Statistics Basics| Assignment

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

Ans:

Descriptive statistics: It is consists of organise and summarising the complete data / Population

Example: Average Weight of students in class

Inferential statistics: In Inferential statistics the sample is choose from population / a some part of population not complete data is used to conclude the about the population.

It consists of using data that has been measured to form conclusion about a population.

With given sample data can we conclude somethings about population

Example: Average Age / Weight of population of India

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

Ans: Sampling is a process of selecting a small group (sample) from a large group (population) for study and analysis.

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

Stratified Sampling: Population is divided into groups (strata) and samples are taken from each group.

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

Ans:

1. **Mean:** Mean is the average of a set of numbers.
   In Which we adding all the values and dividing by the total numbers of values.

   Mean = (Sum of all values) / (Total Number of values)

   Ex:
   Numbers: 1,2,3,4,5
   Mean: (1+2+3+4+5) / 5 = 15/5 = 3

2. **Median:** Median is the middle/physical value of an ordered data set.
   If the number of values is even, the median is the average of the two middle values.
   If the Data is dispersed then arrange the data in ascending order.

Ex: 1) Data is even
Median = (Value at n/2 + Value at (n/2 + 1)) / 2

Data: 4,6,8,12 n = 4(Even)
Middle positions = 4/2 = **2nd** and 4/2 + 1 = **3rd** values
So median = (6 + 8) / 2 = **7**

Ex: 1) Data is Odd
Median = Value at the position (n + 1)/2

Data: Data: 3, 7, 9 n = 3 (odd)
Position = (3 + 1)/2 = 4/2 = **2nd value**
Median = **7**

3. **Mode:** Mode is the value that occurs most frequency in the data.

   Example:
   Numbers: 2, 4, 4, 7, 9
   Mode = 4

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

1)Skewness: Skewness is a measure that shows whether the data is symmetrical or skewed (asymmetrical) around the mean.

1.If data is symmetrical, skewness = 0
2)If the data stretches more to the right or left , it is skewed

Types of Skewness:

1)Positive (Right) Skew: Tail is longer on the right side, tail to the right, mean > median
2) Negative (Left) Skew: Tail is longer on the left side

2)Kurtosis: Kurtosis measures the peakedness or flatness of a data distribution compared to a normal distribution.

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]

Answer:

1) **Mean:**

```
import numpy as np
numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]
np.mean(numbers)
```

Output: 19.6

2)Median:

```
import numpy as np
numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]
np.median(numbers)
```

Output: 19.0

3)Mode:

i)
```
from scipy import stats
stats.mode(numbers)
```

Output: ModeResult(mode=12, count=3)

ii)
```
import statistics
statistics.mode(numbers)
```

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] (Include your Python code and output in the code box below.)

Answer:

```
import numpy as np
list_x = [10, 20, 30, 40, 50]
list_y = [15, 25, 35, 45, 60]

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

cov_matrix = np.cov(x,y, bias = False)
cov_xy = cov_matrix[0,1]

corr_xy = np.corrcoef(x,y)[0,1]
cov_xy , corr_xy
```

