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

Answer:-

Descriptive statistics

1. It means organising and summarising the complete dataset/ population. They make you understand the features and characteristics of data without making predictions or generalisation.
2. Key tools of Descriptive statistics are:-
    i. Mean
    ii. Median
    iii. Mode
    iv. Range
    v. Percentage and percentile
    vi. Inter Quartile Range
    vii. Variance
    viii. Standard Deviation
3. For example :- Marks of students in maths of a class
   Marks = { 45, 56, 78, 89, 65, 78, 89}
   Average of marks(mean):- (45+56+78+89+65+78+89)/7 = 71.428
   Median :- 78
   Standard Deviation :- 15.49

Inferential Statistics

1. It means using a sample of data to conclude a prediction or generalization on a large datasets.
2. Key tools of inferential statistics are:-
   Probability distribution
   - Hypothesis testing (e.g., t-tests, chi-square tests)
   - Regression analysis
   - ANOVA
   - Confidence intervals

For example:- From a sample of 100 people you have made predictions for the whole large population.

Out of 100 people, 60 people drink tea and 40 people prefer drinking coffee which gives a prediction that out of the whole population around 60% people drink tea and 40% drink coffee.

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

Answer:-

Sampling is a process of selecting a subset(sample) from a larger group to analyse and draw conclusions for the entire population.

Since it is very quite difficult to study the large population, sampling makes data collection more efficient and manageable.
Example:- It is not possible to analyze the weight of all people living in India, so to resolve that problem we'll be using sampling to analyze data.

Random Sampling :-

It means the sampling where each individual has an equal and independent chance of being selected.

Example:- In a lucky draw of 30 people each person has an equal chance of being selected.

Stratified Sampling :-

Stratified sampling is a method where the population is divided into distinct groups(strata) which share similar characteristics and then random sampling is taken from each group.

Example:-

- Imagine a college has 1000students 600 females and 400 males.
- You want to select a sample of 100 from a college. As per stratified sampling there will be randomly selected 60 females and 40 males in the sample.

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

Answer:-
Mean:- Sum of all datasets divided by the number of values. But mean is affected by outliers that's why the median is most appropriate in case of outliers.
Median:- Mid value of the dataset where data is arranged in order.
Mode:- Mode is the number that occurred most frequently in a dataset.

Example:-
A = [1, 3, 2, 4, 6, 4 , 6, 6, 6, 4, 8, 9, 7]

Mean = (1+3+2+4+6+4+6+6+6+4+8+9+7)/13 = 5.07
Median = 6
Mode = 6

measures of central tendency are important due to following reasons:-

1. They provide a single value that summarizes a datasets representing its central value.
2. It helps in understanding and analyzing the data.
3. Helps in identifying the outliers.

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

Answer:-

1. **Skewness** measures the **asymmetry** of a data distribution.

- A perfectly symmetrical distribution (like the normal distribution) has **zero skewness**.
- Skewness shows whether the **tail** of the data is longer on one side

Type of Skewness:-

1. **Positive Skew** (Right-skewed)

The tail is longer on the **right** side. Most values are on the lower end.
Graph Shape:- Long tail to the right

2. **Negative Skew** (Left-skewed)

The tail is longer on the **left** side. Most values are higher.
Graph Shape:- Long tail to the left

3. **Zero Skew**

Symmetrical distribution (e.g., bell curve)
Graph Shape:- Even on both sides

What Does a Positive Skew Imply?

- Most of the data values are low, but a few high outliers pull the mean to the right.
- Mean > Median > Mode
- Common in income distributions, house prices, etc**.**
- Example:- If most of the people earn between Rs 30-50thousand and some earns Rs 3-4lac, the income data is positively skewed.

2. Kurtosis

- Kurtosis measures the tailedness or peakedness of a distribution. It tells us how much of the data is in the tails and center compared to a normal distribution.
- It helps in detecting the extreme values or outliers.
- Positive kurtosis data has more outliers than normal.

