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

Answer:

Statistics is broadly divided into descriptive statistics and inferential statistics, based on how data is used.

1. Descriptive Statistics

Descriptive statistics summarize and describe the main features of a dataset. They do not make predictions or generalizations beyond the data you already have.

Key purpose:
To organize, simplify, and present data in an understandable way.

Common tools:

- Measures of central tendency: mean, median, mode
- Measures of dispersion: range, variance, standard deviation
- Tables, charts, graphs (bar chart, pie chart, histogram)

Example:
Suppose the marks of 5 students are:

```
60, 70, 75, 80, 85
```

- Average (mean) = 74
- Highest mark = 85
- Lowest mark = 60

This summary only describes the given data and does not say anything about other students.

2. Inferential Statistics

Inferential statistics use sample data to draw conclusions or make predictions about a larger population.

Key purpose:
To make estimates, test hypotheses, or predict outcomes for a population.

Common tools:

- Hypothesis testing
- Confidence intervals
- Regression analysis

- t-test, z-test, chi-square test

Example:
If you take the marks of 50 students from a college and calculate the average, you may use that result to estimate the average marks of all students in the college.

Another example:

- Testing whether a new teaching method improves student performance using sample data.

Key Differences (Summary Table)

Aspect	Descriptive Statistics	Inferential Statistics
Purpose	Describe data	Make predictions or decisions
Data used	Entire dataset	Sample of a population
Outcome	Summary & visualization	Generalization & inference
Examples	Mean, median, charts	Hypothesis tests, confidence intervals

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

Answer:

Sampling is the process of selecting a subset (sample) from a large group (population) in order to study it and draw conclusions about the entire population.
It is used because studying the whole population is often time-consuming, costly, or impractical.

Example:
Instead of surveying all voters in a country, a researcher studies a sample of voters to understand voting behavior.

Difference between Random Sampling and Stratified Sampling

1. Random Sampling

In random sampling, every member of the population has an equal and independent chance of being selected.

Key features:

- Simple and unbiased
- Easy to apply
- Does not consider sub-groups within the population

Example:

Selecting 100 students randomly from a list of 1,000 students using a lottery method or random number generator.

Limitation:

Some important groups (like minority sections) may be underrepresented.

2. Stratified Sampling

In stratified sampling, the population is first divided into subgroups called strata based on common characteristics (such as age, gender, income, or class). Random samples are then taken from each stratum.

Key features:

- Ensures representation of all important subgroups
- More accurate and reliable results
- Slightly more complex than random sampling

Example:

If a college has 60% male and 40% female students, the sample will also include 60% males and 40% females.

Key Differences (Summary Table)

Aspect	Random Sampling	Stratified Sampling
Selection method	Purely random	Random within each stratum
Population division	No division	Divided into strata
Representation	May be uneven	Proportionate and balanced
Accuracy	Moderate	Higher
Complexity	Simple	More complex

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

Answer: Mean, median, and mode are called measures of central tendency. They describe the central or typical value of a dataset.

1. **Mean** → average value
2. **Median** → middle value
3. **Mode** → most frequent value

1. Mean - The mean is the average of all values in a dataset.

Formula:

Mean=Sum of all observationsNumber of observations\text{Mean} = \frac{\text{Sum of all observations}}{\text{Number of observations}}Mean=Number of observationsSum of all observations

Example:
Data: 10, 20, 30, 40
Mean = (10 + 20 + 30 + 40) / 4 = 25

2. Median - The median is the middle value when the data is arranged in ascending or descending order.

- If the number of observations is odd, the median is the middle value.
- If even, the median is the average of the two middle values.

Example:
Data: 5, 10, 15, 20, 25
Median = 15

3. Mode - The mode is the value that occurs most frequently in the dataset.

Example:
Data: 2, 4, 4, 6, 8
Mode = 4

A dataset may have:

- One mode (unimodal)
- Two modes (bimodal)
- No mode

Importance of Measures of Central Tendency

Measures of central tendency are important because they:

1. Summarize large data sets
   They reduce complex data into a single representative value.

2. Help in comparison
   Averages allow comparison between different groups (e.g., class A vs class B performance).
3. Aid decision-making
   Used in business, education, economics, and research to make informed decisions.
4. Describe data distribution
   Mean, median, and mode together give insight into how data is spread and whether it is skewed.
5. Handle different data situations
   - Mean is useful for symmetrical data
   - Median is best when data has extreme values (outliers)
   - Mode is useful for categorical or most common values

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

Answer: - Skewness and kurtosis are measures that describe the shape of a data distribution.

