Q 1: What is the difference between descriptive statistics and inferential statistics? Explain with examples.

Ans:- Descriptive statistics summarize and describe the basic features of a dataset, such as mean, median, and standard deviation. Inferential statistics use sample data to make conclusions or predictions about a larger population, such as hypothesis testing and confidence intervals.

Descriptive Statistics:

- Summarize and describe the basic features of a dataset.
- Examples:
  - Calculating the average score of a class of students.
  - Finding the median income of a group of employees.
  - Determining the standard deviation of stock prices.

Inferential Statistics:

- Use sample data to make conclusions or predictions about a larger population.
- Examples: - Using a sample of voters to predict the outcome of an election.
  - Estimating the average lifespan of a product based on a sample of test results.
  - Conducting a hypothesis test to determine if there's a significant difference between two groups.

Q 2: What is sampling in statistics? Explain the differences between random and stratified sampling.

Ans:- Sampling is a statistical technique used to select a subset of individuals or data from a larger population, with the goal of making inferences about the characteristics of the population.

Types of Sampling are two type-

Random Sampling:-

Random sampling is a method where every member of the population has an equal chance of being selected. This method aims to minimize bias by ensuring that the sample is representative of the population.

Stratified Sampling:-

Stratified sampling involves dividing the population into distinct subgroups or strata, and then sampling from each stratum. This method ensures that each subgroup is represented in the sample, which can be particularly useful when there are significant differences between subgroups.

Differences Between Random and Stratified Sampling-

- Random Sampling: Does not guarantee representation of all subgroups within the population. It relies on chance to include diverse members.
- Stratified Sampling: Ensures that all identified subgroups are represented in the sample, reducing the risk of bias and increasing the precision of estimates for subgroups.

Q.3 Define mean, median, and mode. Explain why these measures of central tendency are important.

Ans:- Definition of Mean:-

The mean, also known as the arithmetic mean, is the average value of a set of numbers. It is calculated by summing all the values and dividing by the number of values.

Definition of Median:-The median is the middle value of a dataset when the numbers are arranged in ascending or descending order. If there is an even number of observations, the median is the average of the two middle numbers.
Definition of Mode:-
The mode is the value that appears most frequently in a dataset. A dataset may have one mode (unimodal), more than one mode (bimodal or multimodal), or no mode at all if all values are unique.

Importance of Measures of Central Tendency-
These measures are important because they provide a single value that can represent the entire dataset, giving us an idea of the "typical" value. They are crucial for:-
- Summarizing data:- They help in understanding the central position of the data.
- Comparing datasets:- Measures of central tendency allow for comparison between different groups or datasets.
- Decision-making:- They are used in various fields like economics, social sciences, and business for decision-making purposes.
Q.4 Explain skewness and kurtosis. What does a positive skew imply about the data?
Ans:- Definition of Skewness-
Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. In simpler terms, it measures how much the distribution deviates from being symmetrical.
Definition of Kurtosis-
Kurtosis is a measure of the "tailedness" or "peakedness" of the probability distribution. It indicates how outlier-prone a distribution is.
Types of Skewness-
- Positive Skew (Right Skew):- The distribution has a longer tail on the right side. This means most of the data points are concentrated on the left side of the distribution.
- Negative Skew (Left Skew):- The distribution has a longer tail on the left side. This means most of the data points are concentrated on the right side of the distribution.
- Zero Skew:- The distribution is symmetrical, like the normal distribution.
Types of Kurtosis:-
- Leptokurtic (High Kurtosis): Distributions with heavy tails or more outliers than the normal distribution.
- Platykurtic (Low Kurtosis): Distributions with light tails or fewer outliers than the normal distribution.
- Mesokurtic: Distributions with kurtosis similar to the normal distribution.
Importance of Skewness and Kurtosis
Understanding skewness and kurtosis is crucial for:
- Statistical Analysis: Many statistical tests assume normality. Skewness and kurtosis help determine if these assumptions are met.
- Data Interpretation: Knowing the shape of the distribution helps in understanding the data's behavior and potential outliers.
Importance of Skewness and Kurtosis:-

Understanding skewness and kurtosis is crucial for:
- Statistical Analysis: Many statistical tests assume normality. Skewness and kurtosis help determine if these assumptions are met.
- Data Interpretation: Knowing the shape of the distribution helps in understanding the data's behavior and potential outliers.

