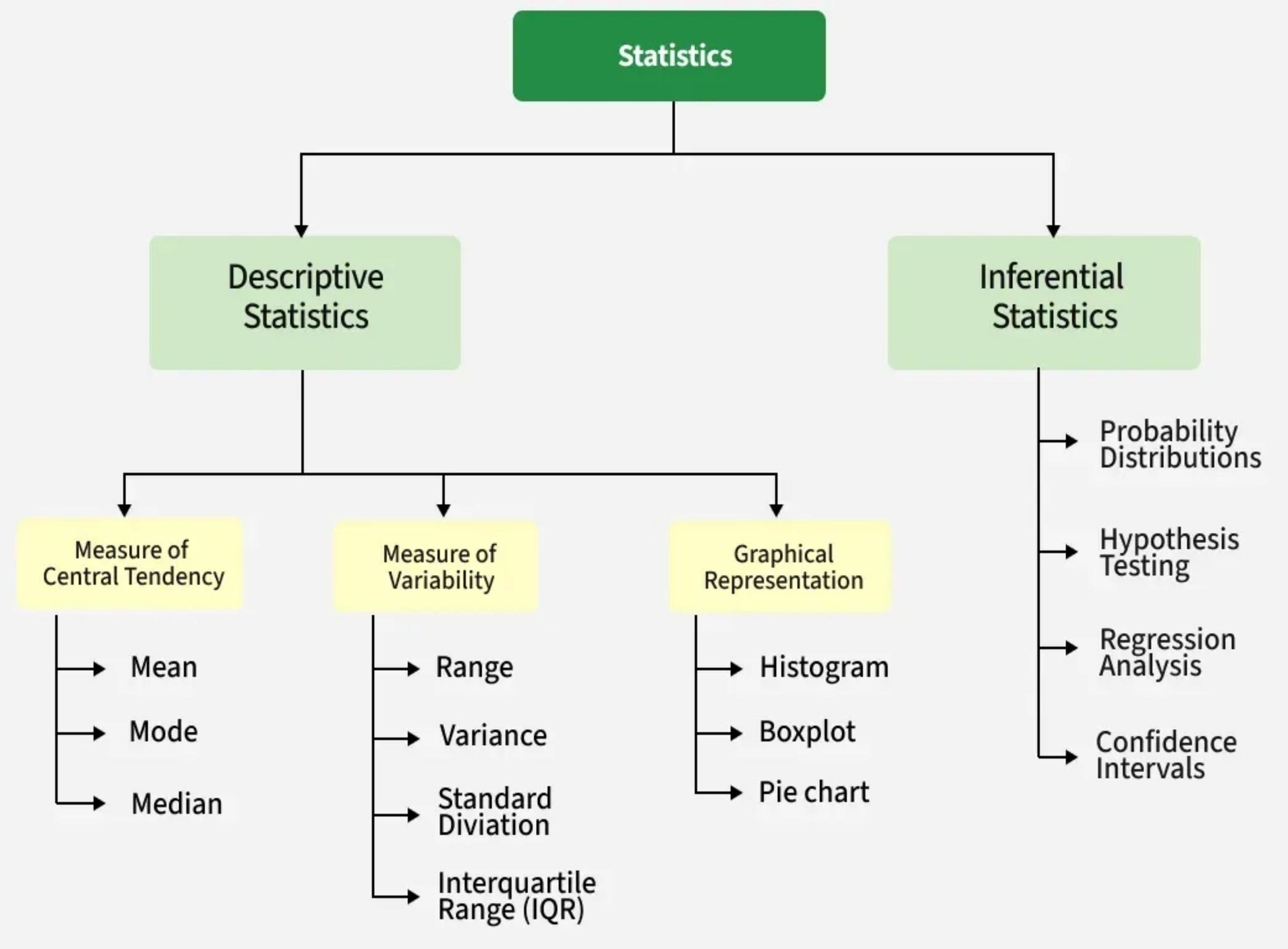
STATISTICS ASSIGNMENT

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

Answer :

