

Exercise 07 – December 11, 2024

1. A researcher wants to compare the growth of plants under three types of fertilizers (A, B, and C). The heights of the plants after 30 days (in cm) are:

| Fertilizer A | Fertilizer B | Fertilizer C |
|--------------|--------------|--------------|
| 15 | 20 | 25 |
| 16 | 22 | 27 |
| 14 | 19 | 26 |
| 15 | 21 | 28 |
| 17 | 20 | 24 |

$k=3$

① $H_0 = \mu_A = \mu_B = \mu_C$
 $H_1 = \text{see if there's any difference}$

Does the type of fertilizer (A, B, or C) significantly affect plant growth (with $\alpha = 0.05$)? Perform a one-way ANOVA to determine if fertilizer type affects plant growth. Create a null hypothesis and alternative hypothesis first.

Solution:

State the Hypotheses

Null Hypothesis (H_0):

The mean plant heights are the same for all three fertilizers:

$$\mu_A = \mu_B = \mu_C$$

Alternative Hypothesis (H_1):

At least one fertilizer produces a different mean plant height.

② $Df_{\text{between}} = k - 1$
 $3 - 1$

$$Df_{\text{within}} = N - k$$

$$Df_{\text{total}} = 12 + 2 = 14 = 15 - 1$$

use f distribution = 12

$$F_{\text{crit}} = 3.88$$

SUMMARY

| Groups | Count | Sum | Average | Variance |
|--------------|-------|-----|---------|----------|
| Fertilizer A | 5 | 77 | 15,4 | 1,3 |
| Fertilizer B | 5 | 102 | 20,4 | 1,3 |
| Fertilizer C | 5 | 130 | 26 | 2,5 |

ANOVA

| Source of Variation | SS | df | MS | F | P-value | F crit |
|---------------------|-------|----|-------|-----------|------------|-------------|
| Between Groups | 281,2 | 2 | 140,6 | 82,705882 | 9,5762E-08 | 3,885293835 |
| Within Groups | 20,4 | 12 | 1,7 | | | |
| Total | 301,6 | 14 | | | | |

Group Means and Overall Mean

Calculate the grand mean (\bar{X})

$$\bar{X} = \frac{15+16+14+15+17+20+22+19+21+20+25+27+26+28+24}{15} = 20.6$$

① $H_0 = \mu_A = \mu_B = \mu_C$
 $H_1 = \text{see if there's any difference}$

② $Df_{\text{between}} = k - 1$
 $3 - 1$

$Df_{\text{within}} = N - k$

Df_{total}
 $= 12 + 2 = 14 = 15 - 3$
 use F distribution $= 17$

$F_{\text{crit}} = 3.88$

③ $\bar{x}_1 = \frac{15 + 16 + 14 + 15 + 17}{5} = \frac{77}{5}$

$\bar{x}_2 = \frac{102}{5}$

$\bar{x}_3 = 26$

$\frac{G}{N} = \frac{309}{15} = 20.6$

$SS_{\text{total}} = \sum (x - \bar{x})^2 = (15 - 20.6)^2 + \dots + (24 - 20.6)^2 = 301.6$

$SS_{\text{between}} = 5 \times ((15.4 - 20.6)^2 + (20.4 - 20.6)^2 + (26 - 20.6)^2)$
 $= 281.2$

$SS_{\text{within}} = \text{total} - \text{between} = 301.6 - 281.2 = 20.4$

④ $MS_{\text{between}} = SS_b / df_b$
 $= 281.2 / 2 = 140.6$

$MS_{\text{within}} = SS_w / df_w$
 $= 20.4 / 12$
 $= 1.7$

$FFF = \frac{140.6}{1.7} = 82.71$

⑤ $F = 82.71$ $82.71 > 3.88$
 $F_{crit} = 3.88$ F is bigger than F_{crit}

Calculate the group means

$$\bar{X}_A = \frac{15+16+14+15+17}{5} = 15.4$$

$$\bar{X}_B = \frac{20+22+19+21+20}{5} = 20.4$$

$$\bar{X}_C = \frac{25+27+26+28+24}{5} = 26.0$$

Hence we reject the H_0

if $F > F_{crit}$
we reject

if $F \leq F_{crit}$
we fail to reject

Sum of Squares

Total Sum of Squares:

$$SS_{total} = (15-20.6)^2 + (16-20.6)^2 + \dots + (24-20.6)^2 = 89.2$$

Between-Groups Sum of Squares:

$$SS_{between} = 5 \times ((15.4-20.6)^2 + (20.4-20.6)^2 + (26.0-20.6)^2) = 71.6$$

Within-Groups Sum of Squares:

$$SS_{within} = SS_{total} - SS_{between} = 89.2 - 71.6 = 17.6$$

Calculate Mean Squares and F-Statistic

Degree of Freedom (df):

$$df_{between} = k-1 = 3-1 = 2$$

$$df_{within} = N-k = 15-3 = 12$$

Mean Squares (MS):

$$MS_{between} = SS_{between} / df_{between} = 71.6 / 2 = 35.8$$

$$MS_{within} = SS_{within} / df_{within} = 17.6 / 12 = 1.47$$

Calculate F-Statistic:

$$F = \frac{MS_{between}}{MS_{within}} = \frac{35.8}{1.47} \approx 24.35$$

Decision

Critical F-Value:

From an F-distribution table with $df_{between} = 2$ and $df_{within} = 12$ at $\alpha = 0.05$, the critical value is $F_{critical} = 3.89$

Compare F:

Since $F = 24.35 > 3.89$, reject the null hypothesis.

OR

Calculate the p-Value

The p-value is the probability of observing an F-value as extreme as the calculated value ($F = 24.35$) under the null hypothesis.

Using $df_{\text{between}} = 2$ and $df_{\text{within}} = 12$, the p-value can be found using an F-distribution table or statistical software.

For $F = 24.35$: Using statistical software or a table, we find that $p\text{-value} < 0.001$

Decision Rule

Compare the p-value to $\alpha = 0.05$:

- If $p \leq \alpha$, reject the null hypothesis (H_0).
- If $p > \alpha$, fail to reject the null hypothesis.