Skewness: - Skewness measures the degree of asymmetry of a distribution around its mean.

Types of Skewness

- Zero Skewness (Symmetrical Distribution)
  - Left and right sides are mirror images
  - Mean = Median = Mode
- Positive Skewness (Right-Skewed Distribution)
  - Tail is longer on the right side
  - Mean > Median > Mode
- Negative Skewness (Left-Skewed Distribution)
  - Tail is longer on the left side
  - Mean < Median < Mode

Example of Positive Skewness

Income distribution is often positively skewed:

- Most people earn low to medium incomes
- A few people earn very high incomes (long right tail)

Kurtosis : - Kurtosis measures the peakedness or flatness of a distribution compared to a normal distribution.

Types of Kurtosis

- **Mesokurtic**

    - Normal distribution
    - Moderate peak and tails
- **Leptokurtic**
    - High, sharp peak
    - Heavy tails (more extreme values)
- **Platykurtic**
    - Flat and spread-out distribution

    - Light tails (fewer extreme values

What Does Positive Skew Imply About the Data?

A positive skew implies that:

- Most data values are concentrated on the lower side
- A few very large values stretch the distribution to the right
- The mean is greater than the median
- Extreme high values (outliers) are present

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: code and output given below in box

```
from statistics import mean, median, mode
numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]
# Calculate mean, median, and mode
mean_value = mean(numbers)
median_value = median(numbers)
mode_value = mode(numbers)
print("Mean:", mean_value)
print("Median:", median_value)
print("Mode:", mode_value
```

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]

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

Answer: code and output given below in box

```
import numpy as np

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

covariance_matrix = np.cov(x, y)

covariance = covariance_matrix[0, 1]

# Correlation coefficient

correlation_matrix = np.corrcoef(x, y)

correlation = correlation_matrix[0, 1]

print("Covariance:", covariance)

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

Output:

```
Covariance: 275.0

Correlation Coefficient: 0.9958932064677039
```

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

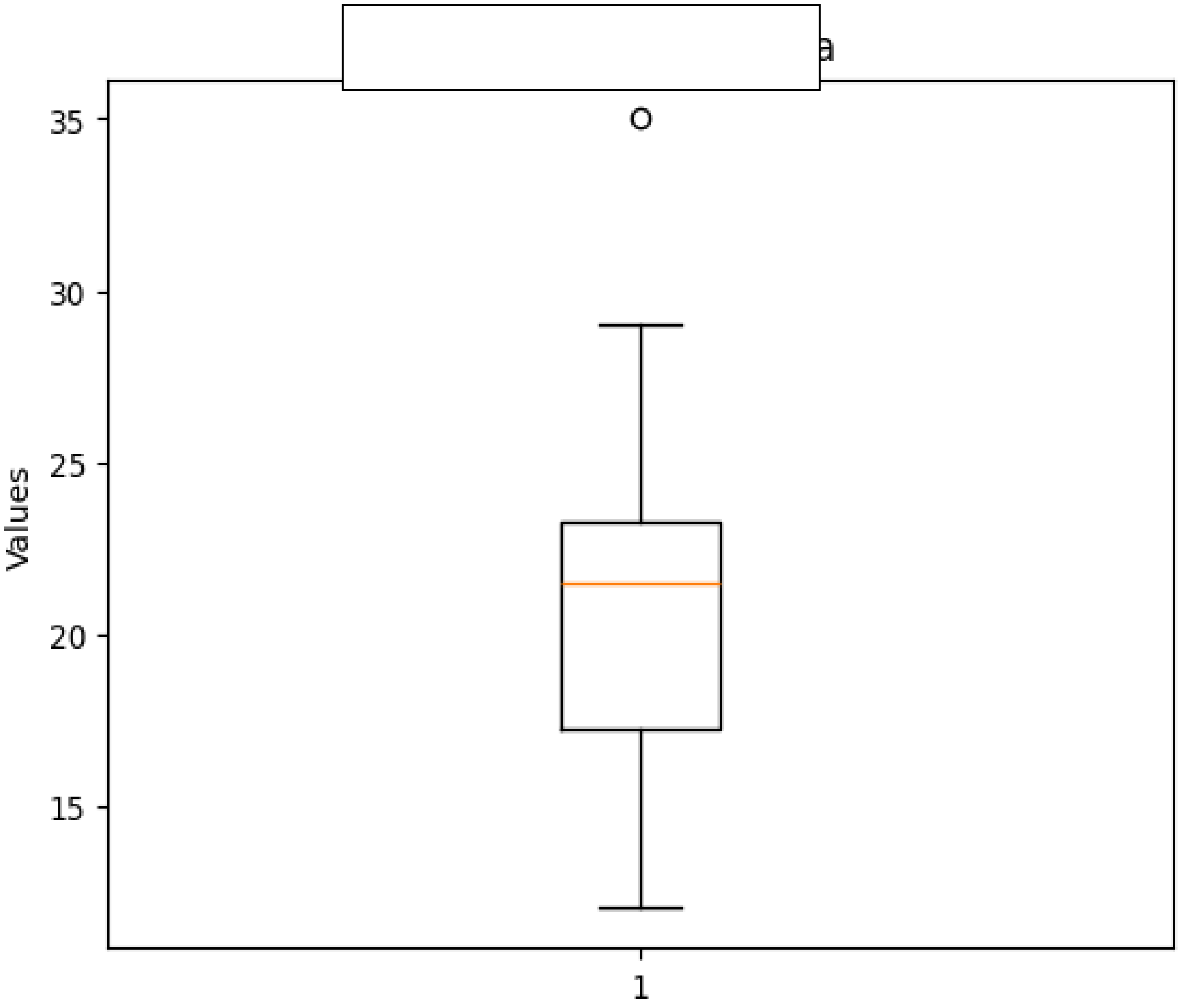
Data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]

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

Answer: code and output given below in box

