In Q1 to Q11, only one option is correct, choose the correct option:

1. Which of the following methods do we use to find the best fit line for data in Linear Regression?
   A) Least Square Error B) Maximum Likelihood
   C) Logarithmic Loss D) Both A and B
   Answer : A

2. Which of the following statement is true about outliers in linear regression?
   A) Linear regression is sensitive to outliers B) linear regression is not sensitive to outliers
   C) Can't say D) none of these
   Answer: A

3. A line falls from left to right if a slope is _____?
   A) Positive B) Negative
   C) Zero D) Undefined
   Answer: B

4. Which of the following will have symmetric relation between dependent variable and independent variable?
   A) Regression B) Correlation
   C) Both of them D) None of these
   Answer: C

5. Which of the following is the reason for over fitting condition?
   A) High bias and high variance B) Low bias and low variance
   C) Low bias and high variance D) none of these

6. If output involves label then that model is called as:
   A) Descriptive model B) Predictive modal
   C) Reinforcement learning D) All of the above
   Answer: B

7. Lasso and Ridge regression techniques belong to _____?
   A) Cross validation B) Removing outliers
   C) SMOTE D) Regularization
   Answer: D

8. To overcome with imbalance dataset which technique can be used?
   A) Cross validation B) Regularization
   C) Kernel D) SMOTE
   Answer: D

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____to make graph?
   A) TPR and FPR B) Sensitivity and precision
   C) Sensitivity and Specificity D) Recall and precision
   Answer: C

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

    Answer: A&B

11. Pick the feature extraction from below:
    A) Construction bag of words from a email
    B) Apply PCA to project high dimensional data

C) Removing stop words
D) Forward selection

Answer: A, B &C

In Q12, more than one options are correct, choose all the correct options:

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?
A) We don't have to choose the learning rate.
B) It becomes slow when number of features is very large.
C) We need to iterate.
D) It does not make use of dependent variable.

Answer: A&B

Q13 and Q15 are subjective answer type questions, Answer them briefly.

1. Explain the term regularization?
2. Which particular algorithms are used for regularization?
3. Explain the term error present in linear regression equation?

Q13 Explain the term regularization?

Ans:

Regularization refers to techniques that are used to calibrated machine learning models in order to minimize the adjusted loss function and prevent over fitting or under fitting. Another explanation the term regularization refers to a set of techniques that regularizes learning from particular features for traditional algorithms (or) neurons in the case of neural network algorithms

Q14 Which particular algorithms are used for regularization?

Ans: 1. Ridge regression (or) L2 Norm

2. LASSO (Least absolute shrinkage and selection operator) regression (or) L1 Norm

3. Elastic-net regression (or) Dropout.

Q15 Explain the term error present in linear regression equation?

Ans:

With in a linear regression model Tracking a stock's price over time is the difference between the expected price at a particular time and the price that was actually observed. Linear regression is a form of analysis that relates to current trends experienced by a particular security (or) index by producing a relationship between a dependent and independent variables, such as the price of security and the passage of time, resulting in a trend line that can be used as a predictive model

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

   Answer: 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

   Answer: 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

   Answer: 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

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

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

   Answer : b
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

   Answer: b
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

   Answer: 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

Answer: c

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

1. What do you understand by the term Normal Distribution?
2. How do you handle missing data? What imputation techniques do you recommend?
3. What is A/B testing?
4. Is mean imputation of missing data acceptable practice?
5. What is linear regression in statistics?
6. What are the various branches of statistics?

Q10. What do you understand the term Normal Distribution?
Ans:

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "bell curve".

Q11. How do you handle missing data? What imputation techniques do you recommend?
Ans :

Handling missing data:

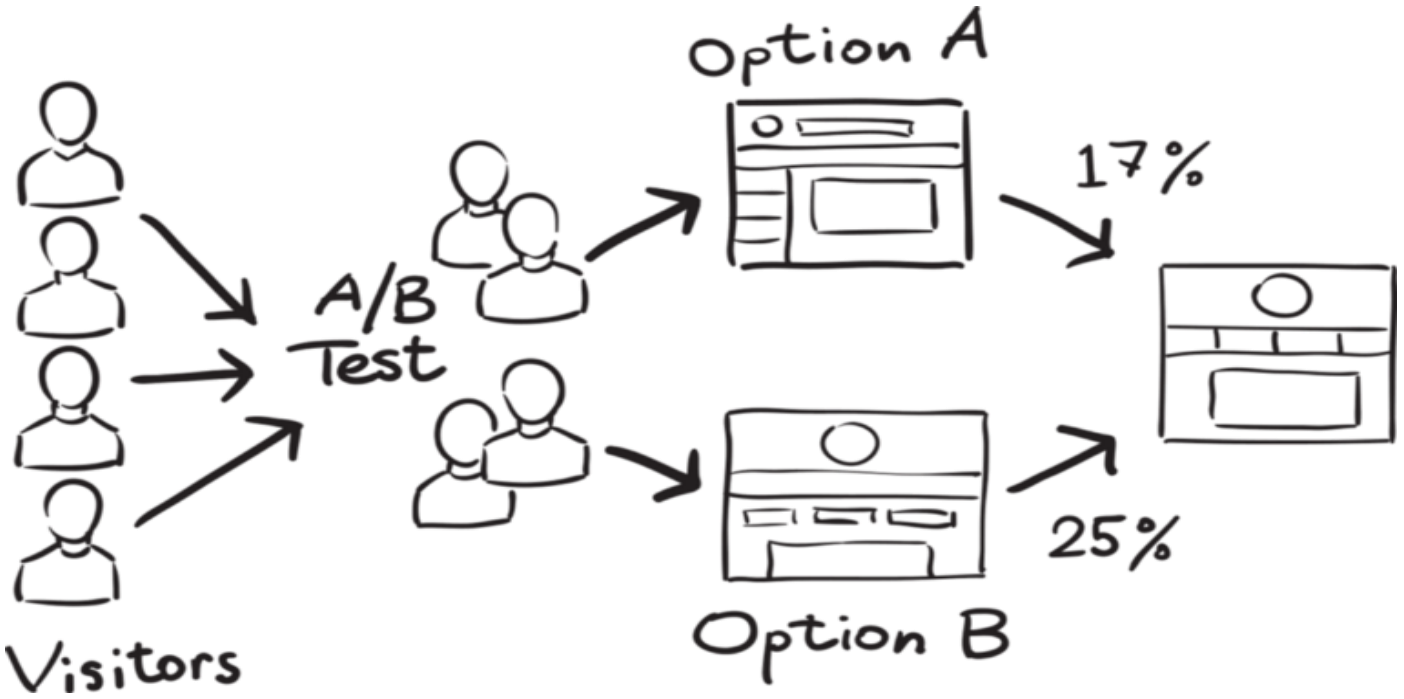
- Listwise or case deletion
- Pairwise deletion
- Mean substitution
- Regression imputation
- Last observation carried forward
- Maximum likelihood
- Expectation-Maximization
- Multiple imputation.

Imputation techniques:

- Mean imputation
- Substitution
- Hot deck imputation
- Cold deck imputation
- Regression imputation
- Stochastic regression imputation
- Interpolation and extrapolation

Q12. What is A/B testing?
Ans:

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.
For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.
In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The **population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

Q13. Is mean imputation of missing data acceptable practice?
Ans:

Mean imputation reduces the variance of the imputed variables. Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval. Mean imputation does not preserve relationships between variables such as correlations.

Yet, there is no established cutoff from the literature regarding an acceptable percentage of missing data in a data set for valid statistical inferences. For example, Schafer (1999) asserted that a missing rate of 5% or less is inconsequential.

Q14.What is linear regression in statistics?
Ans:

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Q15. What are the various branches of statistics?
Ans:

There are THREE types:

1. Data Collection
2. Descriptive Statistics
3. Inferential Statistics

FLIP ROBO

PYTHON – WORKSHEET 1

Q1 to Q8 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following operators is used to calculate remainder in a division?
   A) # B) &
   C) % D) $
   Answer: C
2. In python 2//3 is equal to?
   A) 0.666 B) 0
   C) 1 D) 0.67
   Answer: B
3. In python, 6<<2 is equal to?
   A) 36 B) 10
   C) 24 D) 45
   Answer: A
4. In python, 6&2 will give which of the following as output?
   A) 2 B) True
   C) False D) 0
   Answer: A
5. In python, 6|2 will give which of the following as output?
   A) 2 B) 4
   C) 0 D) 6
   Answer: D
6. What does the finally keyword denotes in python?
   A) It is used to mark the end of the code
   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.
   C) the finally block will be executed no matter if the try block raises an error or not.
   D) None of the above
   Answer: B
7. What does raise keyword is used for in python?
   A) It is used to raise an exception. B) It is used to define lambda function
   C) it's not a keyword in python. D) None of the above
   Answer: A
8. Which of the following is a common use case of yield keyword in python?
   A) in defining an iterator B) while defining a lambda function
   C) in defining a generator D) in for loop.
   Answer: C

Q9 and Q10 have multiple correct answers. Choose all the correct options to answer your question.

9. Which of the following are the valid variable names?
   A) _abc B) 1abc
   C) abc2 D) None of the above
   Answer: A & C
10. Which of the following are the keywords in python?
    A) yield B) raise
    C) look-in D) all of the above
    Answer: A&B

Q11 to Q15 are programming questions. Answer them in Jupyter Notebook.

11. Write a python program to find the factorial of a number.
12. Write a python program to find whether a number is prime or composite.
13. Write a python program to check whether a given string is palindrome or not.
14. Write a Python program to get the third side of right-angled triangle from two given sides.
15. Write a python program to print the frequency of each of the characters present in a given string.