Data Mining
Chapter 4&5 Exercises

Omid Naeej Nejad
610301189

November 25, 2024

1 Exercise 1

Constructing the 95% Confidence Interval

- $\alpha = 0.05 \rightarrow t_{\alpha/2} = 1.96$.
- Calculate the margin of error (ME):

$$ME = t_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{3}{\sqrt{30}} \approx 1.96 \times 0.548 \approx 1.074$$

- Construct the confidence interval:

$$(\mu - ME, \mu + ME) = (15 - 1.074, 15 + 1.074) = \mathbf{(13.926,\ 16.074)}$$

2 Exercise 2

2.1

Yes, we can conclude that the new drug significantly reduces blood pressure. A p-value of 0.03 indicates that there is a 3% chance of observing a more extreme effect if the null hypothesis (that the drug has no effect) were true. Since this p-value is typically below the conventional significance level of 0.05, we reject the null hypothesis and conclude that the drug is effective.

$$H_0 : \mu = 0$$
$$H_1 : \mu < 0$$

where μ is the true mean change in blood pressure after taking the drug (in this case, a reduction). If $\mu = 0$, it means there is no change, indicating the drug is not effective.

2.2

The margin of error in this experiment represents the range within which the true average reduction in blood pressure is likely to occur. It provides a measure of the uncertainty or variability associated with the estimated average reduction.
In this case, with a margin of error of ± 2 mmHg, we can be reasonably confident (95%) that the true average reduction lies between 3 mmHg (5 mmHg - 2 mmHg) and 7 mmHg (5 mmHg + 2 mmHg).

2.3

Margin of error: Increasing the sample size reduces the margin of error $\Rightarrow$ the confidence interval becomes narrower, and we can be more certain about the true value of the average reduction in blood pressure.
P-value: Increasing the sample size generally increases the statistical power of the test $\rightarrow$ the p-value is likely to decrease

3 Exercise 3

95% Confidence Interval of the proportion

- $\alpha = 0.05 \rightarrow Z_{\alpha/2} = 1.96$.
- Calculate the margin of error (ME):

$$p = \frac{180}{1200} = 0.15$$

$$ME = Z_{\alpha/2} \times \sqrt{\frac{p.(1-p)}{n}} = 1.96 \times \sqrt{\frac{0.15 \times 0.85}{1200}} \approx 1.96 \times 0.0103 \approx 0.0202$$

- Construct the confidence interval:

$$(p - ME, p + ME) = (0.15 - 0.0202, 15 + 0.0202) = \mathbf{(0.1298,\ 0.1702)}$$

4 Exercise 4

Hypothesis Test For Proportion

$$H_0 : p \geq 0.2 \quad vs \quad H_1 : p < 0.2$$

$$n = 150, \quad x = 30, \quad \hat{p} = \frac{x}{n} = \frac{30}{150} = 0.2$$

Calculate the Test Statistic The test statistic (Z-score) is given by:

$$Z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \quad \pi_0 = 0.2$$

$$Z = \frac{0.2 - 0.2}{\sqrt{\frac{0.2\times(1-0.2)}{150}}} = 0$$

Compare the p-value to the significance level ($\alpha = 0.05$)

$$p - value = p(Z \leq Z_{data}) = \Phi(0) = 0.5$$

Since the p-value (0.5) is greater than α (0.05), so we fail to reject the null hypothesis $\rightarrow$ the population proportion of deactivated users is not less than 20%.

5 Exercise 5

$$H_0 : \mu_1 = \mu_2 \quad \text{(No difference in means)}$$
$$H_1 : \mu_1 \neq \mu_2 \quad \text{(Difference in means)}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(35.2 - 34.8)}{\sqrt{\frac{6.1^2}{1500} + \frac{5.9^2}{700}}} \approx 1.465$$

$$t_{\alpha/2} = 1.96$$

Since $|t| \not> t_{\alpha/2} \rightarrow$ fail to reject the null hypothesis, so the partition is valid.

6 Exercise 6

Null Hypothesis (H_0): The distribution of education levels is independent of the group (experimental vs. control).
Alternative Hypothesis (H_1): The distribution of education levels is not independent of the group.

