Chapter 11: Inference for Distributions of Categorical Data

11.1 Chi-Square Goodness-of-Fit Tests

Introduction

<table>
<thead>
<tr>
<th>Input/Desired Result</th>
<th>Test Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Categorical Variable</td>
<td>Chi-Square Goodness of Fit (GOF)</td>
</tr>
<tr>
<td>Two Categorical Variables</td>
<td>Chi-Square Test for Homogeneity</td>
</tr>
<tr>
<td>Study Relationship between Two Categorical Variables</td>
<td>Chi-Square Test for Association/Independence</td>
</tr>
</tbody>
</table>

Categorical Data

Mars, Incorporated makes milk chocolate candies. Here's what the company's Consumer Affairs Department says about the color distribution of its M&M’S Milk Chocolate Candies: On average, the new mix of colors of M&M’S Milk Chocolate Candies will contain 13 percent of each of browns and reds, 14 percent yellows, 16 percent greens, 20 percent oranges and 24 percent blues.

WHY? Chi-Square Goodness-of-Fit Tests

The one-way table below summarizes the data from a sample bag of M&M’S Milk Chocolate Candies.

<table>
<thead>
<tr>
<th>Color</th>
<th>Blue</th>
<th>Orange</th>
<th>Green</th>
<th>Yellow</th>
<th>Red</th>
<th>Brown</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>9</td>
<td>8</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>6</td>
<td>60</td>
</tr>
</tbody>
</table>

The sample proportion of blue M & M’s is \( \hat{p} = \frac{9}{60} = 0.15 \).

Since the company claims that 24% of all M&M’S Milk Chocolate Candies are blue, we might believe that something fishy is going on. We could use the one-sample z test for a proportion from Chapter 9 to test the hypotheses

\[
H_0: p = 0.24 \\
H_1: p \neq 0.24
\]

However, performing a one-sample z test for each proportion would be pretty inefficient and would lead to the problem of multiple comparisons.

WHY?

Section 11.1

Chi-Square Goodness-of-Fit Tests

- COMPUTE expected counts, conditional distributions, and contributions to the chi-square statistic
- CHECK the Random, Large sample size, and Independent conditions before performing a chi-square test
- PERFORM a chi-square goodness-of-fit test to determine whether sample data are consistent with a specified distribution of a categorical variable
- EXAMINE individual components of the chi-square statistic as part of a follow-up analysis
**Hypothesis:**

**H₀:** The company’s stated color distribution for M&M’S Milk Chocolate Candies is correct.

**H₁:** The company’s stated color distribution for M&M’S Milk Chocolate Candies is not correct.

**WHY? Chi-Square Goodness-of-Fit Tests**

More important, performing one-sample z tests for each color wouldn’t tell us how likely it is to get a random sample of 60 candies with a color distribution that differs as much from the one claimed by the company as this bag does (taking all the colors into consideration at one time).

For that, we need a new kind of significance test, called a chi-square goodness-of-fit test.

**Calculating Expected Values**

For random samples of 60 candies, the average number of blue M&M’s should be $(0.24)(60) = 14.40$. This is our expected count of blue M&M’s.

Using this same method, we can find the expected counts for the other color categories:

<table>
<thead>
<tr>
<th>Color</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>9</td>
<td>14.40</td>
</tr>
<tr>
<td>Orange</td>
<td>9</td>
<td>12.00</td>
</tr>
<tr>
<td>Green</td>
<td>12</td>
<td>9.60</td>
</tr>
<tr>
<td>Yellow</td>
<td>15</td>
<td>8.40</td>
</tr>
<tr>
<td>Red</td>
<td>10</td>
<td>7.80</td>
</tr>
<tr>
<td>Brown</td>
<td>6</td>
<td>7.80</td>
</tr>
</tbody>
</table>

The Chi-Square Statistic

To see if the data give convincing evidence against the null hypothesis, we compare the observed counts from our sample with the expected counts assuming $H₀$ is true. If the expected counts differ from the expected counts, that’s the evidence we were seeking.

Big Chi-Square values means the data is far from what we are expecting, giving us evidence against the null (reject).

The chi-square statistic is a measure of how far the observed counts are from the expected counts.

