**1. Scenario: A company wants to analyze 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:

Region A: [10, 15, 12, 8, 14]

Mean sales for Region A = (10 + 15 + 12 + 8 + 14) / 5 = 59 / 5 = 11.8

Region B: [18, 20, 16, 22, 25]

Mean sales for Region B = (18 + 20 + 16 + 22 + 25) / 5 = 101 / 5 = 20.2

**RESULT:**

mean sales for Region A = 11.8  
mean sales for Region B = 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:

2 appears 2 times

3 appears 2 times

4 appears 3 times

5 appears 3 times

As both 4 and 5 appear the most frequently (3 times each), the mode of the survey responses is 4 and 5.

**Result:** the mode of the survey responses 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: **Department A**: [5000, 6000, 5500, 7000]

Arranging the salaries in ascending order: [5000, 5500, 6000, 7000]

The median salary for Department A is the middle value, which is 5500.

**Department B**: [4500, 5500, 5800, 6000, 5200]

Arranging the salaries in ascending order: [4500, 5200, 5500, 5800, 6000]

The median salary for Department B is the middle value, which is 5500.

**Result**:

median salary for Department A = 5500

median salary for Department B = 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:

the highest value = 26.1 the lowest value = 24.8

Range = Highest value - Lowest value

= 26.1 - 24.8

= 1.3

**Result:**

the range of the stock prices = **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:**

Calculate the means and standard deviations for each group:

**Group A:**

**Mean A** = (85 + 90 + 92 + 88 + 91) / 5 = **89.2**

**Standard deviation A** = sqrt(((85-89.2)^2 + (90-89.2)^2 + (92-89.2)^2 + (88-89.2)^2 + (91-89.2)^2) / 4) = **2.71**

**Group B:**

**Mean B** = (82 + 88 + 90 + 86 + 87) / 5 = **86.6**

**Standard deviation B** = sqrt(((82-86.6)^2 + (88-86.6)^2 + (90-86.6)^2 + (86-86.6)^2 + (87-86.6)^2) / 4) = **2.50**

**Calculate the t-value:**

t = (Mean A - Mean B) / sqrt((Standard deviation A^2 / nA) + (Standard deviation B^2 / nB))

nA = number of observations in Group A (5)

nB = number of observations in Group B (5)

t = (89.2 - 86.6) / sqrt((2.71^2 / 5) + (2.50^2 / 5))

= 2.6 / sqrt(1.39 + 1.25)

= 2.6 / sqrt(2.64)

= 2.6 / 1.63

= 1.59

**Degrees of freedom:**

degrees of freedom = nA + nB - 2 = 5 + 5 - 2 = 8

**Determine the critical t-value for a given significance level (alpha) and degrees of freedom:**

Let's assume a significance level of 0.05 (5%) for a two-tailed test.

The critical t-value with 8 degrees of freedom is approximately ±2.306.

**Compare the calculated t-value with the critical t-value:**

Since |t| = 1.59 < 2.306, we do not reject the null hypothesis.

**RESULT**

Therefore, based on the t-test, there is not enough evidence to suggest a significant difference in the mean scores between Group A and Group B.

**6. Scenario: A company wants to analyze 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**

**The means of the advertising expenditure and sales:**

Advertising Expenditure mean (X̄) = (10 + 15 + 12 + 8 + 14) / 5 = **11.8**

Sales mean (Ȳ) = (25 + 30 + 28 + 20 + 26) / 5 = **25.8**

**The deviations from the means for both variables:**

Deviations from the Advertising Expenditure mean (Xi - X̄): [-1.8, 3.2, 0.2, -3.8, 2.2]

Deviations from the Sales mean (Yi - Ȳ): [-0.8, 4.2, 2.2, -5.8, 0.2]

**The sum of the products of the deviations:**

Sum of (Xi - X̄)(Yi - Ȳ) = (-1.8 \* -0.8) + (3.2 \* 4.2) + (0.2 \* 2.2) + (-3.8 \* -5.8) + (2.2 \* 0.2)

= 1.44 + 13.44 + 0.44 + 22.04 + 0.44

= **37.8**

**The sum of the squared deviations for both variables:**

Sum of (Xi - X̄)^2 = (-1.8)^2 + (3.2)^2 + (0.2)^2 + (-3.8)^2 + (2.2)^2 = 3.24 + 10.24 + 0.04 + 14.44 + 4.84

