Statistics Basics| Assignment

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

Answer:
Descriptive statistics summarize and describe the features of a dataset (e.g., mean, median, mode, standard deviation, histograms). Example: Reporting the average test score (mean = 78) and the distribution of scores for a class.

Inferential statistics use sample data to make conclusions or predictions about a population, often with uncertainty quantified (confidence intervals, hypothesis tests). Example: Using a sample of 100 voters to estimate the proportion of voters in a city who will vote for a candidate and providing a margin of error.

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

Answer:
Sampling is selecting a subset of individuals or observations from a population to estimate characteristics of the whole population.

Random sampling: Every member of the population has an equal chance of being chosen. It's simple and helps avoid selection bias. Example: drawing 100 names from a hat.

Stratified sampling: The population is divided into strata (subgroups) based on a characteristic (e.g., age groups), and samples are drawn from each stratum, often proportionally. This ensures representation across important subgroups and can increase precision. Example: sampling students separately from each grade level.

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

Answer:
Mean: The arithmetic average of values (sum divided by count). Useful for symmetric distributions and when all observations are meaningful.
Median: The middle value when data are ordered (or average of two middle values). Robust to outliers; gives a better center for skewed data.
Mode: The most frequently occurring value(s). Useful for categorical data or to find common/typical values.

These measures summarize a dataset with a single value that represents its center; choosing which to use depends on distribution shape and presence of outliers.

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

Answer:
Skewness measures the asymmetry of a distribution. Positive skew (right skew) means the right tail is longer: most observations lie to the left and a few large values pull the tail to the right. Median < Mean typically for positively skewed data.

Kurtosis measures the 'tailedness' or concentration of values in the tails and peak. High kurtosis (leptokurtic) indicates heavy tails and a sharp peak; low kurtosis (platykurtic) indicates light tails and a flatter peak. Note: many statistical packages report excess kurtosis (kurtosis - 3) so normal distribution has excess kurtosis 0.

Positive skew implies there are some large outliers on the right side; the mean is pulled to the right of the median.

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]

Answer:
Python code used (standard library):

```
import statistics as stats
numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]
mean_val = stats.mean(numbers)
median_val = stats.median(numbers)
mode_val = stats.multimode(numbers)
```

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

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

Answer:
Python code used:

```
list_x = [10, 20, 30, 40, 50]
list_y = [15, 25, 35, 45, 60]
def covariance(xs, ys): ...  # sample covariance
def pearson_corr(xs, ys): ...
```

Output:
Sample covariance = 275.0
Pearson correlation coefficient = 0.9958932064677041

Interpretation: Correlation is close to 1, indicating a strong positive linear relationship. The last y value (60) is slightly larger relative to x which affects covariance.

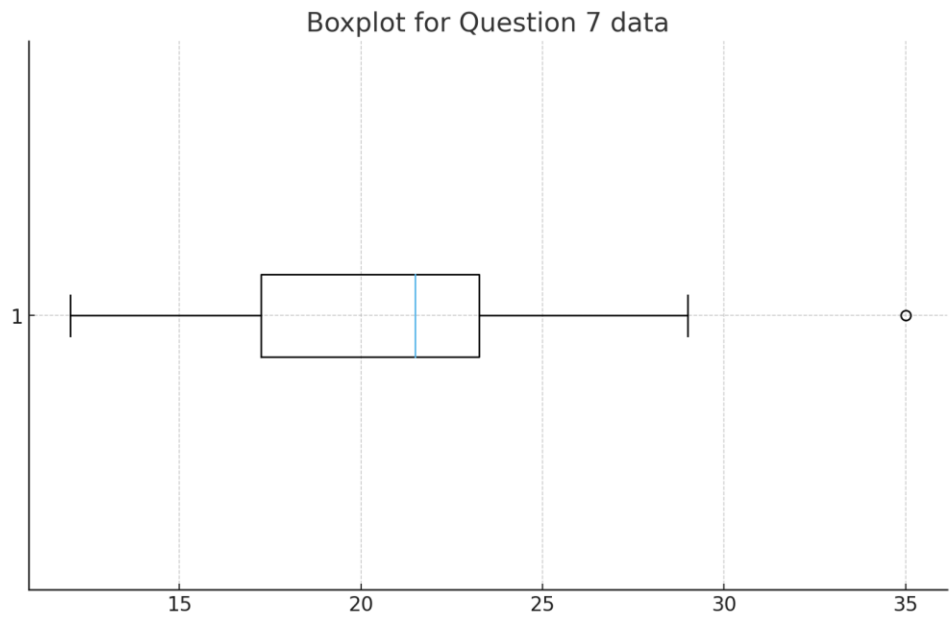
Question 7: Write a Python script to draw a boxplot for the following numeric list and identify its outliers. Explain the result:
data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]

