Assignment Code: DS-AG-005

Statistics Basics| **Assignment**

Instructions: Carefully read each question. Use Google Docs, Microsoft Word, or a similar tool to create a document where you type out each question along with its answer. Save the document as a PDF, and then upload it to the LMS. Please do not zip or archive the files before uploading them. Each question carries 20 marks.

Total Marks: 200

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

Answer:

1. Descriptive Statistics

- **Definition: Summarizes and organizes collected data so we can understand it easily.**
- **Purpose: To describe what the data shows.**
- **Tools: Mean, Median, Mode, Standard Deviation, Graphs, Charts, Tables.**

Example:- Marks of 50 students in a class:

- **Average marks = 68**
- **Highest marks = 95**
- **Lowest marks = 32**

- **Standard deviation = 10**

***Here, we only describe the collected data, no prediction about a larger population.**

2. Inferential Statistics

- **Definition: Makes predictions or conclusions about a population based on a sample.**
- **Purpose: To generalize findings beyond the data we have.**
- **Tools: Hypothesis testing, Confidence intervals, Regression, Correlation.**

Example:- You survey 100 voters in a city of 10,000:

- **60% of sample prefer Party A → You estimate that roughly 60% of the entire city supports Party A.**

***Here, we infer population trends from a sample.**

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

Answer:

Sampling :- Selecting a small group (sample) from a large population to study and make conclusions.

- **Why? Studying the whole population is time-consuming, costly, or impossible.**

Example:
A school has 2000 students. Instead of checking marks of all students, you randomly select 200 students and study their performance.

1. Random Sampling

- **Every individual in the population has an equal chance of being selected.**

- **Simple and unbiased method.**

Example: Pick 200 students randomly from 2000 students' roll numbers.

2. Stratified Sampling

- **Population is divided into groups/strata (like gender, class, age).**

- **A proportional sample is selected from each group to ensure representation.**

Example:

- **2000 students → 1000 boys & 1000 girls.**

1

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

Answer:

These are statistical tools that indicate the center or average of a dataset.

1. Mean (Arithmetic Average) :- Sum of all observations divided by the number of observations.

- **Formula: Mean=Sum of observations/Number of observations**
- **Example: Marks = 10, 20, 30 → Mean = (10+20+30)/3 = 20**

2. Median :-Middle value of data when arranged in ascending or descending order.

- **Example: 10, 20, 30 → Median = 20**
  If data is even: 10, 20, 30, 40 → Median = (20+30)/2 = 25

3. Mode :- Most frequently occurring value in the dataset.

- **Example: 2, 4, 4, 6, 7 → Mode = 4**

Importance:-

1. **Summarizes data by representing it with a single value.**
2. **Shows the central point of the data distribution.**
3. **Helps in comparison between different datasets (e.g., average marks of classes).**
4. **Each measure has specific uses:**
   - **Mean: Useful for overall average (salary, marks).**
   - **Median: Useful when there are extreme values (income distribution).**
   - **Mode: Useful to find the most common value (fashion size, exam scores).**

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

Answer:

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

- **Types:**
  - **Positive skew (right skew): Tail stretches to the right. Most data values are clustered on the left, with few high values on the right.**
  - **Negative skew (left skew): Tail stretches to the left. Most data values are clustered on the right, with few low values on the left.**
- **Example:**
  - **Positive skew: Income of people in a city → most earn low/average, few earn very high.**
  - **Negative skew: Age at retirement → most retire around 60, few retire early.**
- **Implication of Positive Skew:**
  - **Mean > Median > Mode**

    - **Data is concentrated at lower values with some extremely high values pulling the tail to the right.**

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

- **Types:**
    - **Leptokurtic: High peak, heavy tails (more extreme values).**
    - **Platykurtic: Flat distribution, light tails (less extreme values).**
    - **Mesokurtic: Normal peak (similar to normal distribution).**
- **Example:**
    - **Leptokurtic: Test scores where most students score around average but some score extremely high/low.**
    - **Platykurtic: Uniform distribution like rolling a fair die.**

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:

Paste your code and output inside the box below:

2

```
#code:-

# Import required modules

from statistics import mean, median, mode


# Given list of numbers

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


# Compute mean, median, and mode

mean_value = mean(numbers)

median_value = median(numbers)

mode_value = mode(numbers)


# Print results

print("Numbers:", numbers)

print("Mean:", mean_value)

print("Median:", median_value)

print("Mode:", mode_value)


##########################################################
OUTPUT:-
Numbers: [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]

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:

Paste your code and output inside the box below:

```
# Import required modules
import numpy as np

# Given datasets
list_x = [10, 20, 30, 40, 50]
list_y = [15, 25, 35, 45, 60]

# Convert lists to numpy arrays
x = np.array(list_x)
y = np.array(list_y)

# Compute covariance
cov_matrix = np.cov(x, y)  # covariance matrix
cov_xy = cov_matrix[0, 1]  # covariance between x and y

