Chapter 22
Two Categorical Variables: The Chi-Square Test

Two-Way Tables
- When there are two categorical variables, the data are summarized in a two-way table.
- The number of observations falling into each combination of the two categorical variables is entered into each cell of the table.
- Relationships between categorical variables are described by calculating appropriate percents from the counts given in the table.

Case Study
Health Care: Canada and U.S.

Quality of life

<table>
<thead>
<tr>
<th>Quality of life</th>
<th>Canada</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much better</td>
<td>75</td>
<td>541</td>
</tr>
<tr>
<td>Somewhat better</td>
<td>71</td>
<td>498</td>
</tr>
<tr>
<td>About the same</td>
<td>96</td>
<td>779</td>
</tr>
<tr>
<td>Somewhat worse</td>
<td>50</td>
<td>282</td>
</tr>
<tr>
<td>Much worse</td>
<td>19</td>
<td>65</td>
</tr>
<tr>
<td>Total</td>
<td>311</td>
<td>2165</td>
</tr>
</tbody>
</table>

Case Study
Health Care: Canada and U.S.

Compare the Canadian group to the U.S. group in terms of feeling much better:

<table>
<thead>
<tr>
<th>Quality of life</th>
<th>Canada</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much better</td>
<td>75</td>
<td>541</td>
</tr>
<tr>
<td>Somewhat better</td>
<td>71</td>
<td>498</td>
</tr>
<tr>
<td>About the same</td>
<td>96</td>
<td>779</td>
</tr>
<tr>
<td>Somewhat worse</td>
<td>50</td>
<td>282</td>
</tr>
<tr>
<td>Much worse</td>
<td>19</td>
<td>65</td>
</tr>
<tr>
<td>Total</td>
<td>311</td>
<td>2165</td>
</tr>
</tbody>
</table>

We have that 75 Canadians reported feeling much better, compared to 541 Americans.
The groups appear greatly different, but look at the group totals.
**Case Study**

Health Care: Canada and U.S.

Compare the Canadian group to the U.S. group in terms of feeling much better:

<table>
<thead>
<tr>
<th>Quality of life</th>
<th>Canada</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much better</td>
<td>24%</td>
<td>25%</td>
</tr>
<tr>
<td>Somewhat better</td>
<td>23%</td>
<td>23%</td>
</tr>
<tr>
<td>About the same</td>
<td>16%</td>
<td>13%</td>
</tr>
<tr>
<td>Much worse</td>
<td>6%</td>
<td>3%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Change the counts to percents

Now, with a fairer comparison using percents, the groups appear very similar in terms of feeling much better.

**Health Care: Canada and U.S.**

Is there a relationship between the explanatory variable (Country) and the response variable (Quality of life)?

Look at the **conditional distributions** of the response variable (Quality of life), given each level of the explanatory variable (Country).

### Conditional Distributions

- If the conditional distributions of the second variable are **nearly the same** for each category of the first variable, then we say that there is **not an association** between the two variables.
- If there are significant differences in the conditional distributions for each category, then we say that there is a **association** between the two variables.

### Hypothesis Test

- In tests for two categorical variables, we are interested in whether a relationship observed in a single sample reflects a real relationship in the population.
- Hypotheses:
  - Null: the percentages for one variable are the same for every level of the other variable (no difference in conditional distributions). (No real relationship).
  - Alt: the percentages for one variable vary over levels of the other variable. (Is a real relationship).

### Multiple Comparisons

- Problem of how to do many comparisons at the same time with some overall measure of confidence in all the conclusions.
- Two steps:
  - overall test to test for any differences
  - follow-up analysis to decide which parameters (or groups) differ and how large the differences are.
- Follow-up analyses can be quite complex; we will look at only the overall test for a relationship between two categorical variables.
Hypothesis Test

- **H₀**: no real relationship between the two categorical variables that make up the rows and columns of a two-way table
- To test **H₀**, compare the **observed counts** in the table (the original data) with the **expected counts** (the counts we would expect if **H₀** were true)
  - if the observed counts are far from the expected counts, that is evidence against **H₀** in favor of a real relationship between the two variables

Expected Counts

- The expected count in any cell of a two-way table (when **H₀** is true) is:
  
  \[
  \text{expected count} = \frac{(\text{row total}) \times (\text{column total})}{\text{table total}}
  \]
- The development of this formula is based on the fact that the number of expected successes in \( n \) independent tries is equal to \( n \) times the probability \( p \) of success on each try (expected count = \( np \))
  - Example: find expected count in certain row and column (cell):
    - \( p = \text{proportion in row} = \frac{(\text{row total})}{(\text{table total})} \)
    - \( n = \text{column total} \)
    - expected count in cell = \( np = \frac{(\text{row total})(\text{column total})}{(\text{table total})} \)

Case Study

Health Care: Canada and U.S.

