1. Scenario: A company wants to analyse the sales performance of its products in different regions. They have collected the following data:
Region A: [10, 15, 12, 8, 14]
Region B: [18, 20, 16, 22, 25]
Calculate the mean sales for each region.
ANS :-->To calculate the mean sales for each region, you need to sum up the sales values for each region and divide by the number of data points in that region.

Let's calculate the mean sales for each region:
Region A: [10, 15, 12, 8, 14]
Mean sales for Region A = (10 + 15 + 12 + 8 + 14) / 5
= 59 / 5
= 11.8
Therefore, the mean sales for Region A is 11.8.
Region B: [18, 20, 16, 22, 25]
Mean sales for Region B = (18 + 20 + 16 + 22 + 25) / 5
= 101 / 5
= 20.2
Therefore, the mean sales for Region B is 20.2.
So, the mean sales for Region A is 11.8, and the mean sales for Region B is 20.2.

—--

2. Scenario: A survey is conducted to measure customer satisfaction on a scale of 1 to 5. The data collected is as follows:
[4, 5, 2, 3, 5, 4, 3, 2, 4, 5]
Calculate the mode of the survey responses.
ANS :To calculate the mode of the survey responses, you need to find the value(s) that appear most frequently in the data set.Let's calculate the mode of the survey responses: [4, 5, 2, 3, 5, 4, 3, 2, 4, 5]In this data set, the values 4 and 5 both appear three times, which is more frequently than any other value. Therefore, the mode of the survey responses is 4 and 5.So, the mode of the survey responses [4, 5, 2, 3, 5, 4, 3, 2, 4, 5] is 4 and 5.

—--

3. Scenario: A company wants to compare the salaries of two departments. The salary data for Department A and Department B are as follows:
Department A: [5000, 6000, 5500, 7000]
Department B: [4500, 5500, 5800, 6000, 5200]
Calculate the median salary for each department.
ANS :To calculate the median salary for each department, you need to arrange the salary data in ascending order and find the middle value(s). If there is an odd number of data points, the median is the middle value. If there is an even number of data points, the median is the average of the two middle values.
Let's calculate the median salary for each department:
Department A: [5000, 6000, 5500, 7000]
Arrange the salaries in ascending order: [5000, 5500, 6000, 7000]

Since there is an odd number of data points, the median is the middle value, which is 6000.
Therefore, the median salary for Department A is 6000.
Department B: [4500, 5500, 5800, 6000, 5200]
Arrange the salaries in ascending order: [4500, 5200, 5500, 5800, 6000]
Since there is an odd number of data points, the median is the middle value, which is 5500.
Therefore, the median salary for Department B is 5500.
So, the median salary for Department A is 6000, and the median salary for Department B is 5500.

—---

4. Scenario: A data analyst wants to determine the variability in the daily stock prices of a company. The data collected is as follows:
[25.5, 24.8, 26.1, 25.3, 24.9]
Calculate the range of the stock prices.
ANS :To calculate the range of the stock prices, you need to find the difference between the maximum and minimum values in the data set.
Let's calculate the range of the stock prices: [25.5, 24.8, 26.1, 25.3, 24.9]
Arrange the stock prices in ascending order: [24.8, 24.9, 25.3, 25.5, 26.1]
The minimum value is 24.8, and the maximum value is 26.1.
Range = Maximum value - Minimum value
= 26.1 - 24.8
= 1.3
Therefore, the range of the stock prices is 1.3.

—---

5. Scenario: A study is conducted to compare the performance of two different teaching methods. The test scores of the students in each group are as follows:
Group A: [85, 90, 92, 88, 91]
Group B: [82, 88, 90, 86, 87]
Perform a t-test to determine if there is a significant difference in the mean scores between the two groups.
ANS :

To perform a t-test to compare the mean scores between Group A and Group B, we can follow these steps:

Step 1: Define the null hypothesis (H0) and alternative hypothesis (H1):

- Null hypothesis (H0): There is no significant difference in the mean scores between Group A and Group B.
- Alternative hypothesis (H1): There is a significant difference in the mean scores between Group A and Group B.

