Chapter 11: Inference for Distribution of Categorical Variables: Chi-Square Procedures

Objectives: Students will:
- Explain what is meant by a chi-square goodness of fit test.
- Conduct a chi-square goodness of fit test.
- Given a two-way table, compute conditional distributions.
- Conduct a chi-square test for homogeneity of populations.
- Conduct a chi-square test for association/independence.
- Use technology to conduct a chi-square significance test.

AP Outline Fit:
IV. Statistical Inference: Estimating population parameters and testing hypotheses (30%–40%)
   B. Tests of significance
      6. Chi-square test for goodness of fit, homogeneity of proportions, and independence (one- and two-way tables)

What you will learn:
A. Choose the Appropriate Chi-Square Procedure
   1. For goodness of fit tests, use percents and bar graphs to compare hypothesized and actual distributions.
   2. Distinguish between tests of homogeneity of populations and tests of association/independence.
   3. Organize categorical data in a two-way table. Then use percents and bar graphs to describe the relationship between the categorical variables.
B. Perform Chi-Square Tests
   1. Explain what null hypothesis is being tested.
   2. Calculate expected counts.
   3. Calculate the component of the chi-square statistic for any cell, as well as the overall statistic.
   4. Give the degrees of freedom of a chi-square statistic.
   5. Use the chi-square critical values in Table D to approximate the $P$-values of a chi-square test.
C. Interpret Chi-Square Tests
   1. Locate expected cell counts, the chi-square statistic, and its $P$-value in output from computer software or a calculator.
   2. If the test is significant, use percents, comparison of expected and observed counts, and the components of the chi-square statistic to see which deviations from the null hypothesis are most important.
Chapter 11: Inference for Distribution of Categorical Variables: Chi-Square Procedures

Section 11.1: **Test for Goodness of Fit**

**Knowledge Objectives:** Students will be able to:

- State appropriate hypotheses and compute the expected counts and chi-square test statistic for a chi-square test for goodness of fit
- State and check the Random, 10%, and Large Counts conditions for performing a chi-square test for goodness of fit
- Calculate the degrees of freedom and P-value for a chi-square for goodness of fit
- Perform a chi-square test for goodness of fit
- Conduct a follow-up analysis when the results of a chi-square test are statistically significantly

**Vocabulary:**

- *Chi-square test statistic* – a measure of how far the observed counts are from the expected counts.
- *Chi-square distribution* – defined by a density curve that takes only nonnegative values and is skewed to the right. A particular chi-square distribution is specified by its degrees of freedom

**Key Concepts:**

**Chi-Square Distribution:**

- Total area under a chi-square curve is equal to 1
- It is not symmetric, it is skewed right
- The shape of the chi-square distribution depends on the degrees of freedom (just like t-distribution)
- As the number of degrees of freedom increases, the chi-square distribution becomes more nearly symmetric
- The values of $\chi^2$ are nonnegative; that is, values of $\chi^2$ are always greater than or equal to zero (0); they increase to a peak and then asymptotically approach 0
- Table D in the back of the book gives critical values

**Goodness-of-fit test Conditions:**

- All expected counts are greater than or equal to 1 (all $E_i \geq 1$)
- No more than 20% of expected counts are less than 5

Remember it is the expected counts, not the observed that are critical conditions

**Goodness-of-Fit Test**

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

where

- $O_i$ is observed count for $i$th category
- $E_i$ is the expected count for the $i$th category

<table>
<thead>
<tr>
<th>Reject null hypothesis, if</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-value &lt; $\alpha$</td>
</tr>
<tr>
<td>$\chi^2_0 &gt; \chi^2_{\alpha, k-1}$ (Right-Tailed)</td>
</tr>
</tbody>
</table>
Chapter 11: Inference for Distribution of Categorical Variables: Chi-Square Procedures

Example 1: Are you more likely to have a motor vehicle collision when using a cell phone? A study of 699 drivers who were using a cell phone when they were involved in a collision examined this question. These drivers made 26,798 cell phone calls during a 14 month study period. Each of the 699 collisions was classified in various ways.

<table>
<thead>
<tr>
<th>Sun</th>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>133</td>
<td>126</td>
<td>159</td>
<td>136</td>
<td>113</td>
<td>12</td>
</tr>
</tbody>
</table>

Are accidents equally likely to occur on any day of the week?

a) Do a graphical analysis (with a bar chart) using your calculator

b) Using a chi-square goodness of fit test

- Hypotheses:

- Conditions:

- Calculations:

- Interpretation:
Example 2: Does either the large bag or the small bag of M&M’s fit the distribution of Peanut M&M's?

