Lab 12: Chi-Square Tests

Objective: In this lab, you will learn how to perform three Chi-square tests (the test of goodness of fit, the test of independence, and the test of homogeneity) that are used to analyze categorical responses.

Overview: You will see three **Chi-Square tests**: the tests of **goodness of fit**, **independence**, and **homogeneity**. For all three tests, the data are generally presented in the form of a **contingency table** (a rectangular array of numbers in cells). All three tests are based on the **Chi-Square statistic:**

$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$, where O_i is the observed count and

E_i is the expected count under the corresponding null hypothesis.

Goodness of Fit Test: This test answers the question, "Do the data fit well compared to a specified distribution?" It considers one categorical response, and assesses whether the proportion of sampled observations falling into each category matches well to a specified distribution. The null hypothesis specifies this distribution which describes the population proportion of observations in each category.

Test of Homogeneity: This test answers the question, "Do two or more populations have the same distribution for one categorical variable?" It considers one categorical response, and assesses whether the model for this response is the same in two (or more) populations. The null hypothesis is that the distribution of the categorical variable is the same for the two (or more) populations.

Test of Independence: This test answers the question, "Are two factors (or variables) independent for a population under study?" It considers two categorical variables (sometimes one is a response and the other is explanatory), and assesses whether there appears to be a relationship between these two variables for a single population. The null hypothesis is that the two categorical variables are independent (not related) for the population of interest.

There are a few properties of the Chi-square distribution that you might find useful. The expected value of a Chi-square distribution is its degrees of freedom (mean = $\mu = df$), and its variance is 2 times its degrees of freedom. Thus, its standard deviation is the square root of 2 times the degrees of freedom ($\sigma^2 = 2 * df$ so $\sigma = \sqrt{2 * df}$). This frame of reference can help assess if our observed statistic is unusual under the null hypothesis or somewhat consistent with the null hypothesis.

Formula Card

Chi-Square Tests	
Test of Independence & Test of Homogeneity	Test for Goodness of Fit
Expected Count $E = \text{expected} = \frac{\text{row total} \times \text{column total}}{\text{total } n}$	**Expected Count** $E_i = \text{expected} = np_{i0}$
Test Statistic $X^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ $df = (r-1)(c-1)$	**Test Statistic** $X^2 = \sum \frac{(O-E)^2}{E} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ $df = k - 1$
If Y follows a $\chi^2(df)$ distribution, then $E(Y) = df$ and $Var(Y) = 2(df)$.	

Warm-Up: Check Your Understanding

Fill in the blank with the name of the most appropriate Chi-square test to address the research question.

1. A researcher wants to determine if scoring high or low on an artistic ability test depends on being right or left-handed.

   Answer: Chi-Square test of _____ _ _____ _____

2. A national organization wants to compare the distribution of level of highest education completed (high school, college, masters, doctoral) for Republicans versus Democrats.

   Answer: Chi-Square test of _____ _ _____ _____

3. A preservation society has the percentages of five main types of fish in the river from 10 years ago. After noticing an imbalance recently, they add some fish from hatcheries to the river. How can they determine if they restored the ecosystem from a new sample of fish?

   Answer: Chi-Square test of _____ _ _____ _____

ILP: Is There a Different Pattern in the Distribution of Accidental Deaths in a Certain Region Compared to the Pattern in the Entire United States?

Background: According to the records of the National Safety Council, accidental deaths in the United States during 2002 had the following distribution according to the principal types of accidents.

Motor Vehicle	Falls	Drowning	Fire	Poison	Other
45%	15%	4%	3%	16%	17%

Suppose that an accidental death data set from a particular geographical region yielded the following frequency distribution for the principal types of accidents:

Motor Vehicle	Falls	Drowning	Fire	Poison	Other
442	161	42	33	162	150

Do these data show a significantly different pattern in the distribution of accidental deaths in the particular region compared to the pattern in the entire United States? Use a 5% significance level. (Source: National Safety Council Website, 2005)

Task: Perform a Chi-square goodness of fit test to assess whether the data fit well with the model specified in the null hypothesis.

1. **State the null hypothesis:** H_0: _____

   Determine Alpha: We were told to use a significance level of 5%.

Remember: Your hypotheses and parameter definition should always be a statement about the **population(s)** under study.

2. Checking the Assumptions

a. Find the expected counts and fill them in the table below.

