Statistics Basics| Assignment

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

Answer : DESCRIPTIVE STATISTICS :-
The type of statistics dealing with numbers (numerical facts, figures, or information) to describe phenomena.
These numbers are descriptive statistics. They are used to describe, summarise the characteristics of a sample
or dataset, such as variables mean, standard deviation, or frequency etc.
e.g. Reports of industry production, cricket batting averages, government deficits, Movie Ratings etc.

INFERENTIAL STATISTICS :-
Inferential statistics is a decision, estimate, prediction, or generalisation about a population based on a sample.
Here we deal with the sample/samples and compare them and perform some tests to draw some conclusions
on the population data based on the observations from the sample.

- A population is a collection of all possible individuals, objects, or measurements of interest.

- A sample is a portion, or part, of the population of interest.
- Inferential statistics is used to make inferences from data, whereas descriptive statistics describe what’s

happening in our data.

Example :-

- 3rd and 4th-year students are the main target for restaurant sales.
- You can discount the 1st year students to increase the number count.

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

Answer : - Sampling is a practical way to study a population without having to examine every single element within it. Sampling involves carefully selecting a representative subset of the population to ensure that the characteristics and attributes of the sample reflect those of the entire population as accurately as possible. If the sample is chosen correctly and is truly representative,

The differences between random and stratified sampling are :
Random Sampling: Every individual in the population has an equal chance of being selected. This helps

minimize bias and increase the likelihood that the sample represents the population accurately.

Stratified Sampling: The population is divided into distinct subgroups (strata) based on certain
characteristics, and then a random sample is taken from each subgroup proportionate to its size. This
ensures representation from various subgroups.

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

Answer :
Mean:
The mean is the arithmetic average;
for calculating the mean add up all of the values and divide by the number of observations in your dataset.

Median:

The Median for the sample data arranged in increasing order is defined as :

- if “n” is an odd number then median is the middle value
- if “n” is an even number then median is midway between the two middle values

Mode :

The mode is the value that occurs the most frequently in your data set i.e. has the highest frequency. On a bar chart, the mode is the highest bar. If the data have multiple values that are tied for occurring the most frequently, you have a multimodal distribution. If no value repeats, the data do not have a mode.

Why they are important:

- They show the *central point* of data.

- Mean tells the overall average.

- Median shows the middle value / physical midpoint of data, useful when data has extreme values (outliers).

- Mode shows the most common item, useful in things like surveys or sales.

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

Answer:

Skewness

- Skewness tells us whether data is symmetrical or tilted to one side.

- Positive skew (right skew)**:** the tail of the graph stretches more to the right → most data is on the left, but a few very large values pull the average higher.

Kurtosis

- Kurtosis tells us about the shape of the peak of data distribution.

- High kurtosis (leptokurtic): sharp peak, heavy tails (more extreme values).

- Low kurtosis (platykurtic)**:** flat peak, light tails (fewer extreme values).

- Normal kurtosis (mesokurtic)**:** bell-shaped, like the normal distribution.

Positive Skew means

- Data has a long tail on the right.
- Mean > Median > Mode.
- Shows that a few very large values are affecting the average.

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 :

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

Python code mean:

```
import numpy as np
np.mean(nos)
```

Output

```
np.float64(19.6)
```

Python code median :

```
np.median(nos)
```

Output

```
np.float64(19.0)
```

Python code mode :

```
import statistics
statistics.mode(nos)
```

Output

```
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]

Answer :

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

Input :

```
import numpy as np
```

```
x = np.array(list_x)
y = np.array(list_y)