<table>
<thead>
<tr>
<th></th>
<th>Yellow</th>
<th>Orange</th>
<th>Red</th>
<th>Green</th>
<th>Brown</th>
<th>Blue</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>66</td>
<td>88</td>
<td>38</td>
<td>59</td>
<td>53</td>
<td>96</td>
<td>400</td>
</tr>
<tr>
<td>Sample 2</td>
<td>10</td>
<td>9</td>
<td>4</td>
<td>16</td>
<td>9</td>
<td>7</td>
<td>55</td>
</tr>
<tr>
<td>Peanut</td>
<td>0.15</td>
<td>0.23</td>
<td>0.12</td>
<td>0.15</td>
<td>0.12</td>
<td>0.23</td>
<td>1</td>
</tr>
<tr>
<td>Plain</td>
<td>0.14</td>
<td>0.2</td>
<td>0.13</td>
<td>0.16</td>
<td>0.13</td>
<td>0.24</td>
<td>1</td>
</tr>
</tbody>
</table>

K = 6 classes (different colors)

<table>
<thead>
<tr>
<th></th>
<th>CS(5,.1)</th>
<th>CS(5,.05)</th>
<th>CS(5,.025)</th>
<th>CS(5,.01)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9.236</td>
<td>11.071</td>
<td>12.833</td>
<td>15.086</td>
</tr>
</tbody>
</table>

a) Large Bag

- Hypotheses:
- Conditions:
- Calculations:
- Interpretation:

b) Small Bag

- Hypotheses:
- Conditions:
- Calculations:
Example 3:
A random sample of 80 NHL players from a recent season was selected and their birthdays were recorded. The one-way table summarizes the data on birthdays for these 80 players.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of players</td>
<td>32</td>
<td>20</td>
<td>16</td>
<td>12</td>
</tr>
</tbody>
</table>

Do these data provide convincing evidence that the birthdays of NHL players are not uniformly distributed across the four quarters of the year?

• Hypotheses:

• Conditions:

• Calculations:

• Interpretation:

**Homework:** Problems 1, 3, 9
Chapter 11: Inference for Distribution of Categorical Variables: Chi-Square Procedures

Section 11.2: Inference for Two-Way Tables

Knowledge Objectives: Students will:
State appropriate hypothesis and compute the expected counts and chi-square statistic for a chi-square test based on data in a two-way table
State and check the Random, 10%, and Large Count conditions for a chi-square test based on data in a two-way table
Calculate the degrees of freedom and P-value for a chi-square test based on data in a two-way table
Perform a chi-square test for homogeneity
Perform a chi-square test for independence
Choose the appropriate chi-square test in a given setting

Vocabulary: none new in this edition

Statistical Inference – provides methods for drawing conclusions about a population parameter from sample data
Chi-Squared Test for Independence – used to determine if there is an association between a row variable and a column variable in a contingency table constructed from sample data
Expected Frequencies – row total * column total / table total
Chi-Squared Test for Homogeneity of Proportions – used to test if different populations have the same proportions of individuals with a particular characteristic

Key Concepts:

Chi-Square Test for Homogeneity
- H₀: distribution of response variable is the same for all c populations
- Hₐ: distributions are not the same
Conditions:
- Independent SRS from each of c populations (the same)
- No more than 20% of the expected counts are less than 5 and all individual counts are 1 or greater

The P-value for the chi-square test is the area to the right of $\chi^2$ under the chi-square density curve with $(r - 1)(c - 1)$ degrees of freedom.

Chi-Square Statistic
The chi-square statistic is a measure of how far the observed counts in a two-way table are from the expected counts. The formula for the statistic is

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

The sum is over all $r \times c$ cells in the table.

- Large values of $\chi^2$ are evidence against $H₀$ because they say the observed counts are far from what we would expect if $H₀$ were true.
- Chi-Square tests are one-side (even though $Hₐ$ is many-sided)
Example 1: Market researchers know that background music can influence the mood and purchasing behavior of customers. One study in supermarket in Northern Ireland compared three treatments: no music, French accordion music, and Italian string music. Under each condition, the researchers recorded the numbers of bottles of French, Italian, and other wine purchased. Here is a table that summarizes the data:

<table>
<thead>
<tr>
<th>Wine</th>
<th>None</th>
<th>French</th>
<th>Italian</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>30</td>
<td>39</td>
<td>30</td>
<td>99</td>
</tr>
<tr>
<td>Italian</td>
<td>11</td>
<td>1</td>
<td>19</td>
<td>31</td>
</tr>
<tr>
<td>Other</td>
<td>43</td>
<td>35</td>
<td>35</td>
<td>113</td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td>75</td>
<td>84</td>
<td>243</td>
</tr>
</tbody>
</table>

- Hypotheses:
- Conditions:
- Calculations:
- Interpretation:
Example 2: For a class project, Abby and Mia wanted to know if the gender of an interviewer could affect the responses to a survey question. The subjects in their experiment were 100 males from their school. Half of the males were randomly assigned to be asked, “Would you vote for a female president?” by a female interviewer. The other half of the males were asked the same question by a male interviewer. The table shows the results.

