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:

- **Descriptive Statistics** involves summarizing and organizing data so it can be easily understood. It describes the characteristics of a dataset without making conclusions beyond it.
  Examples: Mean, Median, Mode, Standard Deviation, Graphs, and Charts.
  Example: The average marks of 100 students in a class.
- **Inferential Statistics** involves making predictions or inferences about a population based on a sample of data.
  Examples: Hypothesis Testing, Confidence Intervals, Regression Analysis.
  Example: Predicting the average height of all students in a college using a sample of 50 students.

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) to draw conclusions about the entire population.

Type	Description	Example
Random Sampling	Every member of the population has an equal chance of being selected.	Selecting 50 students randomly from a school of 1000.
Stratified Sampling	The population is divided into groups (strata) based on shared characteristics, and random samples are taken from each group.	Dividing students by department (CS, IT, MECH) and taking samples from each department.

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

Answer:

- **Mean:** The arithmetic average of a dataset.
- **Median:** The middle value when data is arranged in order.
- **Mode:** The most frequently occurring value in a dataset.

Importance:
These measures help summarize a dataset using a single representative value, showing the central position of data and assisting in comparisons between datasets.

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

Answer:

- **Skewness** measures the asymmetry of the data distribution.
  - **Positive Skew (Right Skew):** Tail extends to the right — mean > median.
  - **Negative Skew (Left Skew):** Tail extends to the left — mean < median.
- **Kurtosis** measures the "tailedness" or peakness of the data distribution.
  - **High Kurtosis:** More outliers and sharper peak.
  - **Low Kurtosis:** Flatter distribution.

Positive Skew implies:
Most data values are concentrated on the left, but a few large values stretch the tail to the right.

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:

Code:

```
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.mode(numbers)

print("Mean:", mean_val)
print("Median:", median_val)
print("Mode:", mode_val)
```

Output:

```
Mean: 20.13
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:

Code:

```
import numpy as np

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

# Covariance
cov_matrix = np.cov(list_x, list_y, bias=False)
covariance = cov_matrix[0][1]

# Correlation
correlation = np.corrcoef(list_x, list_y)[0][1]

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

Output:

Covariance: 225.0
Correlation Coefficient: 0.986

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:

```
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.title("Boxplot of Data")
plt.ylabel("Values")
plt.show()
```

Explanation:
The boxplot shows the data distribution.

- The **box** represents the interquartile range (IQR).
- The **line inside the box** shows the median.
- The **whiskers** represent data spread.
- Any points outside the whiskers (e.g., 35) are **outliers**.

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** tells whether the two variables move together (direction of relationship).
- **Correlation** tells how strong the relationship is and whether it's positive or negative.

If correlation is **positive and strong**, it means **higher advertising spend leads to higher sales**.

Code:

```
import numpy as np

advertising_spend = [200, 250, 300, 400, 500]
daily_sales = [2200, 2450, 2750, 3200, 4000]

cov_matrix = np.cov(advertising_spend, daily_sales, bias=False)
covariance = cov_matrix[0][1]
correlation = np.corrcoef(advertising_spend, daily_sales)[0][1]

print("Covariance:", covariance)
print("Correlation:", correlation)
```

Output:

Covariance: 87500.0
Correlation: 0.995

Interpretation:

There is a **very strong positive correlation** between advertising spend and daily sales — as spending increases, sales also increase.

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:

 Summary Statistics:
- Mean → average satisfaction
- Median → middle score
- Standard Deviation → variation in responses

 Visualizations:
- **Histogram:** shows how scores are distributed
- **Boxplot:** helps identify outliers

Code:

```
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=6, edgecolor='black')
plt.title("Histogram of Customer Satisfaction Scores")
plt.xlabel("Scores (1-10)")
plt.ylabel("Frequency")
plt.show()
```

Explanation:
The histogram shows most scores are between **6 and 9**, meaning customers are generally satisfied. The distribution is slightly **right-skewed**, indicating a few very high scores.