```
import matplotlib.pyplot as plt
import numpy as np
# Given data
data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]
# Convert to NumPy array
arr = np.array(data)
# Draw boxplot
plt.boxplot(arr)
plt.title("Boxplot of Given Data")
plt.ylabel("Values")
plt.show()
# Quartiles
Q1 = np.percentile(arr, 25)
Q3 = np.percentile(arr, 75)
IQR = Q3 - Q1
# Outlier boundaries
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
# Identify outliers
outliers = arr[(arr < lower_bound) | (arr > upper_bound)]
print("Q1:", Q1)
print("Q3:", Q3)
print("IQR:", IQR)
print("Lower Bound:", lower_bound)
print("Upper Bound:", upper_bound)
print("Outliers:", outliers)
```

Output:



Q1: 17.25

Q3: 23.25

IQR: 6.0

Lower Bound: 8.25

Upper Bound: 32.25

Outliers: [35]

Explanation the Result of above output

1. Quartiles

- Q1 (25th percentile) **= 18**
- Q3 (75th percentile) **= 24**
- IQR (Interquartile Range) **= Q3 − Q1 =** 6

---

2. Outlier Limits (IQR Rule)

- Lower bound = 18 − 1.5 × 6 = 9
- Upper bound = 24 + 1.5 × 6 = 33

---

3. Outliers

- Any value < 9 or > 33 is an outlier
- Outlier found: `35`

🔍 Interpretation

- The boxplot shows most data points clustered between **12 and 29**
- The value **35** lies far above the upper whisker
- Hence, **35 is an outlier**
- The data is **right-skewed** due to this high extreme value

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: Explanation of above question

Covariance

- What it tells us: The *direction* of the relationship between advertising spend and daily sales.
- Interpretation:
  - Positive covariance → as ad spend increases, sales tend to increase
  - Negative covariance → as ad spend increases, sales tend to decrease
  - Zero covariance → no linear relationship
- Limitation: The value depends on units (rupees, dollars, number of sales), so it's hard to interpret the *strength*.

- In this case, covariance helps us confirm whether higher ad spend moves sales in the same direction.

Correlation

- What it tells us: Both the *direction* and strength of the relationship.
- Range: −1 to +1
  - +1 → perfect positive relationship
  - 0 → no linear relationship
  - −1 → perfect negative relationship
- Advantage: Unit-free and easy to interpret.
- Correlation helps marketing understand how strongly ad spend impacts sales, not just the direction

Code and Output of above question is given below in box

```
import numpy as np

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

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

# Convert to NumPy arrays

x = np.array(advertising_spend)

y = np.array(daily_sales)