Step 2: Calculate the means of Group A and Group B:

- Group A: [85, 90, 92, 88, 91] Mean of Group A (x̄1) = (85 + 90 + 92 + 88 + 91) / 5 = 88.4
- Group B: [82, 88, 90, 86, 87] Mean of Group B (x̄2) = (82 + 88 + 90 + 86 + 87) / 5 = 86.6

Step 3: Calculate the standard deviations of Group A and Group B:

- Group A: Standard deviation of Group A (s1) = √(Σ((x - x̄1)^2) / (n1 - 1)) = √(((85-88.4)^2 + (90-88.4)^2 + (92-88.4)^2 + (88-88.4)^2 + (91-88.4)^2) / (5 - 1)) = √((7.84 + 0.16 + 9.76 + 0.16 + 6.76) / 4) = √(24.68 / 4) = √6.17 ≈ 2.48
- Group B: Standard deviation of Group B (s2) = √(Σ((x - x̄2)^2) / (n2 - 1)) = √(((82-86.6)^2 + (88-86.6)^2 + (90-86.6)^2 + (86-86.6)^2 + (87-86.6)^2) / (5 - 1)) = √((18.04 + 1.36 + 7.84 + 0.16 + 0.04) / 4) = √(27.44 / 4) = √6.86 ≈ 2.62

Step 4: Calculate the t-statistic:

- t = (x̄1 - x̄2) / √((s1^2 / n1) + (s2^2 / n2)) = (88.4 - 86.6) / √((2.48^2 / 5) + (2.62^2 / 5)) = 1.8 / √(6.1456/5 + 6.8644/5) = 1.8 / √(1.22912 + 1.37288) = 1.8 / √2.602 = 1.8 / 1.611 ≈ 1.117

Step 5: Determine the degrees of freedom (df):

- Degrees of freedom (df) = (n1 + n2 - 2) = (5 + 5 - 2) = 8

Step 6: Determine the critical value for the desired significance level (α):

- Assuming a significance level of 0.05, the critical value for a two-tailed t-test with 8 degrees of freedom is approximately ±2.306.

Step 7: Compare the calculated t-statistic with the critical value:

- If the absolute value of the calculated t-statistic is greater than the critical value, we reject the null hypothesis.
- If the absolute value of the calculated t-statistic is less than or equal to the critical value, we fail to reject the null hypothesis.
- In this case, the calculated t-statistic (1.117) is less than the critical value (2.306). Therefore, we fail to reject the null hypothesis. This means that there is no significant difference in the mean scores between Group A and Group B based on the given data.

—---

6. Scenario: A company wants to analyse the relationship between advertising expenditure and sales. The data collected is as follows:
Advertising Expenditure (in thousands): [10, 15, 12, 8, 14]
Sales (in thousands): [25, 30, 28, 20, 26]
Calculate the correlation coefficient between advertising expenditure and sales.
ANs:To calculate the correlation coefficient between advertising expenditure and sales, we'll use the provided data. Let's label the Advertising Expenditure as variable X and Sales as variable Y.

Given data:
X = [10, 15, 12, 8, 14]
Y = [25, 30, 28, 20, 26]
Calculate mean for X and Y mean_x=11.8 mean_y=25.8

$[10-11.8,15-11.8,12-11.8,8-11.8,14-11.8]=[-1.8,3.2,0.2,-3.8,2.2]$

$[25-25.8,30-25.8,28-25.8,20-25.8,26-25.8]=[-0.8,4.2,2.2,-5.8,0.2]$

$[25-25.8,30-25.8,28-25.8,20-25.8,26-25.8]=[-0.8,4.2,2.2,-5.8,0.2]$The product of the deviations is calculated by multiplying each deviation of X (x) with the corresponding deviation of Y (y).

$[-1.8\times-0.8,3.2\times4.2,0.2\times2.2,-3.8\times-5.8,2.2\times0.2]=[1.44,13.44,0.44,22.04,0.4]$

$[-1.8\times-0.8,3.2\times4.2,0.2\times22,-3.8\times-5.8,2.2\times0.2]=[1.44,13.44,0.44,22.04,0.4$

$=1.44+13.44+0.44+22.04+0.4432.8\cdot57.8=37.81890.64\approx37.843.47\approx0.869$

$=32.8\cdot57.8$

$1.44+13.44+0.44+22.04+0.44$

$=37.8/\mathrm{sqrt}(1890.64)$

≈0.869

7. Scenario: A survey is conducted to measure the heights of a group of people. The data collected is as follows:
[160, 170, 165, 155, 175, 180, 170]
Calculate the standard deviation of the heights.
ANS:

To calculate the standard deviation of the heights based on the given data, you can follow these steps:
Step 1: Calculate the mean (average) of the heights.