### Expected Values

<table>
<thead>
<tr>
<th>Color</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>9</td>
<td>14.40</td>
</tr>
<tr>
<td>Orange</td>
<td>9</td>
<td>12.00</td>
</tr>
<tr>
<td>Green</td>
<td>12</td>
<td>9.60</td>
</tr>
<tr>
<td>Yellow</td>
<td>15</td>
<td>8.40</td>
</tr>
<tr>
<td>Red</td>
<td>10</td>
<td>7.80</td>
</tr>
<tr>
<td>Brown</td>
<td>6</td>
<td>7.80</td>
</tr>
</tbody>
</table>

### Calculating Expected Values

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

\[
\chi^2 = \frac{(9 - 14.40)^2}{14.40} + \frac{(9 - 12.00)^2}{12.00} + \frac{(1 - 9.60)^2}{9.60} + \frac{(15 - 8.40)^2}{8.40} + \frac{(10 - 7.80)^2}{7.80} + \frac{(6 - 7.80)^2}{7.80}
\]

\[
\chi^2 = 2.025 + 1.333 + 0.600 + 5.186 + 0.621 + 0.415 = 10.180
\]

We computed the chi-square statistic for our sample of 60 M&M’s to be $\chi^2 = 10.180$. Because all of the expected counts are at least 5, the $\chi^2$ statistic will follow a chi-square distribution with df = 6 (number of categories - 1 = 5 reasonable) and when $H₀$ is true.

**Calculating Expected Values**

The value $\chi^2 = 10.180$ falls between the critical values 9.49 and 11.07. The corresponding areas in the right tail of the chi-square distribution with df = 5 are 0.10 and 0.05.

Since our $P$ value is between 0.05 and 0.10, it is greater than $\alpha = 0.05$. Therefore, we fail to reject $H₀$. We don’t have sufficient evidence to conclude that the company’s claimed color distribution is incorrect.
Theory: The Chi-Square Distributions and P-Values

- The sampling distribution of Chi-squared statistic is not Normal.
- It is a right-skewed distribution that allows only positive values because \( x^2 \) can never be negative.
- When the expected counts are all at least 5, the sampling distribution of \( x^2 \) statistic is close to a chi-square distribution with degrees of freedom equal to the number of categories minus 1.

Conditions for Chi-Square

- **Random:** Samples must be randomly taken.
- **Large Sample Size:** All expected counts must be at least 5.
- **Independent:** All observations are independent. When sampling without replacement, the sample size must be less than 10% of the population size.

Using your TI-Nspire:

1. Enter the observed and expected data in the spreadsheet.
2. Label the 1st column: mmobs and 2nd column: mmexp.

<table>
<thead>
<tr>
<th>Color</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>9</td>
<td>14.40</td>
</tr>
<tr>
<td>Orange</td>
<td>8</td>
<td>12.00</td>
</tr>
<tr>
<td>Green</td>
<td>12</td>
<td>9.80</td>
</tr>
<tr>
<td>Yellow</td>
<td>15</td>
<td>8.40</td>
</tr>
<tr>
<td>Red</td>
<td>10</td>
<td>7.80</td>
</tr>
<tr>
<td>Brown</td>
<td>6</td>
<td>7.80</td>
</tr>
</tbody>
</table>

When Were You Born?

Are births evenly distributed across the days of the week?
The one-way table below shows the distribution of births across the days of the week in a random sample of 140 births from local records in a large city. Do these data give significant evidence that local births are not equally likely on all days of the week?

<table>
<thead>
<tr>
<th>Day</th>
<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>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>
**Parameters & Hypothesis:**
We want to perform a test of

- \( H_0 \): Birth days in this local area are evenly distributed across the days of the week.
- \( H_a \): Birth days in this local area are not evenly distributed across the days of the week.

The null hypothesis says that the proportions of births are the same on all days. In that case, all 7 proportions must be 1/7. So we could also write the hypotheses as

- \( H_0 \): \( p_{\text{Sun}} = p_{\text{Mon}} = \ldots = p_{\text{Sat}} = 1/7 \).
- \( H_a \): At least one of the proportions is not 1/7.

We will use \( \alpha = 0.05 \).

**Assess Conditions:**
- **Random** The data came from a random sample of local births.
- **Large Sample Size** Assuming \( H_0 \) is true, we would expect one-seventh of the births to occur on each day of the week. For the sample of 140 births, the expected count for all 7 days would be 1/7(140) = 20 births. Since 20 ≥ 5, this condition is met.
- **Independent** Individual births in the random sample should occur independently (assuming no twins). Because we are sampling without replacement, there need to be at least 10(140) = 1400 births in the local area. This should be the case in a large city.

**Name Test & (Calculate) Test Statistic**

**Name Test:** Chi-Square goodness-of-fit test.

**Test statistic:**

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

**Obtain P Value**

\( P \)-value = 0.269.