# Compute correlation coefficient
corr_matrix = np.corrcoef(x, y)
corr_xy = corr_matrix[0, 1]  # correlation coefficient
```

```
# Print results

print("List X:", list_x)

print("List Y:", list_y)

print("Covariance:", cov_xy)

print("Correlation Coefficient:", corr_xy)


############################################

Output :-

 List X: [10, 20, 30, 40, 50]

List Y: [15, 25, 35, 45, 60]

Covariance: 275.0

Correlation Coefficient: 0.995893206467704
```

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]

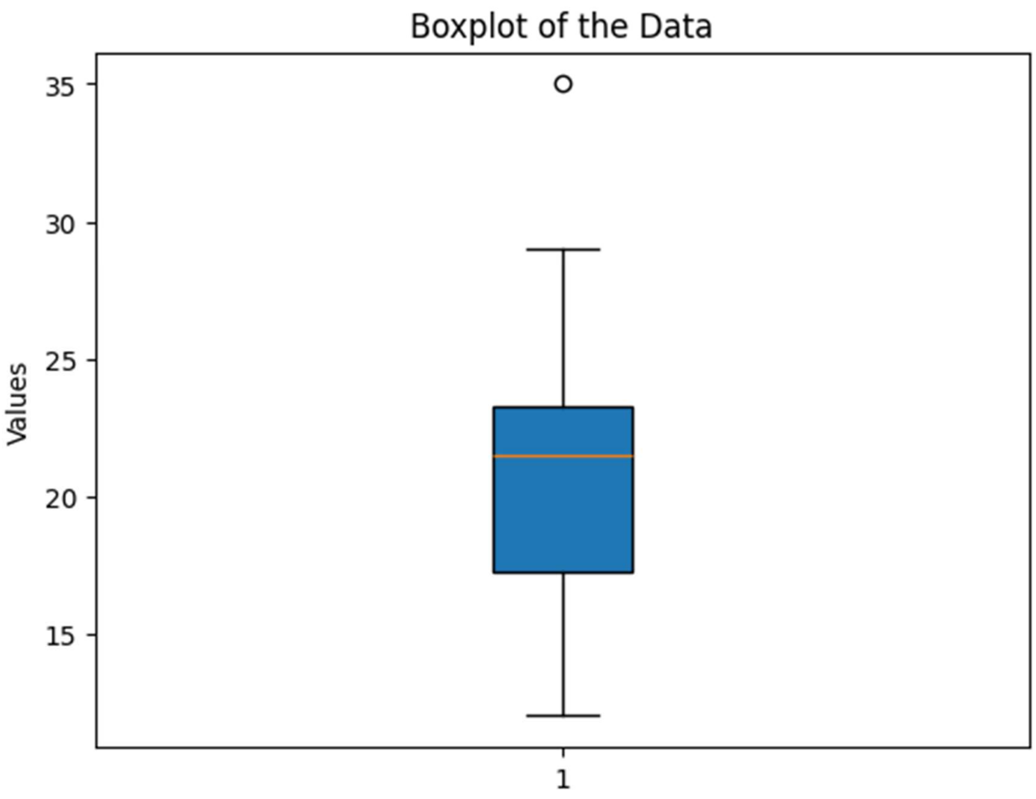
(*Include your Python code and output in the code box below.*)

Answer:

3

```
# Import required libraries

import matplotlib.pyplot as plt


# Given dataset

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

```
# Draw boxplot
plt.boxplot(data, vert=True, patch_artist=True)
plt.title("Boxplot of the Data")
plt.ylabel("Values")
plt.show()

# Identify outliers using IQR method
Q1 = np.percentile(data, 25)  # 1st quartile
Q3 = np.percentile(data, 75)  # 3rd quartile
IQR = Q3 - Q1               # Interquartile range

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Find outliers
outliers = [x for x in data if x < lower_bound or x > upper_bound]

print("Data:", data)
print("Q1:", Q1, "Q3:", Q3, "IQR:", IQR)
print("Lower Bound:", lower_bound, "Upper Bound:", upper_bound)
print("Outliers:", outliers)

###################################
OUTPUT:-
```



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

Q1: 17.25 Q3: 23.25 IQR: 6.0

Lower Bound: 8.25 Upper Bound: 32.25

Outliers: [35]

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.*)

Answer:

Covariance:

- Measures how two variables vary together.
- Positive covariance → when advertising spend increases, sales tend to increase.
- Negative covariance → when advertising spend increases, sales tend to decrease.
- Limitation: Not standardized; hard to compare magnitude.