$\bar{x} = 167.85$

[160−166.43,170−166.43,165−166.43,155−166.43,

175−166.43,180−166.43,170−166.43]

=[−6.43,3.57,−1.43,−11.43,8.57,13.57,3.57]

The squared deviations are:

[(−6.43)2,(3.57)2,(−1.43)2,(−11.43)2,(8.57)2,(13.57)2,(3.57)2]=[41.3449,12.7449,2.0449,130.4649,73.3249,184.3249,12.7449]

=[41.3449,12.7449,2.0449,130.4649,73.3249,184.3249,12.7449]

Variance:
457.99847=65.42834

=65.42834
Standard Deviation:
sqrt(65.42834)≈8.106

≈8.106

8. Scenario: A company wants to analyse the relationship between employee tenure and job satisfaction. The data collected is as follows:
Employee Tenure (in years): [2, 3, 5, 4, 6, 2, 4]
Job Satisfaction (on a scale of 1 to 10): [7, 8, 6, 9, 5, 7, 6]
Perform a linear regression analysis to predict job satisfaction based on employee tenure
.
ANS :step1: calculate mean of X(employee tenure) and Y(job satisfaction) which is To perform a linear regression analysis to predict job satisfaction based on employee tenure, we can use the given data points and apply the following steps:

Step 1: Organize the data
Let's denote the employee tenure as X and job satisfaction as Y. We can organize the data as follows:

Employee Tenure (in years): [2, 3, 5, 4, 6, 2, 4]

Job Satisfaction (on a scale of 1 to 10): [7, 8, 6, 9, 5, 7, 6]

X = [2, 3, 5, 4, 6, 2, 4]
Y = [7, 8, 6, 9, 5, 7, 6]

Step 2: Calculate the means of X and Y
Calculate the mean of X (X̄) and the mean of Y (Ȳ).

X̄ = (2 + 3 + 5 + 4 + 6 + 2 + 4) / 7 = 4
Ȳ = (7 + 8 + 6 + 9 + 5 + 7 + 6) / 7 = 6.857

Step 3: Calculate the deviations from the mean
Calculate the deviations from the mean for both X and Y. Denote the deviations as (Xi - X̄) and (Yi - Ȳ) respectively.

Deviation from the mean for X:
(-2, -1, 1, 0, 2, -2, 0)

Deviation from the mean for Y:
(0.143, 1.143, -0.857, 2.143, -1.857, 0.143, -0.857)

Step 4: Calculate the product of deviations
Calculate the product of deviations for each data point (Xi - X̄) * (Yi - Ȳ).

Product of deviations:
(-2 * 0.143, -1 * 1.143, 1 * -0.857, 0 * 2.143, 2 * -1.857, -2 * 0.143, 0 * -0.857)
(-0.286, -1.143, -0.857, 0, -3.714, -0.286, 0)

Step 5: Calculate the squared deviations
Calculate the squared deviations for X (Xi - X̄)^2 and Y (Yi - Ȳ)^2.

Squared deviations for X:
(4, 1, 1, 0, 4, 4, 0)

Squared deviations for Y:
(0.0204, 1.3069, 0.7344, 4.5919, 3.4489, 0.0204, 0.7344)

Step 6: Calculate the sum of squared deviations
Calculate the sum of squared deviations for X, Y, and the product of deviations.
Sum of squared deviations for X:
4 + 1 + 1 + 0 + 4 + 4 + 0 = 14
Sum of squared deviations for Y:
0.0204 + 1.3069 + 0.7344 + 4.5919 + 3.4489 + 0.0204 + 0.7344 = 10.8573
Sum of the product of deviations:
-0.286 + -1.143 + -0.857 + 0 + -3.714 + -0.286 + 0 = -6.286
Step 7: Calculate the slope (β1)
The slope (β1) of the linear regression line can be calculated using the formula:
β1 = Σ((Xi - X̄) * (Yi - Ȳ)) / Σ((Xi - X̄)^2)

β1 = -6.286 / 14 = -0.448
Step 8: Calculate the intercept (β0)
The intercept (β0) of the linear regression line can be calculated using the formula:
β0 = Ȳ - β1 * X̄
β0 = 6.857 - (-0.448 * 4) = 8.137
Step 9: Write the regression equation
The regression equation can be written as:
Y = β0 + β1 * X
Y = 8.137 - 0.448 * X
Now, you can use this equation to predict job satisfaction (Y) based on employee tenure (X).
—--
9. Scenario: A study is conducted to compare the effectiveness of two different medications. The recovery times of the patients in each group are as follows:
  Medication A: [10, 12, 14, 11, 13]
  Medication B: [15, 17, 16, 14, 18]
  Perform an analysis of variance (ANOVA) to determine if there is a significant difference in the mean recovery times between the two medications.
ANS:To perform an analysis of variance (ANOVA) to compare the mean recovery times between Medication A and Medication B, you can follow these steps:

Step 1: Define the null and alternative hypotheses:
  - Null hypothesis (H0): There is no significant difference in the mean recovery times between Medication A and Medication B.
  - Alternative hypothesis (HA): There is a significant difference in the mean recovery times between Medication A and Medication B.