In this case, $p < 0.001 < 0.05$, so we **reject the null hypothesis**.

Conclusion

The p-value is extremely small ($p < 0.001$), which indicates very strong evidence against the null hypothesis. Therefore, we conclude that the type of fertilizer has a significant effect on plant growth.

OR

The F-statistic ($F=24.35$) is significant at $\alpha = 0.05$. This indicates that the type of fertilizer has a significant effect on plant growth. At least one fertilizer produces a different mean plant height.

2. A researcher wants to determine if there is an association between **plant type** and **fertilizer preference**. The researcher surveys 90 plants and records the following data:

| Fertilizer | Plant Type A | Plant Type B | Plant Type C | Total |
|--------------|--------------|--------------|--------------|-----------|
| Fertilizer X | 10 | 20 | 10 | 40 |
| Fertilizer Y | 15 | 10 | 5 | 30 |
| Fertilizer Z | 5 | 5 | 10 | 20 |
| Total | 30 | 35 | 25 | 90 |

Conduct a Chi-Square test of Independence whether plant type and fertilizer preference are independent at $\alpha = 0.05$.

Solution:

State the Hypotheses

Null Hypothesis (H_0):

Plant type and fertilizer preference are independent.

Alternative Hypothesis (H_1):

Plant type and fertilizer preference are not independent (there is an association)

Calculate the Expected Frequencies

The formula for the expected frequency for a cell is:

$$E_{ij} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

For each cell:

① H_0 = there is no association
 H_1 = there is an association

$$df = (3-1) \times (3-1) = 4$$

$$\alpha = 0.05$$

$$C_v = 9.488$$

$$df = (r-1) \times (c-1) \text{ [independent]}$$

$$df = k-1 \text{ good fit}$$

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\textcircled{2} E_{ij} = \frac{\text{Row total} \times \text{Column total}}{\text{grand total}}$$

$$E_{11} = \frac{40 \times 30}{90}$$

$$E_{...}$$

$$E_{33} = \frac{20 \times 25}{90}$$

$$\textcircled{3} \text{ chi square stats}$$

$$\chi^2 = \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\chi^2 = 11.24$$

$$11.24 > 9.488 \text{ reject } H_0$$

Fertilizer X, Plant Type A:

$$E_{11} = \frac{40 \times 30}{90} = \frac{1200}{90} = 13.33$$

Fertilizer X, Plant Type B:

$$E_{12} = \frac{40 \times 35}{90} = \frac{1400}{90} = 15.56$$

Fertilizer X, Plant Type C:

$$E_{13} = \frac{40 \times 25}{90} = \frac{1000}{90} = 11.11$$

Repeat for all cells to construct the Expected Frequency Table:

| Fertilizer | Plant Type A (E) | Plant Type B (E) | Plant Type C (E) | Total |
|--------------|------------------|------------------|------------------|-------|
| Fertilizer X | 13.33 | 15.56 | 11.11 | 40 |
| Fertilizer Y | 10 | 11.67 | 8.33 | 30 |
| Fertilizer Z | 6.67 | 7.78 | 5.56 | 20 |
| Total | 30 | 35 | 25 | 90 |

Compute the Chi-Square Statistic

The formula for the Chi-Square statistic is:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Calculate for each cell:

Fertilizer X, Plant Type A:

$$\frac{(10-13.33)^2}{13.33} = \frac{(-3.33)^2}{13.33} = \frac{11.09}{13.33} = 0.83$$

Fertilizer X, Plant Type B:

$$\frac{(20-15.56)^2}{15.56} = \frac{(4.44)^2}{15.56} = \frac{19.71}{15.56} = 1.27$$

Fertilizer X, Plant Type C:

$$\frac{(10-13.33)^2}{13.33} = \frac{(-3.33)^2}{13.33} = \frac{11.09}{13.33} = 0.11$$

Repeat for all cells. Summing these values gives:

$$\chi^2 = 0.83 + 1.27 + 0.11 + 2.50 + 0.24 + 1.33 + 0.42 + 0.99 + 3.55 = 11.24$$

Degrees of Freedom

The degrees of freedom (df) for a contingency table is:

$$df = (\text{Number of Rows} - 1) \times (\text{Number of Columns} - 1)$$

Here:

$$df = (3 - 1) \times (3 - 1) = 2 \times 2 = 4$$

Determine the Critical Value

From the Chi-Square distribution table, for df = 4 and $\alpha = 0.05$:

$$\chi^2_{\text{critical}} = 9.488$$

Decision

Compare the test statistics to the critical value:

- If $\chi^2 \leq \chi^2_{\text{critical}}$, fail to reject H_0
- If $\chi^2 > \chi^2_{\text{critical}}$, reject H_0

Here:

$$\chi^2 = 11.24 \text{ and } \chi^2_{\text{critical}} = 9.488$$

Since $11.24 > 9.488$, we **reject the null hypothesis**.

Conclusion

At $\alpha = 0.05$, there is sufficient evidence to conclude that **plant type** and **fertilizer preference** are not independent. There is an association between plant type and fertilizer preference

3. A professor wants to investigate whether the **type of programming language** (Python, Java, C++) and the **study method** (Self-Study, Instructor-Led) affects students' test scores. The professor records the test scores of students after completing a course under each combination of factors.

| Language | Self-Study | Instructor-Led |
|----------|------------|----------------|
| Python | 78, 82, 85 | 90, 88, 92 |
| Java | 72, 75, 74 | 85, 80, 84 |
| C++ | 65, 68, 70 | 78, 75, 80 |

Perform a Two-Way ANOVA to determine if there are significant effects of programming language, study method, or their interaction on test scores.

Create all null hypotheses.

Use $\alpha = 0.05$

Solution:

State Hypotheses:

Main Effect of Programming Language (H_0): Mean test scores are the same across Python, Java, and C++.

Main Effect of Study Method (H_0): Mean test scores are the same for Self-Study and Instructor-Led methods.

