Name – Hritik Maheshwari

Email ID – hritikmaheshwari82@gmail.com

Statistics Basics| Assignment

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

Answer –

- Descriptive Statistics involves methods for summarizing and organizing data using numbers and graphs. It describes the basic features of the data.

  Example: The average score of students in a class (mean = 75).

- Inferential Statistics uses a sample of data to make estimates or test hypotheses about a population.

  Example: Predicting the average height of all students in a college based on a sample of 100 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 larger population to make statistical inferences.

- **Random Sampling:** Every member of the population has an equal chance of being selected.
  Example: Selecting 10 students at random from a class of 100.

- **Stratified Sampling:** The population is divided into subgroups (strata) and random samples are taken from each subgroup.

Example: Dividing students by year (1st, 2nd, 3rd) and randomly selecting students from each year.

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

Answer –

Mean: Average of all data values.

Median: Middle value when data is ordered.

Mode: Most frequently occurring value.

They help understand the center or typical value of the data, assisting in comparisons and decision-making.

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

Answer –

- **Skewness:** Measures the asymmetry of a distribution.
  Positive Skew: Tail on the right; most data is on the left.

- **Kurtosis:** Measures the "tailedness" of the distribution.

  High kurtosis = more outliers.

Positive Skew Implication: The mean is greater than the median; a few high values are pulling the average up.

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

Answer –

```
import statistics

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

mean_val = statistics.mean(numbers)

median_val = statistics.median(numbers)
```

```
mode_val = statistics.mode(numbers)
print("Mean:", mean_val)
print("Median:", median_val)
print("Mode:", mode_val)
```

Output –

Mean: 20.0

Median: 19

Mode: 12

Question 6: Compute the covariance and correlation coefficient between the two datasets.

Answer –

```
import numpy as np

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

cov_matrix = np.cov(list_x, list_y)
correlation = np.corrcoef(list_x, list_y)

print("Covariance:", cov_matrix[0][1])
print("Correlation Coefficient:", correlation[0][1])
```

Output -

Covariance: 250.0

Correlation Coefficient: 0.9938586931957764

Question 7: Draw a boxplot and identify outliers.

Answer –

```
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.grid(True)
plt.show()
```

Explanation:

The boxplot will show Q1, Q2 (median), Q3, and outliers.

Outlier: 35 appears as an outlier (above upper whisker).

Question 8: Relationship between advertising spend and daily sales.

Answer –

- **Covariance** shows the direction of the relationship.
- **Correlation** shows the strength and direction of a linear relationship between two variables.

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)
correlation = np.corrcoef(advertising_spend, daily_sales)
```

```
print("Covariance:", cov_matrix[0][1])
print("Correlation Coefficient:", correlation[0][1])
```

Output –

Covariance: 112500.0

Correlation Coefficient: 0.9970544855015815

Question 9: Summary statistics and histogram for survey scores.

Answer –

Use mean, median, and standard deviation for summary statistics.

Use a histogram to visualize the distribution of survey scores.

Code –

```
import matplotlib.pyplot as plt
import statistics

survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]

mean_score = statistics.mean(survey_scores)
std_dev = statistics.stdev(survey_scores)

print("Mean:", mean_score)
print("Standard Deviation:", std_dev)

plt.hist(survey_scores, bins=6, edgecolor='black')
plt.title("Histogram of Survey Scores")
plt.xlabel("Score")
```

```
plt.ylabel("Frequency")

plt.grid(True)

plt.show()
```

Output –

Mean: 7.4

Standard Deviation: 1.636

The histogram shows a distribution slightly skewed toward higher scores, indicating overall satisfaction is high.