Step 2: Calculate the necessary statistics:
  - Calculate the sample means for each group:
    Mean A = (10 + 12 + 14 + 11 + 13) / 5 = 12
    Mean B = (15 + 17 + 16 + 14 + 18) / 5 = 16

  - Calculate the sum of squares within groups (SSW):
    SSW = Σ(xi - x̄i)^2
        = (10-12)^2 + (12-12)^2 + (14-12)^2 + (11-12)^2 + (13-12)^2
          + (15-16)^2 + (17-16)^2 + (16-16)^2 + (14-16)^2 + (18-16)^2
        = 10 + 0 + 2 + 1 + 1 + 1 + 1 + 0 + 4 + 4
        = 24

  - Calculate the sum of squares between groups (SSB):
    SSB = nA * (Mean A - Grand Mean)^2 + nB * (Mean B - Grand Mean)^2
        = 5 * (12 - 14)^2 + 5 * (16 - 14)^2
        = 5 * 4 + 5 * 4
        = 40

  - Calculate the degrees of freedom:
    dfw = N - k
        = 10 - 2

= 8 (where N is the total number of observations and k is the number of groups)

dfb = k - 1
= 2 - 1
= 1 (where k is the number of groups)

- Calculate the mean squares:
MSW = SSW / dfw
= 20/8
= 2.5

MSB = SSB / dfb
= 40 / 1
= 40

Step 3: Calculate the F-statistic:
F = MSB / MSW
= 40 / 2.5
=16

Step 4: Determine the critical value:
The critical value for the F-distribution depends on the chosen significance level (alpha) and the degrees of freedom (dfb and dfw). Let's assume alpha = 0.05.

Using an F-table or statistical software, the critical value for alpha = 0.05 with dfb = 1 and dfw = 8 is approximately 5.32.

Step 5: Compare the F-statistic with the critical value:
Since F (16)is greater than the critical value (5.32), we can reject the null hypothesis.

Step 6: Interpret the results:
There is sufficient evidence to suggest that there is a significant difference in the mean recovery times between Medication A and Medication B.

```
import scipy.stats as stats

# Define the data for each group
group_A = [10, 12, 14, 11, 13]
group_B = [15, 17, 16, 14, 18]

print("Let assume null hypotheis is two groups are same")
# Perform one-way ANOVA
f_value, p_value = stats.f_oneway(group_A, group_B)

# Print the results
print("F-value:", f_value)
print("p-value:", p_value)
if f_value>p_value:
    print("We reject the Null hypothesis")
else:
    print("We fail to reject null hypothesis")
```

—---

10. Scenario: A company wants to analyse customer feedback ratings on a scale of 1 to 10. The data collected is
as follows:
[8, 9, 7, 6, 8, 10, 9, 8, 7, 8]
Calculate the 75th percentile of the feedback ratings.

```
ratings = [8, 9, 7, 6, 8, 10, 9, 8, 7, 8]

sorted_ratings=sorted(ratings)

index=(75/100)*(len(sorted_ratings)-1)

if index.is_integer():
    percentile_75th=sorted_ratings(int[index])
else:
    lower=int(index)
    upper=lower+1
    percentile_75th=(sorted_ratings[lower]+sorted_ratings[upper])/2
print(percentile_75th)
```

—---

11. Scenario: A quality control department wants to test the weight consistency of a product. The weights of a sample of products are as follows:
[10.2, 9.8, 10.0, 10.5, 10.3, 10.1]
Perform a hypothesis test to determine if the mean weight differs significantly from 10 grams.

ANS :To perform a hypothesis test to determine if the mean weight differs significantly from 10 grams, you can follow these steps:

Step 1: Define the null and alternative hypotheses:
- Null hypothesis (H0): The mean weight is equal to 10 grams.
- Alternative hypothesis (HA): The mean weight differs significantly from 10 grams.

Step 2: Calculate the necessary statistics:
- Calculate the sample mean:
  $\bar{x} = (10.2 + 9.8 + 10.0 + 10.5 + 10.3 + 10.1) / 6 = 10.15$

- Calculate the sample standard deviation (s):
  $s = sqrt(\Sigma(xi - \bar{x})^2 / (n - 1))$
  $= sqrt((10.2 - 10.15)^2 + (9.8 - 10.15)^2 + (10.0 - 10.15)^2$
  $+ (10.5 - 10.15)^2 + (10.3 - 10.15)^2 + (10.1 - 10.15)^2 / (6 - 1))$
  ≈ 0.243

- Calculate the test statistic (t-value):
  $t = (\bar{x} - \mu) / (s / sqrt(n))$

= (10.15 - 10) / (0.234 / sqrt(6))
≈ 1.3859

- Calculate the degrees of freedom:
  df = n - 1 = 6 - 1 = 5 (where n is the sample size)

Step 3: Determine the critical value:
The critical value for a two-tailed t-test with alpha = 0.05 and df = 5 can be obtained from a t-table or statistical software. For this example, the critical value is approximately ±2.571.