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 :-

```
#Mean
import numpy as np
numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]
```

```
np.mean(numbers)
np.float64(19.6)

#Median
np.median(numbers)
np.float64(19.0)

#Mode
import statistics as stat
stat.mode(numbers)
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:-

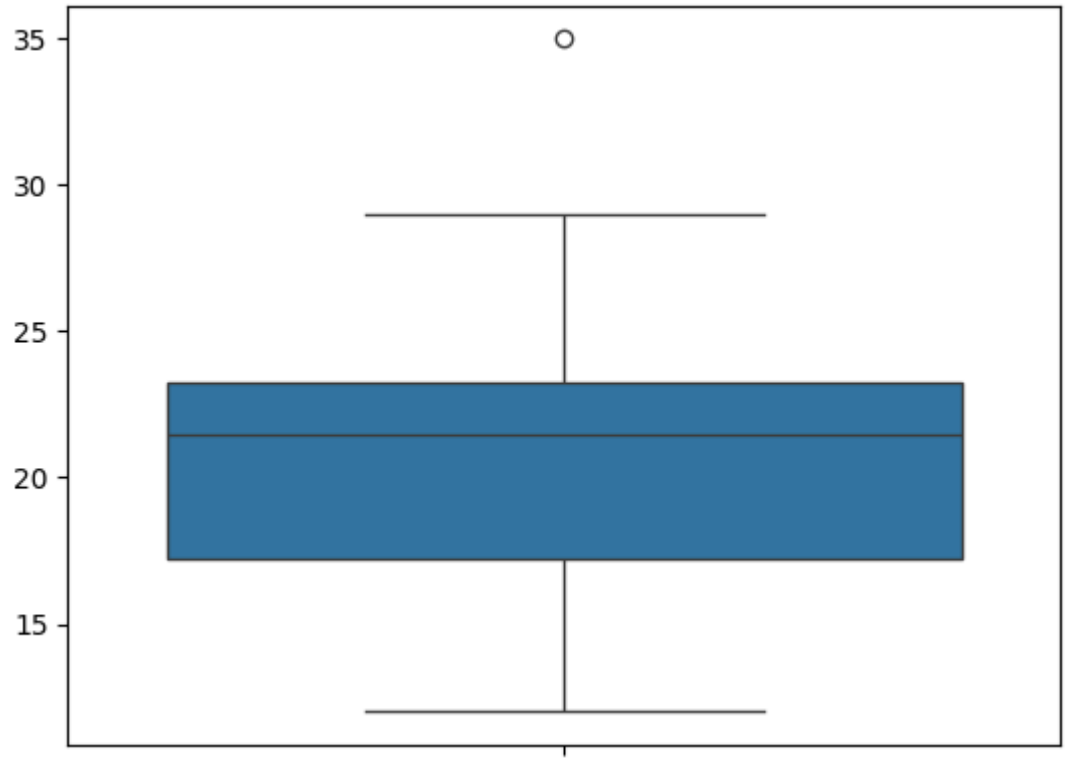
```
import numpy as np
list_x = [10, 20, 30, 40, 50]
list_y = [15, 25, 35, 45, 60]
x = np.array(list_x)
y = np.array(list_y)
cov_xy = np.cov(x, y)
covariance = cov_xy[0, 1]
correlation = np.corrcoef(x, y)[0, 1]
print(covariance)
print(correlation)

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]
Answer:-

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

```
sns.boxplot(data)
```

<Axes: >




```
#lower_bend = Q1-1.5(IQR)
#upper_bend = Q3+1.5(IQR)
#IQR = Q3-Q1
Q0 = np.percentile(data, 0)
Q1 = np.percentile(data, 25)
Q2 = np.percentile(data, 50)
Q3 = np.percentile(data, 75)
IQR = Q3-Q1
lower_bend = Q1-(1.5*IQR)
upper_bend = Q3+(1.5*IQR)
print(lower_bend)
print(upper_bend)
8.25
32.25
```

Any value beyond this upper and lower value is an outlier, therefore 35 is an outlier in the above dataset.

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]

Answers:-

#Covariance :-

- Covariance tells how two variables vary together.
- Positive covariance indicates direct relationship between variables.
- Negative covariance indicates inverse relationship between variables.

#Correlation :-

- Correlation measures the strength and direction of the linear relationship.
- Correlation is dimensionless quantity varies from -1 to 1.
- -1 indicates negative correlation
- 0 indicates no relationship
- 1 indicates positive relationship

Python Code:-

```
advertising_spend = [200, 250, 300, 400, 500]
daily_sales = [2200, 2450, 2750, 3200, 4000]
import numpy as np
ad = np.array(advertising_spend)
sales = np.array(daily_sales)

cov_matrix = np.cov(ad, sales, bias=False)
covariance = cov_matrix[0, 1]

correlation = np.corrcoef(ad, sales)[0, 1]

print(covariance)
print(correlation)
```

```
84875.0
0.9935824101653329
```

- The Positive covariance(84875) shows a positive relationship that is a direct relationship as advertising_spend increases daily_sales also increases.
- The correlation(0.9935824101653329)shows a positive linear relationship.