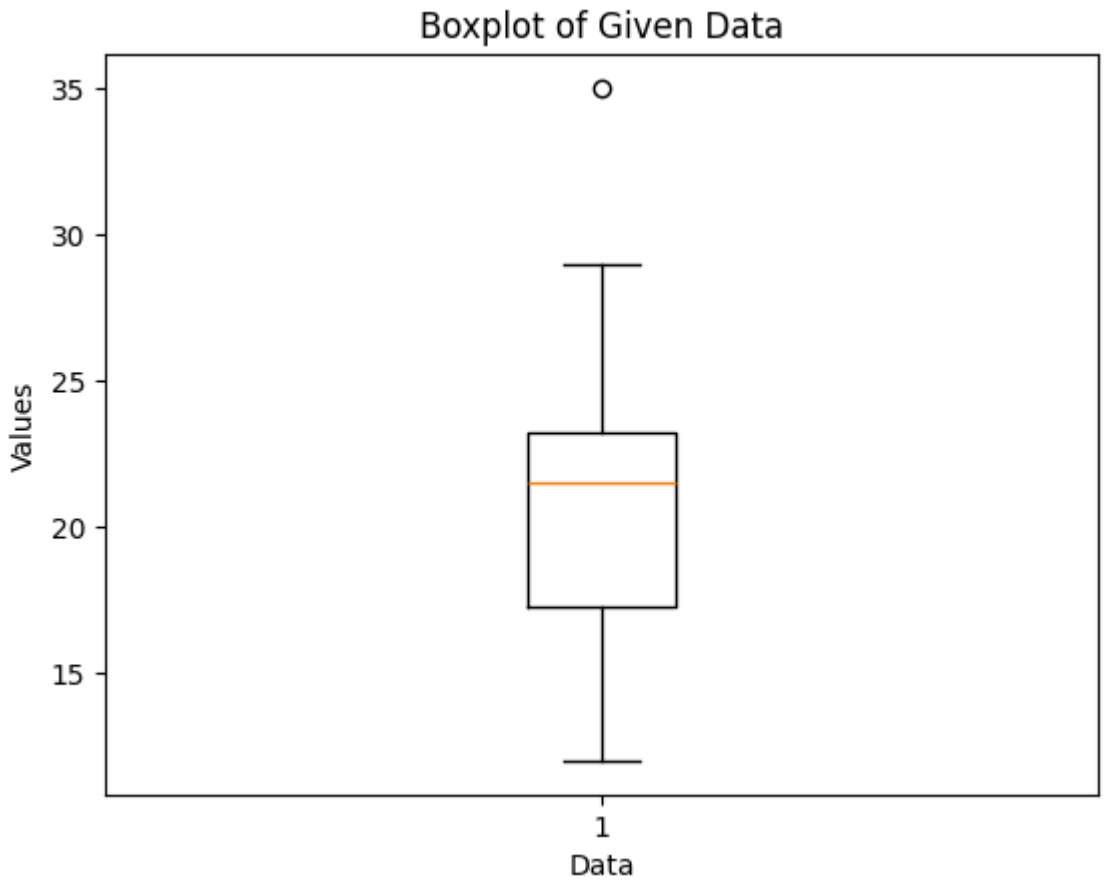
OUTPUT: (275.0, 0.995893206467704)

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] (Include your Python code and output in the code box below.)

Answer:

```
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]

q1 = np.percentile(data, 25)
q3 = np.percentile(data, 75)
iqr = q3 - q1

lower_bound = q1 - 1.5 * iqr
upper_bound = q3 + 1.5 * iqr

outliers = [x for x in data if x < lower_bound or x > upper_bound]

plt.boxplot(data)
plt.title("Boxplot of Given Data")
plt.xlabel("Data")
plt.ylabel("Values")
plt.show()

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:

- Covariance checks whether two variables move together.
- If covariance is positive, it means:
  - When advertising increases → sales also increase.
- If covariance is negative:
  - When advertising increases → sales decrease.
- It shows direction of relationship, but not strength.

Correlation:

- Correlation measures both direction and strength of the relationship.
- Value ranges from -1 to +1
  - +1 = perfect positive relationship
  - 0 = no relationship
  - -1 = perfect negative relationship

Correlation will help us understand whether higher ad spend strongly increases sales.

```
import numpy as np

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

x = np.array(advertising_spend)
y = np.array(daily_sales)

cov_matrix = np.cov(x, y, bias=False)
cov_xy = cov_matrix[0, 1]

corr_xy = np.corrcoef(x, y)[0, 1]

