MATH 254: Introduction to Statistics **Mintaek Lee**

Lesson 7: Tests for Independence and Goodness-of-Fit

Corresponding Workbook Module: 7

Example 7.1 *(The Chi-square test for independence)*
Mintaek is a member of the penguin enthusiasts club. He became curious whether preference on penguin species is independent of the member's gender. He surveys 200 members of the club on their favorite penguin species (out of Emperor, King, and Gentoo penguins) and summarizes the results in the table below. Does the evidence suggest that the preference on penguin species is independent of the member's gender? Use the appropriate statistical methods and set up the hypotheses that best match Mintaek's interests. Use the significance level of $\alpha = 0.05$*.*

	Emperor Penguin	King Penguin	Gentoo Penguin	Total
Men	50	30	40	120
Women	20	30	30	80
Total	70	60	70	200

This question is asking you to check if the preference on penguin species (*a categorical variable taking on values: Emperor, King, or Gentoo*) is independent of the member's gender (*a categorical variable taking on values: men and women*). Note that *preference on penguin species* and *member's gender* are two variables of interest here. In this case, we need to use the Chi-square test for independence. Note that you may only use this test when you have two rows in your data. If you have 3 or more rows in your data, you need to use the Chi-square test for homogeneity instead. For more information, see the workbook page 214.

Step 1: The hypotheses are:

- H_0: preference on penguin species (*a categorical variable taking on values: Emperor, King, or Gentoo*) and the member's gender (*a categorical variable taking on values: men and women*) are independent
- H_A: preference on penguin species is independent of the member's gender

Step 2: Conditions are

- **Assume** that Mintaek randomly selected 200 penguin enthusiasts club members. We are assuming it because the question simply stated he surveyed 200 members.
- We see that all of the cells are larger than 5 (50, 30, 40, 20, 30, and 30).
- Population size is at least 10 times the sample size (assuming the penguin enthusiasts club has at least 2000 members).

Since all conditions are met, we can use the Chi-square test for independence.

Step 3: To calculate the test statistic $\mathcal{X}^2$, we first calculate the expected counts using row totals and column totals. Equation for the expected counts is:

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

We then find the expected counts as:

	Emperor Penguin	King Penguin	Gentoo Penguin	Total
Men	$\frac{120 \times 70}{200} = 42$	$\frac{120 \times 60}{200} = 36$	$\frac{120 \times 70}{200} = 42$	120
Women	$\frac{80 \times 70}{200} = 28$	$\frac{80 \times 60}{200} = 24$	$\frac{80 \times 70}{200} = 28$	80
Total	70	60	70	200

Using the observed and expected counts, we can calculate the test statistic $\mathcal{X}^2$ as:

$$\begin{aligned}
\mathcal{X}^2 &= \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \\
&= \frac{(50-42)^2}{42} + \frac{(30-36)^2}{36} + \frac{(40-42)^2}{42} + \frac{(20-28)^2}{28} + \frac{(30-24)^2}{24} + \frac{(30-28)^2}{28} \\
&= 6.5476
\end{aligned}$$

where $df = (r-1)(c-1) = (2-1)(3-1) = 2$ (r is number of rows and c is number of columns). Make sure that you do NOT include "Total"s when counting for r and c.

At the Chi-square distribution table, look at the row corresponding to $df = 2$. We can see that our Chi-square test statistic $\mathcal{X}^2 = 6.5476$ is between 5.99 and 7.38. Since the upper tail probability p corresponding to $\mathcal{X}^2 = 5.99$ and $\mathcal{X}^2 = 7.38$ are 0.05 and 0.025, respectively, we can guess that the **upper** tail probability p corresponding to our Chi-square test statistic $\mathcal{X}^2 = 6.5476$ is between 0.05 and 0.025. Since Chi-square tests are always one-sided, our P-value would be $0.025 <$ P-value < 0.05.

Step 4: Our P-value was less than $\alpha = 0.05$. So we would reject the null hypothesis that preference on penguin species and member's gender are independent with the significance level of $\alpha = 0.05$.