The Chi-Square test statistic is calculated as:

$$\chi^2_{data} = \sum \sum \frac{(O - E)^2}{E}$$

O : observed frequency and E : expected frequency in a cell

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

$$E(\text{Experimental, Below High School}) = \frac{1200 \times 650}{1550} \approx 503.23$$
$$E(\text{Experimental, High School}) = \frac{1200 \times 520}{1550} \approx 403.23$$
$$E(\text{Experimental, Bachelor's and Above}) = \frac{1200 \times 380}{1550} \approx 293.55$$
$$E(\text{Control, Below High School}) = \frac{350 \times 650}{1550} \approx 146.77$$
$$E(\text{Control, High School}) = \frac{350 \times 520}{1550} \approx 116.77$$
$$E(\text{Control, Bachelor's and Above}) = \frac{350 \times 380}{1550} \approx 86.45$$

$$\chi^2_{data} = \frac{(500 - 503.23)^2}{503.23} + \frac{(400 - 403.23)^2}{403.23} + \cdots + \frac{(80 - 86.45)^2}{86.45}$$
$$\approx 1.97$$

$$p - value = P(\chi^2 > \chi^2_{data}) = P(\chi^2 > 1.97) \approx 0.373$$

Since the calculated p-value is greater than the α, we fail to reject the null hypothesis, so the partition is valid.

7 Exercise 7

Null Hypothesis (H_0): The population means for the three methods are equal. Alternative Hypothesis (H_1): At least one of the population means is different.

$$\text{Mean of Website} = \frac{25 + 30 + 28 + 32}{4} = 28.75$$
$$\text{Mean of Mobile App} = \frac{35 + 40 + 38 + 36}{4} = 37.25$$
$$\text{Mean of In-Person} = \frac{40 + 45 + 50 + 48}{4} = 45.75$$

$$\text{Total Mean} = \frac{25 + 30 + 28 + 32 + 35 + 40 + 38 + 36 + 40 + 45 + 50 + 48}{12} = 37.25$$

Sum of Squares Between Groups (SSB)
The sum of squares between groups measures the variation between the means of each group:

$$SSB = n\left((\overline{X}_{\text{website}} - \overline{X}_{\text{total}})^2 + (\overline{X}_{\text{mobile}} - \overline{X}_{\text{total}})^2 + (\overline{X}_{\text{in-person}} - \overline{X}_{\text{total}})^2\right)$$

$$SSB = 4((28.75 - 37.25)^2 + (37.25 - 37.25)^2 + (45.75 - 37.25)^2) = 578$$

Sum of Squares Within Groups (SSW)
The sum of squares within groups measures the variation within each group:

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \overline{X}_{\text{group}})^2$$
$$= 98.25$$

Degrees of Freedom & Mean Squares
The degrees of freedom between groups is:

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

The degrees of freedom within groups is:

$$df_{\text{within}} = N - k = 12 - 3 = 9$$

$$MSB = \frac{SSB}{df_{\text{between}}} = \frac{578}{2} = 289$$
$$MSW = \frac{SSW}{df_{\text{within}}} = \frac{98.25}{9} \approx 10.917$$

F-statistic

$$F = \frac{MSB}{MSW} = 26.47$$

$$p - value = P(F > F_{data}) = P(F > 26.47) \approx 0.00017$$

Since the calculated p-value is less than the α, we reject the null hypothesis, so the population mean time spent for receiving services differs among the three methods.

8 Exercise 8

Estimated score=20+3×(Number of study hours)

8.1

According to the regression equation, for each additional hour of study, the score increases by 3 points.
So, the student who studied 5 hours more would have an estimated score that is 15 points higher than the other student.

8.2

$$\text{Estimated score} = 20 + 3 \times 10 = 50$$

8.3

The result might not be reliable or accurate for students outside range(5, 15) like this one:

$$\text{Estimated score} = 20 + 3 \times 20 = 80$$

8.4

This means that for each additional hour a student spends studying, the estimated score increases by 3 points.

8.5

The 20 in the equation represents the y-intercept of the regression line or equation predicts that if a student does not study at all, the score would be 20.