Interaction Effect (H_0): There is no interaction between programming language and study method.

Grand Mean (\bar{X}) = 78.9444 (average for all 18 values)

Group Means:

Python : 85.8333
Java : 78.3333
C++ : 72.6667
Self-Study : 74.3333
Instructor-Led : 83.5556

| | |
|---------------------------|-----------|
| Python and Self-Study | : 81.6667 |
| Python and Instructor-Led | : 90.0 |
| Java and Self-Study | : 73.6667 |
| Java and Instructor-Led | : 83.0 |
| C++ and Self-Study | : 67.6667 |
| C++ and Instructor-Led | : 77.6667 |

Compute Sum of Squares:

$$\text{Total} = \sum (x_{ij} - \bar{X})^2$$

Using $\bar{X} = 78.94$, calculate for each observation:

Sum of Squares for Factor Programming Language (A):

$$\begin{aligned} \text{SSA} &= 6 * (85.8333 - 78.9444)^2 + 6 * (78.3333 - 78.94)^2 + 6 * (72.6667 - 78.94)^2 \\ \text{SSA} &= 523.4394 \end{aligned}$$

Sum of Squares for Factor Study Method (B):

$$\begin{aligned} \text{SSB} &= 9 * (74.3333 - 78.9444)^2 + 9 * (83.5556 - 78.9444)^2 \\ \text{SSB} &= 382.7222 \end{aligned}$$

Sum of Squares Within (Error)

$$\begin{aligned} \text{SS Python and Self-Study} &= (78 - 81.6667)^2 + (82 - 81.6667)^2 + (85 - 81.6667)^2 = 24.6664 \\ \text{SS Python and Instructor-Led} &= 8 \\ \text{SS Java and Self-Study} &= 4.6667 \\ \text{SS Java and Instructor-Led} &= 14 \\ \text{SS C++ and Self-Study} &= 12.6667 \\ \text{SS C++ and Instructor-Led} &= 12.6667 \end{aligned}$$

$$\text{SSE} = 24.6667 + 8 + 4.6667 + 14 + 12.6667 + 12.6667 = 76.6664$$

Total Sum of Squares

$$\begin{aligned} \text{SSTotal} &= (78 - 78.9444)^2 + (82 - 78.9444)^2 + \dots + (80 - 78.9444)^2 \\ \text{SSTotal} &= 984.9444 \end{aligned}$$

$$\text{SSInteraction} = \text{SS Total} - \text{SSA} - \text{SSB} - \text{SSE}$$

$$\text{SSInteraction} = 984.9444 - 523.4394 - 382.7222 - 76.6664 = 2.1164$$

Degrees of Freedom:

$$df_A = 2, df_B = 1, df_{\text{interaction}} = 2, df_{\text{within}} = 12, df_{\text{Total}} = 17$$

Mean Squares and FFF-Statistics:

$$MS_A = \text{SS}/df = 523.4394/2 = 261.7197$$

$$MS_B = 191.3611$$

$$MS_{A \times B} = 1.0852$$

$$MS_E = 38.3332$$

$$F_A = MS_A / MS_E = 261.7197 / 38.3332 = 40.9652$$

$$F_B = 59.9045$$

$$F_{A \times B} = 0.1656$$

Decision:

p-value Programming Language for

$F = 40.965$, $df = (2, 12)$ at $\alpha = 0.05$ is 0.00000435

p-value Study Method for

$F = 59.9045$, $df = (1, 12)$ at $\alpha = 0.05$ is 0.00000527

p-value Interaction for

$F = 0.1656$, $df = (2, 12)$ at $\alpha = 0.05$ is 0.84928886

Conclusion:

Significant main effects of programming language on test scores. (p-value < α)

Significant main effects of study method on test scores. p-value < α)

No Significant interaction between language and study methods. p-value > α)

ANOVA

| <i>Source of Variation</i> | <i>SS</i> | <i>df</i> | <i>MS</i> | <i>F</i> | <i>P-value</i> |
|----------------------------|-------------|-----------|-------------|------------|----------------|
| Sample (study) | 523,4444444 | 2 | 261,7222222 | 40,9652174 | 4,3476E-06 |
| Columns (program) | 382,7222222 | 1 | 382,7222222 | 59,9043478 | 5,26602E-06 |
| Interaction | 2,111111111 | 2 | 1,055555556 | 0,16521739 | 0,849605144 |
| Within | 76,66666667 | 12 | 6,388888889 | | |
| Total | 984,9444444 | 17 | | | |

Exercise 06 – November 20, 2024

1. A company claims their light bulbs last 1000 hours on average. A sample of 10 bulbs yields the following lifespans (in hours):

950, 960, 970, 980, 1020, 1030, 990, 1010, 1000, 995

Test whether the mean lifespan differs significantly from 1000 hours using $\alpha = 0.05$

Solution:

Null Hypothesis (H_0): The mean lifespan is 1000 hours ($\mu = 1000$)

Alternative Hypothesis (H_1): The mean lifespan is not 1000 hours ($\mu \neq 1000$)

$$\text{Sample Mean} = \frac{950 + 960 + 970 + 980 + 1020 + 1030 + 990 + 1010 + 1000 + 995}{10} = 990.5$$

Standard Deviation (s): $s \approx 25.87$

Formula for t-statistic:

$$t = \frac{990.5 - 1000}{25.87/\sqrt{10}} \approx \frac{-9.5}{8.18} \approx -1.16$$

Compare t-statistic with Critical Value

Degrees of freedom: $n - 1 = 10 - 1 = 9$.

At $\alpha = 0.05$ (two-tailed), the critical t-value is approximately ± 2.262 (from the t-distribution table).

Since $t = -1.16$ falls within the range $[-2.262, 2.262]$, we fail to reject the null hypothesis.