Step 4: Compare the test statistic with the critical value:
Since the absolute value of the test statistic (1.3859) is less than the critical value (2.571), we fail to reject the null hypothesis.

Step 5: Interpret the results:
Based on the hypothesis test, there is not enough evidence to conclude that the mean weight significantly differs from 10 grams at a significance level of 0.05.

Python code:import scipy.stats as stat

```
import scipy.stats as stat
import numpy as np
'Null hypothesis=>Ho=Mean weight =10 gm'
'Alternate hypothesis is Mean weight is not equal to 10gms'
weights=[10.2, 9.8, 10.0, 10.5, 10.3, 10.1]
t_stats,p_value=stat.ttest_1samp(weights,10)
print(t_stats)
print(p_value)
if p_value < 0.05 :
  print("We accept the Null hypothesis")
else:
  print("We failed to reject Null Hypothesis")
```

```
1.5126584522688367
0.19077595151110102
 We failed to reject Null Hypothesis
```

—--

12. Scenario: A company wants to analyse the click-through rates of two different website designs. The number of clicks for each design is as follows:

Design A: [100, 120, 110, 90, 95]
Design B: [80, 85, 90, 95, 100]

Perform a chi-square test to determine if there is a significant difference in the click-through rates between the two designs.

ANS:To perform a chi-square test to determine if there is a significant difference in the click-through rates between two designs, we need to follow these steps:

Step 1: State the null hypothesis (H0) and the alternative hypothesis (H1):
- Null Hypothesis (H0): There is no significant difference in the click-through rates between Design A and Design B.
- Alternative Hypothesis (H1): There is a significant difference in the click-through rates between Design A and Design B.

Step 2: Define the significance level (α). Let's assume α = 0.05, which is a common choice.

Step 3: Calculate the expected frequencies. To do this, we first need to calculate the overall click-through rate for each design and the total number of clicks for both designs.

- Design A: [100, 120, 110, 90, 95]
  Total clicks in Design A = 100 + 120 + 110 + 90 + 95 = 515
  Click-through rate for Design A = 515 / 5 = 103

- Design B: [80, 85, 90, 95, 100]
  Total clicks in Design B = 80 + 85 + 90 + 95 + 100 = 450
  Click-through rate for Design B = 450 / 5 = 90

- Total clicks for both designs = 515 + 450 = 965

Now we can calculate the expected frequencies for each design by multiplying the click-through rate of each design by the total clicks:
- Expected frequency for Design A = (103 / 100) * 965 = 992.95 (rounded to 2 decimal places)
- Expected frequency for Design B = (90 / 100) * 965 = 868.50 (rounded to 2 decimal places)

Step 4: Calculate the chi-square test statistic. We can use the formula:

$X^2 = \sum ((O - E)^2 / E)$

where:
- X^2 is the chi-square test statistic
- O is the observed frequency (actual number of clicks for each design)
- E is the expected frequency (calculated in Step 3)

Using the given data, we have:
- Observed frequencies for Design A: [100, 120, 110, 90, 95]
- Observed frequencies for Design B: [80, 85, 90, 95, 100]
- Expected frequencies for Design A: [992.95, 992.95, 992.95, 992.95, 992.95]
- Expected frequencies for Design B: [868.50, 868.50, 868.50, 868.50, 868.50]

Calculating the chi-square test statistic:

$X^2 = ((100 - 992.95)^2 / 992.95) + ((120 - 992.95)^2 / 992.95) + ((110 - 992.95)^2 / 992.95) + ((90 - 992.95)^2 / 992.95) + ((95 - 992.95)^2 / 992.95)$
$+ ((80 - 868.50)^2 / 868.50) + ((85 - 868.50)^2 / 868.50) + ((90 - 868.50)^2 / 868.50) + ((95 - 868.50)^2 / 868.50) + ((100 - 868.50)^2 / 868.50)$

Step 5: Determine the degrees of freedom (df). For a chi-square test comparing two designs, the degrees of freedom can be calculated as (number of categories - 1) * (number of groups - 1). In this case, we have (5 - 1) * (2 - 1) = 4.

Step 6: Find the critical value. We need to find the critical value from the chi-square distribution table for the given significance level (α) and degrees of freedom (df). In this example, with α = 0.05 and df = 4, the critical value is approximately 9.488.

Step 7: Compare the calculated chi-square test statistic with the critical value. If the calculated chi-square test statistic is greater than the critical value, we reject the null hypothesis; otherwise, we fail to reject the null hypothesis.

