LAB: INTRO TO STAT ANALYSIS

Zeyi Qian

zeqian@clarku.edu
Office Hours: JC 201, Tuesday 4-5 PM & Thursday 3-4 PM

November 15, 2024

Question 1

Census data for New York City indicate that 29.2% of the under-18 population is white, 28.2% black, 31.5% Latino, 9.1% Asian, and 2% other ethnicities. The New York Civil Liberties Union points out that of the 26,181 police officers, 64.8% are white, 14.5% are black, 19.1% Hispanic, and 1.4% Asian. Do the police officers reflect the ethnic composition of the city's youth? Test an appropriate hypothesis and state your conclusion.

Question 1

Ethnicity	Observed (%)	Observed (N)	Expected (%)	Expected (N)
White	64.8%	16,965	29.2%	7,644
Black	14.5%	3,796	28.2%	7,383
Hispanic	19.1%	5,000	31.5%	8,247
Asian	1.4%	366	9.1%	2,382
Other	0.2%	50	2.0%	523

Question 1

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \frac{(16,965 - 7,644)^2}{7,644} + \frac{(3,796 - 7,383)^2}{7,383} + \cdots$$

$$\chi^2 = 16,513, \quad \text{df} = 4$$

Use excel "$= CHISQ.DIST.RT(ChiSquare, df)$"
Given that the p-value is effectively 0, we reject the null hypothesis H_0.

$$H_0 : \text{The police force represents the population's ethnic distribution.}$$

$$H_a : \text{The police force does not represent the population's ethnic distribution.}$$

Question 2

The table below shows the rank attained by male and female officers in the New York City Police Department (NYPD). Do these data indicate that men and women are equitably represented at all levels of the department?

Position	Male	Female
Officer	21,900	4,281
Detective	4,058	806
Sergeant	3,898	415
Lieutenant	1,333	89
Captain	359	12
Higher Ranks	218	10

Question 2

Hypotheses:

- H_0: There is no association between gender and rank. Men and women are equitably represented at all levels of the department.
- H_a: There is an association between gender and rank. Men and women are not equitably represented at all levels of the department.

Question 2

The test statistic for the chi-square test is:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

We first calculate the total counts:

$$\begin{aligned}
\text{Total Male} &= 21,900 + 4,058 + 3,898 + 1,333 + 359 + 218 = 31,766 \\
\text{Total Female} &= 4,281 + 806 + 415 + 89 + 12 + 10 = 5,613 \\
\text{Grand Total} &= 31,766 + 5,613 = 37,379
\end{aligned}$$

Question 2

The expected count for each cell is calculated as:

$$\text{Expected Count} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

Expected Frequencies

Rank	Expected Male	Expected Female
Officer	22,249.54	3,931.46
Detective	4,133.60	730.40
Sergeant	3,665.34	647.66
Lieutenant	1,208.47	213.53
Captain	315.29	55.71
Higher Ranks	193.76	34.24

Question 2

Chi-Square Calculation

Rank	Male χ^2	Female χ^2
Officer	5.49	31.08
Detective	1.38	7.82
Sergeant	14.77	83.58
Lieutenant	12.83	72.63
Captain	6.06	34.30
Higher Ranks	3.03	17.16
Total	43.56	246.57

Question 2

We calculate the chi-square statistic:

$$\chi^2 = 290.13, \quad \text{df} = 5$$

In Excel, use the formula:

$$= CHISQ.DIST.RT(290.13, 5)$$

The p-value is approximately 0, indicating a significant difference in gender representation. We reject the null hypothesis H_0. There is sufficient evidence to suggest that men and women are not equitably represented across all ranks in the NYPD.

Question 3

We buy a bag of Skittles and M&M's. The bag of Skittles has 24% red out of a total 65 Skittles. The bag of M&M's has 19% red out of a total of 55 M&M's. Construct a 95% confidence interval for the difference in percent red for Skittles versus M&M's. Interpret the interval. Does your interval indicate that is there evidence for a difference in proportions between the two candies?