# Compute covariance matrix (population covariance)
cov_matrix = np.cov(x, y, ddof=0)

# Extract covariance between x and y
cov_xy = cov_matrix[0, 1]

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

print("Covariance:", cov_xy)
print("Correlation coefficient:", corr_xy)
```

Output :

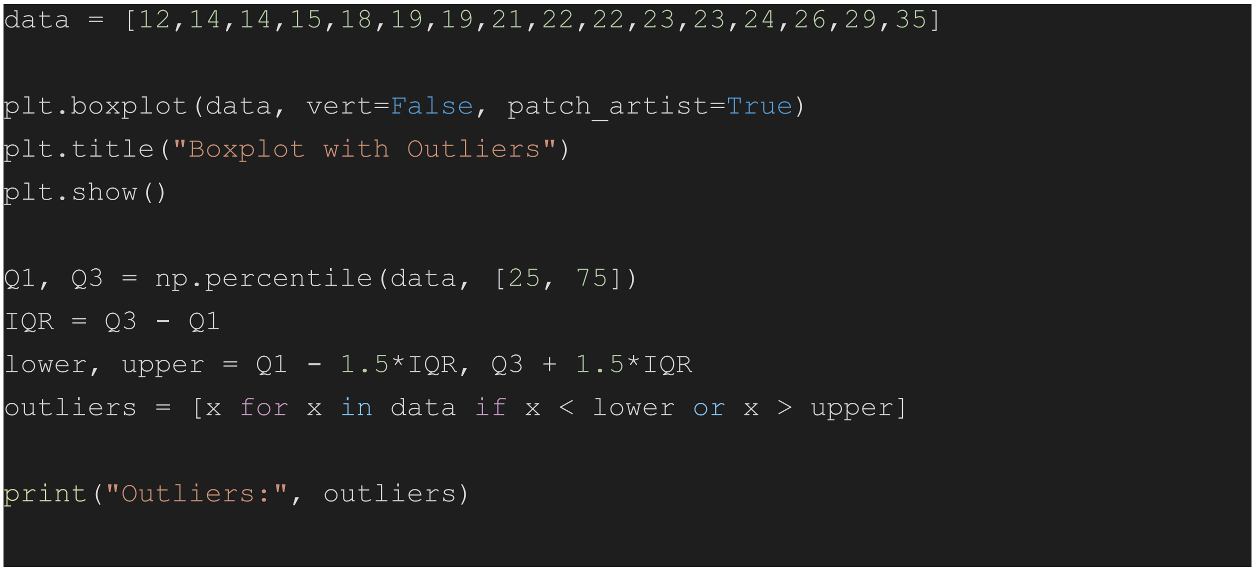
```
Covariance: 220.0
Correlation coefficient: 0.995893206467704
```

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]

Answer :

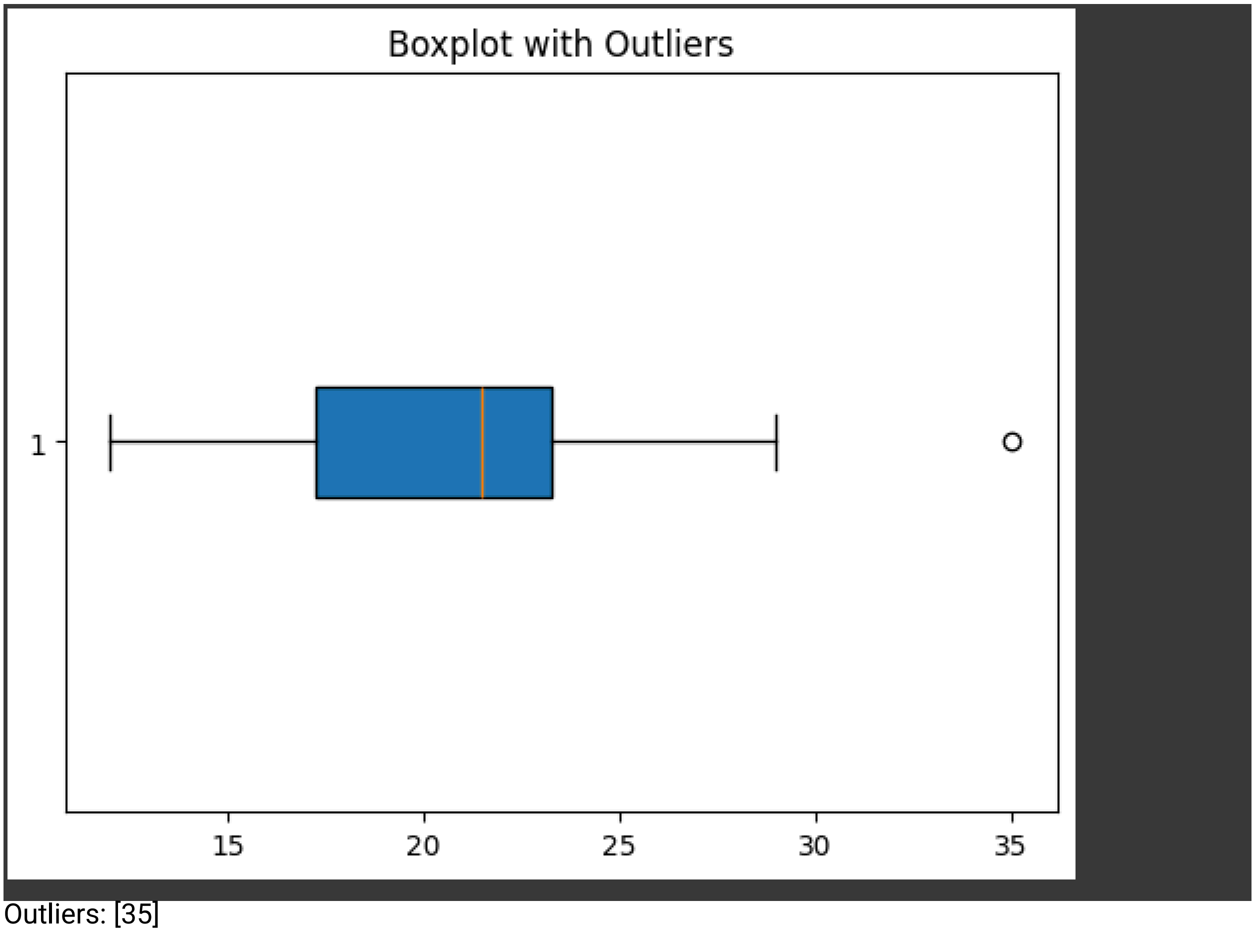
Input :