For the observed data to the right, find the expected value for each cell:

\[
\begin{array}{c|c|c|c}
\text{Quality of life} & \text{Canada} & \text{United States} & \text{Total} \\
\hline
\text{Much better} & 77.37 & 538.63 & 616 \\
\text{Somewhat better} & 71.47 & 497.53 & 569 \\
\text{About the same} & 109.91 & 765.09 & 875 \\
\text{Somewhat worse} & 41.70 & 290.30 & 332 \\
\text{Much worse} & 10.55 & 73.45 & 84 \\
\hline
\text{Total} & 311 & 2165 & 2476
\end{array}
\]

For the expected count of Canadians who feel 'Much better' (expected count for Row 1, Column 1):

\[
\text{expected count} = \frac{616 \times 311}{2476} = 77.37
\]

Chi-Square Statistic

- To determine if the differences between the observed counts and expected counts are statistically significant (to show a real relationship between the two categorical variables), we use the **chi-square statistic**:
  
  \[
  X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}
  \]
  where the sum is over all cells in the table.
Case Study
Health Care: Canada and U.S.

<table>
<thead>
<tr>
<th>Quality of Life</th>
<th>Observed counts in Canada</th>
<th>Expected counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much better</td>
<td>75</td>
<td>77.37</td>
</tr>
<tr>
<td>Somewhat better</td>
<td>71</td>
<td>71.47</td>
</tr>
<tr>
<td>About the same</td>
<td>96</td>
<td>109.91</td>
</tr>
<tr>
<td>Somewhat worse</td>
<td>58</td>
<td>87.70</td>
</tr>
<tr>
<td>Much worse</td>
<td>19</td>
<td>10.55</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \sum \left( \frac{(O - E)^2}{E} \right) - 538.63 \\
= 0.073 + 0.010 + ... \\
= 11.725
\]

Chi-Square Test

- Calculate value of chi-square statistic
  - by hand (cumbersome)
  - using technology (computer software, etc.)
- Find P-value in order to reject or fail to reject \( H_0 \)
  - use chi-square table for chi-square distribution (later in this chapter)
  - from computer output
- If significant relationship exists (small P-value):
  - compare appropriate percents in data table
  - compare individual observed and expected cell counts
  - look at individual terms in the chi-square statistic

Chi-Square Test: Requirements

- The chi-square test is an approximate method, and becomes more accurate as the counts in the cells of the table get larger
- The following must be satisfied for the approximation to be accurate:
  - No more than 20% of the expected counts are less than 5
  - All individual expected counts are 1 or greater
- If these requirements fail, then two or more groups must be combined to form a new ("smaller") two-way table

Uses of the Chi-Square Test

- Tests the null hypothesis
  \( H_0 \): no relationship between two categorical variables
  when you have a two-way table from either of these situations:
  - independent SRSs from each of several populations, with each individual classified according to one categorical variable
    [Example: Health Care case study; two samples (Canadians & Americans); each individual classified according to "Quality of life"]
  - a single SRS with each individual classified according to both of two categorical variables
    [Example: Sample of 2053 subjects, with each classified according to their "Job Grade" (1, 2, 3, or 4) and their "Marital Status" (Single, Married, Divorced, or Widowed)]

Chi-Square Distributions

- Distributions that take only positive values and are skewed to the right
- Specific chi-square distribution is specified by giving its degrees of freedom (similar to t distn)
Chi-Square Test

- Chi-square test for a two-way table with \( r \) rows and \( c \) columns uses critical values from a chi-square distribution with \((r - 1)(c - 1)\) degrees of freedom
- \( P \)-value is the area to the right of \( X^2 \) under the density curve of the chi-square distribution
  - use chi-square table

Table D: Chi-Square Table

- See page 694 in text for Table D ("Chi-square Table")
- The process for using the chi-square table (Table D) is identical to the process for using the \( t \)-table (Table C, page 693), as discussed in Chapter 17
- For particular degrees of freedom \((df)\) in the left margin of Table D, locate the \( X^2 \) critical value \((x^* )\) in the body of the table; the corresponding probability \((p)\) of lying to the right of this value is found in the top margin of the table (this is how to find the \( P \)-value for a chi-square test)

Case Study

Health Care: Canada and U.S.