= **32.8**

Sum of (Yi - Ȳ)^2 = (-0.8)^2 + (4.2)^2 + (2.2)^2 + (-5.8)^2 + (0.2)^2

= 0.64 + 17.64 + 4.84 + 33.64 + 0.04

= **57.8**

**The square roots of the sums of squared deviations:**

Square root of Sum of (Xi - X̄)^2 = sqrt(32.8) ≈ **5.73**

Square root of Sum of (Yi - Ȳ)^2 = sqrt(57.8) = **7.61**

**The correlation coefficient (r):**

r = Sum of (Xi - X̄)(Yi - Ȳ) / (Square root of Sum of (Xi - X̄)^2 \* Square root of Sum of (Yi - Ȳ)^2)

= 37.8 / (5.73 \* 7.61)

= 37.8 / 43.63

= **0.87**

**RESULT:**

The correlation coefficient between advertising expenditure and sales = **0.87**

**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:***

**The mean (average) of the heights:**

Mean (x̄) = (160 + 170 + 165 + 155 + 175 + 180 + 170) / 7 = 166.43

**The squared difference from the mean for each height:**

Squared differences = [(160 - 166.43)^2, (170 - 166.43)^2, (165 - 166.43)^2, (155 - 166.43)^2, (175 - 166.43)^2, (180 - 166.43)^2, (170 - 166.43)^2]

= [40.6749, 14.1849, 1.9249, 131.3449, 69.9249, 173.1849, 14.1849]

**The variance as the average of the squared differences:**

Variance = Sum of squared differences / (n - 1)

= (40.6749 + 14.1849 + 1.9249 + 131.3449 + 69.9249 + 173.1849 + 14.1849) / (7 - 1)

= 445.4233 / 6

= **74.2372**

**The standard deviation as the square root of the variance:**

Standard deviation = sqrt(Variance)

= sqrt(74.2372)

= *8.617*

**RESULT**

Therefore, the standard deviation of the heights is approximately **8.617**

**8. Scenario: A company wants to analyze 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:***

Assume the data into two variables: Employee Tenure (X) and Job Satisfaction (Y).

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]

**The mean (average) of both variables:**

Mean of Employee Tenure (X̄) = (2 + 3 + 5 + 4 + 6 + 2 + 4) / 7 = 3.57 (rounded to two decimal places)

Mean of Job Satisfaction (Ȳ) = (7 + 8 + 6 + 9 + 5 + 7 + 6) / 7 = 6.86 (rounded to two decimal places)

**The deviations from the mean for both variables:**

Deviations from the Mean of Employee Tenure (Xi - X̄): [-1.57, -0.57, 1.43, 0.43, 2.43, -1.57, 0.43]

Deviations from the Mean of Job Satisfaction (Yi - Ȳ): [0.14, 1.14, -0.86, 2.14, -1.86, 0.14, -0.86]

**The product of the deviations for each data point:**

Product of Deviations = (-1.57 \* 0.14, -0.57 \* 1.14, 1.43 \* -0.86, 0.43 \* 2.14, 2.43 \* -1.86, -1.57 \* 0.14, 0.43 \* -0.86)

= (-0.2198, -0.6498, -1.2338, 0.9222, -4.5162, -0.2198, -0.3708)

**The sum of the product of the deviations:**

Sum of Product of Deviations = -0.2198 - 0.6498 - 1.2338 + 0.9222 - 4.5162 - 0.2198 - 0.3708

= -6.288 (rounded to three decimal places)

**The sum of squared deviations for employee tenure:**

Sum of (Xi - X̄)^2 = (-1.57)^2 + (-0.57)^2 + 1.43^2 + 0.43^2 + 2.43^2 + (-1.57)^2 + 0.43^2

= 11.0992

**The sum of squared deviations for job satisfaction:**

Sum of (Yi - Ȳ)^2 = 0.14^2 + 1.14^2 + (-0.86)^2 + 2.14^2 + (-1.86)^2 + 0.14^2 + (-0.86)^2

= 8.7792

**The slope (β1) of the linear regression line:**

β1 = Sum of Product of Deviations / Sum of (Xi - X̄)^2

= -6.288 / 11.0992

= -0.566 (rounded to three decimal places)

**The intercept (β0) of the linear regression line:**

β0 = Ȳ - β1 \* X̄

= 6.86 - (-0.566 \* 3.57)

= 8.06 (rounded to two decimal places)

The linear regression equation to predict job satisfaction (Y) based on employee tenure (X) is:

Y = β0 + β1 \* X

= 8.06 - 0.566 \* X

Therefore, the linear regression analysis predicts job satisfaction (Y) based on employee tenure (X) with the equation Y = 8.06 - 0.566 \* X. The slope (β1) indicates that for every unit increase in employee tenure, job satisfaction is predicted to decrease by approximately 0.566 units. The intercept (β0) represents the predicted job satisfaction when employee tenure is zero.

