Math 101: Elementary Statistics

The $\chi^2$ Distribution

Department of Mathematics and Computer Science
University of the Philippines Baguio

November 21, 2018
Motivation.

Does color affect your appetite?

A study was done at a certain university. When people were given six varieties of gummy bears mixed in a bowl or separated by color, they ate about twice as many from the bowl with the mixed gummy bears as from the bowls that were separated by color. It is thought that when the gummy bears were mixed, people felt that it offered a greater variety of choices, and the variety of choices increased their appetites. In this case one variable color is categorical, and the other variable amount of gummy bears eaten is numerical.
Motivation.

Does color affect your appetite?

A study was done at a certain university. When people were given six varieties of gummy bears mixed in a bowl or separated by color, they ate about twice as many from the bowl with the mixed gummy bears as from the bowls that were separated by color. It is thought that when the gummy bears were mixed, people felt that it offered a greater variety of choices, and the variety of choices increased their appetites. In this case one variable color is categorical, and the other variable amount of gummy bears eaten is numerical.
Motivation.

Does color affect your appetite?

A study was done at a certain university. When people were given six varieties of gummy bears mixed in a bowl or separated by color, they ate about twice as many from the bowl with the mixed gummy bears as from the bowls that were separated by color. It is thought that when the gummy bears were mixed, people felt that it offered a greater variety of choices, and the variety of choices increased their appetites. In this case one variable color is categorical, and the other variable amount of gummy bears eaten is numerical.
Motivation.

Does color affect your appetite?

A study was done at a certain university. When people were given six varieties of gummy bears mixed in a bowl or separated by color, they ate about twice as many from the bowl with the mixed gummy bears as from the bowls that were separated by color. It is thought that when the gummy bears were mixed, people felt that it offered a greater variety of choices, and the variety of choices increased their appetites. In this case one variable color is categorical, and the other variable amount of gummy bears eaten is numerical.
Suppose a college professor distributes a faculty evaluation form to 150 students of his Calculus class. The following are two examples of multiple-choice questions appearing on the questionnaire.
Suppose a college professor distributes a faculty evaluation form to 150 students of his Calculus class. The following are two examples of multiple-choice questions appearing on the questionnaire.

1. In what interval does your current class standing in Calculus lie?
   
a. [1.0, 1.5]   
   b. (1.5, 3]   
   c. below 3.0
Suppose a college professor distributes a faculty evaluation form to 150 students of his Calculus class. The following are two examples of multiple-choice questions appearing on the questionnaire.

1. In what interval does your current class standing in Calculus lie?
   a. [1.0, 1.5]  
   b. (1.5, 3]  
   c. below 3.0

2. Would you be willing to take another course with this teacher?
   a. Yes  
   b. No
The results for these two questions are summarized in the following table:

<table>
<thead>
<tr>
<th></th>
<th>[1.0, 1.5]</th>
<th>(1.5, 3]</th>
<th>Below 3.0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>28</td>
<td>36</td>
<td>11</td>
<td>75</td>
</tr>
<tr>
<td>No</td>
<td>22</td>
<td>44</td>
<td>9</td>
<td>75</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>80</td>
<td>20</td>
<td>150</td>
</tr>
</tbody>
</table>
The results for these two questions are summarized in the following table:

<table>
<thead>
<tr>
<th></th>
<th>[1.0, 1.5]</th>
<th>(1.5, 3]</th>
<th>Below 3.0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>28</td>
<td>36</td>
<td>11</td>
<td>75</td>
</tr>
<tr>
<td>No</td>
<td>22</td>
<td>44</td>
<td>9</td>
<td>75</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>80</td>
<td>20</td>
<td>150</td>
</tr>
</tbody>
</table>

Is it true that students with higher grades tend to rate the teacher differently than students with lower grades?
Let $p_1$, $p_2$, and $p_3$ be the true proportion of students in the $[1.0, 1.5]$ category, $(1.5, 3]$, and Below 3.0 category, respectively who will take another course with the teacher.
Let $p_1$, $p_2$, and $p_3$ be the true proportion of students in the $[1.0, 1.5]$ category, $(1.5, 3]$, and Below 3.0 category, respectively who will take another course with the teacher. The null hypothesis we wish to test is

$$H_0 = p_1 = p_2 = p_3$$
Let $p_1$, $p_2$, and $p_3$ be the true proportion of students in the [1.0, 1.5] category, (1.5, 3], and Below 3.0 category, respectively who will take another course with the teacher. The null hypothesis we wish to test is

$$H_0 = p_1 = p_2 = p_3$$

and the alternative hypothesis is that at least one of $p_1$, $p_2$, and $p_3$ is different.
The $\chi^2$ Distribution