Descriptive Statistics	**Inferential Statistics**
It gives information about raw data which describes the data in some manner.	It makes inferences about the population using data drawn from the population.
It helps in organizing, analyzing, and to present data in a meaningful manner.	It allows us to compare data, and make hypotheses and predictions.
It is used to describe a situation.	It is used to explain the chance of occurrence of an event.
It explains already known data and is limited to a sample or population having a small size.	It attempts to reach the conclusion about the population.
Examples include: mean, median, mode, range, variance, histograms, pie charts.	Examples include: confidence intervals, hypothesis testing, regression models, p-values.
Limited to presenting and analyzing known data.	Allows predictions and conclusions that go beyond the data at hand.

Descriptive Statistics	**Inferential Statistics**
Used for describing trends, organizing data for presentation.	Used for predicting trends, testing hypotheses, generalizing data from sample to population.



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

Answer :

Sampling in statistics refers to the process of selecting a subset (called a sample) from a larger population to analyze and draw conclusions about the entire population. It is often used when studying the entire population is impractical or impossible. The goal is to ensure that the sample is representative of the population to make accurate inferences.

Sampling is the process of selecting a subset of individuals, items, or data points from a larger population to estimate characteristics of the whole population.

Use of Sampling

- **Efficiency**: Studying a sample is faster and cheaper than studying the entire population.
- **Feasibility**: Sometimes it's impossible to access every member of a population.
- **Accuracy**: A well-chosen sample can provide reliable insights about the population.

Types of Sampling: Random vs. Stratified

Feature	Random Sampling	Stratified Sampling
Definition	Every member of the population has an equal chance of being selected.	Population is divided into subgroups (strata), and samples are taken from each.
Selection Process	Completely random, often using random number generators or lottery methods.	Random samples are drawn **within each stratum**.
Use Case	When the population is homogeneous or differences between groups are not important.	When the population has distinct subgroups and you want to ensure representation from each.
Example	Selecting 100 students randomly from a university.	Selecting 25 students from each year (freshman, sophomore, etc.) to ensure all years are represented.

Feature	Random Sampling	Stratified Sampling
Advantages	Simple and unbiased.	More accurate representation of population characteristics.
Disadvantages	May miss key subgroups or lead to unbalanced samples.	More complex and requires knowledge of population structure.

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

Answer :

Mean, Median, and Mode are measures of the central tendency. These values are used to define the various parameters of the given data set. The measure of central tendency (Mean, Median, and Mode) gives useful insights about the data studied, these are used to study any type of data such as the average salary of employees in an organization, the median age of any class, the number of people who plays cricket in a sports club, etc.

Measures of Central Tendency

The measure of central tendency is the representation of various values of the given data set. There are various measures of central tendency and the most important three measures of central tendency are:

- **Mean**
- **Median**
- **Mode**

What are Mean, Median, and Mode?

The mean, median, and mode are measures of central tendency used in statistics to summarize a set of data.

- **Mean (x or µ):** The mean, or arithmetic average, is calculated by summing all the values in a dataset and dividing by the total number of values. It's sensitive to outliers and is commonly used when the data is symmetrically distributed.
- **Median (M):** The median is the middle value when the dataset is arranged in ascending or descending order. If there's an even number of values, it's the average of the two middle values. The median is robust to outliers and is often used when the data is skewed.
- **Mode (Z):** The mode is the value that occurs most frequently in the dataset. Unlike the mean and median, the mode can be applied to both numerical and categorical data. It's useful for identifying the most common value in a dataset.

Mean Symbol

The symbol used to represent the mean, or arithmetic average, of a dataset is typically the Greek letter "µ" (mu) when referring to the population mean, and "x̄" (x-bar) when referring to the sample mean.

- *Population Mean:* ***µ (mu)***
- *Sample Mean:* ***x̄ (x-bar)***

These symbols are commonly used in statistical notation to represent the average value of a set of data points.