**Inherited Traits**

Biologists wish to cross pairs of tobacco plants having genetic makeup \( Gg \), indicating that each plant has one dominant gene \( G \) and one recessive gene \( g \) for color. In other words, the biologists predict that 25% of the offspring will be green, 50% will be yellow-green and 25% will be albino.

<table>
<thead>
<tr>
<th>Offspring color</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td>Yellow-green</td>
<td>50</td>
<td>42</td>
</tr>
<tr>
<td>Albino</td>
<td>11</td>
<td>21</td>
</tr>
</tbody>
</table>

Do these data differ significantly from what the biologists have predicted? Carry out an appropriate test at the \( \alpha = 0.05 \) level to help answer this question.
Parameters & Hypothesis:

$H_0$: The biologists’ predicted color distribution for tobacco plant offspring is correct.
That is, $p_{green} = 0.25$, $p_{yellow-green} = 0.5$, $p_{albino} = 0.25$
$H_a$: The biologists’ predicted color distribution isn’t correct. That is, at least one of the stated proportions is incorrect.
We will use $\alpha = 0.05$.

Make a Decision & State Conclusion

Because the $P$-value, 0.0392, is less than $\alpha = 0.05$, we will reject $H_0$. We have convincing evidence that the biologists’ hypothesized distribution for the color of tobacco plant offspring is incorrect.

Assess Conditions

• **Random** The data came from a random sample of local births.
• **Large Sample Size** We check that all expected counts are at least 5. Assuming $H_0$ is true, the expected counts for the different colors of offspring are green: (0.25)(84) = 21; yellow-green: (0.50)(84) = 42; albino: (0.25)(84) = 21
• **Independent** Individual offspring inherit their traits independently from one another. Since we are sampling without replacement, there would need to be at least $10(84) = 840$ tobacco plants in the population. This seems reasonable to believe.

WARNING!!!!

In order to perform a Chi-Square Test, the data must be counts!!!!

Name Test, (Calculate) Test Statistic & Obtain P-value

Name: Chi-square goodness-of-fit test.

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

\[
= \frac{(23 - 21)^2}{21} + \frac{(50 - 42)^2}{50} + \frac{(11 - 2)^2}{21}
\]

\[
= 6.476
\]

11.2 Inference for Relationships
Chapter 11

Section 11.2
Inference for Relationships

- COMPUTE expected counts, conditional distributions, and contributions to the chi-square statistic
- CHECK the Random, Large sample size, and Independent conditions before performing a chi-square test
- PERFORM a chi-square test for homogeneity to determine whether the distribution of a categorical variable differs for several populations or treatments
- PERFORM a chi-square test for association/independence to determine whether there is convincing evidence of an association between two categorical variables
- EXAMINE individual components of the chi-square statistic as part of a follow-up analysis
- INTERPRET computer output for a chi-square test based on a two-way table

Expected Counts and the Chi-Square Statistic

The problem of how to do many comparisons at once with an overall measure of confidence in all our conclusions is common in statistics. This is the problem of multiple comparisons. Statistical methods for dealing with multiple comparisons usually have two parts:

1. An overall test to see if there is good evidence of any differences among the parameters that we want to compare.
2. A detailed follow-up analysis to decide which of the parameters differ and to estimate how large the differences are.

Introduction

<table>
<thead>
<tr>
<th>Input/Desired Result</th>
<th>Test Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Categorical Variable</td>
<td>Chi-Square Goodness of Fit (GOF)</td>
</tr>
<tr>
<td>Two Categorical Variables</td>
<td>Chi-Square Test for Homogeneity</td>
</tr>
<tr>
<td>Study Relationship between</td>
<td>Chi-Square Test for Association/Independence</td>
</tr>
<tr>
<td>Two Categorical Variables</td>
<td></td>
</tr>
</tbody>
</table>

Example: Comparing Conditional Distributions

Market researchers suspect that background music may affect the mood and buying behavior of customers. One study in a supermarket compared three randomly assigned 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>Music</th>
<th>Wine</th>
<th>None</th>
<th>French</th>
<th>Italian</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>French</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>Italian</td>
<td>11</td>
<td>1</td>
<td>19</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>43</td>
<td>35</td>
<td>35</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>84</td>
<td>75</td>
<td>84</td>
<td>243</td>
</tr>
</tbody>
</table>

Are the distributions of wine purchases under the three music treatments similar or different?
Hypothesis:

\( H_0 \): There is no difference in the distribution of a categorical variable for several populations or treatments.