A contingency table containing $r$ rows and $c$ columns is referred to as an $r \times c$ table. The values in parentheses are the expected frequencies and the actual values obtained are the observed frequencies. The row and column totals are called marginal frequencies.

<table>
<thead>
<tr>
<th></th>
<th>[1.0, 1.5]</th>
<th>(1.5, 3]</th>
<th>Below 3.0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>28 (25)</td>
<td>36 (40)</td>
<td>11 (10)</td>
<td>75</td>
</tr>
<tr>
<td>No</td>
<td>22 (25)</td>
<td>44 (40)</td>
<td>9 (10)</td>
<td>75</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>80</td>
<td>20</td>
<td>150</td>
</tr>
</tbody>
</table>
A contingency table containing $r$ rows and $c$ columns is referred to as an $r \times c$ table. The values in parentheses are the expected frequencies and the actual values obtained are the observed frequencies. The row and column totals are called marginal frequencies.
Let $E_{ij}$ and $O_{ij}$ represent the expected frequencies and observed frequencies, respectively, in the $ith$ row and $jth$ column.
The $\chi^2$ Statistic

Let $E_{ij}$ and $O_{ij}$ represent the expected frequencies and observed frequencies, respectively, in the $ith$ row and $jth$ column. Then

$$E_{ij} = \text{expected number of cases under } H_0$$

$$= \frac{(\text{column total}) \times (\text{row total})}{\text{grand total}}$$
The $\chi^2$ Statistic

Let $E_{ij}$ and $O_{ij}$ represent the expected frequencies and observed frequencies, respectively, in the $ith$ row and $jth$ column. Then

$$E_{ij} = \text{expected number of cases under } H_0$$

$$= \frac{(\text{column total}) \times (\text{row total})}{\text{grand total}}$$

and the **chi-square test statistic**, denoted by $\chi^2$ is defined as

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
The $\chi^2$ Statistic

Let $E_{ij}$ and $O_{ij}$ represent the expected frequencies and observed frequencies, respectively, in the $ith$ row and $jth$ column. Then

$$E_{ij} = \text{expected number of cases under } H_0$$
$$= \frac{(\text{column total}) \times (\text{row total})}{\text{grand total}}$$

and the **chi-square test statistic**, denoted by $\chi^2$ is defined as

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where $r$ is the total number of rows and $c$ is the total number of columns.
The $\chi^2$ Statistic

Using the previous example, the solution for the $\chi^2$ test statistic is as follows.

$$\chi^2 = \frac{(28 - 25)^2}{25} + \frac{(36 - 40)^2}{40} + \frac{(11 - 10)^2}{10}$$

$$+ \frac{(22 - 25)^2}{25} + \frac{(44 - 40)^2}{40} + \frac{(9 - 10)^2}{10}$$

$$= \frac{9}{25} + \frac{16}{40} + \frac{1}{10}$$

$$+ \frac{9}{25} + \frac{16}{40} + \frac{1}{10}$$

$$= \frac{1}{25} + \frac{1}{40} + \frac{1}{10}$$

$$= 1.72$$
Using the previous example, the solution for the $\chi^2$ test statistic is as follows.

$$\chi^2 = \frac{(28 - 25)^2}{25} + \frac{(36 - 40)^2}{40} + \frac{(11 - 10)^2}{10}$$

$$+ \frac{(22 - 25)^2}{25} + \frac{(44 - 40)^2}{40} + \frac{(9 - 10)^2}{10}$$

$$= 0.36 + 0.40 + 0.10 + 0.36 + 0.40 + 0.10$$
The \( \chi^2 \) Statistic

Using the previous example, the solution for the \( \chi^2 \) test statistic is as follows.

\[
\chi^2 = \frac{(28 - 25)^2}{25} + \frac{(36 - 40)^2}{40} + \frac{(11 - 10)^2}{10} + \frac{(22 - 25)^2}{25} + \frac{(44 - 40)^2}{40} + \frac{(9 - 10)^2}{10}
\]

\[
= 0.36 + 0.40 + 0.10 + 0.36 + 0.40 + 0.10
\]

\[
= 1.72
\]
The test for independence is used to determine whether two variables are related or not.
The test for independence is used to determine whether two variables are related or not.

