STATISTICS

1. True
2. a) Central Limit Theorem.
3. b) Modeling bounded count data.
4. c) The square of a standard normal random variable follows what is called chi-squared distribution.
5. c) Poisson .
6. b) False .
7. b) Hypothesis.
8. a) 0 .
9. c) Outliers cannot conform to the regression relationship.
10. The normal distribution is a symmetric continuous probability distribution for real-valued random variables also called the Gaussian Distribution. When it comes to statistics, it is important as it seems like a bell curve. This means that when plotted on a graph, data that are closer to the mean appear more often than data that is further away from the mean. According to the central limit theorem, averages of many samples tend towards a normal distribution as sample size increases. Linear combinations of independent normal deviates are also normally distributed. It has found use in different disciplines such as natural and social sciences among others.
11. Handling missing data is crucial for accurate analysis. some common techniques:
    Deletion:
    istwise Deletion: Remove entire records with any missing values.
    Pairwise Deletion: Analyze data using available pairs of variables.
    Imputation:
    Mean, Median, or Mode Imputation: Replace missing values with the mean, median, or mode of the variable.
    Regression Imputation: Predict missing values using regression models.
    K-Nearest Neighbors (KNN) Imputation: Use nearby data points to estimate missing values.
    Multiple Imputation: Generate multiple datasets with imputed values and combine results. Remember that the choice depends on the type of missingness (MCAR, MAR,

or MNAR) and the specific context .

12. A/B testing, also known as split testing, is a marketing experiment where you split your audience to test variations on a campaign and determine which performs better. For instance, you can show version A of a webpage, email, or design to one group and version B to another. It helps compare two versions of content (like subject lines, web pages, or app designs) to see which is more successful .

13. Of course! Mean imputation stands for a simple method in solving the problem of the missed data. The concept is that the missing values will be replaced by the average of the remaining observations. It maintains a sample size, but there is a possibility that it could have a bias and can underestimate the variability. Look for better results with more efficient ways of imputation.

14. Nowadays, one of the statistical methods used for modeling the relationship between a dependent variable and one or more independent variables is called linear regression. It is the symbolic representation of the line, which serves as the dependent variables' response to the independent variable's change. This is used for the innovation of prediction of data, thus various sectors have been implementing it.

15. The two main divisions of statistics are:
    Descriptive Statistics: This branch deals with the organization, summarization, and presentation of data. It enables us to understand patterns and features in a dataset. Examples include calculation of averages, e.g. mean; measures of variability; creating charts .
    Inferential Statistics: This is where information obtained from the sample is used to infer or predict a larger population. Inferential statistics encompasses techniques such as hypothesis testing, regression analysis, and confidence intervals.
    Remember, descriptive statistics summarizes data while inferential statistics uses data to make broader conclusions.

PYTHON WORKSHEET

1. C)%

2. A)0.666

3. C)24

4. A)2

5. none

6. C) The finally block will be executed no matter if the try block raises an error or not.
7. A) It is used to raise an exception.
8. C)In defining a generator.
9. _abc ,abc2
10. Yield, raise.
11. #codes to find the factorial of a number.

```
num=input("Enter a number")

factorial = 1
if num < 0:

   print("Sorry, factorial does not exist for negative numbers")

elif num == 0:

   print("The factorial of 0 is 1")

else:

   for i in range(1, num + 1):

      factorial *= i


print(f"The factorial of {num} is {factorial}")
```

12.# Wheathe a number is prime or composite.

```
num = input("enter a number")

if num == 1:

   print(f"{num} is not a prime number")

elif num > 1:
```

```
    for i in range(2, num):
        if (num % i) == 0:
            print(f"{num} is not a prime number")
            print(f"{i} times {num // i} is {num}")
            break
    else:
        print(f"{num} is a prime number")
else:
    print(f"{num} is not a prime number")
```

13.#wheather the given string ia palendrome or not.

```
def isPalindrome(str):
    for i in range(0, int(len(str) / 2)):
        if str[i] != str[len(str) - i - 1]:
            return False
    return True


s = "malayalam"
if isPalindrome(s):
    print("Yes")
else:
    print("No")
```

14. #to get the third side of a right angle triangle.

```
import math


a = float(input("Enter base: "))
```

```
b = float(input("Enter height: "))
x = float(input("Enter angle (in degrees): "))


c = math.sqrt(a**2 + b**2 - 2*a*b*math.cos(math.radians(x)))
print("Hypotenuse =", c)
```

15.# charactor frequency.

```
def character_frequency(string):
    freq_dict = {}
    for char in string:
        if char.isalpha():  # Consider only alphabetic characters
            freq_dict[char] = freq_dict.get(char, 0) + 1
    for char, count in freq_dict.items():
        print(f"'{char}': {count}")
```

MACHINE LEARNING:

1. A) Least Square Error
2. A) Linear regression is sensitive to outliers
3. B) Negative
4. B) Correlation
5. C) Low bias and high variance
6. B) Predictive model
7. D) Regularization
8. D) SMOTE

9. A) TPR and FPR

10. B) False

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

12. A) We don't have to choose the learning rate. AND D) It does not make use of dependent variable.

13. **Regularization** is a technique that would avoid overfitting in the model and improve generalization. This is done by adding a penalty term to the loss function during training time. There exist two types of standard regularization methods:
    L1 Regularization (Lasso):
    Loss function adds absolute values of coefficients.
    This will promote sparsity by driving some coefficients to exactly zero. Useful in feature selection.
    L2 Regularization (Ridge):
    Loss function adds squared values of coefficients.
    Penalizes large coefficients, promoting smoother models.Helps prevent multicollinearity.

14. Here are the commonly used algorithms for regularization:
    L2 Regularization (Ridge Regression):
    Ridge regression is an extension of linear regression in which an L2 penalty term is added to the cost function.It keeps the magnitude of the model weights (coefficients) as small as possible.Useful for avoiding model complexity and preventing overfitting12.
    L1 Regularization (Lasso Regression):
    Lasso Regression adds an L1 penalty term to the cost function.This approach encourages sparsity by driving some coefficients to an exact value of zero. Very useful for feature selection and significantly reduces model complexity .
    Elastic Net:
    Elastic Net is a mix of both the L1 and L2 regularization techniques. It balances the benefits brought forth by both Lasso and Ridge regression. Very useful when dealing with multicollinearity and high numbers of features.

15. In linear regression, the error term means deviation of the value of the observed dependent variable from that predicted by the regression equation—residual.