```
import matplotlib.pyplot as plt
import numpy as np
```



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

plt.boxplot(data, vert=False, patch_artist=True)
plt.title("Boxplot with Outliers")
plt.show()

Q1, Q3 = np.percentile(data, [25, 75])
IQR = Q3 - Q1
lower, upper = Q1 - 1.5*IQR, Q3 + 1.5*IQR
outliers = [x for x in data if x < lower or x > upper]

print("Outliers:", outliers)
```

Output :



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]

Answer :

Input :

```
import numpy as np

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

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

print("Covariance:", cov_xy)
print("Correlation:", corr_xy)
```

Output :

```
Covariance: 67900.0
Correlation: 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]

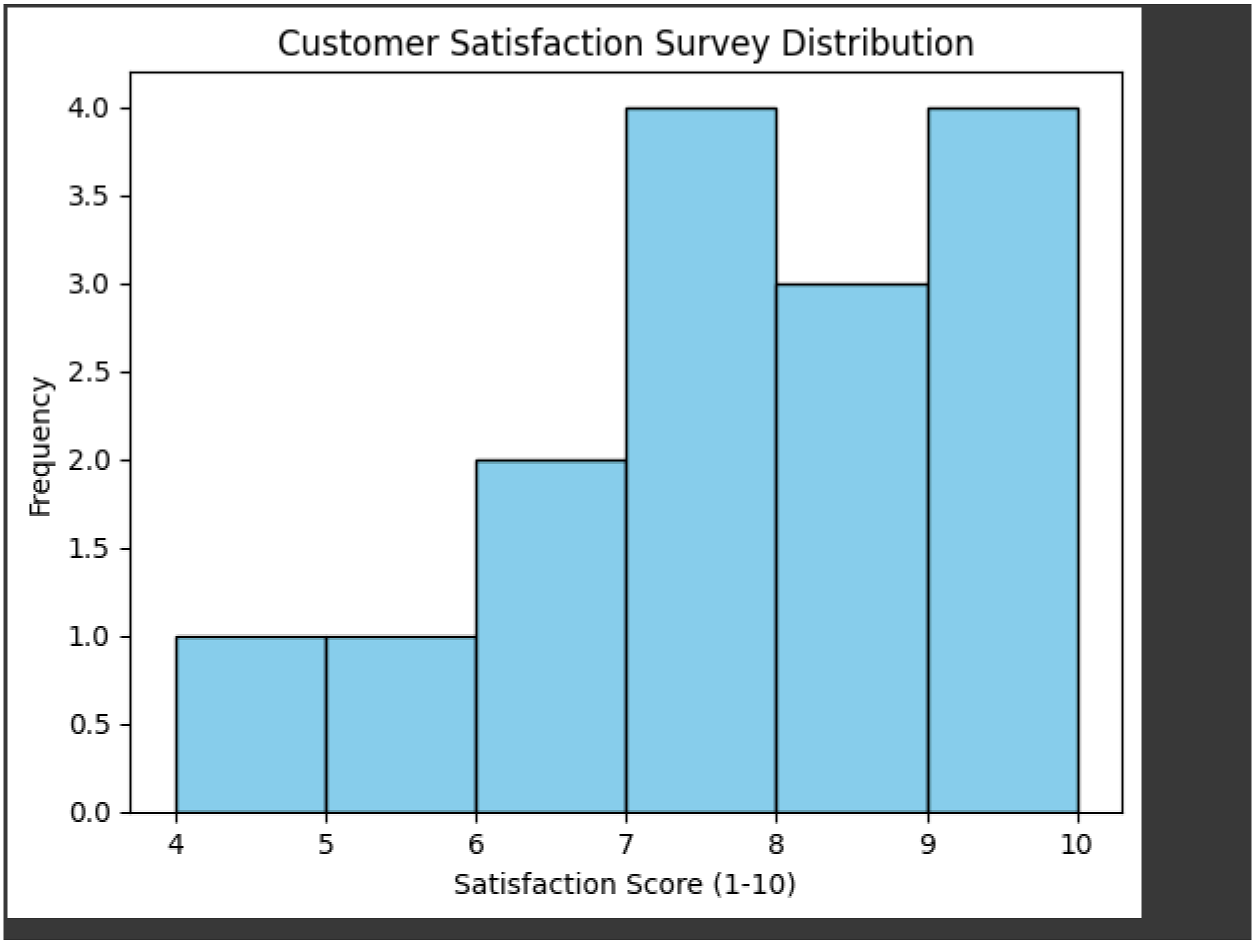
Answer :

Input :

```
import matplotlib.pyplot as plt
import numpy as np

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

mean = np.mean(survey_scores)
median = np.median(survey_scores)
std_dev = np.std(survey_scores)

print("Mean:", mean)
print("Median:", median)
print("Standard Deviation:", std_dev)
print("Min:", min(survey_scores))
print("Max:", max(survey_scores))

plt.hist(survey_scores, bins=6, color="skyblue", edgecolor="black")
plt.title("Customer Satisfaction Survey Distribution")
plt.xlabel("Satisfaction Score (1-10)")
plt.ylabel("Frequency")
plt.show()
```

Output :

```
Mean: 7.333333333333333
Median: 7.0
Standard Deviation: 1.577621275493231
Min: 4
Max: 10
```


Customer Satisfaction Survey Distribution
Frequency
Satisfaction Score (1-10)