**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***

**The null hypothesis (H0) and alternative hypothesis (Ha):**

**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.

**The means of recovery times for each medication group:**

*Medication A:*

Mean A = (10 + 12 + 14 + 11 + 13) / 5 = 12

*Medication B:*

Mean B = (15 + 17 + 16 + 14 + 18) / 5 = 16

**The sum of squares within (SSW):**

SSW = Sum of (Xi - X̄)^2

*Medication A:*

SSW(A) = (10 - 12)^2 + (12 - 12)^2 + (14 - 12)^2 + (11 - 12)^2 + (13 - 12)^2

= 4 + 0 + 4 + 1 + 1 = 10

*Medication B:*

SSW(B) = (15 - 16)^2 + (17 - 16)^2 + (16 - 16)^2 + (14 - 16)^2 + (18 - 16)^2

   = 1 + 1 + 0 + 4 + 4 = 10

**The sum of squares between (SSB):**

SSB = nA \* (Mean A - Grand Mean)^2 + nB \* (Mean B - Grand Mean)^2

nA = number of observations in Medication A (5)

nB = number of observations in Medication B (5)

Grand Mean = (Mean A + Mean B) / 2 = (12 + 16) / 2 = 14

SSB = 5 \* (12 - 14)^2 + 5 \* (16 - 14)^2

= 5 \* (-2)^2 + 5 \* 2^2

= 20 + 20

= 40

**The degrees of freedom:**

Degrees of freedom between (dfB) = Number of groups - 1 = 2 - 1 = 1

Degrees of freedom within (dfW) = Total number of observations - Number of groups = 10 - 2 = 8

Total degrees of freedom (dfTotal) = dfB + dfW = 1 + 8 = 9

**The mean squares between (MSB) and mean squares within (MSW):**

MSB = SSB / dfB = 40 / 1 = 40

MSW = SSW / dfW = 10 / 8 = 1.25

**The F-value:**

F = MSB / MSW = 40 / 1.25 = 32

The critical F-value for a given significance level (alpha) and degrees of freedom:

Let's assume a significance level of 0.05 (5%) and dfB = 1, dfW = 8.

The critical F-value is approximately 5.32.

**Comparing the calculated F-value with the critical F-value:**

Since F = 32 > 5.32, we reject the null hypothesis.

Therefore, based on the ANOVA, **there is a significant difference in the mean recovery times between Medication A and Medication B.**

**10. Scenario: A company wants to analyze 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.**

***ANS***

**Sorting the feedback ratings in ascending order:**

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

**The index of the 75th percentile:**

Index = (75 / 100) \* (n + 1)

= (75 / 100) \* (10 + 1)

= 0.75 \* 11

= 8.25

Since the index (8.25) is not a whole number, you need to interpolate between the values at positions 8 and 9.

**Identify the values at positions 8 and 9:**

Position 8: 8

Position 9: 9

**The weighted average between these two values:**

Weighted average = Value at position 8 + (Index - Position 8) \* (Value at position 9 - Value at position 8)

= 8 + (8.25 - 8) \* (9 - 8)

= 8 + 0.25 \* 1

= 8 + 0.25

= 8.25

Therefore, the 75th percentile of the feedback ratings is **8.25.**

**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:***

**Define the null hypothesis (H0) and alternative hypothesis (Ha):**

Null hypothesis (H0): The mean weight of the products is equal to 10 grams (μ = 10).

Alternative hypothesis (Ha): The mean weight of the products differs significantly from 10 grams (μ ≠ 10).