Correlation Coefficient (r):

- Standardized version of covariance.
- Ranges from -1 to 1.
  - $r \approx 1$ → strong positive relationship
  - $r \approx -1$ → strong negative relationship
  - $r \approx 0$ → no linear relationship

- **Helps marketing team understand strength and direction of the relationship between ad spend and sales.**

CODE:-

```
# Import required library
import numpy as np

# Given data
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)

# Compute covariance
cov_matrix = np.cov(x, y)
cov_xy = cov_matrix[0, 1]

# Compute correlation coefficient
corr_matrix = np.corrcoef(x, y)
corr_xy = corr_matrix[0, 1]

# Print results
print("Covariance:", cov_xy)
print("Correlation Coefficient:", corr_xy)

###############################
```

OUTPUT:-

```
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.*)

4

Answer:

Summary Statistics

1. **Mean → Average satisfaction score.**

2. **Median → Middle value; useful if there are extreme scores.**

3. **Mode → Most common satisfaction score.**

4. **Standard Deviation (SD) → Measures how spread out the scores are.**

5. **Range / Min / Max → Gives the overall spread of scores.**

Visualizations:-

1. **Histogram → Shows how frequently each score occurs; visualizes distribution.**
2. **Boxplot → Helps identify median, quartiles, and any outliers.**

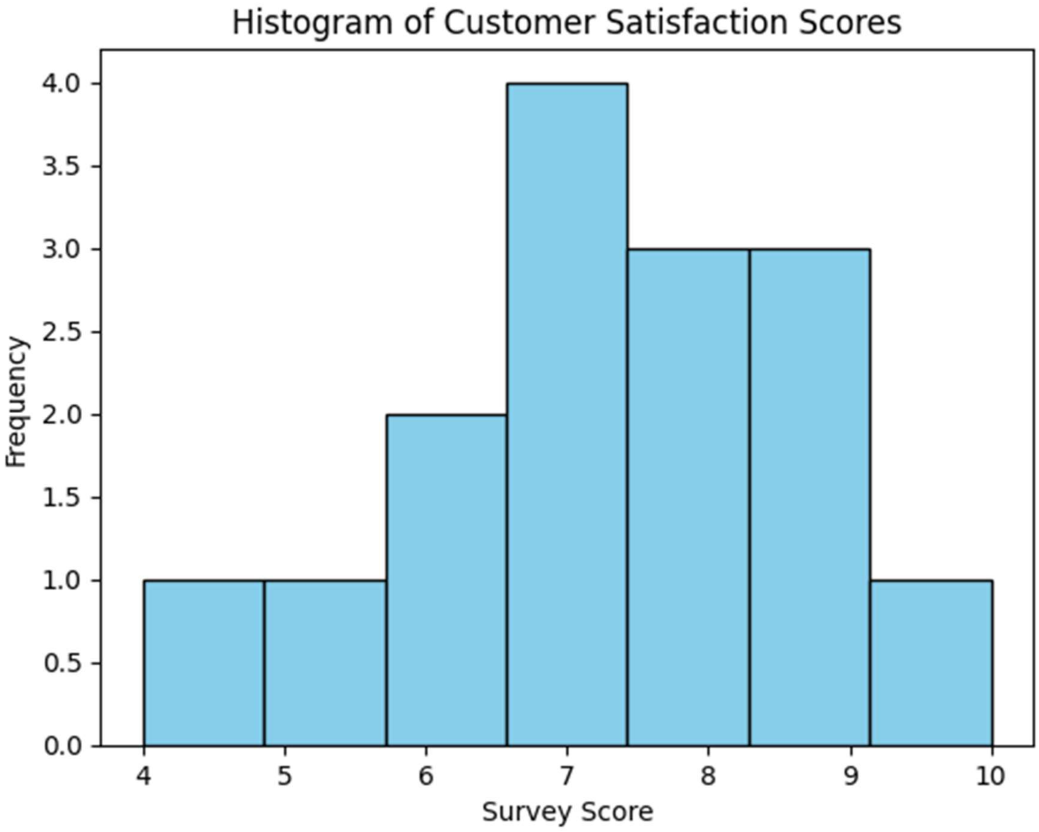
Why?

- **Helps marketing team see the overall trend of customer satisfaction before product launch.**
- **Can identify if most customers are satisfied (high scores) or if there are concerns (low scores).**

CODE:-

```
# Import required library
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, edgecolor='black', color='skyblue')
plt.title("Histogram of Customer Satisfaction Scores")
plt.xlabel("Survey Score")
plt.ylabel("Frequency")
plt.xticks(range(4, 11))  # Set x-axis labels from 4 to 10
plt.show()

#######################
```

OUTPUT:-


Histogram of Customer Satisfaction Scores
Frequency
Survey Score