Chi-square Procedures

Recall Properties

1. The total area under $\chi^2$ curve is equal to 1.

2. The $\chi^2$ curve starts at 0 on the horizontal axis and extends indefinitely to the right, approaching the horizontal axis as it does so.

3. The $\chi^2$ curve is not symmetrical.

4. As the number of df gets larger, $\chi^2$ curves look increasingly like normal curves.

Finding $\chi^2$ for a Specified AUC

Example: $\alpha = 0.05$

Df = 7

Solution:
The $\chi^2$ value for $\alpha = 0.05$ is 14.067, $\chi^2_R$
The $\chi^2$ value for $1-\alpha = 0.95$ is 2.167, $\chi^2_L$

Note: To find the left value look at the table in the book; not on the formula card!
Three Commonly Used Chi-Sq Tests

1. Goodness-Of-Fit Test
   a. Based on a simple frequency distribution involving one variable.
   b. NOTE: If the frequency distribution of the variable is reported in classes, then use the class mark to represent the class.
   c. Used to test if the frequency distribution fits a theoretical distribution or some specified pattern

2. Test of Independence
   a. Based on frequency in a contingency table of two or more variables of a single sample
   b. Used to test if the two variables are associated or not; alternatively tests if the outcome of one variable has an effect on the outcome of the other variable(s)

3. Test of Homogeneity of Proportions
   a. Based on proportions of a variable in a contingency table. The variable is sampled multiple times from different populations.
   b. Used to test if the proportions are the same
Goodness-Of-Fit Test

I. Assumptions:
1. All expected frequencies [NOT observed frequencies (O) – see step 2 for calculation of expected frequencies] are at least 1
2. At most 20% of the expected frequencies are < 5.0

II. Steps:
1. State the null and alternative hypotheses
   $H_0$: The variable under consideration has the specified distribution.
   $H_a$: The variable under consideration does not have the specified distribution.

2. Calculate the expected frequencies using the formula
   $E = np$
   Where: $n$ is the sample size and $p$ is the probability for the category given in the null hypothesis.

3. Check whether the expected frequencies (from step 2) satisfy the assumptions listed above in I. If they do not, then do not use this procedure. Alternative methods to use are outside the scope of this class. If assumptions are met continue to step 4.

4. Prepare to compute the $\chi^2$ test. Start by choosing the significance level (generally provided to you; if not default is generally $\alpha = 0.05$)

5. Obtain the critical value of $\chi^2$ with $df=k-1$, where $k$ is the number of categories in the distribution.
6. Compute the test statistic: \( \chi^2 = \sum_{i}^{k} \left[ \frac{(O_i - E_i)^2}{E_i} \right] \)

7. If the value of the test statistic falls in the rejection region, reject \( H_0 \); otherwise, do not reject \( H_0 \).

8. State conclusion in words.

**Example**

Observing that the proportion of blue M&Ms in his bowl of candy appeared to be less than that of the other colors, R. Fricker, Jr. decided to compare the color distribution in randomly chosen bags of M&Ms to the theoretical distribution reported by M&M/MARS consumer affairs. Fricker published his findings in the article, “The Mysterious Case of the Blue M&Ms” (Chance, Vol 9(4), pp 19-22). For his experiment Fricker bought three bags of M&Ms from local stores and counted the number of each color. The average number of each color in the three bags was distributed as shown in the “Observed Frequency” column.

Q: Do the data provide sufficient evidence to conclude that the color distribution of M&Ms observed by Fricker differs from the distribution of colors reported by M&M/MARS consumer affairs? Use a 95% confidence level.
### M&M Colors

<table>
<thead>
<tr>
<th>M&amp;M Colors</th>
<th>Observed Frequency (O)</th>
<th>Expected Frequency (E=np)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>152</td>
<td>(509)(0.30) = 152.7</td>
</tr>
<tr>
<td>Yellow</td>
<td>114</td>
<td>(509)(0.20) = 101.8</td>
</tr>
<tr>
<td>Red</td>
<td>106</td>
<td>(509)(0.20) = 101.8</td>
</tr>
<tr>
<td>Orange</td>
<td>51</td>
<td>(509)(0.10) = 50.9</td>
</tr>
<tr>
<td>Green</td>
<td>43</td>
<td>(509)(0.10) = 50.9</td>
</tr>
<tr>
<td>Blue</td>
<td>43</td>
<td>(509)(0.10) = 50.9</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>1.00</strong></td>
<td><strong>n=509</strong></td>
</tr>
</tbody>
</table>

### Solution

We have \( n = 509 \)

\( Df = (k-1) = 5 \)

Critical Chi-Square value at \( \alpha = 0.05 \) with \( df=5 \) is 11.070

All expected frequencies are 1 or greater.

At most 20% of the expected frequencies are less than 5.

Ho: The color distribution of the M&Ms is as reported by the company.

Ha: The color distribution of the M&Ms is different from that reported by the company.