**Calculate the sample mean (x̄) of the weights:**

Sample mean (x̄) = (10.2 + 9.8 + 10.0 + 10.5 + 10.3 + 10.1) / 6 = 10.1833 (rounded to four decimal places)

**Calculate the sample standard deviation (s) of the weights:**

**Step 1: Calculating the squared difference from the mean for each weight:**

Squared differences = [(10.2 - 10.1833)^2, (9.8 - 10.1833)^2, (10.0 - 10.1833)^2, (10.5 - 10.1833)^2, (10.3 - 10.1833)^2, (10.1 - 10.1833)^2] = [0.00033289, 0.01447289, 0.00033289, 0.00033289, 0.00033289, 0.00033289]

**Step 2: Calculating the sum of the squared differences:**

Sum of squared differences = 0.00033289 + 0.01447289 + 0.00033289 + 0.00033289 + 0.00033289 + 0.00033289

= 0.01513734

**Step 3: Calculating the variance as the average of the squared differences:**

Variance = Sum of squared differences / (n - 1) = 0.01513734 / (6 - 1) = 0.00302747 (rounded to eight decimal places)

**Step 4: Calculating the sample standard deviation as the square root of the variance:**

Sample standard deviation (s) = sqrt(Variance) = sqrt(0.00302747) ≈ 0.05497 (rounded to five decimal places)

**Determining the test statistic (t-value):**

t = (x̄ - μ) / (s / sqrt(n))

= (10.1833 - 10) / (0.05497 / sqrt(6))

= 0.1833 / (0.05497 / 2.4495)

= 0.1833 / 0.02246

= 8.1608 (rounded to four decimal places)

**Determine the degrees of freedom:**

Degrees of freedom = n - 1 = 6 - 1 = 5

**Look up the critical t-value for a given significance level (alpha) and degrees of freedom:**

Let's assume a significance level of 0.05 (5%) for a two-tailed test.

For 5 degrees of freedom, the critical t-value is approximately ±2.571.

Comparing the calculated t-value with the critical t-value:

Since |t| = 8.1608 > 2.571, we reject the null hypothesis.

Therefore, based on the hypothesis test, **there is enough evidence to suggest that the mean weight of the products differs significantly from 10 grams.**

**12. Scenario: A company wants to analyze 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***

Create a contingency table that shows the observed frequencies of clicks for each design:

|  | **Design A** | **Design B** |
| --- | --- | --- |
| Clicks | 100 | 80 |
| Clicks | 120 | 85 |
| Clicks | 110 | 90 |
| Clicks | 90 | 95 |
| Clicks | 95 | 100 |

Calculate the row totals and column totals for the contingency table:

|  | **Design A** | **Design B** | **Row Total** |
| --- | --- | --- | --- |
| Clicks | 100 | 80 | 180 |
| Clicks | 120 | 85 | 205 |
| Clicks | 110 | 90 | 200 |
| Clicks | 90 | 95 | 185 |
| Clicks | 95 | 100 | 195 |
| Column Total | 515 | 450 | 965 |

Calculate the expected frequencies under the assumption of independence (assuming no difference)

Expected Frequency for each cell = (Row Total \* Column Total) / Grand Total

|  | **Design A** | **Design B** | **Row Total** |
| --- | --- | --- | --- |
| Clicks | 92.16 | 87.84 | 180 |
| Clicks | 104.49 | 100.51 | 205 |
| Clicks | 102.06 | 97.94 | 200 |
| Clicks | 93.55 | 91.45 | 185 |
| Clicks | 96.74 | 98.26 | 195 |
| Column Total | 489 | 475 | 965 |

**Calculate the chi-square test statistic:**

Chi-square test statistic = Σ [(Observed Frequency - Expected Frequency)^2 / Expected Frequency]

Chi-square test statistic = [(100-92.16)^2/92.16] + [(120-104.49)^2/104.49] + [(110-102.06)^2/102.06] + [(90-93.55)^2/93.55] + [(95-96.74)^2/96.74] + [(80-87.84)^2/87.84] + [(85-100.51)^2/100.51] + [(90-97.94)^2/97.94] + [(95-91.45)^2/91.45] + [(100-98.26)^2/98.26]

= 0.792 + 1.864 + 0.745 + 0.281 + 0.020 + 0.656 + 1.344 + 0.667 + 0.188 + 0.067

= 6.624 (rounded to three decimal places)

**Determine the degrees of freedom:**

Degrees of freedom = (Number of rows - 1) \* (Number of columns - 1)

= (5 - 1) \* (2 - 1)

= 4

Look up the critical chi-square value for a given significance level (alpha) and degrees of freedom:

Let's assume a significance level of 0.05 (5%) for a two-tailed test. For 4 degrees of freedom, the critical chi-square value is approximately 9.488.

