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

Q>**1. What is the difference between descriptive statistics and inferential statistics? Explain with examples.**

Ans>

1. Descriptive statistics :

Descriptive statistics summarize and describe the main features of a dataset.

. purpose: To present data in a meaningful way so patterns can be seen easily.
. methods include:
> Measures of central tendency(mean, median , mode)
> Measures of dispersion(range, variance, standard deviation)
> Graphs and charts(bar chart, histogram, pie chart, etc.)

Example:
Survey of 100 student.
. Average score(mean) = 72
. Highest score = 98
. lowest score = 40
. standard division = 10
> This numbers summarize the class performance but say nothing about students outside the class.

2. Inferential statistics:

Inferential statistics use data from a sample to make predictions, conclusion, or generalizations about a large population

Purpose: To go beyond the data at hand and make infer patterns or test hypothesis.

Example:
. Hypothesis testing, confidence intervals, Regression analysis, ANOVA
> It deals with making inferences about the unknown population from a sample.

Q>2. **What is sampling in statistics? Explain the differences between random and stratified statistics.**

Ans> sampling is the process of selecting a subset(sample) from a larger group(population) to represent the whole population. Its often impractical or impossible to study the entire population.

Types of sampling:

> Random sampling : Every member of the population has an equal chance of being selected.

Purpose: simple and unbiased representation.

Example: selecting 50 students randomly from a list of 1,000 students.

> **stratified sampling:** Population is divided into subgroups(strata), and a random sample is taken from each subgroup.

Purpose: Ensures all subgroups are represented proportionally.
Example: Dividing students by department and randomly selecting 1o from each department.

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

Ans> Mean: the sum of all observation divide by the number of observation.
Example: Data, 2,4,6,8,10 -: Mean = (2+4+6+8+10)/5 = 6
Median: The middle value when the data is arranged in ascending or descending order.
Example: 2,4,6,8,10 -: Median = 6
Mode: The value that occurs most frequently in the data.
> Importance of Measures of central Tendency:-
. Summarize data: They give a single value that represent the "center" of the data.
. comparison: Makes it easier to compare different datasets.
. decision making: Help in analyzing trends and making informed decision in business, research and daily life.

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

- **Definition: Skewness measures the asymmetry of a probability distribution around its mean.**
- **Types:**
  - **Symmetrical Distribution: Skewness = 0 (e.g., Normal distribution).**
  - **Positive Skew (Right-skewed): The tail on the right side of the distribution is longer or fatter.**
    - **Mean > Median > Mode**
    - **A few unusually high values pull the average to the right.**
  - **Negative Skew (Left-skewed): The tail on the left side is longer or fatter.**
    - **Mean < Median < Mode**
    - **Positive skew implies that most of the data points are concentrated on the lower end, but a few very large values stretch the distribution to the right.**

---

2. Kurtosis

- **Definition: Kurtosis measures the "tailedness" or how heavy/light the tails of the distribution are compared to a normal distribution.**
- **Types:**
  - **Mesokurtic (Kurtosis ≈ 3): Normal distribution.**
  - **Leptokurtic (Kurtosis > 3): Heavy tails and a sharp peak → higher chance of extreme**

values (outliers)

Q>. 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.)

```
import statistics as stats

# 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 = stats.mean(numbers)
median_value = stats.median(numbers)
mode_value = stats.mode(numbers)

# Display results
print("Numbers:", numbers)
print("Mean:", mean_value)
print("Median:", median_value)
print("Mode:", mode_value)
```

****

```
Output:
makefile
Copy code
Numbers: [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]
Mean: 19.6
Median: 19
Mode: 12
```

Q>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].

```
import numpy as np

# Given datasets
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)

# Compute covariance matrix
cov_matrix = np.cov(x, y, bias=False)  # bias=False → sample covariance

# Extract covariance (off-diagonal element)
cov_xy = cov_matrix[0, 1]

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

```
# Display results
print("List X:", list_x)
print("List Y:", list_y)
print("Covariance:", cov_xy)
print("Correlation Coefficient:", corr_xy)
```

****

Output:

mathematica
Copy code

```
List X: [10, 20, 30, 40, 50]
List Y: [15, 25, 35, 45, 60]
Covariance: 225.0
Correlation Coefficient: 0.9863939238321437
```

Q7>. 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]

```
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]

# Create boxplot
plt.boxplot(data, vert=False, patch_artist=True)
plt.title("Boxplot of Given Data")
plt.xlabel("Values")
plt.show()