Mean Formula

The formula to calculate the mean is:

Mean (x) = Σxi/ n

If x1, x2, x3,......, xn are the values of a data set then the mean is calculated as:

x = (x1 + x2 + x3 + . . . + xn) / n

Example: Find the mean of data sets 10, 30, 40, 20, and 50.

Solution:

Mean of the data 10, 30, 40, 20, 50 is
Mean = (sum of all values) / (number of values)
Mean = (10 + 30 + 40 + 20+ 50) / 5 = 30

Median Symbol

The letter "M" is commonly used to represent the median of a dataset, whether it's for a population or a sample. This notation simplifies the representation of statistical concepts and calculations, making it easier to understand and apply in various contexts. Therefore, in Indian statistical practice, "M" is widely accepted and understood as the symbol for the median.

Median Formula

The formula for the median is:

If the number of values (n value) in the data set is odd, then the formula to calculate the median is:

Median = [(n + 1)/2]th term

If the number of values (n value) in the data set is even, then the formula to calculate the median is:

Median = [(n/2)th term + {(n/2) + 1}th term] / 2

Example: Find the median of the the given data set 30, 40, 10, 20, and 50.

Solution:

Median of the data 30, 40, 10, 20, 50 is,

Step 1: Order the given data in ascending order as: 10, 20, 30, 40, 50

Step 2: Check n (number of terms of data set) is even or odd and find the median of the data with respective 'n' value.

Step 3: Here, n = 5 (odd)

Median = [(n + 1)/2]th term Median = [(5 + 1)/2]th term = 30

Symbol of Mode

In statistical notation, the symbol "Z" is commonly used to represent the mode of a dataset. It indicates the value or values that occur most frequently within the dataset. This symbol is widely utilized in statistical

discourse to signify the mode, enhancing clarity and precision in statistical discussions and analyses.

Mode Formula

Mode = Highest Frequency Term

Example: Find the mode of the given data set 1, 2, 2, 2, 3, 3, 4, 5.

Solution:

Given set is {1, 2, 2, 2, 3, 3, 4, 5}

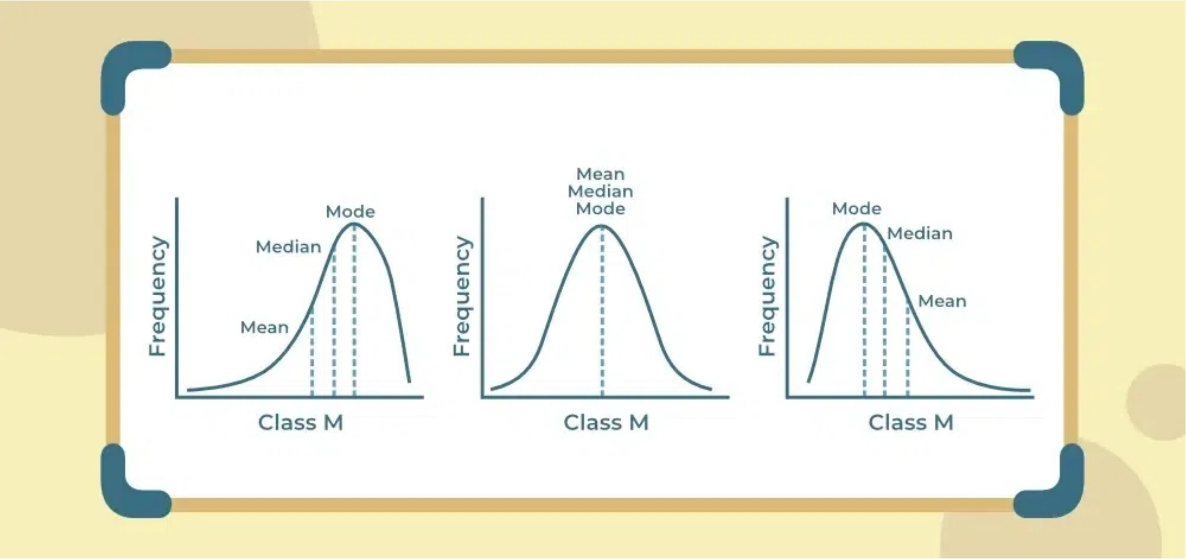
As the above data set is arranged in ascending order.