Procedure:
1. State the null and alternative hypothesis.
   - \( H_0 \): The two variables are independent.
   - \( H_a \): The two variables are not independent.
The test for independence is used to determine whether two variables are related or not.

Procedure:

1. State the null and alternative hypothesis.
   
   \( \text{Ho:} \) The two variables are independent.
   
   \( \text{Ha:} \) The two variables are not independent.

2. Choose the level of significance and solve for \( \chi^2_{\alpha,(r-1)(c-1)} \). Here, \( (r - 1)(c - 1) \) is the degree of freedom.
The test for independence is used to determine whether two variables are related or not.

Procedure:
1. State the null and alternative hypothesis.
   - **Ho**: The two variables are independent.
   - **Ha**: The two variables are not independent.
2. Choose the level of significance and solve for $\chi^2_{\alpha,(r-1)(c-1)}$. Here, $(r-1)(c-1)$ is the degree of freedom.
3. Compute the $\chi^2$ test statistic.
The **test for independence** is used to determine whether two variables are related or not.

**Procedure:**

1. State the null and alternative hypothesis.
   - Ho: The two variables are independent.
   - Ha: The two variables are not independent.

2. Choose the level of significance and solve for $\chi^2_{\alpha,(r-1)(c-1)}$. Here, $(r - 1)(c - 1)$ is the degree of freedom.

3. Compute the $\chi^2$ test statistic.

4. Decision Rule: Reject Ho if $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$. 
EXAMPLE: Using the previous example, at 0.05 level of significance, test the hypothesis that the two variables are not independent.
**EXAMPLE:** An analysis of the medical records of a certain corporation reveals the following information about 732 of its employees who applied for coverage under the company’s extended health-coverage plans:

<table>
<thead>
<tr>
<th></th>
<th>Heavy Smoker and Drinker</th>
<th>Heavy Smoker and Nondrinker</th>
<th>Nonsmoker but Heavy Drinker</th>
<th>Nonsmoker and Nondrinker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>172</td>
<td>108</td>
<td>124</td>
<td>16</td>
</tr>
<tr>
<td>Female</td>
<td>84</td>
<td>146</td>
<td>62</td>
<td>20</td>
</tr>
</tbody>
</table>

Using a 5% level of significance, test the null hypothesis that the sex of the employee is independent of whether or not the employee is a heavy smoker or a drinker.
A researcher wants to test whether a person’s music preference is related to his intelligence as measured by IQ. She then takes a random sample and for each subject, determine his/her music preference, and classify his/her IQ into different categories (high, medium, low). The observed frequencies are presented in a contingency table shown below.

<table>
<thead>
<tr>
<th></th>
<th>Classical</th>
<th>Pop</th>
<th>Rock</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>40</td>
<td>47</td>
<td>83</td>
<td>170</td>
</tr>
<tr>
<td>Medium</td>
<td>26</td>
<td>59</td>
<td>104</td>
<td>189</td>
</tr>
<tr>
<td>Low</td>
<td>17</td>
<td>25</td>
<td>79</td>
<td>121</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>83</strong></td>
<td><strong>131</strong></td>
<td><strong>266</strong></td>
<td><strong>480</strong></td>
</tr>
</tbody>
</table>
A researcher wants to test whether a person’s music preference is related to his intelligence as measured by IQ. She then takes a random sample and for each subject, determine his/her music preference, and classify his/her IQ into different categories (high, medium, low). The observed frequencies are presented in a contingency table shown below.