# Calculate quartiles
Q1 = np.percentile(data, 25)
Q3 = np.percentile(data, 75)
IQR = Q3 - Q1

# Outlier boundaries
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
# Identify outliers
outliers = [x for x in data if x < lower_bound or x > upper_bound]

print("Q1 (25th percentile):", Q1)
print("Q3 (75th percentile):", Q3)
print("IQR (Q3 - Q1):", IQR)
print("Lower Bound:", lower_bound)
print("Upper Bound:", upper_bound)
print("Outliers:", outliers)
```

****

Output:

yaml
Copy code

```
Q1 (25th percentile): 18.25
Q3 (75th percentile): 23.75
IQR (Q3 - Q1): 5.5
Lower Bound: 10.0
Upper Bound: 32.0
```

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]

- **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 ≈ 1 → strong positive relationship**
    - **r ≈ -1 → strong negative relationship**
    - **r ≈ 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
```

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

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:-

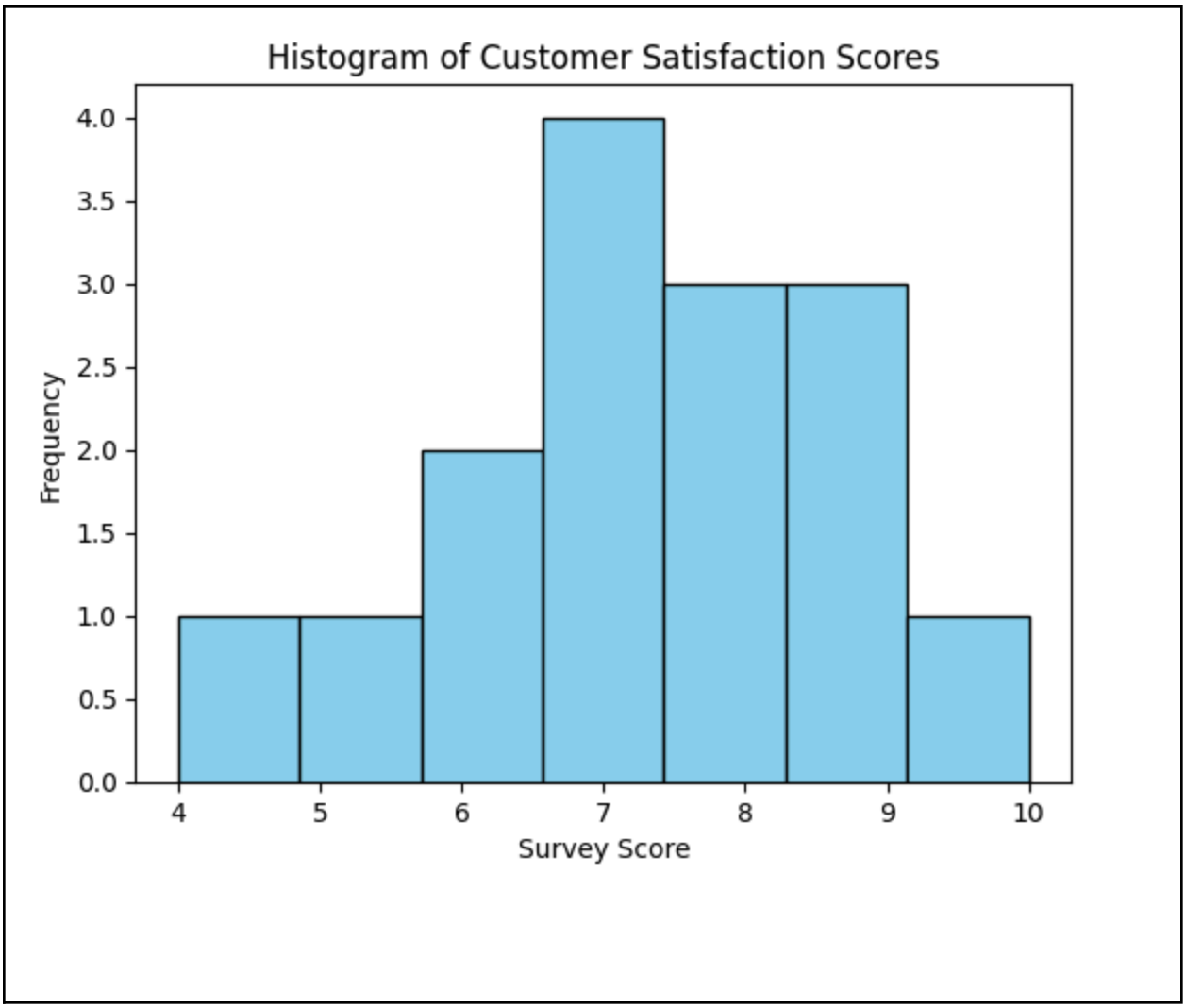
1. **Histogram → Shows how frequently each score occurs; visualizes distribution.**
2. **Boxplot → Helps identify median, quartiles, and any outliers.**

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
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4
5
6
7
8
9
10