12 Chi-squared ($\chi^2$) Tests for Goodness-of-fit and Independence

The chi-squared tests are for $H_0$: “The frequency distribution of events observed in a sample is with a particular distribution” against $H_A$: “Not $H_0$”. We consider two of its forms: the test for goodness-of-fit of counts for one categorical variable to a distribution and the test for independence of two categorical variables.

Each uses a chi-square statistic of the form

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

This is a measure of .

If expected counts are all at least , and under a suitable $H_0$, then $X^2$ fits a $\chi^2$ distribution.

The Chi-Square Distributions

(Background: if $Z_1, \cdots, Z_\nu$ are independent, $N(0, 1)$ random variables, then $X^2 = \sum_{i=1}^{\nu} Z_i^2 \sim \chi^2_\nu$.)

A $\chi^2$ distribution is specified by its degrees of freedom, $\nu$. Here are some of its properties:

- $X^2 \geq 0$ (it’s a measure of distance)
- $X^2 = 0 \implies$ observed and expected counts are
- Large $X^2 \implies$ observed counts aren’t
- Each $\chi^2_\nu$ density function is skewed
- e.g. Here’s $\chi^2_6$:

- The $\chi^2$ table gives, in row and column , the point $\chi^2_{\nu, \alpha}$ with area $\alpha$ to its right.
  e.g. $\chi^2_{6,0.05} =$ (draw)
The Chi-Square Test For Goodness-of-Fit

Recall the z-test for a population proportion, \( H_0 : \pi = \pi_0 \) vs. \( H_A : \pi \neq \pi_0 \), for which an outcome takes one of two values, success or failure. The chi-square test for goodness-of-fit generalizes to the case of an outcome taking any of \( k \) values of a categorical variable, testing \( H_0: \) “These categorical data came from the specified distribution” vs. \( H_A: \) “These categorical data came from a different distribution.”

E.g. The Nice family gives trick-or-treaters a scoop of ______ M&Ms. The Naughty family gives ______ M&Ms. Anna, Teresa, Margaret, Monica, Andrew, Mary, and Philip return from trick-or-treating, and their father says, “Where did you get the M&Ms?” They know they visited only one of the Nice and Naughty homes, but can’t remember which one. Their father says, “Throw away the M&Ms.” The children ______. Their mother (a ______) says, “Let’s figure out their source.” She investigates and finds these color distributions:

<table>
<thead>
<tr>
<th></th>
<th>Brown</th>
<th>Yellow</th>
<th>Green</th>
<th>Red</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nice supply</td>
<td>20%</td>
<td>25%</td>
<td>40%</td>
<td>15%</td>
<td>100%</td>
</tr>
<tr>
<td>Naughty supply</td>
<td>50%</td>
<td>20%</td>
<td>10%</td>
<td>20%</td>
<td>100%</td>
</tr>
<tr>
<td>Anna, ..., &amp; Philip (sample)</td>
<td>12</td>
<td>15</td>
<td>17</td>
<td>6</td>
<td>( n = )</td>
</tr>
</tbody>
</table>

From which family did the kids get their M&Ms?

Test \( H_0: \) “The kids got M&Ms from the Nice family” vs. \( H_A: \) “They did not”.

**Expected Counts**

Let \( k = \# \) category values = ______. If \( n \) is the sample size and \( \pi_i \) is the expected proportion in category \( i \) under \( H_0 \), the expected count of each type is \( E_i = \) ______. The test statistic is

\[
X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},
\]

whose value for the M&Ms is \( \chi^2 = \) ______.

Under \( H_0 \), \( X^2 \sim \chi^2_\nu \), where \( \nu = k - 1 = \) ______. The P-value is \( P(X^2 > \) ______) = ______.

Conclusion:

Next, test \( H_0: \) “The kids got M&Ms from the Naughty family” vs. \( H_A: \) “They did not”. Here \( \chi^2 = \) ______.

The P-value is \( P(X^2 > \) ______) = ______.

Conclusion:
The Chi-Square Test for Independence

The *chi-square test for independence* tests $H_0$: “Categorical variables $A$ and $B$ are independent” against $H_A$: “There is __________ between $A$ and $B$.”

e.g. Here is a *contingency table* of __________ that relates the education level and smoking status of a SRS of 459 French men. Are education and smoking related?