Question 3

Step 1: State the Hypotheses
We are testing whether there is a difference in the proportion of red candies between Skittles and M&M's:

- Null hypothesis (H_0): $p_1 = p_2$
- Alternative hypothesis (H_a): $p_1 \neq p_2$

Step 2: Calculate the Difference in Sample Proportions The sample proportions are:

$$\hat{p}_1 = 0.24, \quad \hat{p}_2 = 0.19$$

The difference in sample proportions is:

$$\hat{p}_1 - \hat{p}_2 = 0.24 - 0.19 = 0.05$$

Question 3

Step 3: Calculate the Standard Error The standard error (SE) for the difference in proportions is given by:

$$\text{SE} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Substituting the values:

$$\text{SE} = \sqrt{\frac{0.24 \times 0.76}{65} + \frac{0.19 \times 0.81}{55}}$$

$$\text{SE} = \sqrt{\frac{0.1824}{65} + \frac{0.1539}{55}}$$

$$\text{SE} = \sqrt{0.002805 + 0.002798} = \sqrt{0.005603} = 0.0749$$

Question 3

Step 4: Find the Critical Value For a 95% confidence level, the critical value for the standard normal distribution is:

$$z^* = 1.96$$

Step 5: Construct the Confidence Interval The confidence interval for the difference in proportions is:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times \text{SE}$$

$$0.05 \pm 1.96 \times 0.0749$$

$$0.05 \pm 0.1468$$

$$(-0.0968, 0.1968)$$

Question 3

Step 6: Interpret the Confidence Interval The 95% confidence interval for the difference in proportions of red candies between Skittles and M&M's is $(-0.097, 0.197)$.
Since this interval includes 0, there is **no evidence** to suggest a significant difference in the proportion of red candies between Skittles and M&M's at the 95% confidence level.

Question 4

In 2010, the United Nations claimed that there was a higher rate of illiteracy in men than in women from the country of Qatar. A humanitarian organization went to Qatar to conduct a random sample. The results revealed that 45 out of 234 men and 42 out of 251 women were classified as illiterate on the same measurement test. Do these results indicate that the United Nations findings were correct? Test an appropriate hypothesis (performing the 4 steps we have discussed) and state your conclusions. You can use 95% confidence level.

Question 4

Step 1: State the Hypotheses We are testing whether the proportion of illiterate men is higher than the proportion of illiterate women:

- Null hypothesis (H_0): $p_{\text{men}} = p_{\text{women}}$
- Alternative hypothesis (H_a): $p_{\text{men}} > p_{\text{women}}$

Step 2: Calculate the Sample Proportions The sample proportions are:

$$\hat{p}_{\text{men}} = \frac{45}{234} = 0.1923, \quad \hat{p}_{\text{women}} = \frac{42}{251} = 0.1673$$

The difference in sample proportions is:

$$\hat{p}_{\text{men}} - \hat{p}_{\text{women}} = 0.1923 - 0.1673 = 0.025$$

Question 4

Step 3: Calculate the Standard Error The standard error (SE) for the difference in proportions is given by:

$$\text{SE} = \sqrt{\frac{\hat{p}_{\text{men}}(1-\hat{p}_{\text{men}})}{n_{\text{men}}} + \frac{\hat{p}_{\text{women}}(1-\hat{p}_{\text{women}})}{n_{\text{women}}}}$$

Substituting the values:

$$\text{SE} = \sqrt{\frac{0.1923 \times 0.8077}{234} + \frac{0.1673 \times 0.8327}{251}}$$

$$\text{SE} = \sqrt{\frac{0.1553}{234} + \frac{0.1394}{251}}$$

$$\text{SE} = \sqrt{0.000664 + 0.000555} = \sqrt{0.001219} = 0.0349$$

Question 4

Step 4: Test the Hypothesis To test the hypothesis, we compute the test statistic:

$$z = \frac{\hat{p}_{\text{men}} - \hat{p}_{\text{women}}}{\text{SE}} = \frac{0.025}{0.0349} = 0.7163$$

For a one-tailed test at the 95% confidence level, the critical value of z is 1.645.
Step 5: Make a Decision Since the test statistic $z = 0.7163$ is less than the critical value of $z = 1.645$, we fail to reject the null hypothesis.