2. A fitness coach measures the weight of 8 clients before and after a 6-week training program.

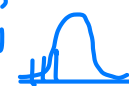
| Client | Before (kg) | After (kg) | Difference (d) |
|--------|-------------|------------|----------------|
| 1 | 85 | 82 | -3 |
| 2 | 78 | 75 | -3 |
| 3 | 90 | 85 | -5 |
| 4 | 76 | 74 | -2 |
| 5 | 88 | 85 | -3 |
| 6 | 81 | 78 | -3 |
| 7 | 79 | 76 | -3 |
| 8 | 92 | 89 | -3 |

Conduct a paired t-test to determine if the training program significantly reduced weight. Use $\alpha = 0.05$

Solution:

Null Hypothesis (H_0): The training program has no effect on weight ($\mu_d = 0$).

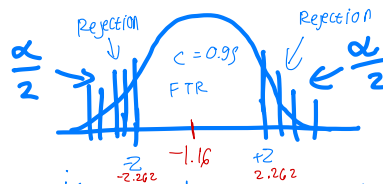
Alternative Hypothesis (H_1): The training program reduces weight ($\mu_d < 0$).

$\alpha = 0.05$

 the mean difference is less than zero

FTR =
Fail to Reject

← opposite

→ whenever our alternative hypothesis is not equal to something it's two tailed

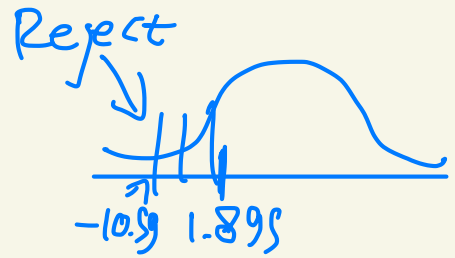


if z value greater than critical then reject null hypothesis
 if z value less than critical then fail to reject null hypothesis

$$\alpha = 0.05$$

Null hypothesis = $\mu_d = 0$

Null alternative = $\mu_d < 0$



$$\mu_d = \frac{-3 + -3 + -5 + -2 + -3 + -3 + -3 + -3}{8} = -3.125$$

$$s = \sqrt{\frac{(-3 + 3.125)^2 + \dots + (-3 + 3.125)^2}{8 - 1}}$$

$$= 0.8345$$

$$t = \frac{-3.125 - 0}{\frac{\sqrt{546}}{28} / \sqrt{8}} = -10.59$$

$$df = 8 - 1$$

$$= 7$$

$$\alpha = 0.05$$

$$t_{\text{-table}} = 1.895$$

Calculate Mean and Standard Deviation of Differences (d):

$$\bar{d} = \frac{\sum d}{n} = \frac{-3 - 3 - 5 - 2 - 3 - 3 - 3 - 3}{8} = \frac{-25}{8} = -3.125$$

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n - 1}} \approx 0.835$$

Formula for t-statistic:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{-3.125}{0.835 / \sqrt{8}} = \frac{-3.125}{0.295} \approx -10.59$$



Degrees of Freedom (df): $df = n - 1 = 8 - 1 = 7$

Critical t-value for $\alpha = 0.05$ (one-tailed): 1.895

Since $-10.59 < 1.895$, we reject H_0

Conclusion: The training program significantly reduced weight.

$$A_x > B_x$$

$$A_x = B_x$$

3. A nutritionist wants to test if a new diet plan (Group A) significantly improves weight loss compared to a standard diet plan (Group B).

The following data was collected:

| Group | Sample Size (n) | Mean Weight Loss (x) | Standard Deviation (s) |
|--------------------|-----------------|----------------------|------------------------|
| Group A (New) | 25 | 8 kg | 2 |
| Group B (Standard) | 25 | 6 kg | 2.5 |

Perform an independent t-test to determine if the new diet plan significantly improves weight loss at a significant level of $\alpha = 0.05$

Solution:

State Hypotheses

- Null Hypothesis (H_0): The mean weight loss for both groups is equal ($\mu_A = \mu_B$)
- Alternative Hypothesis (H_1): The new diet plan leads to greater weight loss ($\mu_A > \mu_B$)

Calculate the t-statistic:

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Substitute the values:

$$t = \frac{8 - 6}{\sqrt{\frac{2^2}{25} + \frac{2.5^2}{25}}}$$

First, calculate the variances divided by sample sizes:

$$\frac{2^2}{25} = 0.16, \quad \frac{2.5^2}{25} = 0.25$$

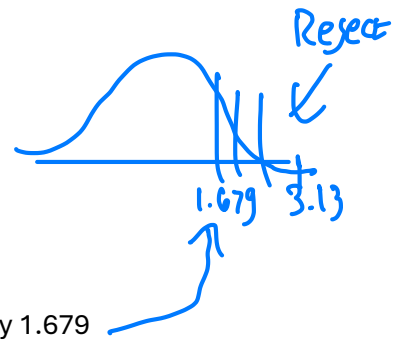
Sum them:



$$\sqrt{0.16 + 0.25} = \sqrt{0.41} \approx 0.64$$

Now calculate t:

$$t = \frac{8 - 6}{0.64} = \frac{2}{0.64} \approx 3.13$$



Degrees of Freedom and Critical t-value

Degrees of freedom: $df = n_A + n_B - 2 = 25 + 25 - 2 = 48$

At $\alpha = 0.05$ (one-tailed), the critical t-value for $df = 48$ is approximately 1.679

Compare the t-statistic to the critical value $t = 3.13 > 1.679$

Since the calculated t-value exceeds the critical t-value, we reject the null hypothesis.

The new diet plan leads to significantly greater weight loss than the standard diet plan.