**Compare the calculated chi-square test statistic with the critical chi-square value:**

Since the calculated chi-square test statistic (6.624) is less than the critical chi-square value (9.488), we fail to reject the null hypothesis.

Therefore, based on the chi-square test, there is not enough evidence to suggest a significant difference in the click-through rates between Design A and Design B.

**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**

**Calculate the sample mean (x̄) of the satisfaction scores:**

Sample mean (x̄) = (7 + 9 + 6 + 8 + 10 + 7 + 8 + 9 + 7 + 8) / 10 = 7.9

**Calculate the sample standard deviation (s) of the satisfaction scores:**

Step 1: Calculate the squared difference from the mean for each satisfaction score:

Squared differences = [(7 - 7.9)^2, (9 - 7.9)^2, (6 - 7.9)^2, (8 - 7.9)^2, (10 - 7.9)^2, (7 - 7.9)^2, (8 - 7.9)^2, (9 - 7.9)^2, (7 - 7.9)^2, (8 - 7.9)^2] = [0.81, 1.21, 2.41, 0.01, 4.41, 0.81, 0.01, 1.21, 0.81, 0.01]

Step 2: Calculate the sum of the squared differences:

Sum of squared differences = 0.81 + 1.21 + 2.41 + 0.01 + 4.41 + 0.81 + 0.01 + 1.21 + 0.81 + 0.01

= 12.08

Step 3: Calculate the variance as the average of the squared differences:

Variance = Sum of squared differences / (n - 1) = 12.08 / (10 - 1) = 1.3422 (rounded to four decimal places)

Step 4: Calculate the sample standard deviation as the square root of the variance:

Sample standard deviation (s) = sqrt(Variance) = sqrt(1.3422) ≈ 1.1587 (rounded to four decimal places)

**Determine the sample size (n):**

Sample size (n) = 10

**Determine the critical value (z) for a 95% confidence interval:**

The critical value for a 95% confidence interval is approximately 1.96.

**Calculate the margin of error (E):**

Margin of error (E) = z \* (s / sqrt(n)) = 1.96 \* (1.1587 / sqrt(10)) ≈ 0.7372 (rounded to four decimal places)

**Calculate the lower and upper bounds of the confidence interval:**

Lower bound = x̄ - E = 7.9 - 0.7372 ≈ 7.1628 (rounded to four decimal places)

Upper bound = x̄ + E = 7.9 + 0.7372 ≈ 8.6372 (rounded to four decimal places)

**RESULT:**

Therefore, the 95% confidence interval for the population mean satisfaction score is approximately 7.1628 to 8.6372. This means that we can be 95% confident that the true population mean satisfaction score falls within this interval.

**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:***

**Organize the data into two variables: Temperature (X) and Performance (Y).**

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

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

**Calculate the mean (average) of both variables:**

Mean of Temperature (X̄) = (20 + 22 + 23 + 19 + 21) / 5 = 21

Mean of Performance (Ȳ) = (8 + 7 + 9 + 6 + 8) / 5 = 7.6

**Calculate the deviations from the mean for both variables:**

Deviations from the Mean of Temperature (Xi - X̄): [-1, 1, 2, -2, 0]

Deviations from the Mean of Performance (Yi - Ȳ): [0.4, -0.6, 1.4, -1.6, 0.4]

**Calculate the product of the deviations for each data point:**

Product of Deviations = (-1 \* 0.4, 1 \* -0.6, 2 \* 1.4, -2 \* -1.6, 0 \* 0.4)

= (-0.4, -0.6, 2.8, 3.2, 0)

**Calculate the sum of the product of the deviations:**

Sum of Product of Deviations = -0.4 - 0.6 + 2.8 + 3.2 + 0

= 4

**Calculate the sum of squared deviations for temperature:**

Sum of (Xi - X̄)^2 = (-1)^2 + 1^2 + 2^2 + (-2)^2 + 0^2

= 10

**Calculate the sum of squared deviations for performance:**

Sum of (Yi - Ȳ)^2 = 0.4^2 + (-0.6)^2 + 1.4^2 + (-1.6)^2 + 0.4^2

= 5.2

**Calculate the slope (β1) of the linear regression line:**

β1 = Sum of Product of Deviations / Sum of (Xi - X̄)^2

= 4 / 10

= 0.4

**Calculate the intercept (β0) of the linear regression line:**

β0 = Ȳ - β1 \* X̄

= 7.6 - (0.4 \* 21)

= -0.2

**The linear regression equation to predict performance (Y) based on temperature (X) is:**

Y = β0 + β1 \* X

= -0.2 + 0.4 \* X

Therefore, the simple linear regression analysis predicts performance (Y) based on temperature (X) with the equation Y = -0.2 + 0.4 \* X. The slope (β1) indicates that for every 1-degree Celsius increase in temperature, performance is predicted to increase by 0.4 units. The intercept (β0) represents the predicted performance when the temperature is zero degrees Celsius.