cov_xy, corr_xy
```

OUTPUT: (84875.0, 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] (Include your Python code and output in the code box below.)

Answer:

Summary Statistics & Visualization Explanation:
To understand customer satisfaction scores (1–10 scale) before launching a new product, we need to analyze the **distribution** of the data.

📌 **Summary Statistics to Use**

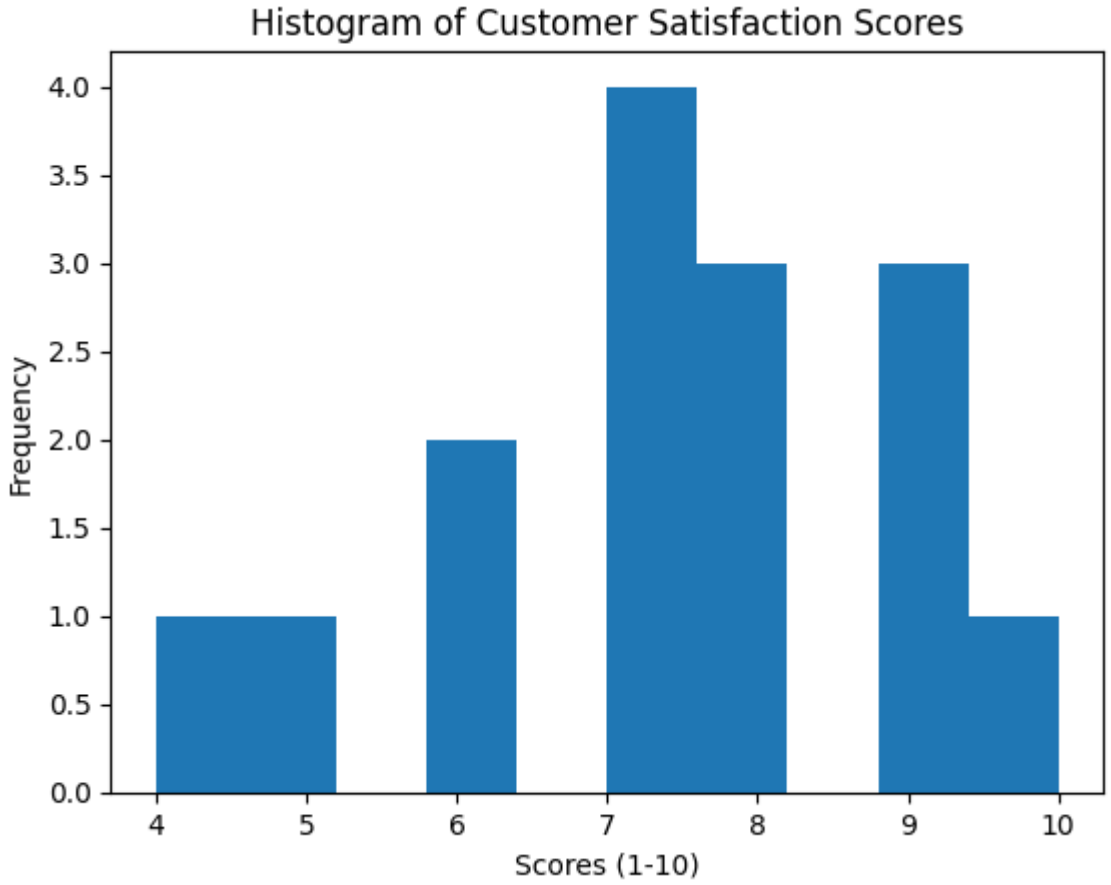
Statistic	Why We Use It
Mean (Average)	To understand the overall satisfaction level
Median	To find the central (middle) satisfaction response
Mode	To see the most common rating
Standard Deviation	To measure how spread out the scores are (variation)
Minimum & Maximum	To see range of customer opinions
Quartiles & IQR	To detect concentration and outliers

📊 **Visualizations to Use**

Plot	Purpose
Histogram	Shows distribution and frequency of scores
Boxplot	Shows spread & outliers (if any)
Bar Chart	(If showing frequency of each score category)

```
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)
plt.title("Histogram of Customer Satisfaction Scores")
plt.xlabel("Scores (1-10)")
plt.ylabel("Frequency")
plt.show()
```

Output:


Histogram of Customer Satisfaction Scores
Frequency
Scores (1-10)
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