Ayiman Bekowei

Ds0523

Python Programming

1). (C) %

2). (B) 0

3). (C) 24

4). (A) 2

5). (D) 6

6). (B) It encloses the lines of code which will be executed if any error occurs while executing the lines of code in the try block.

7). (A) It is used to raise an exception.

8). (C) in defining a generator

9). (A)_abc & (C)abc2

10). (A) yield & (B) raise

Statistic Worksheet.

1). (A) true

2). (A) Central Limit theorem

3). (B) Modelling bounded count data

4). (C)

5). (C)

6). (B)

7). (B)

8). (A)

9). (C)

10). A normal distribution, also known as a Gaussian distribution, is a continuous probability distribution that is symmetric around its mean, forming a bell-shaped curve. The normal distribution is characterized by two parameters: the mean (μ) and

the standard deviation (σ). The shape of the curve is determined by these parameters.

11). Deletion: Listwise Deletion (Complete Case Analysis): Remove entire rows with missing values. This approach is simple but may lead to a loss of valuable information if missing values are not completely at random.

Imputation: Mean, Median, or Mode Imputation: Replace missing values with the mean, median, or mode of the observed values in the variable. This is a simple method but may not be suitable if the data has outliers or if the missingness is not random.

Forward Fill or Backward Fill: Propagate the last known value forward or the next known value backward to fill missing values in time-series or sequential data.

Linear Regression Imputation: Predict missing values using a linear regression model based on other variables. This method assumes a linear relationship between the variable with missing values and other variables.

K-Nearest Neighbors (KNN) Imputation: Impute missing values based on the values of their k-nearest neighbors in the feature space. This method considers multiple variables and is effective when there is a pattern in the missing data.

Multiple Imputation: Generate multiple imputed datasets, each with different imputed values, and perform the analysis on each dataset. This accounts for uncertainty due to missing values.

Advanced Techniques: Predictive Modeling: Use machine learning algorithms to predict missing values based on other features. This can be effective when there is a complex relationship between variables.

Autoencoders: Train an autoencoder neural network to reconstruct missing values. Autoencoders can capture nonlinear relationships in the data.

Domain-Specific Imputation: Expert Knowledge: Use domain knowledge to impute missing values. For example, if missing data is related to a specific condition, a domain expert may provide a reasonable estimate.

12). A/B testing, also known as split testing, is a method used in marketing and product development to compare two versions of a webpage, email campaign, app feature, or other elements to determine which one performs better. The goal of A/B testing is to identify changes that can improve a given metric, such as click-through rate, conversion rate, or user engagement.

13). Mean imputation, where missing values are replaced with the mean of the observed values for a variable, is a straightforward and easy-to-implement method for handling missing data. However, its acceptability depends on the context, nature of the data, and the assumptions made.

14). Linear regression is a statistical method used to model the relationship between a dependent variable (also called the response or outcome variable) and one or more independent variables (also called predictors or features). The relationship is modeled as a linear equation, and the method aims to find the best-fitting linear line (or hyperplane, in the case of multiple predictors) through the data.

15). Descriptive Statistics: Involves methods for summarizing and describing the main features of a dataset. Measures such as mean, median, mode, variance, and standard deviation fall under descriptive statistics.

Inferential Statistics: Concerned with making predictions or inferences about a population based on a sample of data. It includes techniques such as hypothesis testing, confidence intervals, and regression analysis.

Probability Theory: The study of randomness and uncertainty. Probability theory provides the foundation for statistical inference and modeling uncertain events.

Biostatistics: Applies statistical methods to biological and medical research. It involves the design and analysis of experiments, clinical trials, and epidemiological studies.

Econometrics: Applies statistical methods to economic data. It is used to model and analyze economic relationships, forecast economic trends, and test economic theories.

Social Statistics: Focuses on the analysis of social phenomena using statistical methods. It includes the study of demographics, sociology, and social research.

Psychometrics: Applies statistical methods to the measurement of psychological traits and constructs. It is commonly used in educational testing and personality assessment.

Statistical Computing: Involves the development and application of computational methods for statistical analysis. This includes programming languages, software, and algorithms used in statistical computations.

Environmental Statistics: Applies statistical methods to environmental science and research. It involves the analysis of environmental data, including climate, pollution, and ecology. Spatial Statistics: Focuses on the analysis of spatial or geographic data. It includes techniques for studying patterns, trends, and relationships in spatial datasets.

Bayesian Statistics: Based on Bayesian probability theory, this branch deals with statistical inference using Bayesian methods. It involves updating probability estimates based on prior knowledge and new evidence.

Quality Control and Six Sigma: Applied in industries to monitor and control the quality of processes and products. Six Sigma methodologies aim to reduce variability and improve overall quality.

Machine Learning

1). (A) Least Square error

2). (A) Linear regression is sensitive to outliers

3). (B) Negatives

4). (C) Both of them

5). (C) low bias and high variance

6). (B) Predictive model

7). (D) Regularization

8). (D) Smote

9). (A) TPR

10). (B) False

11). (B) Apply PCA to project high dimensional data

12). (D)

13). Regularization is a technique used in machine learning and statistical modeling to prevent overfitting and improve the generalization of a model to new, unseen data. It involves adding a penalty term to the objective function or cost function that the model is trying to optimize. The penalty term discourages the model from becoming too complex or fitting the training data too closely.

14). Linear Regression is commonly used however, Logistic Regression, Support Vector Machines, Neural Networks, Decision Trees, K-Nearest Neighbors are also part of regularization.

15). In the context of linear regression, the term "error" refers to the difference between the predicted values (obtained from the linear regression equation) and the actual observed values in the dataset. These differences are also known as residuals.