Question 5

The mean age of 2397 patients without cardiac disease was 69.8 years (SD=8.7 years), while for the 450 patients with cardiac disease, the mean and standard deviation of the ages were 74.0 and 7.9, respectively. Use the 4-step approach to test whether there is any significant difference in the mean ages of the two groups at the 90% confidence level.

Question 5

Step 1: State the Hypotheses The null hypothesis (H_0) and the alternative hypothesis (H_A) are as follows:

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

where μ_1 is the mean age of patients without cardiac disease, and μ_2 is the mean age of patients with cardiac disease.

Step 2: Set the Significance Level We are testing at the 90% confidence level, so the significance level (α) is 0.10.

Question 5

Step 3: Calculate the Test Statistic We will use the two-sample t-test formula for the test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- $\bar{x}_1 = 69.8$ is the mean age for patients without cardiac disease,
- $\bar{x}_2 = 74.0$ is the mean age for patients with cardiac disease,
- $s_1 = 8.7$ is the standard deviation for patients without cardiac disease,
- $s_2 = 7.9$ is the standard deviation for patients with cardiac disease,
- $n_1 = 2397$ is the number of patients without cardiac disease,
- $n_2 = 450$ is the number of patients with cardiac disease.

Question 5

Calculate the variances:

$$\frac{8.7^2}{2397} = \frac{75.69}{2397} \approx 0.0315$$
$$\frac{7.9^2}{450} = \frac{62.41}{450} \approx 0.1387$$

Calculate the standard error:

$$SE = \sqrt{0.0315 + 0.1387} = \sqrt{0.1702} \approx 0.4125$$

Calculate the t-statistic:

$$t = \frac{69.8 - 74.0}{0.4125} = \frac{-4.2}{0.4125} \approx -10.17$$

Question 5

Step 4: To determine whether to reject the null hypothesis, we compare the calculated t-statistic to the critical t-value. For a 90% confidence level (two-tailed test), the critical z-value is approximately 1.645.

Since the absolute value of the calculated t-statistic (10.17) is much greater than the critical value (1.645), we reject the null hypothesis.

Question 6

Environmentalists concerned about the impact of high-frequency radio transmissions on birds found that there was no evidence of a higher mortality rate among hatchlings in nests near cell towers. They based this conclusion on a test using alpha=0.05. Would they have made the same decision at alpha=0.10? How about alpha=0.01? Explain.

Question 6

- Null hypothesis (H_0): No higher mortality rate among hatchlings near cell towers.
- Alternative hypothesis (H_A): Higher mortality rate among hatchlings near cell towers.
- Test was conducted at $\alpha = 0.05$, and they found no evidence of a higher mortality rate.

Question 6

Decision at $\alpha = 0.05$ The environmentalists **failed to reject the null hypothesis** because the p-value was greater than 0.05.

- Since the p-value was greater than 0.05, they did not reject H_0.
- Conclusion: There was no significant evidence of higher mortality rates among hatchlings near cell towers.

Question 6

Decision at $\alpha = 0.10$ At a significance level of $\alpha = 0.10$, we are more lenient in rejecting the null hypothesis.

- If the p-value is between 0.05 and 0.10, the environmentalists would **reject the null hypothesis**.
- Since they failed to reject at $\alpha = 0.05$, the p-value must be greater than 0.05. If it is still less than 0.10, they would reject H_0.

Question 6

Decision at $\alpha = 0.01$ At a more stringent significance level of $\alpha = 0.01$, the environmentalists would require stronger evidence to reject H_0.

- If the p-value is between 0.05 and 0.01, they would **fail to reject the null hypothesis** at $\alpha = 0.01$.
- Since they failed to reject at $\alpha = 0.05$, the p-value must be greater than 0.05, so they would still fail to reject H_0 at $\alpha = 0.01$.