<table>
<thead>
<tr>
<th>Education</th>
<th>Smoking status</th>
<th>Nonsmoker</th>
<th>Former</th>
<th>Moderate</th>
<th>Heavy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td></td>
<td>56</td>
<td>54</td>
<td>41</td>
<td>36</td>
<td>187</td>
</tr>
<tr>
<td>Secondary</td>
<td></td>
<td>37</td>
<td>43</td>
<td>27</td>
<td>32</td>
<td>139</td>
</tr>
<tr>
<td>University</td>
<td></td>
<td>53</td>
<td>28</td>
<td>36</td>
<td>16</td>
<td>133</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>125</td>
<td>104</td>
<td>84</td>
<td></td>
<td>459</td>
</tr>
</tbody>
</table>

Test $H_0$: “Education and smoking __________” vs. $H_A$: “There’s __________ between education and smoking”.

**Expected Counts**

Under $H_0$, $P(\text{Primary and Nonsmoker}) = $ __________, so the expected count in the Primary / Nonsmoker cell is __________.

More generally, let

- $O_{ij} =$ __________ count in row $i$ and column $j$
- $O_i =$ _____ $i$ total, $O_J =$ __________ $j$ total
- $O_\cdot =$ __________ total
- $I =$ #______, $J =$ #_________

Then, under $H_0$, the *expected cell count* in row $i$ and column $j$ is $E_{ij} = \frac{O_i O_J}{O_\cdot}$ = $\frac{(\text{row total})(\text{column total})}{\text{table total}}$. Here are the 12 expected counts:

<table>
<thead>
<tr>
<th>Education</th>
<th>Smoking status</th>
<th>Nonsmoker</th>
<th>Former</th>
<th>Moderate</th>
<th>Heavy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td></td>
<td>50.9</td>
<td>42.4</td>
<td>34.2</td>
<td></td>
<td>187</td>
</tr>
<tr>
<td>Secondary</td>
<td></td>
<td>44.2</td>
<td>37.9</td>
<td>31.5</td>
<td>25.4</td>
<td>139</td>
</tr>
<tr>
<td>University</td>
<td></td>
<td>42.3</td>
<td>36.2</td>
<td>30.1</td>
<td></td>
<td>133</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>146</td>
<td>125</td>
<td>104</td>
<td>84</td>
<td>459</td>
</tr>
</tbody>
</table>

The chi-square statistic is $X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$. For the smokers, its value $\chi^2$ has 12 terms:
The table sum is $\chi^2 = 13.3$. The required degrees of freedom is $\nu = (\# \text{rows} - 1)(\# \text{columns} - 1) = \underline{\hspace{2cm}}$, and the $P$-value is $P(X_0^2 > 13.3) = \underline{\hspace{2cm}}$.

Conclusion:

**R for $\chi^2$ tests**

```r
rm(list=ls()) # Remove all variables to start with a clean slate.

# Test goodness-of-fit of kids’ sample of M&Ms to Nice distribution.
kids.sample = c(12,15,17,6)
Nice.population = c(.20, .25, .40, .15)
chisq.test(x=kids.sample, p=Nice.population)

# Make comparative bar plots.
colors = c("Brown", "Yellow", "Green", "Red")
layout(matrix(data=1:2, nrow=2, ncol=1)) # Allow two graphs in one plot.
barplot(height=kids.sample, names.arg=colors, main="M&M’s sample")
barplot(height=Nice.population, names.arg=colors, main="Nice population")
layout(1) # Return to one graph per plot.

# Do it again for the Naughty population.
Naughty.population = c(.50, .20, .10, .20)

layout(matrix(data=1:2, nrow=2, ncol=1)) # Allow two graphs in one plot.
barplot(height=kids.sample, names.arg=colors, main="M&M’s sample")
barplot(height=Naughty.population, names.arg=colors, main="Naughty population")
layout(1) # Return to one graph per plot.

chisq.test(x=kids.sample, p=Naughty.population)

# Test independence of education and smoking.

# matrix(data, nrow, ncol, byrow=FALSE) fills an nrow by ncol matrix,
# by column, from the vector data.
(French.men = matrix(data = c(56,37,53, 54,43,28, 41,27,36, 36,32,16),
      nrow=3, ncol=4, byrow=FALSE))
chisq.test(French.men)
```