<https://www.meracalculator.com/math/t-distribution-critical-value-table.php>

Exercise 05 – October 23-24, 2024

1. The following data set represents the scores of 5 students in a quiz:

Scores: 70, 85, 78, 90, 88

Find the standard deviation from those data.

$$\bar{x} = \frac{\text{Sum}}{n} = \frac{70+85+78+90+88}{5} = 82.2$$

Solution:

1. Mean (average):

$$\text{Mean} = \frac{70 + 85 + 78 + 90 + 88}{5} = 82.2$$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

2. Deviation of each score from the mean:

- $70 - 82.2 = -12.2$
- $85 - 82.2 = 2.8$
- $78 - 82.2 = -4.2$
- $90 - 82.2 = 7.8$
- $88 - 82.2 = 5.8$

$$s = \sqrt{\frac{(70-82.2)^2 + \dots + (88-82.2)^2}{5}} = 7.33$$

3. Squared deviations:

- $(-12.2)^2 = 148.84$
- $(2.8)^2 = 7.84$
- $(-4.2)^2 = 17.64$
- $(7.8)^2 = 60.84$
- $(5.8)^2 = 33.64$

if the question don't mention anything about taken from a sample then divide with N

4. Variance (average of squared deviations):

$$\text{Variance} = \frac{148.84 + 7.84 + 17.64 + 60.84 + 33.64}{5} = 53.76$$

5. Standard deviation (square root of the variance):

$$\text{Standard deviation} = \sqrt{53.76} \approx 7.33$$

So, the standard deviation of the quiz scores is approximately 7.33.

2. Suppose a survey indicates that 30% of people prefer coffee over tea. If you randomly select 100 people, what is the probability that fewer than 25 people prefer coffee? Use z-table

$$P(X < 25)$$

$$\mu = n \cdot p \quad \sigma = \sqrt{n \cdot p \cdot q}$$

$$\lambda = n \cdot q$$

Solution:

1. $n = 100$, $p = 0.30$, and $q = 0.70$.

2. Check conditions:

- $n \cdot p = 100 \cdot 0.30 = 30$,
- $n \cdot (1 - p) = 100 \cdot 0.70 = 70$.

Both are greater than 5, so the normal approximation can be used.

3. Calculate the mean and standard deviation:

- $\mu = 100 \cdot 0.30 = 30$,
- $\sigma = \sqrt{100 \cdot 0.30 \cdot 0.70} = \sqrt{21} \approx 4.58$.

4. Apply the continuity correction:

- You want $P(X < 25)$, so with the continuity correction, calculate $P(X \leq 24.5)$.

5. Standardize:

$$Z = \frac{24.5 - 30}{4.58} = \frac{-5.5}{4.58} \approx -1.20$$

6. Find the z-score in the z-table:

- For $Z = -1.20$, the z-table gives $P(Z \leq -1.20) \approx 0.1151$.

The probability that fewer than 25 people prefer coffee is approximately 0.1151 (or 11.51%).

$$P(X < k) - 0.5$$

$$P(X > k) + 0.5$$

$$P(X \leq k) + 0.5$$

$$P(X \geq k) - 0.5$$

3. You are conducting an experiment with 100 trials ($n = 100$), and the probability of success in each trial is $p = 0.4$. You want to find the probability that at least 45 successes will occur.

Solution:

1. Check the conditions:

- $n \cdot p = 100 \cdot 0.4 = 40$,
- $n \cdot (1 - p) = 100 \cdot 0.6 = 60$. Both conditions are satisfied.

2. Calculate the mean and standard deviation:

- $\mu = 100 \cdot 0.4 = 40$,
- $\sigma = \sqrt{100 \cdot 0.4 \cdot 0.6} = \sqrt{24} \approx 4.9$.

3. **Apply continuity correction:** We want $P(X \geq 45)$. Using the continuity correction, we calculate $P(X \geq 44.5)$.
4. **Convert to a z-score:**

$$Z = \frac{44.5 - 40}{4.9} = \frac{4.5}{4.9} \approx 0.92$$

5. **Find the probability using the z-table:** From the z-table, $P(Z \leq 0.92) \approx 0.8212$.

Since we are looking for $P(X \geq 45)$, we calculate $P(Z \geq 0.92)$:

$$P(Z \geq 0.92) = 1 - 0.8212 = 0.1788$$

The probability of having at least 45 successes is approximately **0.1788**, or 17.88%

Exercise 04 – October 09, 2024

1. Find the percentage returns from an investment over 5 consecutive years, were:

Year 1: 10%

Year 2: 15%

Year 3: -5%

Year 4: 8%

Year 5: 12%

$$\sqrt[5]{1.10 \times 1.15 \times 0.95 \times 1.08 \times 1.12} = 1.078 - 1 = 0.078 \times 100 = 7.8\%$$

Solution:

First, convert the percentages to decimal form and add 1 to each to account for negative growth:

1.10, 1.15, 0.95, 1.08, 1.12

Now, apply the geometric mean formula:

$$GM = \sqrt[5]{1.10 \times 1.15 \times 0.95 \times 1.08 \times 1.12}$$

First, multiply the values together:

$$1.10 \times 1.15 \times 0.95 \times 1.08 \times 1.12 = 1.422$$

Now, take the 5th root:

$$GM = \sqrt[5]{1.422} \approx 1.073$$

Convert back to a percentage:

$$GM = (1.073 - 1) \times 100 = 7.3\%$$

2. Create a box plot to compare the distribution of data from two different groups, each containing an odd number of data points. Interpret the box plots to compare the central tendency, spread, and potential outliers between the groups.

You are given the following data sets for two groups:

Group A: 7, 9, 12, 13, 14, 15, 16

Group B: 5, 7, 8, 10, 12, 15, 18

Tasks:

- Calculate the five-number summary (minimum, 1st quartile Q1, median, 3rd quartile Q3, and maximum) for each group.
- Draw the box plots for both groups on the same axis, labeling the minimum, Q1, median, Q3, and maximum values.
- Compare the distributions of the two groups based on the box plots:
 - Which group has a higher median?
 - Are there any outliers?