<table>
<thead>
<tr>
<th>Response to question</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>30</td>
<td>39</td>
<td>69</td>
</tr>
<tr>
<td>No</td>
<td>8</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Maybe</td>
<td>12</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

a) State the appropriate null and alternative hypotheses.

b) Show the calculation for the expected count in the Male/Yes cell. Then provide a complete table of expected counts.

c) Calculate the value of the chi-square test statistic.

d) Calculate the p-value and give your interpretation

AP Tip: Writing out an entire $\chi^2$ summation will be very time consuming (something you don’t have on the test)
To demonstrate to the AP reader that you have an understanding of $\chi^2$ statistic write out statistic, definition, first and last terms and what’s its sum is

z-Test vs $\chi^2$ test:
- We use the $\chi^2$ test to compare any number of proportions
- The results from the $\chi^2$ test for 2 proportions will be the same as a z-test for 2 proportions
- z-Test is recommended to compare two proportions because it gives you a choice of a one-side test and is related to the confidence interval for $p_1 - p_2$. 
Chapter 11: Inference for Distribution of Categorical Variables: Chi-Square Procedures

The Chi-Square Test of Association/Independence

Use the chi-square test of association/independence to test the null hypothesis

\( H_0: \) There is no association between two categorical variables.

when you have a two-way table from a single SRS, with each individual classified according to both of two categorical variables.

This test assesses whether this observed association is statistically significant. That is, is the relationship in the sample sufficiently strong for us to conclude that it is due to a relationship between the two variables and not merely to chance.

Example 3: Many popular businesses, like McDonald’s, are franchises. Some contracts with franchises include a right to exclusive territory (another McDonald’s can’t open in that area). How does the presence of an exclusive territory clause in the contract relate to the survival of the business? A study designed to address this question collected data from a sample of 170 new franchise firms. Here are the observed count data:

<table>
<thead>
<tr>
<th>Exclusive Territory</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>108</td>
<td>15</td>
<td>123</td>
</tr>
<tr>
<td>No</td>
<td>34</td>
<td>13</td>
<td>47</td>
</tr>
<tr>
<td>Total</td>
<td>142</td>
<td>28</td>
<td>170</td>
</tr>
</tbody>
</table>

- Hypotheses:

- Conditions:

- Calculations:

- Interpretation:
Chapter 11: Inference for Distribution of Categorical Variables: Chi-Square Procedures

Example 4: A random sample of 1526 winter visitors to Yellowstone National Park was asked two questions:
1. Do you belong to an environmental club (like the Sierra Club)?
2. What is your experience with a snowmobile: own, rent, or never used?

The two-way table summarizes the results.

<table>
<thead>
<tr>
<th>Snowmobile experience</th>
<th>Environmental club status</th>
<th>Not a member</th>
<th>Member</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never used</td>
<td>445</td>
<td>212</td>
<td>657</td>
<td></td>
</tr>
<tr>
<td>Renter</td>
<td>497</td>
<td>77</td>
<td>574</td>
<td></td>
</tr>
<tr>
<td>Owner</td>
<td>279</td>
<td>16</td>
<td>295</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1221</td>
<td>305</td>
<td>1526</td>
<td></td>
</tr>
</tbody>
</table>

Do these data provide convincing evidence of an association between environmental club status and type of snowmobile use in the population of winter visitors to Yellowstone National Park?

- Hypotheses:

- Conditions:

- Calculations:

- Interpretation:

Homework: Problems 29, 31, 35, 43, 49
Chapter 11: Inference for Distribution of Categorical Variables: Chi-Square Procedures

Chapter 11: Review

Objectives: Students will be able to:
- Summarize the chapter
- Define the vocabulary used
- Know and be able to discuss all sectional knowledge objectives
- Complete all sectional construction objectives
- Successfully answer any of the review exercises

Explain what is meant by a chi-square goodness of fit test.
Conduct a chi-square goodness of fit test.
Given a two-way table, compute conditional distributions.
Conduct a chi-square test for homogeneity of populations.
Conduct a chi-square test for association/independence.
Use technology to conduct a chi-square significance test.