It means that we do have sufficient evidence to suggest that there is a relationship between preference on penguin species and member's gender.

Example 7.2 *(The Chi-square test for goodness-of-fit)*
Suppose a seller from eBay is selling \$20 mystery boxes. The seller claims that 30% of boxes come with items worth \$40, 40% of boxes come with items worth \$20, and 30% of boxes come empty (thus, worth \$0). Curious to test his luck, Mintaek spends his entire savings to buy 150 of those mystery boxes. A couple of weeks later, he receives all 150 boxes and obtains the following outcomes.

Value of the box	\$40	\$20	\$0	Total
Observed Count	35	58	57	150

Does the evidence suggest that the seller is a scammer and that Mintaek got ripped off? Use the appropriate statistical methods and set up the hypotheses that best match Mintaek's interests. Use the significance level of $\alpha = 0.05$.

This question is asking you to check if proportions in each category (proportions from Mintaek's purchase) are the same as a given set of proportions (eBay seller's claim). So we need to use the Chi-square test for goodness-of-fit. For more information, see the workbook page 216.

Step 1: The hypotheses are:

- H_0: The data are consistent with a specified distribution (the seller's claim).
- H_A: The data are NOT consistent with a specified distribution (the seller's claim).

Step 2: Conditions are

- **Assume** that the seller randomly selected 150 boxes from their inventory. We are assuming it because we don't know for sure if it is true.
- We see that all of the cells are larger than 5 (35, 58, and 57).
- Population size is at least 10 times the sample size (assuming the eBay seller has more than 1500 mystery boxes on their inventory).

Since all conditions are met, we can use the Chi-square test for goodness-of-fit.

Step 3: To calculate the test statistic $\mathcal{X}^2$, we first calculate the expected counts assuming the seller's claim is true. Remember that $n = 150$.

Value of the box	\$40	\$20	\$0
Observed Count	35	58	57
Expected Count	$150 \times 0.3 = 45$	$150 \times 0.4 = 60$	$150 \times 0.3 = 45$

Using the observed and expected counts, we can calculate the test statistic $\mathcal{X}^2$ as:

$$\mathcal{X}^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \frac{(35-45)^2}{45} + \frac{(58-60)^2}{60} + \frac{(57-45)^2}{45} = 5.4889$$

where $df = c - 1 = 3 - 1 = 2$ (c is number of columns).

At the Chi-square distribution table, look at the row corresponding to $df = 2$. We can see that our Chi-square test statistic $\mathcal{X}^2 = 5.4889$ is between 4.61 and 5.99. Since the upper tail probability p corresponding to $\mathcal{X}^2 = 4.67$ and $\mathcal{X}^2 = 5.99$ are 0.1 and 0.05, respectively, we can guess that the **upper** tail probability p corresponding to our Chi-square test statistic $\mathcal{X}^2 = 5.4889$ is between 0.1 and 0.05. Since Chi-square tests are always one-sided, our P-value would be $0.05 <$ P-value < 0.1.

Step 4: Our P-value was greater than $\alpha = 0.05$. So we would fail to reject the null hypothesis that the data are consistent with a specified distribution (the seller's claim) under the significance level of $\alpha = 0.05$.

It means that we do not have sufficient evidence to suggest that the distribution of mystery boxes that Mintaek purchased are different from the eBay seller's claim.

Example 7.3 *(The Chi-square test for homogeneity)*
Mintaek likes to watch birds in a park in his downtime. One day, he was curious on whether proportions of three birds (geese, ducks, and swans) are different between Julia Davis Park, Ann Morrison Park, and Esther Simplot Park. He watched a total of 600 birds in those three parks and counted how many of them were geese, ducks, and swans for each park. Assume that there are only three kinds of birds in those parks.

