STATISTICS WORKSHEET-1

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

Answer: a) True

Ques.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?

Answer: a) Central Limit Theorem

Ques.3. Which of the following is incorrect with respect to use of Poisson distribution?

Answer: c) Modeling contingency tables

Ques.4. Point out the correct statement.

Answer: d) All of the mentioned

Ques.5. _____ random variables are used to model rates.

Answer: c) Poisson

Ques.6. Usually replacing the standard error by its estimated value does change the CLT.

Answer: b) False

Ques.7. Which of the following testing is concerned with making decisions using data?

Answer: b) Hypothesis

Ques.8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

Answer: a) 0

Ques.9. Which of the following statement is incorrect with respect to outliers?

Answer: c) Outliers cannot conform to the regression relationship

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

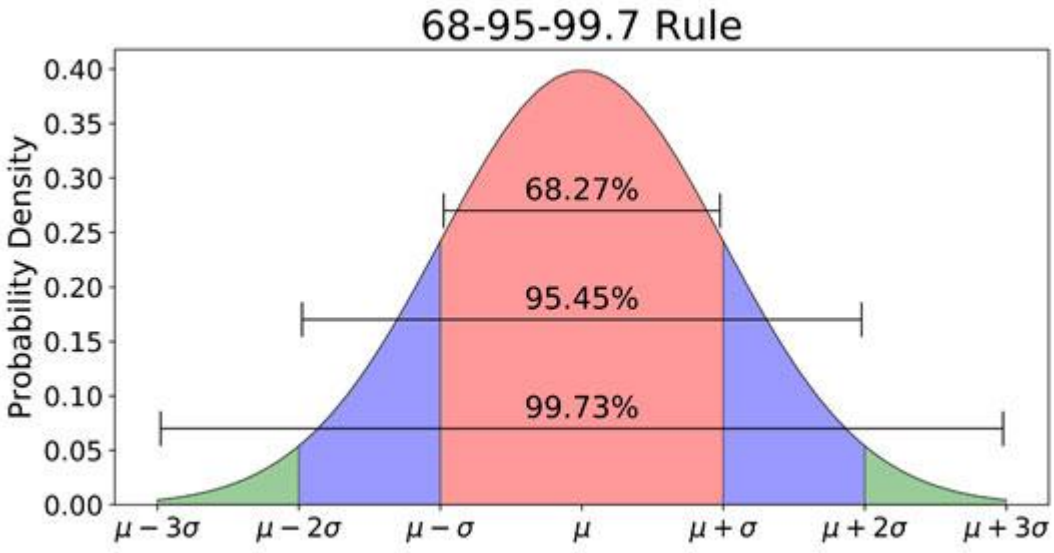
Answer:

Normal Distribution:

The term Normal Distribution is an extremely important continuous probability distribution. It is also called as Gaussion Distribution. It is the most widely known and used of all distributions. Why we should use Normal Distribution? Many things actually are normally distribution, or very close to it. For example, height and intelligence are approximately normally distributed. The normal distribution is easy to work with mathematically.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

The above image is the main definition of Normal Distribution, from which 2 parameters are needed in python:
μ - Mean, the middle of the distribution.
σ - Standard deviation
All this help us to find the Bell shape graph which define the heights and all value of **x**. (Refer below snap)



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

Answer:

When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data. The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low.

"***More understanding we have why your data is missing, Better imputation you can do***".

Below four are the method which help us to understand the missing data:

MCAR - Missing Completely At Random
MAR - Missing At Random
NMAR - Not Missing At Random
SM - Structured Missing

Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation.

Common Methods:
Mean or Median Imputation. When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations.

Ques.12: What is A/B Testing?

Answer:

A/B testing is also known as split-run testing, is a user experience research methodology. It compares the performance of two versions of content to see which one appeals more to visitors/viewers.