In our case, we compare the calculated chi-square test statistic from Step 4 with the critical value from Step 6.
If the calculated chi-square test statistic > critical value, we reject the null hypothesis.
If the calculated chi-square test statistic <= critical value, we fail to reject the null hypothesis.

```python
import numpy as np
from scipy.stats import chi2_contingency

design_A=np.array([100,120,110,90,95]) #observed data for design A
design_B=np.array([80,85,90,95,100]) #observed data for design B
contingency_table=np.array([design_A,design_B]) #observed data for both A,B (crosstab or contingency table)
X2,p_value,dof,expected_freq=chi2_contingency(contingency_table)
print("X2 staatistics->",X2)
print("p_value->",p_value)
print("Degree of Freedom->",dof)
if p_value<0.05:
  print("We failed to reject Null hypotheise that there is no significant difference between click through rates for design A and Design B")
else:
  print("We accepted the Null hypothesis ")
```

```
X2 staatistics-> 6.110658166925435
p_value-> 0.19103526314060293
Degree of Freedom-> 4
We accepted the Null hypothesis
```

—---

13. Scenario: A survey is conducted to measure customer satisfaction with a product on a scale of 1 to 10. The data collected is as follows:

[7, 9, 6, 8, 10, 7, 8, 9, 7, 8]

Calculate the 95% confidence interval for the population mean satisfaction score.

ANS :To calculate the 95% confidence interval for the population mean satisfaction score based on the given data, we can follow these steps:

Step 1: Calculate the sample mean ($\bar{x}$):

- Given data: [7, 9, 6, 8, 10, 7, 8, 9, 7, 8]
- Sample mean: $\bar{x}$ = (7 + 9 + 6 + 8 + 10 + 7 + 8 + 9 + 7 + 8) / 10 = 79 / 10 = 7.9

Step 2: Calculate the sample standard deviation (s):

- Subtract the sample mean from each data point, square the differences, sum them up, divide by (n-1), and take the square root.
- Given data: [7, 9, 6, 8, 10, 7, 8, 9, 7, 8]
- Subtracting the sample mean from each data point gives: [7 - 7.9, 9 - 7.9, 6 - 7.9, 8 - 7.9, 10 - 7.9, 7 - 7.9, 8 - 7.9, 9 - 7.9, 7 - 7.9, 8 - 7.9]
  = [-0.9, 1.1, -1.9, 0.1, 2.1, -0.9, 0.1, 1.1, -0.9, 0.1]
- Squaring the differences gives: [0.81, 1.21, 3.61, 0.01, 4.41, 0.81, 0.01, 1.21, 0.81, 0.01]
- Summing up the squared differences: 0.81 + 1.21 + 3.61 + 0.01 + 4.41 + 0.81 + 0.01 + 1.21 + 0.81 + 0.01 = 13.19
- Divide by (n-1) to get the sample variance: 13.19 / (10 - 1) = 13.19 / 9 = 1.4656
- Take the square root to get the sample standard deviation: s = √1.4656 = 1.21

Step 3: Determine the critical value (z-score) for a 95% confidence interval. Since the sample size is small (n < 30) and the population standard deviation is unknown, we use the t-distribution instead. With a sample size of 10 and a 95% confidence level, the degrees of freedom is (n-1) = 9. Using a t-table or statistical software, the critical value for a 95% confidence interval and 9 degrees of freedom is approximately 2.262.

Step 4: Calculate the margin of error:

- Margin of Error (E) = Critical Value * (Sample Standard Deviation / √Sample Size)
- E = 2.262 * (1.21 / √10) ≈ 0.850

Step 5: Calculate the confidence interval:

- Lower Limit = Sample Mean - Margin of Error
- Upper Limit = Sample Mean + Margin of Error
- Lower Limit = 7.9 - 0.850 ≈ 7.05
- Upper Limit = 7.9 + 0.850 ≈ 8.75

Step 6: State the 95% confidence interval:

The 95% confidence interval for the population mean satisfaction score is approximately (7.05, 8.75).

Therefore, we can say with 95% confidence that the true population mean satisfaction score falls between 7.05 and 8.75 based on the given data.

```
import scipy.stats as st
Import numpy as np
data=np.array( [7, 9, 6, 8, 10, 7, 8, 9, 7, 8])
st.t.interval(confidence=0.90, df=len(data)-1,
              loc=np.mean(data),
              scale=st.sem(data))
```

```
Output
(7.205994644530591, 8.59400535546941)
```

—---

14. Scenario: A company wants to analyze the effect of temperature on product performance. The data collected is as follows:

Temperature (in degrees Celsius): [20, 22, 23, 19, 21]

Performance (on a scale of 1 to 10): [8, 7, 9, 6, 8]

Perform a simple linear regression to predict performance based on temperature.