The test statistic is:

\[
\chi^2 = \frac{(152-152.7)^2}{152.7} + \frac{(114-101.8)^2}{101.8} + \frac{(106-101.8)^2}{101.8} + \frac{(51-50.9)^2}{50.9} + \frac{(43-50.9)^2}{50.9} + \frac{(43-50.9)^2}{50.9} = 4.0910
\]

The value of the test statistic does not fall in the rejection region, so we **do not reject** \( H_0 \). The test results are not statistically significant at the 5% level.

The data do not provide sufficient evidence to conclude that the color distribution of M&Ms differs from that reported by M&M / MARS consumer affairs.
Contingency Tables

Contingency Tables are created by simultaneously grouping data from two or more variables into frequency distribution. That is, contingency tables or n-way tables represent the frequency distribution of a bivariate (n-variate) data.

Association (Independence)

Association is the relationship of two variables with each other. In other words, there exists an association between two variables of a population if knowing the value of one allows us to obtain information about the other.

Conditional Distribution: Each column of a contingency table can provides the conditional distribution of the variable shown on the rows of the table by the class level represented by that column.

Marginal Distribution: The total column provides the (unconditional) distribution of the variable shows as the rows of the contingency table. This is called the marginal distribution.

If there is no association between two variables, the conditional and marginal distributions will be the same for all class levels.

The normal charting technique to show association is a segmented (stacked) bar chart.
Chi-Square Independence Test

Assumptions:
1) All *expected frequencies* (calculated in Step 2 below) are 1 or greater.
2) At most 20% of the *expected frequencies* are < 5.

Steps:
1) State the null and alternative hypotheses
   \[ H_0: \text{The two variables under consideration are not associated} \]
   \[ H_a: \text{The two variables under consideration are associated.} \]

2) Calculate the *expected frequencies* using the formula

\[ E = \frac{(RC)}{n} \]

Where:
- \( R \) = Row total
- \( C \) = Column total
- \( n \) = sample size

Place each expected frequency below its corresponding observed frequency in the contingency table. You will have as many expected frequencies as you have cells in the table.

3) Check whether the expected frequencies satisfy the assumptions. If they do not, then **do not use** this procedure. Alternative methods to use are outside the scope of this class. If assumptions are met continue to step 4.
4) Prepare to compute the χ² test. Start by choosing the significance level (generally provided to you; if not default is generally α = 0.05)

5) Obtain the critical value of χ² with $df = (r-1)(c-1)$, where $r$ and $c$ are the number of possible values for the two variables under consideration. Alternatively, $r$ is the number of rows; and $c$ is the number of columns in the contingency table.

6) Compute the test statistic: $\chi^2 = \sum_{i}^{k} \left[ \frac{(O_i - E_i)^2}{E_i} \right]$ 

Where:
- $k$ is $r \times c$
- $O$ is the observed frequencies
- $E$ is the expected frequencies

7) If the value of the test statistic falls in the rejection region, reject Ho; otherwise, do not reject Ho.

8) State conclusion in words.
Example

<table>
<thead>
<tr>
<th>Siskel’s \ Ebert’s</th>
<th>Thumbs Down</th>
<th>Mixed</th>
<th>Thumbs Up</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thumbs Down</td>
<td>24</td>
<td>8</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>Mixed</td>
<td>8</td>
<td>13</td>
<td>11</td>
<td>32</td>
</tr>
<tr>
<td>Thumbs Up</td>
<td>10</td>
<td>9</td>
<td>64</td>
<td>83</td>
</tr>
<tr>
<td>Column Totals</td>
<td>42</td>
<td>30</td>
<td>88</td>
<td>160</td>
</tr>
</tbody>
</table>

At the 1% significance level do the data provide sufficient evidence to conclude that an association exists between the ratings of Siskel and Ebert?

Ho: Siskel’s and Ebert’s ratings are not associated (no association) - OR - independent
Ha: Siskel’s and Ebert’s ratings are associated - OR - not independent

Table of Observed and Expected Frequencies

<table>
<thead>
<tr>
<th>Siskel’s \ Ebert’s</th>
<th>Thumbs Down</th>
<th>Mixed</th>
<th>Thumbs Up</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thumbs Down</td>
<td>24</td>
<td>8</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>11.81</td>
<td>8.44</td>
<td>24.75</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>8</td>
<td>13</td>
<td>11</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>8.4</td>
<td>6</td>
<td>17.6</td>
<td></td>
</tr>
<tr>
<td>Thumbs Up</td>
<td>10</td>
<td>9</td>
<td>64</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>21.79</td>
<td>15.56</td>
<td>45.65</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>30</td>
<td>88</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>21.79</td>
<td>15.56</td>
<td>45.65</td>
<td></td>
</tr>
</tbody>
</table>

For example: The expected frequency for Thumbs Down from both Siskel and Ebert (E11) is: (45)(42) / 160 = 11.81

E21: (32)(42) / 160 = 8.4

All numbers in red are calculated in the same manner using the appropriate row and column totals.
The test statistic will be:

\[ \chi^2 = \frac{(24 - 11.81)^2}{11.81} + \frac{(8 - 8.44)^2}{8.44} + \frac{(13 - 24.75)^2}{24.75} \]

\[ + \frac{(8 - 8.4)^2}{8.4} + \frac{(13 - 6)^2}{6} + \frac{(11 - 17.6)^2}{17.6} \]

\[ + \frac{(10 - 21.79)^2}{21.79} + \frac{(9 - 15.56)^2}{15.56} + \frac{(64 - 45.65)^2}{45.65} \]

\[ = 45.365 \]

Critical chi-square value at the 1% level with \( df=(r-1)(c-1)=4 \) is 13.277.