Q. Implement a Python program to compute the mean, median, and mode of a given list of numbers.
numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]

```
Import statistics

def calculate_mean_median_mode():
    numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]

    mean_value = statistics.mean(numbers)
    median_value = statistics.median(numbers)
    mode_value = statistics.multimode(numbers)

    print(f"Mean: {mean_value}")
    print(f"Median: {median_value}")
    print(f"Mode: {mode_value}")

calculate_mean_median_mode()
```

out put :-
Mean: 19.6
Median: 19
Mode: [12, 19, 24]

Q.6 Compute the covariance and correlation coefficient between the following two datasets provided as lists in Python:
list_x = [10, 20, 30, 40, 50]
list_y = [15, 25, 35, 45, 60]

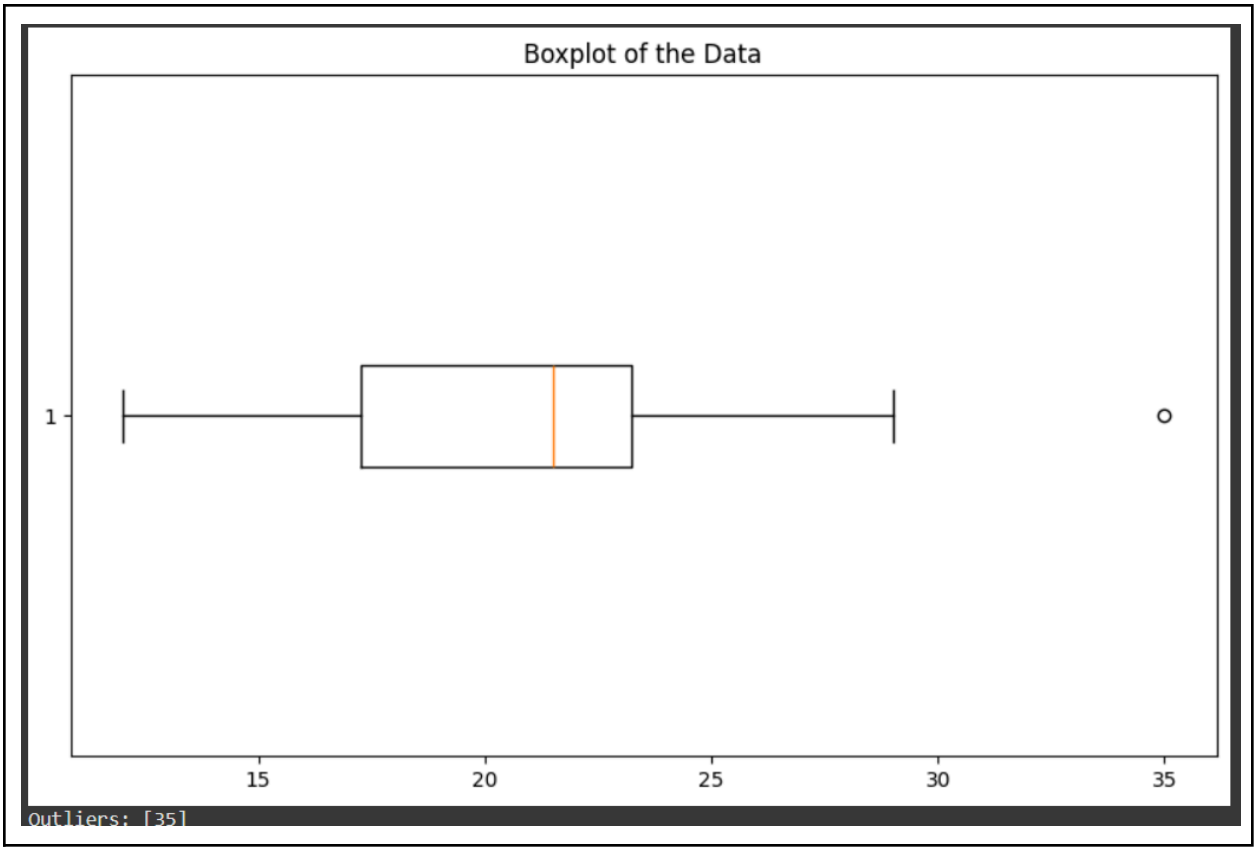
```python
def calculate_covariance_correlation():
    list_x = np.array([10, 20, 30, 40, 50])
    list_y = np.array([15, 25, 35, 45, 60])

    covariance_matrix = np.cov(list_x, list_y)
    covariance = covariance_matrix[0, 1]

    correlation_matrix = np.corrcoef(list_x, list_y)
    correlation_coefficient = correlation_matrix[0, 1]
```

```
    print(f"Covariance: {covariance}")
    print(f"Correlation Coefficient: {correlation_coefficient}")

calculate_covariance_correlation()
```