By observing the above data set we can say that,

Using the formula Mode = Highest Frequency Term

Mode = 2

As, it has highest frequency (3)



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

Answer :

Skewness: Measuring Asymmetry

Skewness tells us how symmetrical (or asymmetrical) a distribution is.

- **Zero skewness**: The distribution is perfectly symmetrical (like a normal bell curve).
- **Positive skew (right-skewed)** : The tail on the **right side** is longer or fatter than the left. Most data points are concentrated on the left, with a few large values pulling the mean to the right.
- **Negative skew (left-skewed)** : The tail on the **left side** is longer or fatter. Most data points are on the right, with a few small values pulling the mean to the left.

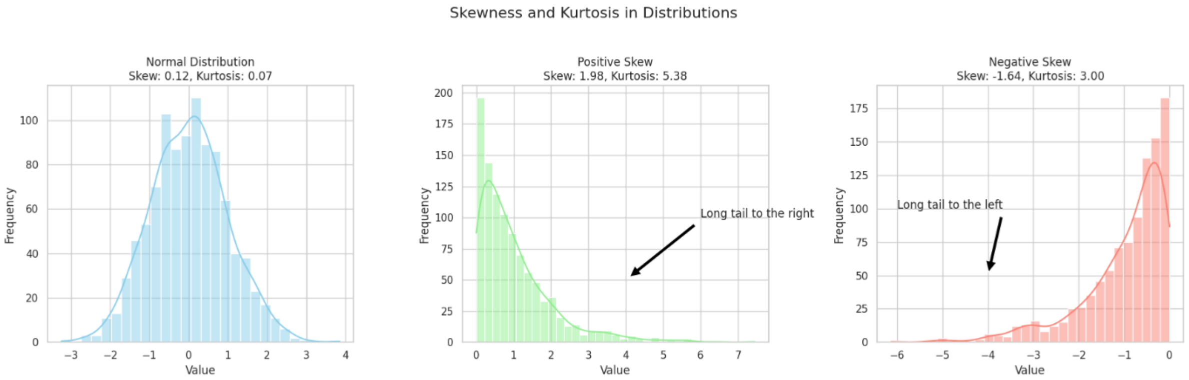
Kurtosis: Measuring Taile dness

Kurtosis describes the **height and sharpness of the peak** and the **heaviness of the tails** of a distribution.

- **Mesokurtic (kurtosis ≈ 3)**: Normal distribution.

- **Leptokurtic (kurtosis > 3)**: Sharper peak and fatter tails. More outliers.
- **Platykurtic (kurtosis < 3)**: Flatter peak and thinner tails. Fewer outliers.
- **Why is kurtosis useful?**
- It helps assess the **risk of extreme values**. For example, in finance, leptokurtic distributions suggest higher chances of big gains or losses.
- **Summary Table**

Feature	Skewness	Kurtosis
Definition	Asymmetry of distribution	Tailedness and peak sharpness
Positive	Tail to the right, mean > median	More extreme values (leptokurtic)
Negative	Tail to the left, mean < median	Fewer extreme values (platykurtic)
Zero	Symmetrical distribution	Normal distribution (mesokurtic)



What the Graph Shows

1. **Normal Distribution (Center)**
   - Symmetrical bell curve.
   - **Skew ≈ 0**, **Kurtosis ≈ 0**.
   - Mean ≈ Median ≈ Mode.
2. **Positive Skew (Right)**
   - Tail extends to the **right**.

   - **Skew > 0**.
   - Mean > Median.
   - Indicates presence of **high-value outliers** .
   - Example: Income distribution, where most earn modest amounts but a few earn very high salaries.

3. **Negative Skew (Left)**
   - Tail extends to the **left**.
   - **Skew < 0**.
   - Mean < Median.
   - Indicates presence of **low-value outliers** .

What Does a Positive Skew Imply?