	MotorVehicle	Falls	Drowning	Fire	Poison	Other	Total
Null %	45%	15%	4%	3%	16%	17%	100%
Observed	442	161	42	33	162	150	990
Expected							

b. Do all cells have expected counts greater than 5? **Yes** **No**

3. Compute the Test Statistic and Calculate the *p*-value

Test-Statistic

a. Complete the calculation of the test statistic based on your table above by calculating the contribution of "other" accidental deaths to the total test statistic.

$$X^2 = \frac{(442-445.5)^2}{445.5} + \frac{(161-148.5)^2}{148.5} + \frac{(42-39.6)^2}{39.6}$$
$$+ \frac{(33-29.7)^2}{29.7} + \frac{(162-158.4)^2}{158.4} + \frac{(.....)^2}{.....} = 3.663$$

Calculate the *p*-Value:

b. Based on the output below, answer the following:

```
        Chi-squared test for given probabilities

                    data:  .Table
  X-squared = 3.6635, df = 5, p-value = 0.598805
```

i. The *p*-value is _____ .

ii. The expected value of the test statistic assuming H_0 is true is _____

iii. The large *p*-value we obtained is consistent with the fact that our observed test statistic value is

greater than **less than** the expected test statistic value (under the null hypothesis).

4. Evaluate the *p*-value and Conclusion:

Evaluate the *p*-value:

What is your decision at a 5% significance level? **Reject H_0** **Fail to Reject H_0**

Remember: Reject H_0 ⇔ Results statistically significant
Fail to Reject H_0 ⇔ Results not statistically significant

Conclusion:

What is your conclusion *in the context of the problem*?

ILP: Comparison of the Distribution of Academic Degrees: Males Versus Females

Background: How do women and men compare in the pursuit of academic degrees? The table presents counts (in thousands) from the Statistical Abstract of degrees earned in 1996 categorized by the level of the degree and the sex of the recipient.

	Bachelor	Master	Professional	Doctorate
Female	642	227	32	18
Male	522	179	45	27

Task: Perform a Chi-square test of homogeneity. Use a 1% significance level.

Hypothesis Test:

1. State the null hypothesis:

H_0: _____

_____

Determine Alpha: We were told to use a significance level of 1%.

2. Checking the Assumptions

Sex * Degree Crosstabulation

			Degree				
			Bachelor	Master	Professional	Doctorate	Total
Sex	Female	Count	642	227	32	18	919
		Expected Count	632.2	220.5	41.8	24.4	919.0
	Male	Count	522	179	45	27	773
		Expected Count	531.8	185.5	35.2	20.6	773.0
	Total	Count	1164	406	77	45	1692
		Expected Count	1164.0	406.0	77.0	45.0	1692.0

a. Show how the expected count 531.8 (first cell for males) was computed.

b. Based on the table, do the assumptions appear to be met to perform the test? (Are all expected counts greater than 5?) **Yes** **No**

3. Compute the Test-Statistic and Calculate the *p*-value

```
   Pearson's Chi-squared test

data:  .Table
X-squared = 9.5135, df = 3, p-value = 0.02319
```

Test-Statistic

a. Based on the output above, the test-statistic is ____ = _____

Calculate the *p*-value

b. Based on the output above, the *p*-value is _____ .

4. **Evaluate the *p*-value and Conclusion**
   Evaluate the *p*-value:
   What is your decision at a 1% significance level? **Reject H_0** **Fail to Reject H_0**

   Conclusion:
   What is your conclusion at a 1% significance level *in context of the problem*?

5. Would your decision and conclusion change if the significance level was:

   - 5% instead of 1%?
   - 3% instead of 1%?
   - 2.3% instead of 1%?
   - 2% instead of 1%?

   Based on your answers, the *p*-value represents the _____ significance level at which the results would be statistically significant.

Cool-Down: Name That Scenario

Match each research question with the appropriate Chi-Square test that should be used to answer the question

1. Is student status (in-state versus out-of-state) associated with one's eventual graduation outcome (graduating versus not graduating)?

   Answer: Chi-Square test of _____ _____ _____

2. To test a theory that people have no preference among four different outdoor activities, you ask 100 people to select among jogging, bicycling, hiking, or swimming.

   Answer: Chi-Square test of _____ _____ _____

3. A biostatistician would like to determine if the ratio of the blood type in the storage for transfusions should be different in Hawaii from the main land. She collected a sample of blood types of 10,000 people in Hawaii and that of 100,000 people in the mainland. She wishes to see if the breakdown of blood types (A, B, AB and 0) is the same for both populations.

   Answer: Chi-Square test of _____ _____ _____