Statistics Basics| Assignment

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

Answer

Descriptive Statistics summarize and organize data using numbers, tables, or graphs.
Example: Average marks of students in a class.

Inferential Statistics make predictions or inferences about a population based on a sample.
Example: Predicting election results from a survey of 1,000 people.

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

Sampling is the process of selecting a subset of individuals from a population to estimate characteristics of the whole group.

- **Random Sampling**: Every individual has an equal chance of selection.
  Example: Drawing 10 students randomly from a list of 100.
- **Stratified Sampling**: The population is divided into subgroups (strata), and samples are drawn from each.
  Example: Selecting students by year (1st, 2nd, etc.) to ensure representation from each group.

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

- **Mean**: Arithmetic average
- **Median**: Middle value in sorted data
- **Mode**: Most frequent value

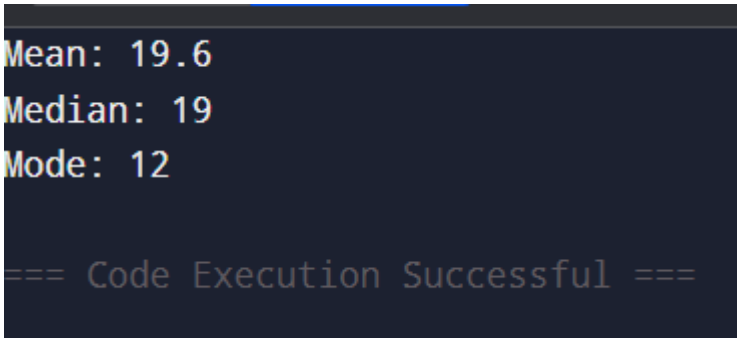
These measures help summarize data with a central value, aiding comparison and interpretation.

Question 4: Explain skewness and kurtosis. What does a positive skew imply about the data?

- **Skewness** measures asymmetry in data.
  - **Positive skew**: Tail on the right, mean > median (e.g., income data)
- **Kurtosis** measures "tailedness."
  - High kurtosis: More outliers;
  - Low kurtosis: Fewer outliers

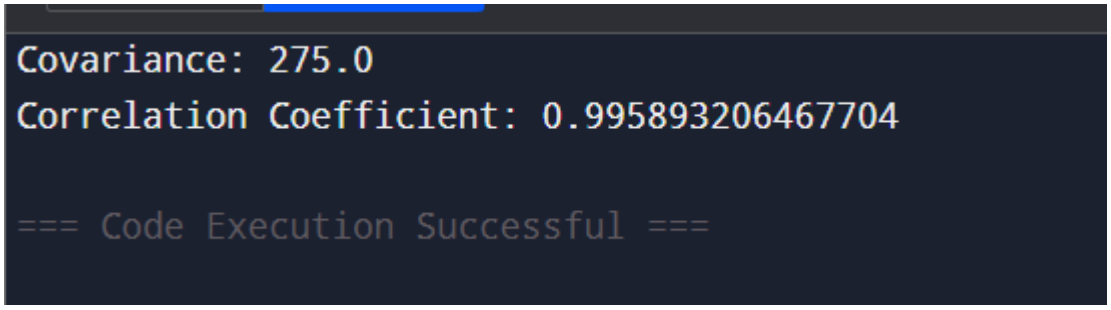
Question 5: Implement a Python program to compute the mean, median, and mode of a given list of numbers. numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28] (Include your Python code and output in the code box be

```
import statistics

numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]

mean_val = statistics.mean(numbers)
median_val = statistics.median(numbers)
mode_val = statistics.mode(numbers)

print("Mean:", mean_val)
print("Median:", median_val)
print("Mode:", mode_val)
```



Question 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] (Include your Python code and output in the code box below.)

```
import numpy as np


list_x = [10, 20, 30, 40, 50]

list_y = [15, 25, 35, 45, 60]


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

correlation = np.corrcoef(list_x, list_y)[0][1]


print("Covariance:", covariance)

print("Correlation Coefficient:", correlation)
```



Question 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] (Include your Python code and output in the code box be

```
import matplotlib.pyplot as plt


data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]


plt.boxplot(data)

plt.title("Boxplot with Outliers")
```

plt.show()

Question 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] (Include your Python code and output in the code box below.)

advertising_spend = [200, 250, 300, 400, 500]

daily_sales = [2200, 2450, 2750, 3200, 4000]

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

corr = np.corrcoef(advertising_spend, daily_sales)[0][1]

print("Covariance:", cov)

print("Correlation Coefficient:", corr)

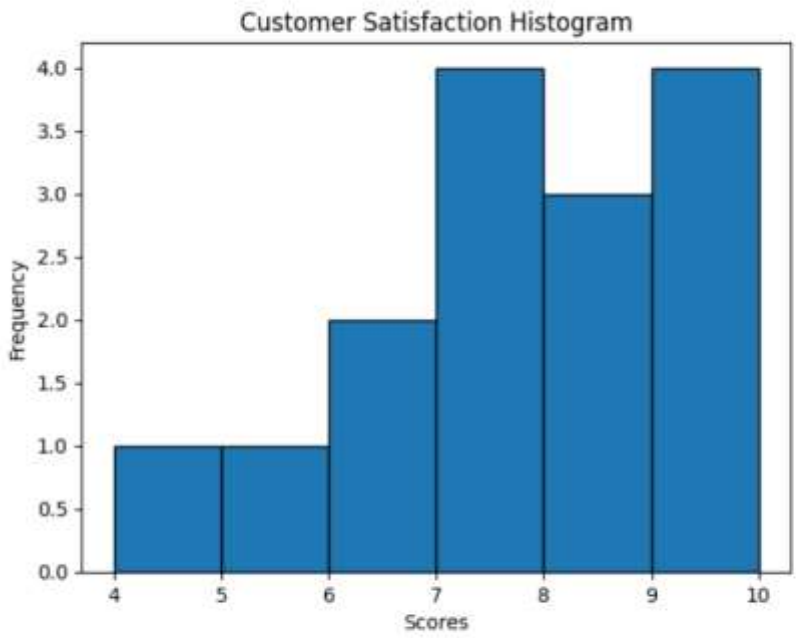
```
 from typing import NewType
+import numpy as np # Import numpy
+
 advertising_spend = [200, 250, 300, 400, 500]
 daily_sales = [2200, 2450, 2750, 3200, 4000]

 cov = np.cov(advertising_spend, daily_sales)[0][1]
 corr = np.corrcoef(advertising_spend, daily_sales)[0][1]

 print("Covariance:", cov)
 print("Correlation Coefficient:", corr)
-
```

```
Covariance: 84875.0
Correlation Coefficient: 0.9935824101653329
```

Question 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] (Include your Python code and output in the code box below.)

```
import matplotlib.pyplot as plt

import statistics


survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]


# Histogram

plt.hist(survey_scores, bins=6, edgecolor='black')

plt.title("Customer Satisfaction Histogram")

plt.xlabel("Scores")

plt.ylabel("Frequency")

plt.show()


# Summary

print("Mean:", statistics.mean(survey_scores))

print("Standard Deviation:", statistics.stdev(survey_scores))
```

[+]




```
Mean: 7.333333333333333
Standard Deviation: 1.632993161855452
```