ANS: to perform the linear regression we have to find equation y=mx+c or y=ax+b
Where b=cov(x,y)/var(x) and where a is intercept and b is slope a=cov(x,y)/var(x)=13.6926 and b=y-ax i.e 0.9615
So equation is y=13.6926+0.9615

```
import numpy as np
from sklearn.linear_model import LinearRegression
# Temperature data
temperature = np.array([20, 22, 23, 19, 21])

# Performance data
performance = np.array([8, 7, 9, 6, 8]).reshape(-1,1)
# Create a linear regression object
regression_model = LinearRegression()

# Fit the model with the data
regression_model.fit(performance,temperature)
# Get the regression coefficients
intercept = regression_model.intercept_
slope = regression_model.coef_[0]

print(intercept)
print(slope)
```

```
13.692307692307693
0.9615384615384616
```

```
--------------------------------------------------------------------------
```

15. Scenario: A study is conducted to compare the preferences of two groups of participants. The preferences are measured on a Likert scale from 1 to 5. The data collected is as follows:

Group A: [4, 3, 5, 2, 4]

Group B: [3, 2, 4, 3, 3]

Perform a Mann-Whitney U test to determine if there is a significant difference in the median preferences between the two groups.

ANS :To perform a Mann-Whitney U test to determine if there is a significant difference in the median preferences between Group A and Group B, you can follow these steps:

Step 1: State the null hypothesis (H0) and the alternative hypothesis (H1):
- Null hypothesis (H0): There is no significant difference in the median preferences between Group A and Group B.
- Alternative hypothesis (H1): There is a significant difference in the median preferences between Group A and Group B.

Step 2: Rank the data:
Rank the combined data from both groups in ascending order. Assign ranks based on their position in the combined ranking, ignoring group membership ties.

Group A: [4, 3, 5, 2, 4] => Ranked: [3, 2, 5, 1, 3]
Group B: [3, 2, 4, 3, 3] => Ranked: [2, 1, 4, 2, 2]

Step 3: Calculate the sum of ranks for each group:
Sum the ranks for each group.

Group A: Sum of ranks = 3 + 2 + 5 + 1 + 3 = 14
Group B: Sum of ranks = 2 + 1 + 4 + 2 + 2 = 11

Step 4: Calculate the Mann-Whitney U statistic:
U = n1 * n2 + (n1 * (n1 + 1)) / 2 - sum of ranks for Group A
Where n1 is the number of participants in Group A (5) and n2 is the number of participants in Group B (5).

U = (5 * 5) + (5 * (5 + 1)) / 2 - 14 = 25 + 15 - 14 = 26

Step 5: Calculate the expected value of U under the null hypothesis:
U_expected = n1 * n2 / 2 = 5 * 5 / 2 = 12.5

Step 6: Calculate the variance of U under the null hypothesis:
Var(U) = (n1 * n2 * (n1 + n2 + 1)) / 12 = (5 * 5 * (5 + 5 + 1)) / 12 = 125 / 12 ≈ 10.42

Step 7: Calculate the Z-score:
Z = (U - U_expected) / sqrt(Var(U)) = (26 - 12.5) / sqrt(10.42) ≈ 2.36

Step 8: Determine the p-value:
Lookup the p-value corresponding to the Z-score in the Mann-Whitney U test table or use statistical software. For a two-tailed test, the p-value would be approximately 0.018.

Step 9: Make a decision:
Compare the p-value to the significance level (alpha) to make a decision. If the p-value is less than alpha (commonly 0.05), reject the null hypothesis. Otherwise, fail to reject the null hypothesis.

In this case, since the p-value (0.018) is less than the commonly used significance level of 0.05, we reject the null hypothesis. Therefore, we conclude that there is a significant difference in the median preferences between Group A and Group B based on the Mann-Whitney U test.

—--

16. Scenario: A company wants to analyze the distribution of customer ages. The data collected is as follows:

[25, 30, 35, 40, 45, 50, 55, 60, 65, 70]

Calculate the interquartile range (IQR) of the ages.

ANS :To calculate the interquartile range (IQR) of the ages, you need to follow these steps:

Step 1: Arrange the ages in ascending order:
[25, 30, 35, 40, 45, 50, 55, 60, 65, 70]

Step 2: Determine the first quartile (Q1):
To find Q1, you need to calculate the median of the lower half of the data. In this case, the lower half consists of the first five ages: [25, 30, 35, 40, 45].
Since there is an odd number of data points, the median is the middle value, which is 35.
Therefore, Q1 = 35.

Step 3: Determine the third quartile (Q3):
To find Q3, you need to calculate the median of the upper half of the data. In this case, the upper half consists of the last five ages: [50, 55, 60, 65, 70].
Again, since there is an odd number of data points, the median is the middle value, which is 60.
Therefore, Q3 = 60.

