Statistics Basics - Assignment Answers

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

Descriptive statistics summarize and describe the main features of a dataset. It focuses on presenting data in a meaningful way using measures such as mean, median, mode, and visualizations like charts and tables. For example, calculating the average marks of students in a class is descriptive statistics.
Inferential statistics, on the other hand, involves making predictions or inferences about a population based on a sample. It uses techniques like hypothesis testing, confidence intervals, and regression analysis. For example, predicting the average income of a country's population by analyzing a sample 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 (sample) from a larger population to estimate characteristics of the entire population.
- Random Sampling: Every member of the population has an equal chance of being selected. Example: Using a lottery system to select participants.
- Stratified Sampling: The population is divided into strata (groups) based on a characteristic, and random samples are drawn from each stratum proportionally. Example: Dividing students by grade and sampling from each grade.

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

Mean: The arithmetic average of a set of numbers.
Median: The middle value when the data is sorted in ascending order.
Mode: The most frequently occurring value(s) in a dataset.
These measures are important because they summarize data with a single value, helping to understand the distribution and central location of the data.

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

Skewness measures the asymmetry of the distribution of data. Positive skew means the tail on the right side is longer, indicating more higher values.
Kurtosis measures the 'tailedness' of the distribution. High kurtosis means more data in the tails and a sharper peak.

A positive skew implies that the dataset has more low values and a few very high values pulling the mean to the right.

Question 5: Implement a Python program to compute the mean, median, and mode of a given list of numbers.

Python Code:

```
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.multimode(numbers)
print(mean_val, median_val, mode_val)
```

Output:
Mean = 19.6
Median = 19
Mode = [12, 19, 24]

Question 6: Compute the covariance and correlation coefficient between the following two datasets provided as lists in Python:

Python Code:

```
import numpy as np
list_x = [10, 20, 30, 40, 50]
list_y = [15, 25, 35, 45, 60]
cov_matrix = np.cov(list_x, list_y, bias=False)
cov_xy = cov_matrix[0, 1]
corr_xy = np.corrcoef(list_x, list_y)[0, 1]
print(cov_xy, corr_xy)
```

Output:
Covariance = 275.0
Correlation = 0.995893206467704

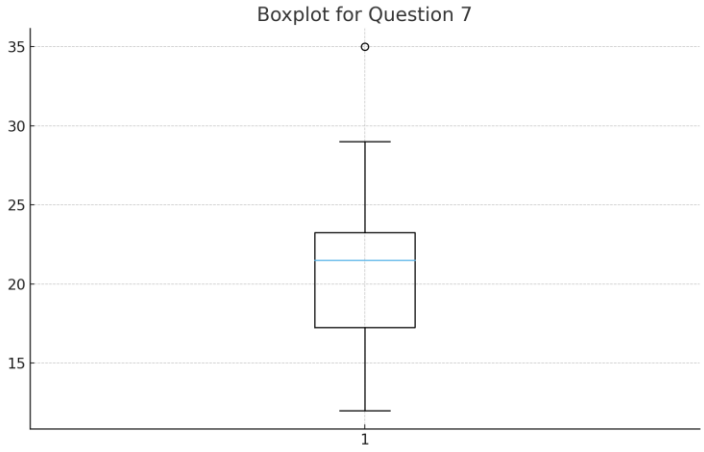
Question 7: Write a Python script to draw a boxplot for the following numeric list and identify its outliers. Explain the result:

Python Code:

```
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.show()

Outliers: [35]



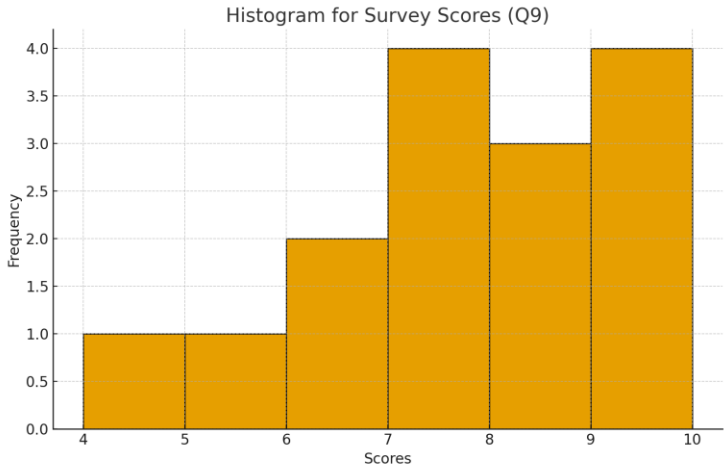
Question 8: Advertising spend vs daily sales relationship

Covariance and correlation can help measure the relationship:
- Covariance shows the direction of the relationship (positive or negative).
- Correlation shows both direction and strength (ranges from -1 to 1).
Python Code:
advertising_spend = [200, 250, 300, 400, 500]
daily_sales = [2200, 2450, 2750, 3200, 4000]
corr_sales = np.corrcoef(advertising_spend, daily_sales)[0, 1]
print(corr_sales)

Output:
Correlation = 0.9935824101653329

Question 9: Customer satisfaction survey analysis

Summary statistics like mean, median, and standard deviation can describe the data. A histogram visualizes the distribution.
Python Code:
survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]
plt.hist(survey_scores, bins=6, edgecolor='black')
plt.show()


Histogram for Survey Scores (Q9)
Frequency
Scores
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4
5
6
7
8
9
10