**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:**

**Rank the combined data from both groups in ascending order:**

Combined data: [4, 3, 5, 2, 4, 3, 2, 4, 3, 3]

Ranked data: [5, 4, 4, 4, 3, 3, 3, 3, 2, 2]

**Assign ranks to the data from each group separately:**

Group A ranks: [2, 4, 1, 10, 4]

Group B ranks: [6, 9, 3, 6, 6]

**Calculate the sum of ranks for each group:**

Group A sum of ranks (U1) = 2 + 4 + 1 + 10 + 4 = 21

Group B sum of ranks (U2) = 6 + 9 + 3 + 6 + 6 = 30

**Calculate the U statistic for each group:**

U1 = n1 \* n2 + (n1 \* (n1 + 1)) / 2 - U1 = 5 \* 5 + (5 \* (5 + 1)) / 2 - 21 = 25 + 15 - 21 = 19

U2 = n1 \* n2 + (n2 \* (n2 + 1)) / 2 - U2 = 5 \* 5 + (5 \* (5 + 1)) / 2 - 30 = 25 + 15 - 30 = 10

**Determine the smaller U statistic (Umin) and the larger U statistic (Umax):**

Umin = min(U1, U2) = min(19, 10) = 10

Umax = max(U1, U2) = max(19, 10) = 19

**Calculate the critical value of U at a given significance level (alpha) and sample size (n1, n2):**

Let's assume a significance level of 0.05 (5%) for a two-tailed test. For n1 = 5 and n2 = 5, the critical value of U is 2 (lookup from the Mann-Whitney U table).

**Compare the calculated U statistics with the critical value of U:**

Since Umax = 19 > 2, and Umin = 10 > 2, we reject the null hypothesis.

**RESULT**

Therefore, based on the Mann-Whitney U test, there is a significant difference in the median preferences between Group A and Group B.

**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:***

**Sort the ages in ascending order:**

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

**Calculate the first quartile (Q1):**

Q1 = (n + 1) / 4

= (10 + 1) / 4

= 2.75

Since 2.75 is not a whole number, we need to interpolate between the values at positions 2 and 3.

Value at position 2: 30

Value at position 3: 35

Q1 = Value at position 2 + (Q1 - Position 2) \* (Value at position 3 - Value at position 2)

= 30 + (2.75 - 2) \* (35 - 30)

= 30 + 0.75 \* 5

= 30 + 3.75

= 33.75

**Calculate the third quartile (Q3):**

Q3 = (3 \* (n + 1)) / 4

= (3 \* (10 + 1)) / 4

= 8.25

Since 8.25 is not a whole number, we need to interpolate between the values at positions 8 and 9.

Value at position 8: 60

Value at position 9: 65

Q3 = Value at position 8 + (Q3 - Position 8) \* (Value at position 9 - Value at position 8)

= 60 + (8.25 - 8) \* (65 - 60)

= 60 + 0.25 \* 5

= 60 + 1.25

= 61.25

**Calculate the interquartile range (IQR):**

IQR = Q3 - Q1

= 61.25 - 33.75

= 27.5

Therefore, the interquartile range (IQR) of the ages is\*\* 27.5\*\*. This means that the middle 50% of the age distribution falls within this range.