Solution:

Group A:

Minimum: 7

Q1: Median of the lower half (7,9,12) → Q1 = 9

Median: Middle value (13)

Q3: Median of the upper half (14,15,16) → Q3 = 15

Maximum: 16

Five-number summary for Group A: 7, 9, 13, 15, 16

Group B:

Minimum: 5

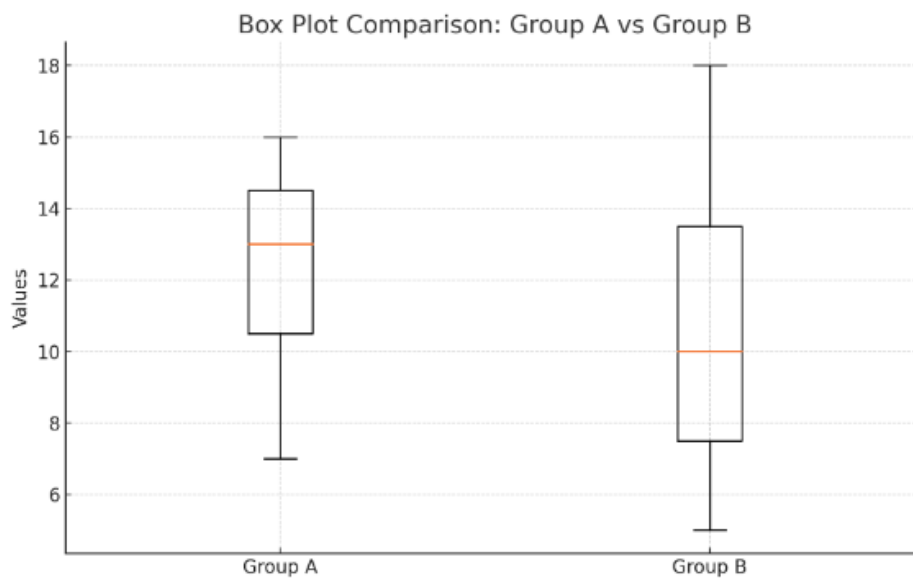
Q1: Median of the lower half (5,7,8) → Q1 = 7

Median: Middle value (10)

Q3: Median of the upper half (12,15,18) → Q3 = 15

Maximum: 18

Five-number summary for Group B: 5, 7, 10, 15, 18



Comparison Between Groups

Median: Group A has a higher median (13) compared to Group B (10).

Outliers: Neither group has extreme outliers based on the data provided.

3. A card is drawn from a standard deck of 52 cards, and then a coin is flipped. What is the probability of drawing a "King" from the deck and flipping a "Tail"?

Solution:

Probability of drawing a "King" from a deck of cards:

There are 4 Kings in a deck of 52 cards, so:

$$P(\text{King}) = \frac{4}{52} = \frac{1}{13}$$

Probability of flipping a "Tail":

$$P(\text{Tail}) = \frac{1}{2}$$

$$P(A \text{ and } B) = P(A) \times P(B)$$

$$= \frac{4}{52} \times \frac{1}{2}$$

$$= \frac{1}{26}$$

Since the two events are independent, the probability of both events happening together is:

$$P(\text{King and Tail}) = P(\text{King}) \times P(\text{Tail}) = \frac{1}{13} \times \frac{1}{2} = \frac{1}{26}$$

4. Two departments at a company recorded the number of sales made by their top 10 salespeople in a month. The number of sales made are as follows:

Department X Sales: 12, 14, 17, 19, 21, 24, 26, 28, 30, 32

Department Y Sales: 13, 16, 18, 20, 23, 25, 27, 29, 31, 33

Please, construct a back-to-back stem-and-leaf display for the two departments' sales data.

Solution:

Back-to-Back Stem-and-Leaf Display:

| Department X (Leaf) | Stem | Department Y (Leaf) |
|---------------------|------|---------------------|
| 2 4 7 9 | 1 | 3 6 8 |
| 1 4 6 8 | 2 | 0 3 5 7 9 |
| 0 2 | 3 | 1 3 |

$$\binom{n}{x} p^x q^{n-x}$$

5. Calculate the probability of getting exactly 3 heads when flipping a fair coin 5 times (where getting heads is considered a success)

Solution:

N = 5 (number of trials)

x = 3 (number of successes)

$\pi = 0.5$ (probability of success, i.e., getting heads)

Using the formula:

$$P(x = 3) = \frac{N!}{x!(N-x)!} \pi^x (1 - \pi)^{N-x}$$

$$P(x = 3) = \frac{5!}{3! 2!} 0.5^3 0.5^2$$

$$P(x = 3) = \frac{5 \times 4}{2 \times 1} 0.5^5 = 10 \times \frac{1}{32} = \frac{10}{32} = \frac{5}{16}$$

Thus, the probability of getting exactly 3 heads in 5 flips of a fair coin is $\frac{5}{16}$ or approximately 0.3125

$$\begin{aligned} n &= 5 \\ p &= \frac{1}{2} \\ x &= 3 \\ q &= \frac{1}{2} \end{aligned}$$

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

$$\binom{5}{3} = 5C_3$$

$$10 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = 0.3125$$

6. In a basketball game, a player has a free throw success rate of 80%. If the player takes 15 free throws, what is the probability that they make at least 12 successful free throws?

Solution:

To find the probability of making at least 12 successful free throws, we need to calculate

$$P(x \geq 12) = P(x = 12) + P(x = 13) + P(x = 14) + P(x = 15)$$

N = 15 (number of trials)

$\pi = 0.8$ (probability of success)

$$\begin{aligned} n &= 15 \\ p &= 0.8 \\ q &= 0.2 \\ x &= 12 \end{aligned}$$

$X \quad Y \quad xy \quad x^2 \quad y^2$

For $x = 12$

$$P(x = 12) = \frac{15!}{12! 3!} 0.8^{12} 0.2^3 \approx 0.227$$

For $x = 13$

$$P(x = 13) = \frac{15!}{13! 2!} 0.8^{13} 0.2^2 \approx 0.236$$

For $x = 14$

$$P(x = 14) = \frac{15!}{14! 1!} 0.8^{14} 0.2^1 \approx 0.137$$

For $x = 15$

$$P(x = 15) = \frac{15!}{15!} 0.8^{15} 0.2^0 \approx 0.035$$

So, the probability that the player makes at least 12 successful free throws is approximately $0.227 + 0.236 + 0.137 + 0.035 = 0.635$.

