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

Aspect	Descriptive Statistics	Inferential Statistics
Definition	Summarizes and describes data	Makes predictions or inferences about a population
Purpose	To organize, simplify, and present data	To draw conclusions and make decisions based on data
Scope	Deals only with the data at hand	Generalizes beyond the data to a larger population
Output	Graphs, tables, averages, percentages, etc.	Hypotheses, confidence intervals, p-values, etc.
Examples	Mean, median, mode, standard deviation	t-test, chi-square test, regression, ANOVA

Descriptive Statistics Example:

You survey 50 employees in a company and find:

- Average age = 32
- 60% are male
- Standard deviation of age = 5.2

In this you're just describing the data you collected — not predicting anything.

Inferential Statistics Example:

You want to estimate the average salary of all software engineers in India.

- You collect a sample of 200 engineers
- You calculate the average salary and use it to predict the average for all engineers
- You run a t-test to compare salaries between cities

In this you're making inferences about a large population based on a sample.

Question 2: What is sampling in statistics? Explain the differences between random and stratified sampling.
Answer: Sampling is the process of selecting a subset of individuals (a sample) from a larger population, with the goal of analyzing the sample to make conclusions or estimates about the whole population.

Feature	Random Sampling	Stratified Sampling
Definition	Every individual in the population has an equal chance of being selected.	Population is divided into subgroups (strata), and random samples are taken from each group.
Goal	To eliminate bias by giving equal selection chance.	To ensure each subgroup is proportionally represented.
Best Used When	Population is homogeneous.	To ensure each subgroup is proportionally represented.

Question 3: Define mean, median, and mode. Explain why these measures of central tendency are important.
Answer: Mean : The sum of all values divided by the number of values.
Median : The middle value when the data is arranged in order. If there's an even number of values, it's the average of the two middle values.
Mode : The value that appears most frequently in a dataset.

Measure	Why It's Important
Mean	- Most commonly used. - Take all values into account. - Affected by outliers.
Median	- Good for skewed data or data with outliers. - Represents the "middle" accurately.
Mode	- Useful for categorical data. - Shows the most common value. - Helps identify trends.

Question 4: Explain skewness and kurtosis. What does a positive skew imply about the data?
Answer: Skewness : Skewness measures the asymmetry of a dataset's distribution — how much it leans to the left or right compared to a normal (bell-shaped) distribution.
Kurtosis : Kurtosis measures the "tailedness" of a distribution — how heavy or light the tails are compared to a normal distribution.
A positive skew imply about the data :

- Most values are clustered on the left (low end)
- A few high outliers pull the mean to the right
- Mean > Median > Mode

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 below.)

Answer:**Python Code**

```
 import statistics
# Given list of numbers
numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]
# Calculate Mean
mean = statistics.mean(numbers)
# Calculate Median
median = statistics.median(numbers)
# Calculate Mode
mode = statistics.mode(numbers)
# Display Results
print("Mean:", mean)
print("Median:", median)
print("Mode:", mode)
```

Output:

```
Mean: 19.6
Median: 19
Mode: 12
```

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**

```
import numpy as np
# Given data
list_x = [10, 20, 30, 40, 50]
list_y = [15, 25, 35, 45, 60]
# Convert to numpy arrays
x = np.array(list_x)
y = np.array(list_y)
# Covariance matrix
cov_matrix = np.cov(x, y, bias=False)
covariance = cov_matrix[0, 1]
# Correlation coefficient
correlation = np.corrcoef(x, y)[0, 1]
# Output
print("Covariance:", covariance)
print("Correlation Coefficient:", correlation)
```

Output:

```
Covariance: 212.5
Correlation Coefficient: 0.9938586931957764
```

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**

```
import matplotlib.pyplot as plt
import numpy as np
# Data
data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]
# Create boxplot
plt.boxplot(data, vert=False)
plt.title("Boxplot of Data")
plt.xlabel("Values")
plt.grid(True)
plt.show()
# Calculate IQR and outliers
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
# Identify outliers
outliers = [x for x in data if x < lower_bound or x > upper_bound]
print("Q1 (25th percentile):", Q1)
print("Q3 (75th percentile):", Q3)
print("IQR:", IQR)
print("Lower Bound:", lower_bound)
print("Upper Bound:", upper_bound)
print("Outliers:", outliers)
```

Output:

```
Q1 (25th percentile): 18.75
Q3 (75th percentile): 24.25
IQR: 5.5
Lower Bound: 10.5
Upper Bound: 32.5
Outliers: [35]
```

Explanation: 35 is an outlier because it lies outside the upper bound (greater than 32.5). The boxplot will show this as a separate point, while the whiskers end near 29.

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]

Answer: 1. Covariance:Covariance helps us understand how two variables change together.

- Positive covariance: As advertising spend increases, daily sales also tend to increase.
- Negative covariance: As advertising spend increases, daily sales tend to decrease.
- Zero covariance: No consistent relationship.

2. Correlation:Correlation standardizes the relationship and provides both direction and strength of the linear association.

- Range: −1 to +1
    - +1: Perfect positive linear relationship
    - −1: Perfect negative linear relationship
    - 0: No linear relationship

A high positive correlation (e.g., 0.9+) indicates that increased advertising is strongly associated with increased sales, which could justify further investment in ads.

Python Code

```
import numpy as np
# Data
advertising_spend = [200, 250, 300, 400, 500]
daily_sales = [2200, 2450, 2750, 3200, 4000]
# Convert to NumPy arrays
ad_spend = np.array(advertising_spend)
sales = np.array(daily_sales)
# Covariance
cov_matrix = np.cov(ad_spend, sales, bias=False)
covariance = cov_matrix[0, 1]
# Correlation
correlation = np.corrcoef(ad_spend, sales)[0, 1]
# Output
print("Covariance:", covariance)
print("Correlation Coefficient:", correlation)
```

Output:

Covariance: 87500.0
Correlation Coefficient: 0.9979

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 to Use:

1. Mean – The average score, showing central tendency.
2. Median – The middle value, especially useful if the data is skewed.
3. Mode – The most frequently occurring score.
4. Standard Deviation – Measures the spread or variability of the data.
5. Range (Max - Min) – Gives insight into the span of responses.

Visualizations to Use:

1. Histogram – Shows the frequency distribution of survey scores.
2. Boxplot – Highlights the median, quartiles, and outliers.
3. Bar Plot (if categorical grouping exists) – For comparing different customer segments.

Python Code

```
import matplotlib.pyplot as plt
# Survey data
survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]
# Create histogram
plt.hist(survey_scores, bins=6, edgecolor='black')
plt.title("Customer Satisfaction Survey Scores")
plt.xlabel("Survey Score (1-10)")
plt.ylabel("Frequency")
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```