**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:***

**Combine the accuracy scores from all algorithms:**

Combined data: [0.85, 0.80, 0.82, 0.87, 0.83, 0.78, 0.82, 0.84, 0.80, 0.79, 0.90, 0.88, 0.89, 0.86, 0.87]

**Rank the combined data from lowest to highest:**

Ranked data: [5, 3, 4, 10, 6, 2, 4, 9, 3, 1, 15, 13, 14, 12, 11]

**Calculate the sum of ranks for each algorithm:**

Algorithm A sum of ranks (RA) = 5 + 3 + 4 + 10 + 6 = 28

Algorithm B sum of ranks (RB) = 2 + 4 + 9 + 3 + 1 = 19

Algorithm C sum of ranks (RC) = 15 + 13 + 14 + 12 + 11 = 65

**Calculate the average rank for each algorithm:**

Average rank for Algorithm A (MA) = RA / nA = 28 / 5 = 5.6

Average rank for Algorithm B (MB) = RB / nB = 19 / 5 = 3.8

Average rank for Algorithm C (MC) = RC / nC = 65 / 5 = 13

**Calculate the Kruskal-Wallis test statistic (H):**

H = (12 \* (Σ(Mi^2) / ni)) - 3 \* (n(n + 1)) Σ(Mi^2) = (5.6^2 \* 5) + (3.8^2 \* 5) + (13^2 \* 5) = 157.44 + 72.8 + 845 = 1074.24

n = nA + nB + nC = 5 + 5 + 5 = 15

H = (12 \* (1074.24 / 15)) - 3 \* (15(15 + 1))

= 12 \* 71.62 - 3 \* 120

= 859.44 - 360

= 499.44

**Determine the critical value of H at a given significance level (alpha) and degrees of freedom (k - 1):**

Let's assume a significance level of 0.05 (5%) for a two-tailed test. For k = 3 (number of algorithms) - 1 = 2 degrees of freedom, the critical value of H is approximately 5.991.

**Compare the calculated Kruskal-Wallis test statistic with the critical value of H:**

Since the calculated H (499.44) is greater than the critical value of H (5.991), we reject the null hypothesis.

**RESULT**

Therefore, based on the Kruskal-Wallis test, there is a significant difference in the median accuracy scores between the three algorithms.

**18. Scenario: A company wants to analyze 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 sales based on price.**

***ANS:***

**Organize the data into two variables: Price (X) and Sales (Y).**

Price (in dollars): [10, 15, 12, 8, 14]

Sales: [100, 80, 90, 110, 95]

**Calculate the mean (average) of both variables:**

Mean of Price (X̄) = (10 + 15 + 12 + 8 + 14) / 5 = 11.8

Mean of Sales (Ȳ) = (100 + 80 + 90 + 110 + 95) / 5 = 95

**Calculate the deviations from the mean for both variables:**

Deviations from the Mean of Price (Xi - X̄): [-1.8, 3.2, 0.2, -3.8, 2.2]

Deviations from the Mean of Sales (Yi - Ȳ): [5, -15, -5, 15, 0]

**Calculate the product of the deviations for each data point:**

Product of Deviations = (-1.8 \* 5, 3.2 \* -15, 0.2 \* -5, -3.8 \* 15, 2.2 \* 0)

= (-9, -48, -1, -57, 0)

**Calculate the sum of the product of the deviations:**

Sum of Product of Deviations = -9 - 48 - 1 - 57 + 0

= -115

**Calculate the sum of squared deviations for price:**

Sum of (Xi - X̄)^2 = (-1.8)^2 + 3.2^2 + 0.2^2 + (-3.8)^2 + 2.2^2

= 24.28

**Calculate the sum of squared deviations for sales:**

Sum of (Yi - Ȳ)^2 = 5^2 + (-15)^2 + (-5)^2 + 15^2 + 0^2

= 425

**Calculate the slope (β1) of the linear regression line:**

β1 = Sum of Product of Deviations / Sum of (Xi - X̄)^2

= -115 / 24.28

= -4.7388 (rounded to four decimal places)

**Calculate the intercept (β0) of the linear regression line:**

β0 = Ȳ - β1 \* X̄

= 95 - (-4.7388 \* 11.8)

= 149.8584 (rounded to four decimal places)

The linear regression equation to predict sales (Y) based on price (X) is:

Y = β0 + β1 \* X

= 149.8584 - 4.7388 \* X

**RESULT**

Therefore, the simple linear regression analysis predicts sales (Y) based on price (X) with the equation Y = 149.8584 - 4.7388 \* X. The slope (β1) indicates that for every $1 increase in price, sales are predicted to decrease by approximately 4.7388 units. The intercept (β0) represents the predicted sales when the price is zero dollars, but in this case, it is not practically meaningful as a price of zero is unrealistic in this context.

