constant.
- **Model Evaluation**: Evaluation of a linear regression model involves assessing the goodness of fit (how well the model fits the data) using metrics like R-squared, adjusted R-squared, and residual analysis.
- **Applications**: Linear regression is widely used in various fields such as economics, social sciences, biology, engineering, and business to understand relationships between variables, make predictions, and infer causal relationships (in some cases).

Linear regression forms the basis for more advanced regression techniques and is a fundamental tool in statistical analysis, hypothesis testing, and making data-driven decisions based on empirical data.

15.What are the various branches of statistics?

Ans : Statistics, as a field of study and application, branches out into several subfields or branches, each focusing on different aspects of data analysis, inference, and interpretation. Some of the main branches of statistics include:

1. **Descriptive Statistics**: This branch involves methods for summarizing and describing data sets, including measures of central tendency (mean, median, mode), measures of dispersion (variance, standard deviation), and graphical representations (histograms, box plots, etc.).
2. **Inferential Statistics**: Inferential statistics involves making inferences and predictions about populations based on sample data. It includes hypothesis testing, confidence intervals, and methods for estimating parameters of populations.
3. **Probability Theory**: Probability theory is the mathematical foundation of statistics, dealing with the study of random events and their likelihood or uncertainty. It provides the basis for understanding the behavior of random variables and distributions.
4. **Biostatistics and Medical Statistics**: These branches focus on the application of statistical methods to biological and medical data,

including clinical trials, epidemiological studies, and health-related research.

5. **Econometrics**: Econometrics applies statistical methods to economic data to test economic theories, forecast economic trends, and analyze economic relationships.
6. **Statistical Modeling**: Statistical modeling involves the development and application of mathematical models to describe relationships between variables and make predictions. This includes linear regression, logistic regression, time series analysis, and more complex models like hierarchical models and machine learning algorithms.
7. **Multivariate Statistics**: This branch deals with the analysis of data sets with more than two variables, focusing on relationships between multiple variables and dimension reduction techniques.
8. **Bayesian Statistics**: Bayesian statistics is an approach to statistical inference that uses Bayes' theorem to update probabilities based on new data and prior knowledge. It emphasizes the use of prior beliefs and incorporates uncertainty in a systematic way.
9. **Nonparametric Statistics**: Nonparametric statistics includes methods that do not assume specific probability distributions for the variables being studied. They are useful when assumptions of parametric statistics (such as normality) are not met or when dealing with ordinal or categorical data.
10. **Spatial Statistics**: Spatial statistics deals with the analysis of data distributed across space or geographic regions, including spatial autocorrelation, spatial interpolation, and spatial regression.
11. **Time Series Analysis**: Time series analysis focuses on analyzing data collected over time to identify patterns, trends, and seasonal variations. It includes methods for forecasting future values based on historical data.

These branches of statistics often overlap and complement each other, depending on the specific application and goals of analysis. They collectively form a diverse and robust toolkit for understanding data, making decisions based on data, and drawing meaningful conclusions in various fields of science, industry, and research.