Benefits of A/B testing:
* Increase Revenue 7 Conversions
* Rapid Iteration
* Learn What Works
* Uses Actual Site Visitors
* Data Driver Decision Making

Ques.13: Is mean imputation of missing data acceptable practice?

Answer:

The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Ques.14: What is liner regression in statistics?

Answer:

Liner Regression is used to predict the relationship between two variables by applying a liner equation to observed data. There are two types of variables,

1. Independent Variable 2. Dependent Variable

In Simple words, it minimize the sum by choosing the appropriate parameters a and b. The resulting line is called the least square line or sample regression line.

Ques.15: What are the various branches of statistics?

Answer: There are two branches of the statistics:

1. Descriptive Statistics 2.Inferential Statistics

Descriptive statistics describe, show, and summarize the basic features of a data set found in a given study, presented in a summary that describes the data sample and its measurements. It helps analysts to understand the data better.Descriptive statistics represent the available data sample and does not include theories, inferences, probabilities, or conclusions. That's a job for inferential statistics.

Inferential statistics helps study a sample of data and make conclusions about its population. A sample is a smaller data set drawn from a larger data set called the population. If the sample does not represent the population, one cannot make accurate estimations related to the latter. The purpose of studying inferential statistics is to infer the behavior of a population.

PYTHON – WORKSHEET- 1

Ques.1. Which of the following operators is used to calculate remainder in a division?

Answer: C) %

Ques.2. In python 2//3 is equal to?

Answer: B) 0

Ques.3. In python, 6<<2 is equal to?

Answer: C) 24

Ques.4. In python, 6&2 will give which of the following as output?

Answer:A) 2

Ques.5. In python, 6|2 will give which of the following as output?

Answer: D) 6

Ques.6. What does the finally keyword denotes in python?

Answer: C) the finally block will be executed no matter if the try block raises an error or not.

Ques.7. What does raise keyword is used for in python?

Answer: A) It is used to raise an exception.

Ques.8. Which of the following is a common use case of yield keyword in python?

Answer: C) in defining a generator

Ques.9. Which of the following are the valid variable names?

Answer: A) _abc C) abc2

Ques.10. Which of the following are the keywords in python?

Answer: A) yield B) raise

MACHINE LEARNING

Ques.1. Which of the following methods do we use to find the best fit line for data in Linear Regression?

Answer: A) Least Square Error

Ques.2. Which of the following statement is true about outliers in linear regression?

Answer: A) Linear regression is sensitive to outliers

Ques.3. A line falls from left to right if a slope is _____?

Answer: B) Negative

Ques.4. Which of the following will have symmetric relation between dependent variable and independent variable?

Answer: B) Correlation

Ques.5. Which of the following is the reason for over fitting condition?

Answer: C) Low bias and high variance

Ques.6. If output involves label then that model is called as:

Answer: B) Predictive modal

Ques.7. Lasso and Ridge regression techniques belong to _____?

Answer: D) Regularization

Ques.8. To overcome with imbalance dataset which technique can be used?

Answer: A) Cross validation

Ques.9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?

Answer: A) TPR and FPR

Ques.10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.

Answer: A) True

Ques.11. Pick the feature extraction from below:

Answer: B) Apply PCA to project high dimensional data

Ques.12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?

Answer: A) We don't have to choose the learning rate.
B) It becomes slow when number of features is very large.

Ques.13. Explain the term regularization?

Answer:

Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting. The following are the types of regularizations:

1. Ridge Regression (L2 Norm)
2. Lasso (L1 Norm)
3. Dropout.

Ques.14. Which particular algorithms are used for regularization?

Answer:

Following algorithms are used for regularization:
1. Ridge Regression.
2. LASSO (Least Absolute Shrinkage and Selection Operator) Regression.
3. Elastic-Net Regression.

Ques.15. Explain the term error present in linear regression equation?

Answer:

In a linear regression equation, Error is the difference between the actual value and Predicted value. A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.