Since 45.365 lies in the rejection region we reject Ho. The test results are statistically significant at the 1% level. We thus conclude that at the 1% significance level, the data provide sufficient evidence to conclude that there is an association (relationship) between Siskel’s and Ebert’s ratings.
Chi-Square Test of Homogeneity of Proportions

Assumptions:
1) All expected proportions (calculated in Step 2 below) are 1 or greater.
2) At most 20% of the expected proportions are < 5.

Steps:
1. State the null and alternative hypotheses
   Ho: $p_1 = p_2 = \ldots = p_s$ (where $s = \# \ of \ samples$)
   Ha: At least one proportion is different

2. Calculate the expected proportion using the formula
   \[ E = \frac{(RC)}{n} \]
   Where:
   \( R = \) Row total
   \( C = \) Column total
   \( n = \) sample size

   Place each expected frequency below its corresponding observed frequency in the contingency table. You will have as many expected frequencies as you have cells in the table.

3. Check whether the expected frequencies satisfy the assumptions. If they do not, then do not use this procedure. Alternative methods to use are outside the scope of this class. If assumptions are met continue to step 4.
4. Prepare to compute the \( \chi^2 \) test. Start by choosing the significance level (generally provided to you; if not default is generally \( \alpha = 0.05 \))

5. Obtain the critical value of \( \chi^2 \) with \( df = (r-1)(c-1) \), where \( r \) and \( c \) are the number of possible values for the two variables under consideration. Alternatively, \( r \) is the number of rows; and \( c \) is the number of columns in the contingency table.

6. Compute the test statistic:
   \[
   \chi^2 = \sum_{i}^{k} \left[ \frac{(O_i - E_i)^2}{E_i} \right]
   \]

Where:
- \( k \) is \( r \times c \)
- \( O \) is the \textit{observed proportions}
- \( E \) is the \textit{expected proportions}

7. If the value of the test statistic falls in the rejection region, reject \( H_0 \); otherwise, do not reject \( H_0 \).

8. State conclusion in words.
Example
A psychologist selected 100 people from each of the four income groups and asked them if they were “very happy”. The percent for each group who responded “yes” and the number from the survey are show in below. At the 95% CL, test the claim that there is no difference in the proportions.

<table>
<thead>
<tr>
<th>Household Income</th>
<th>Less than $30,000</th>
<th>$30,000 to $74,999</th>
<th>$75,000 to $99,999</th>
<th>$100,000 or more</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>24%</td>
<td>33%</td>
<td>38%</td>
<td>49%</td>
<td>144</td>
</tr>
<tr>
<td>No</td>
<td>76%</td>
<td>67%</td>
<td>62%</td>
<td>51%</td>
<td>256</td>
</tr>
<tr>
<td>Column Totals</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>400</td>
</tr>
</tbody>
</table>

Ho: $p_1 = p_2 = p_3 = p_4$
Ha: At least one proportion is different

Table of Observed and Expected Frequencies

<table>
<thead>
<tr>
<th>Household Income</th>
<th>Less than $30,000</th>
<th>$30,000 to $74,999</th>
<th>$75,000 to $99,999</th>
<th>$100,000 or more</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>24</td>
<td>33</td>
<td>38</td>
<td>49</td>
<td>144</td>
</tr>
<tr>
<td>No</td>
<td>76</td>
<td>67</td>
<td>62</td>
<td>51</td>
<td>256</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>400</td>
</tr>
</tbody>
</table>

For example: The expected proportion for (E11) is: $(144)(100) / 400 = 11.81$
E21: $(256)(100) / 400 = 64$

All numbers in red are calculated in the same manner using the appropriate row and column totals.
The test statistic will be:

\[
\chi^2 = \frac{(24 - 36)^2}{36} + \frac{(33 - 36)^2}{36} + \frac{(38 - 36)^2}{36} + \frac{(49 - 36)^2}{36} \\
+ \frac{(76 - 64)^2}{64} + \frac{(67 - 64)^2}{64} + \frac{(62 - 64)^2}{64} \\
+ \frac{(51 - 64)^2}{64} = 14.15
\]

Critical chi-square value at the 5% level with \( df=(r-1)(c-1)=3 \) is 7.815.

Since 14.15 lies in the rejection region we **reject Ho**. The test results are statistically significant at the 5% level. We thus conclude that at the 5% significance level, the data provide sufficient evidence to conclude that there is at least one proportion that is different. Hence the incomes seem to make a difference in the proportion of people that are happy.