<table>
<thead>
<tr>
<th>Music Preference</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical</td>
<td>40</td>
<td>26</td>
<td>17</td>
<td>83</td>
</tr>
<tr>
<td>Pop</td>
<td>47</td>
<td>59</td>
<td>25</td>
<td>131</td>
</tr>
<tr>
<td>Rock</td>
<td>83</td>
<td>104</td>
<td>79</td>
<td>266</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>170</strong></td>
<td><strong>189</strong></td>
<td><strong>121</strong></td>
<td><strong>480</strong></td>
</tr>
</tbody>
</table>

Use a 5% level of significance to test the hypothesis that the IQ level and the music preference of an individual are independent.
Test for Independence

Solution:

1. a. Ho: Music preference and IQ are independent.
   b. Ha: Music preference and IQ are not independent.

2. At $\alpha = 0.05$, $\chi^2 = 9.488$.

3. Computation:

<table>
<thead>
<tr>
<th>Music Preference</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical</td>
<td>40 (29.4)</td>
<td>26 (32.7)</td>
<td>17 (20.9)</td>
<td>83</td>
</tr>
<tr>
<td>Pop</td>
<td>47 (46.4)</td>
<td>59 (51.6)</td>
<td>25 (33.0)</td>
<td>131</td>
</tr>
<tr>
<td>Rock</td>
<td>83 (94.2)</td>
<td>104 (104.7)</td>
<td>79 (67.1)</td>
<td>266</td>
</tr>
<tr>
<td>Total</td>
<td>170</td>
<td>189</td>
<td>121</td>
<td>480</td>
</tr>
</tbody>
</table>

$$\chi^2 = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 12.38.$$ 

4. Decision: Since 12.38 > 9.488, reject Ho. There is sufficient evidence at 0.05 level of significance that music preference and intelligence are not independent.
Four hundred computer science majors were asked to indicate what they use their personal computers for. The results of the survey are shown below.

<table>
<thead>
<tr>
<th></th>
<th>Word Processing</th>
<th>Games</th>
<th>Spreadsheet Analysis</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshman</td>
<td>24</td>
<td>31</td>
<td>18</td>
<td>25</td>
</tr>
<tr>
<td>Sophomore</td>
<td>27</td>
<td>29</td>
<td>16</td>
<td>27</td>
</tr>
<tr>
<td>Junior</td>
<td>25</td>
<td>32</td>
<td>15</td>
<td>28</td>
</tr>
<tr>
<td>Senior</td>
<td>26</td>
<td>34</td>
<td>17</td>
<td>26</td>
</tr>
</tbody>
</table>

Using a 5% level of significance, test the null hypothesis that the principal use of a computer by a computer science major is independent of the year level standing of the individual.
Remarks:

1. The test is valid if at least 80% of the cells have expected frequencies of at least 5 and no cell has an expected frequency of less than 1.
Remarks:

1. The test is valid if at least 80% of the cells have expected frequencies of at least 5 and no cell has an expected frequency of less than 1.

2. If many expected frequencies are very small, researchers commonly combine categories of variables to obtain a table having larger cell frequencies. Generally, one should not pool categories unless there is a natural way to combine them.

\[ \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{|O_{ij} - E_{ij}| - 0.5}{E_{ij}} \right)^2 \]
Remarks:

1. The test is valid if at least 80% of the cells have expected frequencies of at least 5 and no cell has an expected frequency of less than 1.

2. If many expected frequencies are very small, researchers commonly combine categories of variables to obtain a table having larger cell frequencies. Generally, one should not pool categories unless there is a natural way to combine them.

3. For a $2 \times 2$ contingency table, a correction called Yate’s correction for continuity is applied. The formula then becomes

$$
\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}.
$$
**Pizza Preferences.** A pizza shop owner wishes to determine if the type of pizza a person orders is related to his/her age. The data obtained from the sample are shown here. At $\alpha = 0.10$, is the age of the purchaser related to the type of pizza ordered?

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Hawaiian</th>
<th>Pepperoni</th>
<th>Mushroom</th>
<th>5-Cheese</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-19</td>
<td>12</td>
<td>21</td>
<td>39</td>
<td>71</td>
</tr>
<tr>
<td>20-29</td>
<td>18</td>
<td>76</td>
<td>52</td>
<td>87</td>
</tr>
<tr>
<td>30-39</td>
<td>24</td>
<td>50</td>
<td>40</td>
<td>47</td>
</tr>
<tr>
<td>40-49</td>
<td>52</td>
<td>30</td>
<td>12</td>
<td>28</td>
</tr>
</tbody>
</table>