\( H_a \): There is a difference in the distribution of a categorical variable for several populations or treatments.

**Expected Counts and the Chi-Square Statistic**

To find the expected counts, we start by assuming that \( H_0 \) is true. We can see from the two-way table that 99 of the 243 bottles of wine bought during the study were French wines.

If the specific type of music that's playing has no effect on wine purchases, the proportion of French wine sold under each music condition should be \( \frac{99}{243} = 0.407 \).

The overall proportion of Italian wine bought during the study was \( \frac{31}{243} = 0.128 \). So the expected counts of Italian wine bought under each treatment are:

\[
\begin{array}{ccc}
\text{Wine} & \text{French} & \text{Italian} \\
\text{None} & 30 & 30 \\
\text{French} & 30 & 30 \\
\text{Italian} & 11 & 19 \\
\text{Other} & 43 & 35 \\
\text{Total} & 84 & 75 & 84 & 243
\end{array}
\]

The expected count in any cell of a two-way table when \( H_0 \) is true is

\[
\text{expected count} = \frac{\text{row total} \cdot \text{column total}}{\text{table total}}
\]

Calculating The Chi-Square Statistic

The tables below show the observed and expected counts for the wine and music experiment.

<table>
<thead>
<tr>
<th>Observed Counts</th>
<th>Expected Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wine</td>
<td>Music</td>
</tr>
<tr>
<td>Wine</td>
<td>None</td>
</tr>
<tr>
<td>French</td>
<td>30</td>
</tr>
<tr>
<td>Italian</td>
<td>11</td>
</tr>
<tr>
<td>Other</td>
<td>43</td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
</tr>
</tbody>
</table>

The expected count is

\[
\text{Expected} = \frac{\text{row total} \cdot \text{column total}}{\text{table total}}
\]

For the French wine with no music, the observed count is 30 bottles and the expected count is 34.22. The contribution to the \( \chi^2 \) statistic for this cell is

\[
\frac{(30 - 34.22)^2}{34.22} = 0.52
\]

The \( \chi^2 \) statistic is the sum of nine such terms:

\[
\chi^2 = \sum \left( \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \right)
\]

\[
= \frac{(30 - 34.22)^2}{34.22} + \frac{(39 - 30.56)^2}{30.56} + \cdots + \frac{(35 - 39.06)^2}{39.06}
\]

\[
= 0.52 + 2.33 + \cdots + 0.42 + 18.28
\]
Does Music Influence Purchases?

We use df = 4 because:

\[(\text{wine} - 1)(\text{music} - 1) = (3-1)(3-1) = 4\]

<table>
<thead>
<tr>
<th>P</th>
<th>df</th>
<th>0.025</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>16.42</td>
<td>18.47</td>
</tr>
</tbody>
</table>

The small P-value gives us convincing evidence to reject \( H_0 \) and conclude that there is a difference in the distributions of wine purchases at this store when no music, French accordion music, or Italian string music is played.

Furthermore, the random assignment allows us to say that the difference is caused by the music that’s played.

Follow-up Analysis

Chi-Square Test: None, French, Italian

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

<table>
<thead>
<tr>
<th></th>
<th>None</th>
<th>French</th>
<th>Italian</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>39</td>
<td>30</td>
<td>99</td>
</tr>
<tr>
<td>2</td>
<td>34.22</td>
<td>30.86</td>
<td>34.22</td>
<td></td>
</tr>
<tr>
<td>0.521</td>
<td>2.334</td>
<td>0.521</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>1</td>
<td>19</td>
<td>31</td>
</tr>
<tr>
<td>10.72</td>
<td>9.57</td>
<td>10.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.008</td>
<td>0.672</td>
<td>0.008</td>
<td>6.404</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>43</td>
<td>35</td>
<td>35</td>
<td>113</td>
</tr>
<tr>
<td>39.06</td>
<td>34.86</td>
<td>39.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.397</td>
<td>0.000</td>
<td>0.422</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>94</td>
<td>75</td>
<td>84</td>
<td>243</td>
</tr>
</tbody>
</table>

Chi-Sq = 18.278, df = 4, P-Value = 0.001

Looking at the output, we see that just two of the nine components that make up the chi-square statistic contribute about 14 (almost 77%) of the total \( \chi^2 = 18.28 \).

We are led to a specific conclusion: sales of Italian wine are strongly affected by Italian and French music.

Hypothesis: Chi-Square Test for Homogeneity

\( H_0 \): There is no difference in the distribution of a categorical variable for several populations or treatments.