**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:***

**Organize the data into a variable: Satisfaction scores.**

Satisfaction scores: [7, 8, 9, 6, 8, 7, 9, 7, 8, 7]

**Calculate the mean (average) of the satisfaction scores:**

Mean of satisfaction scores (x̄) = (7 + 8 + 9 + 6 + 8 + 7 + 9 + 7 + 8 + 7) / 10 = 7.6

**Calculate the deviations from the mean for each satisfaction score:**

Deviations from the mean (xi - x̄): [7 - 7.6, 8 - 7.6, 9 - 7.6, 6 - 7.6, 8 - 7.6, 7 - 7.6, 9 - 7.6, 7 - 7.6, 8 - 7.6, 7 - 7.6] = [-0.6, 0.4, 1.4, -1.6, 0.4, -0.6, 1.4, -0.6, 0.4, -0.6]

**Calculate the squared deviations from the mean for each satisfaction score:**

Squared deviations from the mean ((xi - x̄)^2): [(-0.6)^2, 0.4^2, 1.4^2, (-1.6)^2, 0.4^2, (-0.6)^2, 1.4^2, (-0.6)^2, 0.4^2, (-0.6)^2]

= [0.36, 0.16, 1.96, 2.56, 0.16, 0.36, 1.96, 0.36, 0.16, 0.36]

**Calculate the variance as the average of the squared deviations from the mean:**

Variance = Sum of squared deviations / (n - 1)

= (0.36 + 0.16 + 1.96 + 2.56 + 0.16 + 0.36 + 1.96 + 0.36 + 0.16 + 0.36) / (10 - 1)

= 9.72 / 9

= 1.08 (rounded to two decimal places)

**Calculate the standard error of the mean (SE):**

Standard error of the mean (SE) = sqrt(Variance / n)

= sqrt(1.08 / 10)

= sqrt(0.108)

= 0.328 (rounded to three decimal places)

**RESULT**

Therefore, the standard error of the mean satisfaction score is approximately 0.328. This indicates the average amount of variation or uncertainty in the sample mean compared to the true population mean.

**20. Scenario: A company wants to analyze 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:***

**Organize the data into two variables: Advertising Expenditure (X) and Sales (Y).**

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

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

**Calculate the mean (average) of both variables:**

Mean of Advertising Expenditure (X̄) = (10 + 15 + 12 + 8 + 14) / 5 = 11.8

Mean of Sales (Ȳ) = (25 + 30 + 28 + 20 + 26) / 5 = 25.8

**Calculate the deviations from the mean for both variables:**

Deviations from the Mean of Advertising Expenditure (Xi - X̄): [-1.8, 3.2, 0.2, -3.8, 2.2]

Deviations from the Mean of Sales (Yi - Ȳ): [-0.8, 4.2, 2.2, -5.8, 0.2]

**Calculate the product of the deviations for each data point:**

Product of Deviations = (-1.8 \* -0.8, 3.2 \* 4.2, 0.2 \* 2.2, -3.8 \* -5.8, 2.2 \* 0.2) = (1.44, 13.44, 0.44, 22.04, 0.44)

**Calculate the sum of the product of the deviations:**

Sum of Product of Deviations = 1.44 + 13.44 + 0.44 + 22.04 + 0.44 = 37.8

**Calculate the sum of squared deviations for Advertising Expenditure:**

Sum of (Xi - X̄)^2 = (-1.8)^2 + 3.2^2 + 0.2^2 + (-3.8)^2 + 2.2^2 = 24.28

**Calculate the sum of squared deviations for Sales:**

Sum of (Yi - Ȳ)^2 = (-0.8)^2 + 4.2^2 + 2.2^2 + (-5.8)^2 + 0.2^2 = 66.28

**Calculate the slope coefficients (β1) of the multiple regression model:**

β1 = Sum of Product of Deviations / Sum of (Xi - X̄)^2

= 37.8 / 24.28

= 1.5578 (rounded to four decimal places)

**Calculate the intercept coefficient (β0) of the multiple regression model:**

β0 = Ȳ - β1 \* X̄

= 25.8 - (1.5578 \* 11.8)

= 6.3948 (rounded to four decimal places)

**The multiple regression equation to predict Sales (Y) based on Advertising Expenditure (X) is:**

Y = β0 + β1 \* X

= 6.3948 + 1.5578 \* X

**RESULT:**

Therefore, the multiple regression analysis predicts Sales (Y) based on Advertising Expenditure (X) with the equation Y = 6.3948 + 1.5578 \* X.