	Geese	Ducks	Swans	Total
Julia Davis Park	100	50	50	200
Ann Morrison Park	150	100	50	300
Esther Simplot Park	50	30	20	100
Total	300	180	120	600

Does the evidence suggest that the proportions of those three birds are different between the Julia Davis Park, Ann Morrison Park, and Esther Simplot Park? Use the appropriate statistical methods and set up the hypotheses that best match Mintaek's interests. Use the significance level of $\alpha = 0.05$.

This question is asking you to check if proportions of birds are the same (or independent) between three parks. This set up is quite similar to the Example 7.1, except that there are three rows instead of two. In this case, we need to use the Chi-square test for homogeneity. For more information, see the workbook page 218.

Step 1: The hypotheses are:

- H_0: Corresponding categories occur in the same proportions in all groups (proportions of birds in Julia Davis Park, proportions of birds in Ann Morrison Park, and proportions of birds in Esther Simplot Park are the same).
- H_A: At least two corresponding proportions differ (at least two parks have different proportions of birds).

Step 2: Conditions are

- **Assume** that the 600 birds Mintaek counted were selected randomly. We are assuming it because we don't know for sure if it is true.
- We see that all of the cells are larger than 5 (100, 50, 50, 150, 100, 50, 50, 30, and 20).
- Population size is at least 10 times the sample size (there are certainly more than 6000 of those birds in Boise).

Since all conditions are met, we can use the Chi-square test for homogeneity.

Step 3: To calculate the test statistic $\mathcal{X}^2$, we first calculate the expected counts using row totals and column totals. Equation for the expected counts is:

$$\text{Expected Counts} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

We then find the expected counts as:

	Geese	Ducks	Swans	Total
Julia Davis Park	$\frac{200 \times 300}{600} = 100$	$\frac{200 \times 180}{600} = 60$	$\frac{200 \times 120}{600} = 40$	200
Ann Morrison Park	$\frac{300 \times 300}{600} = 150$	$\frac{300 \times 180}{600} = 90$	$\frac{300 \times 120}{600} = 60$	300
Esther Simplot Park	$\frac{100 \times 300}{600} = 50$	$\frac{100 \times 180}{600} = 30$	$\frac{100 \times 120}{600} = 20$	100
Total	300	180	120	500

Using the observed counts and expected counts, we can calculate the test statistic $\mathcal{X}^2$ as:

$$\begin{aligned}\mathcal{X}^2 &= \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \\ &= \frac{(100-100)^2}{100} + \frac{(50-60)^2}{60} + \frac{(50-40)^2}{40} + \frac{(150-150)^2}{150} + \frac{(100-90)^2}{90} + \frac{(50-60)^2}{60} \\ &\quad + \frac{(50-50)^2}{50} + \frac{(30-30)^2}{30} + \frac{(20-20)^2}{20} \\ &= 6.9444\end{aligned}$$

where $df = (r-1)(c-1) = (3-1)(3-1) = 4$ (r is number of rows and c is number of columns). Make sure that you do NOT include totals in r and c.

At the Chi-square distribution table, look at the row corresponding to $df = 4$. We can see that our Chi-square test statistic $\mathcal{X}^2 = 6.9444$ is between 6.74 and 7.78. Since the upper tail probability p corresponding to $\mathcal{X}^2 = 6.74$ and $\mathcal{X}^2 = 7.78$ are 0.15 and 0.10, respectively, we can guess that the **upper** tail probability p corresponding to our Chi-square test statistic $\mathcal{X}^2 = 6.9444$ is between 0.15 and 0.10. Since Chi-square tests are always one-sided, our P-value would be $0.10 <$ P-value < 0.15.

Step 4: Our P-value was greater than $\alpha = 0.05$. So we would fail to reject the null hypothesis that corresponding categories occur in the same proportions in all groups (proportions of birds in Julia Davis Park, proportions of birds in Ann Morrison Park, and proportions of birds in Esther Simplot Park are the same) under the significance level of $\alpha = 0.05$.

It means that we do not have sufficient evidence to suggest that at least two parks have different proportions of birds. Our results indicate proportions of geese, duck, and swan are the same across Julia Davis Park, Ann Morrison Park, and Esther Simplot Park.