\( H_a \): There is a difference in the distribution of a categorical variable for several populations or treatments.

Conditions: Chi-Square Statistic

✓ Random: Random assignment
✓ Large Sample Size: All the expected counts must be at least 5. (Use expected matrix on calculator and copy)
✓ Independent:
  ✓ Individual observations (studied groups are not connected)
  ✓ Less than 10% of population

Using Your TI-Nspire:

1. Create matrix. Menu, 7: Matrix, 1: Create
2. Fill in data.
3. Store matrix. Blue “ctrl” button, “var” and then type name of the table.

Using your TI-Nspire:

Using Your TI-Nspire:
To see the expected counts:
Press Vars.
Select: stat. expmatrix

Assess Conditions
• Random The data came from separate random samples of 96 cell-only and 104 landline users.
• Large Sample Size: Since all expected counts are greater than 5, we met the condition. (Use your calculator!! Then list counts in a table.)

Cell-Only Telephone Users
Random digit dialing telephone surveys used to exclude cell phone numbers. If the opinions of people who have only cell phones differ from those of people who have landline service, the poll results may not represent the entire adult population. The Pew Research Center interviewed separate random samples of cell-only and landline telephone users who were less than 30 years old. Here’s what the Pew survey found about how these people describe their political party affiliation.

<table>
<thead>
<tr>
<th></th>
<th>Cell-only sample</th>
<th>Landline sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat or lean Dem.</td>
<td>49</td>
<td>47</td>
</tr>
<tr>
<td>Refuse to lean either way</td>
<td>15</td>
<td>27</td>
</tr>
<tr>
<td>Republican or lean Rep.</td>
<td>32</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>96</td>
<td>104</td>
</tr>
</tbody>
</table>

Name Test & (Calculate) Test Statistic
Since the conditions are satisfied, we can perform chi-test for homogeneity.

Test statistic
\[ \chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \]

USE
\[ \chi^2 = 3.22 \]

Parameters & Hypothesis
\[ H_0: \text{There is no difference in the distribution of party affiliation in the cell-only and landline populations.} \]
\[ H_a: \text{There is a difference in the distribution of party affiliation in the cell-only and landline populations.} \]

Obtain p-value, Make a Decision & State Conclusion
Because the \( p \)-value, 0.20, is greater than \( \alpha = 0.05 \), we fail to reject \( H_0 \). There is not enough evidence to conclude that the distribution of party affiliation differs in the cell-only and landline user populations.

We will use \( \alpha = 0.05 \).
Cocaine Addiction is Hard to Break
Cocaine addicts need cocaine to feel any pleasure, so perhaps giving them an antidepressant drug will help. A three-year study with 72 chronic cocaine users compared an antidepressant drug called desipramine with lithium (a standard drug to treat cocaine addiction) and a placebo. One-third of the subjects were randomly assigned to receive each treatment. Here are the results:

<table>
<thead>
<tr>
<th>Group</th>
<th>Treatment</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Desipramine</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>Lithium</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>Placebo</td>
<td>20</td>
<td>4</td>
</tr>
</tbody>
</table>

Parameters & Hypothesis

$H_0$: there is no difference in the relapse rate for the three treatments.

$H_a$: there is a difference in the relapse rate for are different the three treatments.

We will use $\alpha = 0.01$.

Name Test, (Calculate) Test Statistic & Obtain P-value

Name: Chi-test for homogeneity

![Chi-test results]

Make a Decision and State Conclusion

Because the $P$-value, 0.0052, is less than $\alpha = 0.01$, we reject $H_0$. We have sufficient evidence to conclude that the true relapse rates for the three treatments are not all the same.

Assess Conditions

- **Random**: The subjects were randomly assigned to the treatment groups.
- **Large Sample Size**: All the expected counts are $\geq 5$ so the condition is met. (Use expected matrix; must copy)
- **Independent**: The random assignment helps create three independent groups. If the experiment is conducted properly, then knowing one subject’s relapse status should give us no information about another subject’s outcome. So individual observations are independent.

The Chi-Square Test for Association/Independence

Another common situation that leads to a two-way table is when a single random sample of individuals is chosen from a single population and then classified according to two categorical variables. In that case, our goal is to analyze the relationship between the variables.
More About the Chi-Square Test for Association/Independence

We often gather data from a random sample and arrange them in a two-way table to see if two categorical variables are associated. The sample data are easy to investigate: turn them into percents and look for a relationship between the variables.