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:-

Summary Statistics to Use:

1. **Mean (Average):**
    - Tell you the average satisfaction level.
2. **Median:**
    - Helps understand the middle score — good when data is skewed.
3. **Mode:**
    - Shows the most common score (useful for ordinal data).
4. **Standard Deviation (SD):**
    - Tells how spread out the scores are — a high SD means opinions vary widely.
5. **Minimum and Maximum:**
    - Show the range of satisfaction.

#Histogram:-

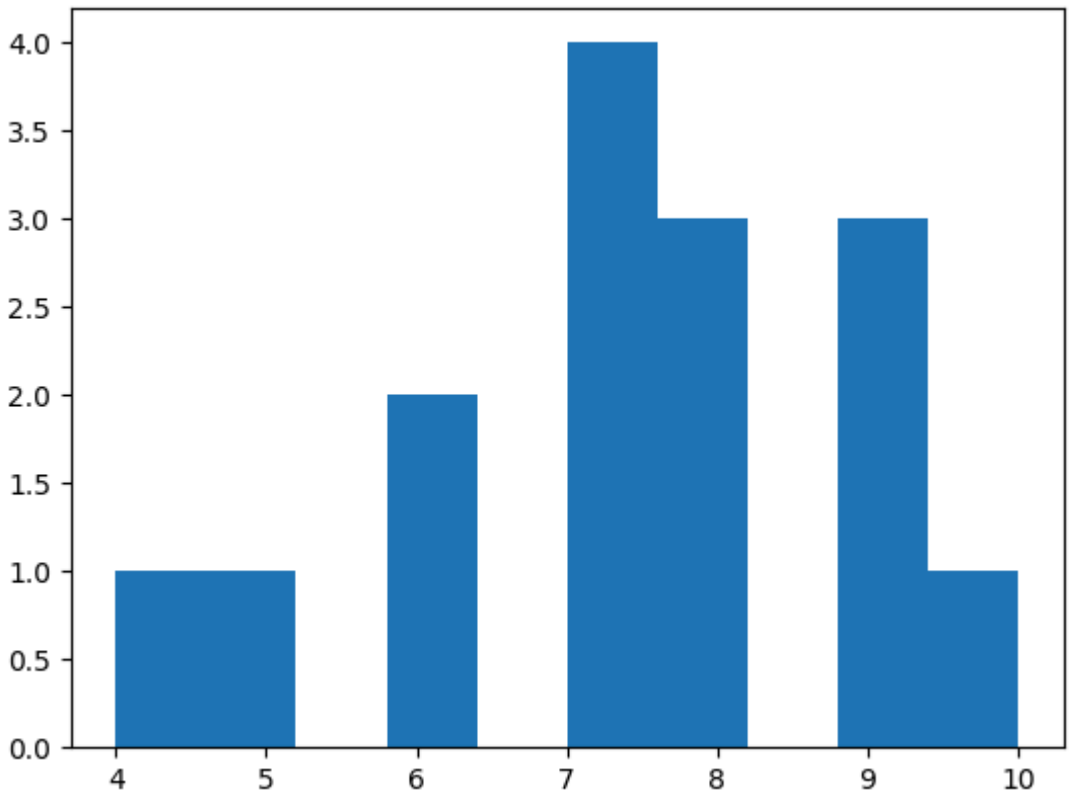
- Displays how frequently each score occurs (ideal for 1–10 rating data).

- Helps you **see the shape** (e.g. skewed, uniform, normal) of the distribution.

Python Code:-

```
import matplotlib.pyplot as plt

import pandas as pd

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

df = pd.DataFrame(survey_scores)

print(df.describe())

plt.hist(survey_scores, bins = 10 )

plt.show()
```

```
       0
count  15.000000
mean    7.333333
std     1.632993
min     4.000000
25%     6.500000
50%     7.000000
75%     8.500000
max    10.000000
```



Min, 25%, 50%, 75%, max is five important summary of statistics used to analyze the data.

Name:- Divya Raghav

Email:- divyaraghav1409@gmail.com