Answer:
Python code used (matplotlib):

```
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, vert=False)
plt.show()
```

Computed summary:
Q1 = 17.25
Q3 = 23.25
IQR = 6.0
Lower bound = 8.25
Upper bound = 32.25
Outliers = [35]

Explanation: The boxplot shows the central 50% between Q1 and Q3. Values outside the whiskers are considered outliers by the 1.5*IQR rule. Here, 35 is an outlier (well above the upper bound). 12 is near the lower side but not below the computed lower bound, so it is not flagged. The presence of 35 indicates a high value pulling the tail to the right; distribution is somewhat right-skewed by the largest value.



<u>Question 8:</u> 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]
```

<u>Answer:</u>
Explanation:
Covariance indicates the direction of the linear relationship: positive covariance means when advertising spend increases, sales tend to increase. However, covariance is scale-dependent so its magnitude isn't easy to interpret across datasets.
Correlation (Pearson) standardizes covariance and gives a value between -1 and 1: values near +1 indicate a strong positive linear relationship, near 0 indicate little linear relationship, and near -1 indicate a strong negative linear relationship.

Python code used:

```
advertising_spend = [200, 250, 300, 400, 500]
daily_sales = [2200, 2450, 2750, 3200, 4000]
cov = covariance(advertising_spend, daily_sales)
corr = pearson_corr(advertising_spend, daily_sales)
```

Output:

Sample covariance = 84875.0
Pearson correlation coefficient = 0.9935824101653329

Interpretation: The positive covariance and correlation (close to 1) indicate a strong positive linear relationship between advertising spend and daily sales in this small sample.

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]

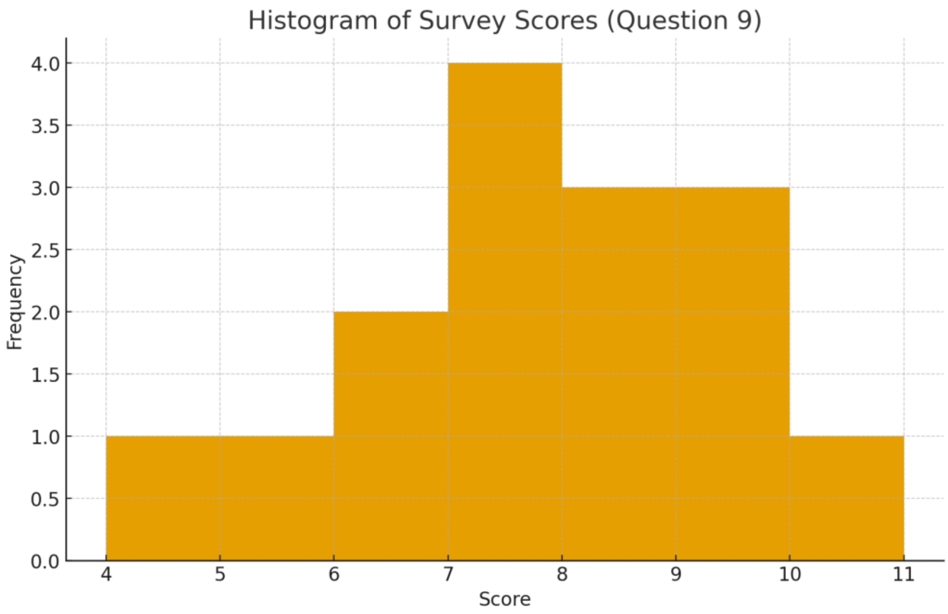
Answer:
Summary statistics and visualizations to use:
- Mean and median to understand central tendency.
- Standard deviation (sample or population) to measure spread.
- Histogram to inspect distribution shape (skewness, modality).
- Boxplot to check for outliers.
- Frequency table for ordinal scales.

Computed values:
Mean = 7.333333333333333
Median = 7
Sample standard deviation = 1.632993161855452
Population standard deviation = 1.5776212754932308

Python code used to build the histogram (matplotlib):

```
import matplotlib.pyplot as plt
survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]
plt.hist(survey_scores, bins=range(4,12))
plt.show()
```

Interpretation: The histogram shows most scores between 6 and 9, indicating generally positive satisfaction with a central tendency around 7-8.



Histogram of Survey Scores (Question 9)
Frequency
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
11
Score