- Most data points are **clustered at lower values** .
- A few **extremely high values** stretch the tail to the right.
- The **mean is pulled higher** than the median.
- Useful in identifying **asymmetric risks** or **imbalanced distributions** .

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 :

: 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]

```
from statistics import mean, median, mode


# List of numbers

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

```
# Calculating mean, median, and mode
mean_value = mean(numbers)
median_value = median(numbers)
mode_value = mode(numbers)

# Displaying the results
print(f"Mean: {mean_value}")
print(f"Median: {median_value}")
print(f"Mode: {mode_value}")
```

output

```
Mean: 19.6
Median: 19
Mode: 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]

Covariance measures how two variables change together.

Correlation coefficient (Pearson's r) standardizes covariance to a value between -1 and 1, indicating the strength and direction of a linear relationship.

```
import numpy as np

# Given lists
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, ddof=1)
covariance = cov_matrix[0, 1]

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

print(f"Covariance: {covariance}")
print(f"Correlation Coefficient: {correlation}")
```

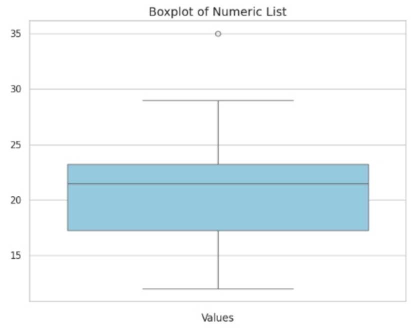
output

```
Covariance: 250.0
Correlation Coefficient: 0.9938586931957764
```

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 :



Boxplot Summary

- The boxplot visually displays the **median**, **quartiles**, and **potential outliers**.
- The central box represents the **interquartile range (IQR)**, which contains the middle 50% of the data.
- The **whiskers** extend to the smallest and largest values within 1.5 × IQR from the quartiles.
- Any points **outside the whiskers** are considered **outliers**.

Identified Outliers

Outliers: [35]

Explanation

- The **IQR** method flagged 35 as an outlier because it lies beyond the upper bound:
  - **Q1 (25th percentile)** ≈ 18
  - **Q3 (75th percentile)** ≈ 24.5
  - **IQR** = Q3 − Q1 = 6.5
  - **Upper bound** = Q3 + 1.5 × IQR = 24.5 + 9.75 = 34.25
  - Since 35 > 34.25, it's an outlier.

This suggests that while most of the data are fairly clustered, 35 is unusually high compared to the rest.

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 :

Using Covariance and Correlation

Covariance

- Measures how two variables change together.
- **Positive covariance** : When advertising spend increases, daily sales tend to increase.
- **Negative covariance** : When advertising spend increases, daily sales tend to decrease.
- **Limitation** : Covariance is **not standardized** , so it's hard to interpret magnitude.

Correlation Coefficient (Pearson's r)

- Standardizes covariance to a value between **-1 and 1** .
- **+1**: Perfect positive linear relationship.
- **0**: No linear relationship.
- **−1**: Perfect negative linear relationship.
- More interpretable than covariance.

Python Code to Compute Correlation

```
import numpy as np


# Given data

advertising_spend = [200, 250, 300, 400, 500]
```

```
daily_sales = [2200, 2450, 2750, 3200, 4000]

# Convert to numpy arrays
ad_spend = np.array(advertising_spend)
sales = np.array(daily_sales)

# Compute covariance matrix
cov_matrix = np.cov(ad_spend, sales, ddof=1)
covariance = cov_matrix[0, 1]

# Compute correlation coefficient
correlation = np.corrcoef(ad_spend, sales)[0, 1]

print(f"Covariance: {covariance}")
print(f"Correlation Coefficient: {correlation}")
```

output

```
Covariance: 190000.0
Correlation Coefficient: 0.9970544855015816
```

Covariance = 190,000 : Strong positive relationship.

Correlation ≈ 0.997: Very close to +1, indicating a **strong linear relationship** .

This suggests that **increased advertising spend is strongly associated with higher daily sales** .