Vocabulary: None new

Goodness of Fit Tests on TI
- Enter Observed values in L1
- Enter Expected values in L2
- Press STAT, highlight TESTS and select D: GOF-Test
- Enter degrees of freedom:
- Calculate

Homogeneity and Independence Tests on TI:
- Press 2nd X\(^{-1}\) (access MATRIX menu)
  - Arrow to EDIT and select 1: [A]
- Enter the number of rows and columns of the matrix
- Enter the cell entries for the observed data and press 2nd QUIT
- Press STAT, highlight TESTS and select C: \(\chi^2\)-Test
- Matrix [A] (and Matrix [B] for expected) are defaults
- Highlight Calculate and press ENTER
- Highlight Draw and the \(\chi^2\) curve will be drawn, the critical area in the tail shaded and the p-value displayed
- If you need the expect counts display Matrix B from the matrix menu

Homework:
Problem 1: The makers of the movie Titanic imply that lower-class passengers were treated unfairly when the lifeboats were being filled. We want to determine whether that portrayal is accurate. The following table contains the survival data by passenger class for the 1316 passengers.

<table>
<thead>
<tr>
<th>Class</th>
<th>Survived</th>
<th>Lost</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>203</td>
<td>122</td>
</tr>
<tr>
<td>Second</td>
<td>118</td>
<td>167</td>
</tr>
<tr>
<td>Third</td>
<td>178</td>
<td>528</td>
</tr>
</tbody>
</table>

Following the outline on the next page, you will use a chi-square test to determine whether there is a relationship between survival and passenger class.

(a) If this table is considered an $r \times c$ table, $r = \text{______}$ and $c = \text{______}$.

(b) State the null and alternative hypotheses that would be appropriate for this test:

(c) Show how to determine the expected number of second class survivors.

(d) In order to validly use the chi-square test, how many expected values could be less than 5? _______
   How many expected values could be less than 1? _______

(e) How many degrees of freedom would be associated with this test? _______
   Show how you determined the degrees of freedom:

(f) Use your calculator to perform this chi-square test.
   (i) Examine the matrix of expected values and write the expected values for each cell next to their observed frequencies in the table. (Round to tenths.)

<table>
<thead>
<tr>
<th>Class</th>
<th>Survived</th>
<th>Lost</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>203</td>
<td>122</td>
</tr>
<tr>
<td>Second</td>
<td>118</td>
<td>167</td>
</tr>
<tr>
<td>Third</td>
<td>178</td>
<td>528</td>
</tr>
</tbody>
</table>

   (ii) In what classes are there more survivors than would be expected under the assumption of the null hypothesis?

   (iii) What is the value of the chi-square statistic?

   (iv) What is the $P$-value associated with the chi-square statistic? _______

   (v) State your conclusion regarding the hypotheses of this test:

   (vi) Examine the matrix of chi-square components that is created by your calculator. Which entry has the greatest contribution to your test statistic? What is the value of this component?

(g) What proportion of all 1316 passengers were third class passengers?

(h) What proportion of survivors were third class passengers?

(i) What proportion of first class passengers survived?
Chapter 11: Inference for Distribution of Categorical Variables: Chi-Square Procedures

Problem 2: It is sometimes said that older people are over-represented on juries. The table below gives the percentage distribution of all people over 21 years of age in Alameda County, CA by age group. The table also shows the age group classification for a sample of 66 people who served on grand juries in this county.

<table>
<thead>
<tr>
<th>Age</th>
<th>Countywide Percentage</th>
<th>Number of jurors</th>
</tr>
</thead>
<tbody>
<tr>
<td>21 to 40</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>41 to 50</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>51 to 60</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>61 or older</td>
<td>19</td>
<td>33</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>66</td>
</tr>
</tbody>
</table>

We would like to perform a chi-square test to determine whether the age distribution of jurors is significantly different from the age distribution of county residents. That is, we want to test the following hypotheses:

$H_0$: 42% of jurors are 21 to 40 years old, 23% of jurors are 41 to 50 years old, 16% of jurors are 51 to 60 years old, and 19% of jurors are 61 or older.

$H_a$: The age distribution of jurors is different from the one above.

(a) Working under the assumption that the $H_0$ is true, write how many of the 66 jurors would you expect next to the observed values in the table below:

<table>
<thead>
<tr>
<th>Age</th>
<th>Countywide Percentage</th>
<th>Number of jurors</th>
</tr>
</thead>
<tbody>
<tr>
<td>21 to 40</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>41 to 50</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>51 to 60</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>61 or older</td>
<td>19</td>
<td>33</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>66</td>
</tr>
</tbody>
</table>

(b) Write a few sentences to describe how the counts you expect if the null hypothesis is true compared to the counts observed in this sample.

(c) Use a chi-square test to determine whether the age distribution of jurors differs significantly from the age distribution of the general population. Show the computations needed to compute the chi-square statistic, and state the degrees of freedom, the P-value, and the conclusion. You do not need to state or check conditions.

<table>
<thead>
<tr>
<th>Age</th>
<th>Countywide Percentage</th>
<th>Number of jurors</th>
</tr>
</thead>
<tbody>
<tr>
<td>21 to 40</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>41 to 50</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>51 to 60</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>61 or older</td>
<td>19</td>
<td>33</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>66</td>
</tr>
</tbody>
</table>