7. A biologist studies the relationship between the number of hours of sunlight a plant receives and its height. The following data shows the hours of sunlight and the corresponding heights of 5 plants:

| Hours of Sunlight (X) | Height (cm) (Y) |
|-----------------------|-----------------|
| 2 | 10 |
| 4 | 15 |
| 6 | 20 |
| 8 | 25 |
| 10 | 30 |

there is a correlation
if r is close to 1, -1

Calculate the Pearson correlation coefficient.

Solution:

$x - \bar{x}$



| | X | Y | x | y | xy | x ² | y ² |
|-------|----|-----|----|-----|-----|----------------|----------------|
| | 2 | 10 | -4 | -10 | 40 | 16 | 100 |
| | 4 | 15 | -2 | -5 | 10 | 4 | 25 |
| | 6 | 20 | 0 | 0 | 0 | 0 | 0 |
| | 8 | 25 | 2 | 5 | 10 | 4 | 25 |
| | 10 | 30 | 4 | 10 | 40 | 16 | 100 |
| Total | 30 | 100 | 0 | 0 | 100 | 40 | 250 |
| Mean | 6 | 20 | 0 | 0 | | | |

$$\frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

$$\frac{5(100) - (30)(100)}{\sqrt{[5(40) - (30)^2][5(250) - (100)^2]}}$$

$$r = 1$$

Calculating

$$r = (\sum xy) / \sqrt{(\sum x^2 \sum y^2)} = 100 / \sqrt{(40 + 250)}$$

$$r = 100/100 = 1$$

Exercise 03 – October 02, 2024

$8P4$ the amount
↓
 nPr
↓
the group chosen

1. You have 8 people, and you need to select and arrange 4 of them in a row for a photo. How many different ways can you arrange them?

Solution:

Here, $n = 8$ (total people) and $r = 4$ (people to arrange).

We apply the permutation formula:

$$P(8, 4) = \frac{8!}{(8-4)!} = \frac{8!}{4!}$$

First, calculate $8!$ and $4!$:

$$8! = 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 40,320$$

$$4! = 4 \times 3 \times 2 \times 1 = 24$$

Now, calculate $P(8, 4)$:

$$P(8, 4) = \frac{40,320}{24} = 1,680$$

So, there are 1,680 ways to arrange 4 people out of 8 in a row.

2. You have 7 books, and you want to choose 4 to take on a trip. How many different ways can you select the books?

$7C4$

Solution:

This is a combination problem where $n = 7$ and $r = 4$:

$$C(7, 4) = \frac{7!}{4!(7-4)!} = \frac{7 \times 6 \times 5 \times 4}{4 \times 3 \times 2 \times 1} = \frac{840}{24} = 35$$

So, there are 35 ways to choose 4 books from a set of 7.

3. A bag contains 10 red balls and 15 blue balls. If you randomly select 5 balls without replacement, what is the probability that exactly 3 of the selected balls are red?

Solution:

$N = 25$ (total balls/ population),

$k = 10$ (total red balls),

$n = 5$ (balls selected/ sample),

$x = 3$ (we want to find the probability of selecting 3 red balls).

Using the hypergeometric formula:

$$p = \frac{{}^k C_x \cdot {}^{(N-k)} C_{(n-x)}}{{}^N C_n}$$

Remember Combination formula:

$${}_n C_r = \frac{n!}{(n-r)! r!}$$

success

total

$$\frac{10C3 \times 15C2}{25C5} = 0.237$$

$$23.7\%$$

$$p = \frac{{}_{10}C_3 \cdot {}_{(25-10)}C_{(5-3)}}{{}_{25}C_5}$$

$${}_{10}C_3 = \frac{10!}{7! 3!} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120$$

$${}_{15}C_2 = \frac{15!}{13! 2!} = \frac{15 \times 14}{2 \times 1} = 105$$

$${}_{25}C_5 = \frac{25!}{20! 5!} = \frac{25 \times 24 \times 23 \times 22 \times 21}{5 \times 4 \times 3 \times 2 \times 1} = 53,130$$

Now, calculate the probability:

$$p = \frac{120 \times 105}{53,130} = \frac{12,600}{53,130} \approx 0.2372$$

So, the probability of drawing exactly 3 red balls is approximately 0.2372 or 23.72%.

Exercise 02 – September 25, 2024

1. Calculate the Trimean for a dataset below

Data set:

10, 12, 15, 18, 21, 24, 27, 30, 33, 36, 39, 42, 45, 48, 50

Solution:

Step 1: Sort the Data

The data is already sorted in ascending order:

10, 12, 15, 18, 21, 24, 27, 30, 33, 36, 39, 42, 45, 48, 50

Step 2: Find the Median

Since there are 15 values (an odd number), the median is the 8th value in the sorted list:

Median = 30

Step 3: Find Q1 and Q3

Q1 (1st Quartile): This is the median of the lower half of the data (excluding the median itself):

Lower half = 10, 12, 15, 18, 21, 24, 27

The median of this lower half (7 values) is the 4th value:

Q1 = 18

Q3 (3rd Quartile): This is the median of the upper half of the data (excluding the median itself):

Upper half = 33, 36, 39, 42, 45, 48, 50

The median of this upper half (7 values) is the 4th value:

Q3 = 42

Step 4: Calculate the Trimean

The formula for the trimean is:

$$\text{Trimean} = \frac{Q1 + 2 \times \text{Median} + Q3}{4}$$

Substitute the values:

$$\text{Trimean} = \frac{18 + 2 \times 30 + 42}{4} = \frac{18 + 60 + 42}{4} = \frac{120}{4} = 30$$

Conclusion

The trimean of this dataset is 30. This measure incorporates the median (with twice the weight) and the quartiles, providing a balanced estimate of central tendency that is robust to outliers.