Our null hypothesis is that there is no association between the two categorical variables. The alternative hypothesis is that there is an association between the variables. For the observational study of anger level and coronary heart disease, we want to test the hypotheses.

The Chi-Square Test for Association/Independence

Hypothesis:

- $H_0$: There is no association between two categorical variables in the population of interest.
- $H_a$: There is an association between two categorical variables in the population of interest.

Or, alternatively

- $H_0^{'}$: Two categorical variables are independent in the population of interest.
- $H_a^{'}$: Two categorical variables are not independent in the population of interest.

The Chi-Square Test for Association/Independence

A study followed a random sample of 8474 people with normal blood pressure for about four years. All the individuals were free of heart disease at the beginning of the study. Each person took the Spielberger Trait Anger Scale test, which measures how prone a person is to sudden anger. Researchers also recorded whether each individual developed coronary heart disease (CHD). This includes people who had heart attacks and those who needed medical treatment for heart disease. Here is a two-way table that summarizes the data:

<table>
<thead>
<tr>
<th>CHD</th>
<th>Low anger</th>
<th>Moderate anger</th>
<th>High anger</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>53</td>
<td>110</td>
<td>27</td>
<td>190</td>
</tr>
<tr>
<td>No</td>
<td>3057</td>
<td>4621</td>
<td>606</td>
<td>8294</td>
</tr>
<tr>
<td>Total</td>
<td>3110</td>
<td>4731</td>
<td>633</td>
<td>8474</td>
</tr>
</tbody>
</table>

Background: Angry People and Heart Disease

<table>
<thead>
<tr>
<th></th>
<th>Low anger</th>
<th>Moderate anger</th>
<th>High anger</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD</td>
<td>3110</td>
<td>4731</td>
<td>633</td>
<td>8474</td>
</tr>
<tr>
<td>CHD</td>
<td>3057</td>
<td>4621</td>
<td>606</td>
<td>8294</td>
</tr>
<tr>
<td>CHD</td>
<td>3097</td>
<td>4621</td>
<td>606</td>
<td>8294</td>
</tr>
</tbody>
</table>

There is a clear trend: the anger score increases, so does the percent who suffer heart disease. A much higher percent of people in the high anger category developed CHD (4.27%) than in the moderate (2.33%) and low (1.70%) anger categories.

Parameters & Hypothesis

$H_0$: There is no association between anger level and heart disease in the population of people with normal blood pressure.

$H_a$: There is an association between anger level and heart disease in the population of people with normal blood pressure.

OR

$H_0$: Anger and heart disease are independent in the population of people with normal blood pressure.

$H_a$: Anger and heart disease are not independent in the population of people with normal blood pressure.

Assess Conditions

- **Random**: The data came from a random sample of 8474 people with normal blood pressure.
- **Large Sample Size**: All the expected counts are at least 5, so this condition is met.
- **Independent**: Knowing the values of both variables for one person in the study gives us no meaningful information about the values of the variables for another person. So individual observations are independent. Because we are sampling without replacement, we need to check that the total number of people in the population with normal blood pressure is at least 10(8474) = 84,740. This seems reasonable to assume.

<table>
<thead>
<tr>
<th></th>
<th>Low anger</th>
<th>Moderate anger</th>
<th>High anger</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD</td>
<td>53</td>
<td>110</td>
<td>27</td>
<td>69.73</td>
<td>159.88</td>
<td>0.19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHD</td>
<td>3057</td>
<td>4621</td>
<td>606</td>
<td>3360.27</td>
<td>4624.92</td>
<td>618.81</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Obtain P-value, Make a Decision and State Conclusion

P-Value: 0.00032

Because the P-value is clearly less than $\alpha = 0.01$, we reject $H_0$ and conclude that anger level and heart disease are associated in the population of people with normal blood pressure.

Which Chi-Square Test Do I Use?!?!?

Instead of focusing on the question asked, it’s much easier to look at how the data were produced.

✔ If the data come from two or more independent random samples or treatment groups in a randomized experiment, then do a chi-square test for homogeneity.

✔ If the data come from a single random sample, with the individuals classified according to two categorical variables, use a chi-square test for association/independence.

<table>
<thead>
<tr>
<th>Chi-Square Test</th>
<th>Key Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goodness of Fit</td>
<td>One variable, one population</td>
</tr>
<tr>
<td>Homogeneity</td>
<td>One variable, two or more populations/groups</td>
</tr>
<tr>
<td>Association/Independence</td>
<td>Two variables, one population</td>
</tr>
</tbody>
</table>