Assignment Code: DS-AG-005

Statistics Basics| Assignment

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

Answer: **Descriptive statistics**

Descriptive statistics are about **summarizing and describing the data you already have**. They don't try to go beyond the dataset or make predictions—they just tell you *what the data looks like*.

Common tools:

- Mean, median, mode
- Range, variance, standard deviation
- Tables, charts, graphs (bar charts, histograms, pie charts)

Example:
Suppose a teacher records the exam scores of **30 students in a class**.

- The **average score** is 72
- The **highest score** is 95 and the **lowest** is 40
- A bar chart shows how many students fall into each score range

All of this is **descriptive statistics** because it only summarizes the scores of those 30 students—nothing more.

Inferential statistics

Inferential statistics go a step further. They use **sample data** to **draw conclusions or make predictions about a larger population**.

Common tools:

- Hypothesis testing (t-test, chi-square test, ANOVA)
- Confidence intervals
- Regression analysis
- Correlation

Example:
Now imagine those same 30 students are a **sample** taken from a school with **1,000 students**.

- You use the sample's average score to **estimate the average score of all 1,000 students**
- You test whether **boys and girls differ significantly in their exam performance**
- You calculate a **95% confidence interval** for the true average score of the entire school

This is **inferential statistics** because you're using a sample to make conclusions about a **population**.

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

Answer: **Sampling in statistics**

Sampling is the process of selecting a **subset (sample)** from a **larger group (population)** in order to study it.
Because studying an entire population is often too costly, time-consuming, or impractical, statisticians use samples to make conclusions about the whole population.

Example:
Instead of surveying **all voters in a country**, a researcher surveys **2,000 voters** to predict election outcomes.

Random sampling

In **random sampling**, **every individual in the population has an equal chance of being selected**.

How it works:

- Names are chosen using a lottery method or random number generator
- Selection is completely unbiased

Example:
A school has 1,000 students. Each student is given a number, and **100 numbers are selected randomly** by a computer.
Those 100 students form the sample.

Key features:

- Simple and unbiased
- Easy to understand
- May accidentally over- or under-represent some groups

Stratified sampling

In **stratified sampling**, the population is first divided into **subgroups (strata)** based on a specific characteristic (such as age, gender, class, or income).
Then **random samples are taken from each subgroup**, usually in proportion to their size.

Example:
A school has:

- 60% female students
- 40% male students

If a sample of 100 students is needed:

- 60 females are randomly selected
- 40 males are randomly selected

This ensures both groups are properly represented.

Key features:

- Ensures representation of all important subgroups
- More accurate when the population is diverse
- Slightly more complex than random sampling

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

Answer: **Mean**

The **mean** is the **average** of a set of numbers.
It is found by adding all the values and dividing by the total number of values.

Formula:
Mean = (Sum of all observations) ÷ (Number of observations)

Example:
For the data: 2, 4, 6, 8
Mean = (2 + 4 + 6 + 8) ÷ 4 = 5

Median

The **median** is the **middle value** of a dataset when the data are arranged in ascending or descending order.

- If the number of observations is **odd**, the median is the middle number.
- If the number of observations is **even**, the median is the average of the two middle numbers.

Example:
Data: 3, 5, 7, 9, 11
Median = 7

Mode

The **mode** is the value that **occurs most frequently** in a dataset.

A dataset may have:

- One mode (unimodal)
- More than one mode (bimodal or multimodal)
- No mode

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

Importance of measures of central tendency

Measures of central tendency are important because they:

1. **Summarize large data sets**
   They represent a large amount of data with a single value, making it easier to understand.

2. **Describe the typical or central value**
   They show where most data values tend to cluster.

3. **Help in comparison**
   Means, medians, or modes can be used to compare different datasets (e.g., average marks of two classes).

4. **Support decision-making**
   They are used in fields like economics, education, business, and health to guide planning and policy decisions.

5. **Handle different types of data**

   - Mean is useful for numerical data
   - Median is best when there are extreme values
   - Mode works well for categorical data

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

Data

Asnwer:**Kurtosis**

Kurtosis measures the **peakedness or flatness** of a distribution compared to a normal distribution.
It also indicates how **heavy or light the tails** of the distribution are.

Types of kurtosis

1. **Leptokurtic**
   - Very sharp peak
   - Heavy tails
   - More extreme values (outliers)
2. **Mesokurtic**
   - Moderate peak
   - Similar to a normal distribution
3. **Platykurtic**
   - Flat peak
   - Light tails
   - Fewer extreme values

What does a positive skew imply about the data?

A **positive skew** implies that:

- Most observations are **smaller values**
- A few **large values** stretch the distribution to the right
- The **mean is pulled upward** by extreme high values
- The data are **not symmetrically distributed**

Question 5: Implement a Python program to compute the mean, median, and mode of a given list of numbers

Answer:from collections import Counter

```
def calculate_mean(numbers):
    return sum(numbers) / len(numbers)


def calculate_median(numbers):
    numbers.sort()
    n = len(numbers)
    mid = n // 2

    if n % 2 == 0:
        return (numbers[mid - 1] + numbers[mid]) / 2
    else:
        return numbers[mid]


def calculate_mode(numbers):
    freq = Counter(numbers)
    max_freq = max(freq.values())
    modes = [num for num, count in freq.items() if count == max_freq]

    if len(modes) == len(numbers):
        return None  # No mode
```

```
    return modes

# Example usage
data = [2, 4, 4, 6, 8, 10]

mean = calculate_mean(data)
median = calculate_median(data)
mode = calculate_mode(data)

print("Mean:", mean)
print("Median:", median)
print("Mode:", mode)
```

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:import math

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

n = len(list_x)
```

```
# Mean of X and Y
mean_x = sum(list_x) / n
mean_y = sum(list_y) / n


# Covariance
covariance = sum((list_x[i] - mean_x) * (list_y[i] - mean_y) for i in range(n)) / n


# Standard deviations
std_x = math.sqrt(sum((x - mean_x) ** 2 for x in list_x) / n)
std_y = math.sqrt(sum((y - mean_y) ** 2 for y in list_y) / n)


# Correlation coefficient
correlation = covariance / (std_x * std_y)


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

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: 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]

# Convert to NumPy array
arr = np.array(data)

# Quartiles
Q1 = np.percentile(arr, 25)
Q3 = np.percentile(arr, 75)
IQR = Q3 - Q1

# Outlier limits
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Identify outliers
outliers = arr[(arr < lower_bound) | (arr > upper_bound)]

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

print("Q1:", Q1)

print("Q3:", Q3)

print("IQR:", IQR)

print("Outliers:", 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]

Answer:import math

Given data

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

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

n = len(advertising_spend)

Means

mean_x = sum(advertising_spend) / n

mean_y = sum(daily_sales) / n

```python
# Covariance
covariance = sum(
    (advertising_spend[i] - mean_x) * (daily_sales[i] - mean_y)
    for i in range(n)
) / n

# Standard deviations
std_x = math.sqrt(sum((x - mean_x) ** 2 for x in advertising_spend) / n)
std_y = math.sqrt(sum((y - mean_y) ** 2 for y in daily_sales) / n)

# Correlation coefficient
correlation = covariance / (std_x * std_y)

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