<table>
<thead>
<tr>
<th>Quality of life</th>
<th>Canada</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much better</td>
<td>75</td>
<td>541</td>
</tr>
<tr>
<td>Somewhat better</td>
<td>71</td>
<td>498</td>
</tr>
<tr>
<td>Most the same</td>
<td>98</td>
<td>779</td>
</tr>
<tr>
<td>Somewhat worse</td>
<td>50</td>
<td>282</td>
</tr>
<tr>
<td>Much worse</td>
<td>19</td>
<td>65</td>
</tr>
</tbody>
</table>

Look in the \( df=4 \) row of Table D; the value \( X^2 = 11.725 \) falls between the 0.02 and 0.01 critical values.
Thus, the \( P \)-value for this chi-square test is between 0.01 and 0.02 (is actually 0.019482).
** \( P \)-value < .05, so we conclude a significant relationship **

Chi-Square Test and Z Test

- For a 2×2 table, the \( X^2 \) with \( df=1 \) is just the square of the \( z \) statistic
  - \( P \)-value for \( X^2 \) will be the same as the two-sided \( P \)-value for \( z \)
  - should use the \( z \) test to compare two proportions, because it gives the choice of a one-sided or two-sided test (and is also related to a confidence interval for the difference in two proportions)

Chi-Square Goodness of Fit Test

- A variation of the Chi-square statistic can be used to test a different kind of null hypothesis: that a single categorical variable has a specific distribution
- The null hypothesis specifies the probabilities \((p_i)\) of each of the \( k \) possible outcomes of the categorical variable
- The chi-square goodness of fit test compares the observed counts for each category with the expected counts under the null hypothesis
Chi-Square Goodness of Fit Test

- **H₀**: \( p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0} \)
- **Hₐ**: proportions are not as specified in H₀
- For a sample of \( n \) subjects, observe how many subjects fall in each category
- Calculate the expected number of subjects in each category under the null hypothesis: expected count = \( np_i \) for the \( p_i \) category

Chi-Square Goodness of Fit Test

- Calculate the chi-square statistic (same as in previous test):
  \[ \chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \]
- The degrees of freedom for this statistic are \( df = k - 1 \) (the number of possible categories minus one)
- Find \( P \)-value using Table D

Chi-Square Goodness of Fit Test

**THE CHI-SQUARE TEST FOR GOODNESS OF FIT**
A categorical variable has \( k \) possible outcomes, with probabilities \( p_1, p_2, p_3, \ldots, p_k \). That is, \( p_i \) is the probability of the \( i \)th outcome. We have \( n \) independent observations from this categorical variable.
To test the null hypothesis that the probabilities have specified values
\( H₀: p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0} \)
use the chi-square statistic
\[ \chi^2 = \sum \frac{(\text{count of outcome } i - np_{i0})^2}{np_{i0}} \]
The \( P \)-value is the area to the right of \( \chi^2 \) under the density curve of the chi-square distribution with \( k - 1 \) degrees of freedom.

Case Study

**Births on Weekends?**


A random sample of 140 births from local records was collected to show that there are fewer births on Saturdays and Sundays than there are on weekdays

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Births</td>
<td>13</td>
<td>23</td>
<td>24</td>
<td>20</td>
<td>27</td>
<td>18</td>
<td>15</td>
</tr>
</tbody>
</table>

Do these data give significant evidence that local births are not equally likely on all days of the week?

Case Study

**Births on Weekends?**

**Null Hypothesis**

\( H₀: \) probabilities are the same on all days
\( Hₐ: \) \( p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = \frac{1}{7} \)
### Case Study
Births on Weekends?

**Expected Counts**

Expected count = \( n \times p_i = 140 \times (1/7) = 20 \)

for each category (day of the week)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed births</td>
<td>13</td>
<td>23</td>
<td>24</td>
<td>20</td>
<td>27</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>Expected births</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

**Chi-square statistic**

\[
X^2 = \sum \frac{(\text{observed count} - 20)^2}{20}
\]

\[
= \frac{(13 - 20)^2}{20} + \frac{(23 - 20)^2}{20} + \ldots + \frac{(15 - 20)^2}{20}
\]

\[
= 2.45 + 0.45 \ldots + 1.25
\]

\[
= 7.60
\]

**P-value, Conclusion**

\( X^2 = 7.60 \)

\( df = k-1 = 7-1 = 6 \)

**P-value** = \( \text{Prob}(X^2 > 7.60) \):

\( X^2 = 7.60 \) is smaller than smallest entry in \( df=6 \) row of Table D, so the \( P \)-value is > 0.25.

**Conclusion:** Fail to reject \( H_0 \) — there is not significant evidence that births are not equally likely on all days of the week.