Step 4: Calculate the interquartile range (IQR):
IQR = Q3 - Q1
Substituting the values, we get:
IQR = 60 - 35
IQR = 25
Therefore, the interquartile range (IQR) of the ages is 25.

```
import numpy as np
data = [25, 30, 35, 40, 45, 50, 55, 60, 65, 70]
q1 = np.percentile(data, 25)
q3 = np.percentile(data, 75)
iqr = q3 - q1 print("Interquartile Range (IQR):", iqr)
```

—---

17. Scenario: A study is conducted to compare the performance of three different machine learning algorithms. The accuracy scores for each algorithm are as follows:

Algorithm A: [0.85, 0.80, 0.82, 0.87, 0.83]
Algorithm B: [0.78, 0.82, 0.84, 0.80, 0.79]
Algorithm C: [0.90, 0.88, 0.89, 0.86, 0.87]

Perform a Kruskal-Wallis test to determine if there is a significant difference in the median accuracy scores between the algorithms.

ANS :`from scipy.stats import kruskal`

```
# Accuracy scores for each algorithm
algorithm_A = [0.85, 0.80, 0.82, 0.87, 0.83]
algorithm_B = [0.78, 0.82, 0.84, 0.80, 0.79]
algorithm_C = [0.90, 0.88, 0.89, 0.86, 0.87]

# Perform the Kruskal-Wallis test
statistic, p_value = kruskal(algorithm_A, algorithm_B, algorithm_C)

# Print the test statistic and p-value
print("Test Statistic (H):", statistic)
print("p-value:", p_value)
```

—--

19. Scenario: A survey is conducted to measure the satisfaction levels of customers with a new product. The data collected is as follows:

[7, 8, 9, 6, 8, 7, 9, 7, 8, 7]

Calculate the standard error of the mean satisfaction score.

ANS :standard error of mean is given as std.deviation/sqrt of n
Where n is sample size
Std.deviation is s=0.9309
Std.error =0.2943

```
import numpy as np

data = [7, 8, 9, 6, 8, 7, 9, 7, 8, 7]

# Calculate the standard error of the mean
sem = np.std(data) / np.sqrt(len(data))

# Print the result
print("Standard Error of the Mean:", sem)
```

—--

18. Scenario: A company wants to analyse the effect of price on sales. The data collected is as follows:
Price (in dollars): [10, 15, 12, 8, 14]
Sales: [100, 80, 90, 110, 95]
Perform a simple linear regression to predict
ANS:# Given data

```
import numpy as np
from sklearn.linear_model import LinearRegression
price = np.array([10, 15, 12, 8, 14]).reshape((-1, 1))
sales = np.array([100, 80, 90, 110, 95])
```

```python
# Create a linear regression model
model = LinearRegression()

# Fit the model to the data
model.fit(price, sales)

# Get the regression coefficients
intercept = model.intercept_
slope = model.coef_[0]

print(intercept)
print(slope)


# Predict sales for a new price value
new_price = np.array([13]).reshape((-1, 1))
predicted_sales = model.predict(new_price)

# Print the predicted sales
print("Predicted sales:", predicted_sales)
```

—--

20. Scenario: A company wants to analyse the relationship between advertising expenditure and sales. The data collected is as follows:

Advertising Expenditure (in thousands): [10, 15, 12, 8, 14]

Sales (in thousands): [25, 30, 28, 20, 26]

Perform a multiple regression analysis to predict sales based on advertising expenditure.

ANS :Multiple linear regression is an extension of simple linear regression that involves more than one independent variable. It allows us to analyze the relationship between multiple predictors and a dependent variable. The general mathematical formulation of multiple linear regression is as follows:

```
y = b0 + b1*x1 + b2*x2 + ... + bn*xn + e
```

where:

- y is the dependent variable (the variable we want to predict)
- x1, x2, ..., xn are the independent variables (predictors)
- b0 is the y-intercept (the value of y when all predictors are zero)
- b1, b2, ..., bn are the regression coefficients (representing the impact of each predictor on y)
- e is the error term (residuals)

To perform multiple linear regression, we can use the least squares method to estimate the regression coefficients. This minimizes the sum of the squared differences between the observed and predicted values. The coefficients can be estimated using matrix operations.

Let's demonstrate with an example using Python and the scikit-learn library:

```python
import numpy as np
from sklearn.linear_model import LinearRegression

# Given data
X = np.array([[10, 20],
              [15, 25],
              [12, 18],
              [8, 15],
              [14, 22]])

y = np.array([100, 80, 90, 110, 95])

# Create a linear regression model
model = LinearRegression()

# Fit the model to the data
model.fit(X, y)

# Get the estimated coefficients
intercept = model.intercept_
coefficients = model.coef_

# Print the estimated coefficients
print("Intercept:", intercept)
print("Coefficients:", coefficients)

# Predict sales for new data
new_data = np.array([[13, 23]])
predicted_sales = model.predict(new_data)

# Print the predicted sales
print("Predicted sales:", predicted_sales)
```