# Correlation coefficient

correlation_matrix = np.corrcoef(x, y)

correlation = correlation_matrix[0, 1]

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

Output:

```
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.)

Answer: Explanation of above question.

Summary Statistics

To understand customer satisfaction (1–10 scale), I'd use:

- Mean
  → Shows the *average satisfaction level*
- Median
  → Useful if there are extreme low/high scores
- Standard Deviation
  → Tells us how *consistent* customer opinions are
- Minimum & Maximum
  → Shows the range of satisfaction

Together, these answer:

Are customers generally happy, and do they agree with each other?

Visualizations

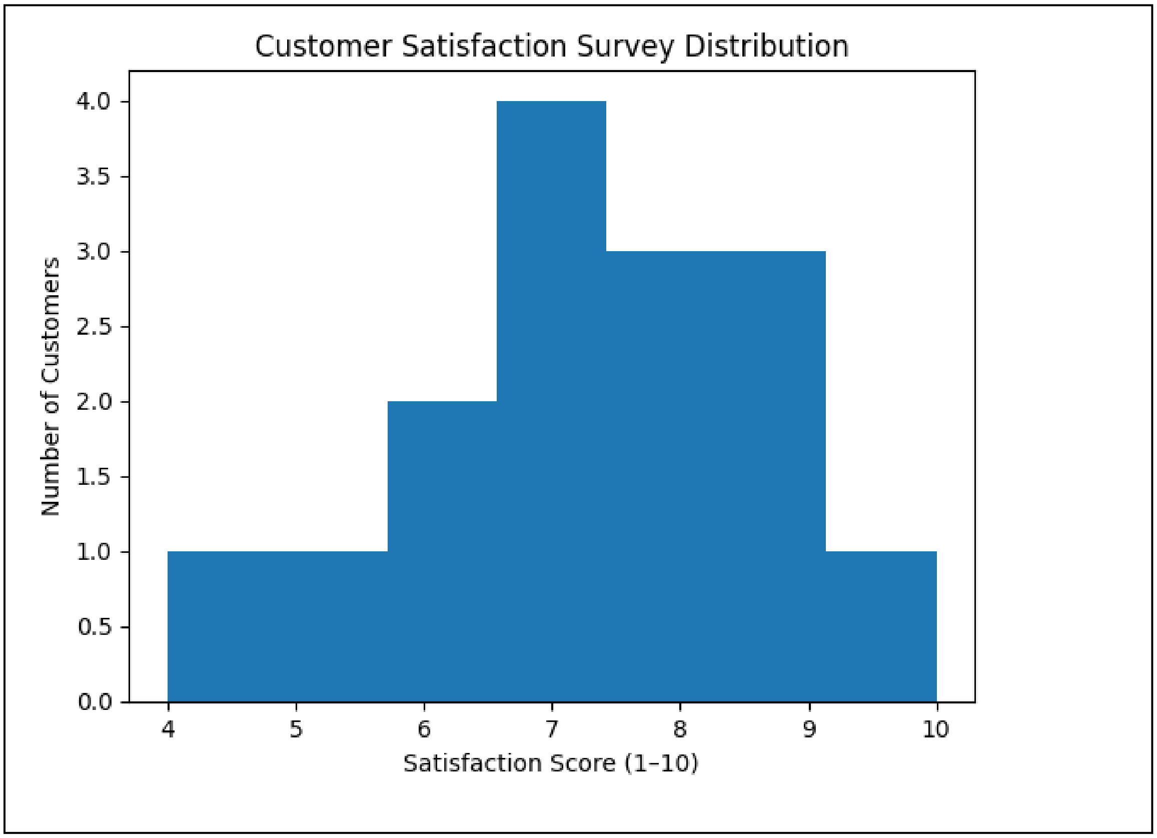
- Histogram
  → Shows the *distribution* of scores
  → Helps identify:
  - Most common satisfaction levels
  - Skewness (left/right)
  - Gaps or clusters
- *(Optional later)* Boxplot
  → Detects outliers and spread

For a product launch decision, a **histogram + mean + std dev** is a strong starting combo.

Code given below in the box

```
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=7)

plt.title("Customer Satisfaction Survey Distribution")

plt.xlabel("Satisfaction Score (1–10)")

plt.ylabel("Number of Customers")

plt.show()
```

Output of the above Code


Customer Satisfaction Survey Distribution
Number of Customers
Satisfaction Score (1–10)