2. Geometric Mean

Suppose the population of a city changes over four years with the following annual growth rates:

Year 1: +5%

Year 2: +10%

$$\begin{aligned} & \sqrt[4]{1.05 \times 1.10 \times 0.97 \times 1.06} \\ & = 1.0426 - 1 = 0.0426 \times 100 = 4.26\% \end{aligned}$$

Year 3: -3%

Year 4: +6%

Calculate the geometric mean of the growth rates to find the average population growth rate over these 4 years

Solution:

Step 1: Convert percentages to growth factors

Convert each percentage into a decimal growth factor by adding 1:

$$\text{Year 1: } 1 + 0.05 = 1.05$$

$$\text{Year 2: } 1 + 0.10 = 1.10$$

$$\text{Year 3: } 1 - 0.03 = 0.97$$

$$\text{Year 4: } 1 + 0.06 = 1.06$$

Step 2: Calculate the product of the growth factors

$$1.05 \times 1.10 \times 0.97 \times 1.06 = 1.181841$$

Step 3: Calculate the geometric mean

Since there are 4 years, the geometric mean is the 4th root of the product:

$$GM = \sqrt[4]{1.181841} \approx 1.0426$$

Step 4: Convert back to percentage

To express this as a percentage growth rate, subtract 1 and multiply by 100:

$$1.0426 - 1 = 0.0426 \text{ or } 4.26\%$$

Conclusion:

The average annual population growth rate over these 4 years is approximately 4.26% per year, even though the individual rates varied each year. The geometric mean smooths out the effect of the fluctuations and gives a meaningful long-term average growth rate.

3. Trimmed Mean

Consider the following dataset of 10 values representing exam scores:

~~65~~, 70, 72, 75, 80, 85, 90, 92, 95, ~~100~~

Calculate the 10% trimmed mean

Solution:

Step 1: Sort the data (if not already sorted)

In this case, the data is already sorted in ascending order:

65, 70, 72, 75, 80, 85, 90, 92, 95, 100

Step 2: Trim 10% from both ends

Since we have 10 values, trimming 10% means removing 10% of the values from each end.

10% of 10 = 1 value from the top and 1 value from the bottom.

So, remove the smallest value (65) and the largest value (100).

$$\frac{70 + 72 + 75 + 80 + 85 + 90 + 92 + 95}{8} = 88.625$$

Step 3: Calculate the mean of the remaining values

The remaining values are:

70, 72, 75, 80, 85, 90, 92, 95

Now, calculate the mean of these 8 values:

Sum = $70 + 72 + 75 + 80 + 85 + 90 + 92 + 95 = 659$

$$\text{Mean} = \frac{659}{8} = 82.375$$

Conclusion:

The 10% trimmed mean of the dataset is 82.375.

By trimming the lowest and highest values (65 and 100), we remove the possible influence of any outliers, providing a more robust average value.

Exercise 01 – September 11, 2024

1. Create a Stem-and-Leaf Display

Data set:

62, 65, 68, 70, 73, 75, 75, 78, 81, 83, 84, 85, 87, 89, 92, 95, 96, 98, 100

Solution:

| Stem | Leaf |
|------|-------------|
| 6 | 2 5 8 |
| 7 | 0 3 5 5 8 |
| 8 | 1 3 4 5 7 9 |
| 9 | 2 5 6 8 |
| 10 | 0 |

2. Construct a Box Plot

Given the following dataset of students' test scores:

Dataset:

55, 60, 62, 63, 65, 66, 68, 70, 72, 75, 77, 78, 80, 85, 88

Tasks:

- Determine the five-number summary (minimum, 25th Quartile, 50th Quartile, 75th Quartile, maximum).
- Draw the box plot based on the five-number summary with whiskers (use $1.5 \times \text{H-spread}$ to identify outliers for step).
- Identify any potential outliers (outside value or/and far out value).

Solution:

- Five-number summary:

Minimum: 55

The 25th percentile is the value between the 4th and 5th values, which is 63

The 50th percentile: 70 (Middle value of the data set.)

The 75th percentile is the value between the 11th and 12th values, which is 78

Maximum: 88

- Box plot:

Whiskers are drawn from the **upper** (75th percentile) and **lower hinges** (25th percentile) (**78** and **63**) to the upper and lower adjacent values (24 and 14)

H-spread = Upper Hinge – Lower Hinge = $78 - 63 = 15$

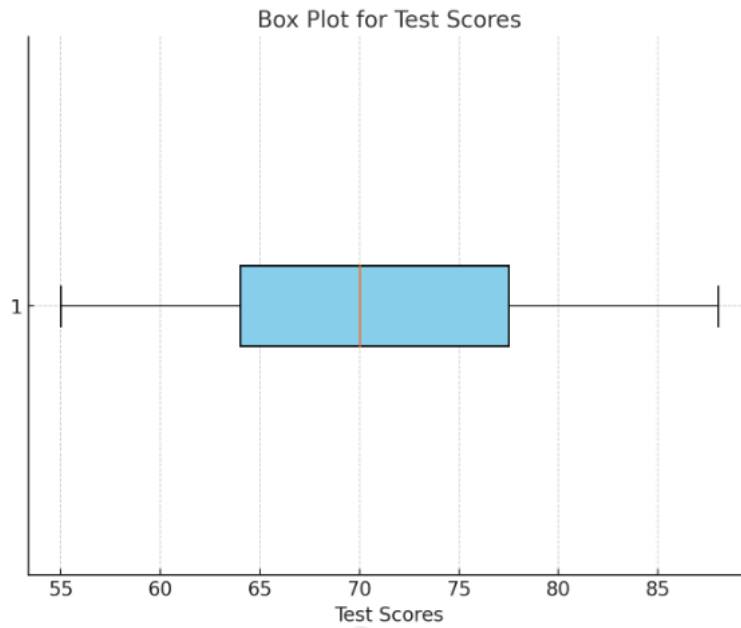
Step = $1.5 * \text{H-spread} = 1.5 * 15 = 22.5$

Upper Inner Fence = Upper Hinge + 1 step = $77.5 + 22.5 = 100$

Lower Inner Fence = Lower Hinge – 1 step = $64 - 22.5 = 41.5$

Upper Adjacent = 88

Lower Adjacent = 55



- c. Since all data points fall within the bounds (41.5 to 100), there are **no outliers** in this dataset.