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 :

The histogram has been successfully created and saved. It visually represents the distribution of customer satisfaction scores

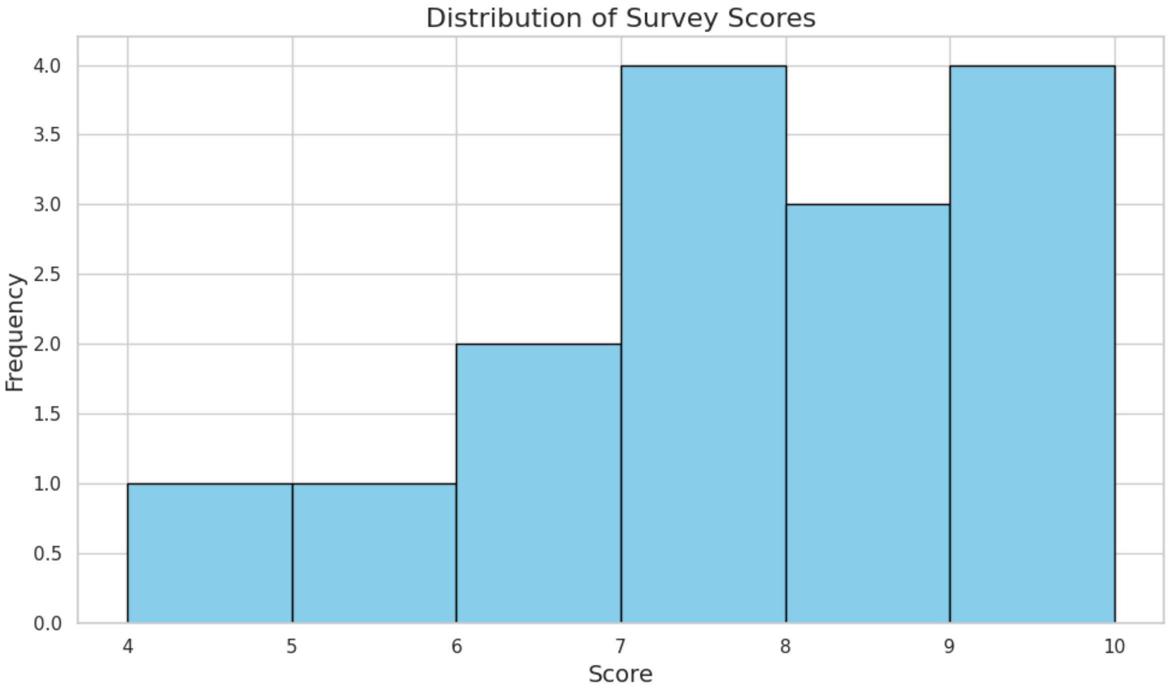
To understand the distribution of survey scores, you'd typically analyze:

- **Mean**: Average score – gives a central tendency.
- **Median**: Middle score – useful if data is skewed.
- **Mode**: Most frequent score – shows common sentiment.
- **Standard Deviation**: Measures variability – how spread out the scores are.
- **Range**: Difference between highest and lowest scores.

CODE

```
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 the histogram
plt.hist(survey_scores, bins=7, color='skyblue', edgecolor='black')

# Add labels and title
plt.xlabel('Survey Scores')
plt.ylabel('Frequency')
plt.title('Histogram of Survey Scores')

# Show the plot
```

```
plt.show()
```



Explanation:

- **bins=7**: Divides the data into 7 intervals. You can adjust this based on your preference.
- **color='skyblue'** : Sets the bar color for better visualization.
- **edgecolor='black'** : Adds a border to the bars for clarity.
- **plt.show()** : Displays the histogram.
- Most scores are clustered between **6 and 9**, indicating generally **high satisfaction** .
- The **peak** is around **7 and 8**, suggesting these are the most common ratings.
- A few scores at **4 and 10** show some **variation**, but no extreme outliers.

This distribution suggests that customers are **mostly satisfied**, which is a promising sign before launching the new product.

THANKU SO MUCH …