out put:-Covariance: 275.0
Correlation Coefficient: 0.995893206467704

Q.7 Write a Python script to draw a boxplot for the following numeric list and identify its outliers. Explain the result:
data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]

```
import matplotlib.pyplot as plt
import numpy as np

def plot_boxplot_identify_outliers():
    data = np.array([12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24,
26, 29, 35])

    Q1 = np.percentile(data, 25)
    Q3 = np.percentile(data, 75)
    IQR = Q3 - Q1

    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers = data[(data < lower_bound) | (data > upper_bound)]

    plt.figure(figsize=(10, 6))
    plt.boxplot(data, vert=False)
    plt.title('Boxplot of the Data')
    plt.show()

    print("Outliers:", outliers)

plot_boxplot_identify_outliers()
```



Q.8 You are working as a data analyst in an e-commerce company. The marketing team wants to know if there is a relationship between advertising spend and daily sales.

● Explain how you would use covariance and correlation to explore this relationship.

● Write Python code to compute the correlation between the two lists:
advertising_spend = [200, 250, 300, 400, 500]
daily_sales = [2200, 2450, 2750, 3200, 4000]

```
import numpy as np

# Given data
advertising_spend = np.array([200, 250, 300, 400, 500])
daily_sales = np.array([2200, 2450, 2750, 3200, 4000])

# Calculate covariance
covariance = np.cov(advertising_spend, daily_sales)[0, 1]

# Calculate correlation coefficient
correlation_coefficient = np.corrcoef(advertising_spend, daily_sales)[0, 1]

print(f"Covariance: {covariance}")
```

print(f"Correlation Coefficient: {correlation_coefficient}")
Out put:- Covariance: 84875.0
Correlation Coefficient: 0.9935824101653329
Interpretation:-

- A correlation coefficient close to 1 indicates a strong positive relationship between advertising spend and daily sales.
- A correlation coefficient close to -1 indicates a strong negative relationship.
- A correlation coefficient around 0 indicates no significant relationship

Q.9 Your team has collected customer satisfaction survey data on a scale of 1-10 and wants to understand its distribution before launching a new product.
● Explain which summary statistics and visualizations (e.g. mean, standard deviation, histogram) you'd use.
● Write Python code to create a histogram using Matplotlib for the survey data: survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]

```
import matplotlib.pyplot as plt
import numpy as np

# Given survey data
survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]

# Calculate summary statistics
mean_score = np.mean(survey_scores)
median_score = np.median(survey_scores)
std_dev = np.std(survey_scores)

# Print summary statistics
print(f"Mean: {mean_score}")
print(f"Median: {median_score}")
print(f"Standard Deviation: {std_dev}")

# Create histogram
plt.figure(figsize=(8, 6))
plt.hist(survey_scores, bins=6, edgecolor='black', align='left',
rwidth=0.8)
plt.xlabel('Satisfaction Score')
plt.ylabel('Frequency')
```

```
plt.title('Histogram of Customer Satisfaction Scores')
plt.xticks(range(4, 11))  # Set x-axis ticks to match the score range
plt.show()
OUT PUT :- Mean: 7.333333333333333
Median: 7.0
Standard Deviation: 1.577621275493231
```