Thursday, April 28, 2016

I can use row relative frequencies or column relative frequencies to informally determine whether there is an association between two categorical variables.

Suppose a random group of people are surveyed about their use of smartphones. The results of the survey are summarized in the tables below.

<table>
<thead>
<tr>
<th>Smartphone Use and Gender</th>
<th>Use a Smartphone</th>
<th>Do Not Use a Smartphone</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>30</td>
<td>10</td>
<td>40</td>
</tr>
<tr>
<td>Female</td>
<td>45</td>
<td>15</td>
<td>60</td>
</tr>
<tr>
<td>Total</td>
<td>75</td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smartphone Use and Age</th>
<th>Use a Smartphone</th>
<th>Do Not Use a Smartphone</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 40 Years of Age</td>
<td>45</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>40 Years of Age or Older</td>
<td>30</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>75</td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

What is the proportion of males that do not use a Smartphone?
\[ \frac{10}{40} = 0.25 = 25\% \]

What is the proportion of Smartphone users who are under the age of 40?
\[ \frac{45}{75} = 0.60 = 60\% \]
Homework

Pregnant women often undergo ultrasound tests to monitor their babies’ health. These tests can also be used to predict the gender of the babies, but these predictions are not always accurate. Data on the gender predicted by ultrasound and the actual gender of the baby for 1,000 babies are summarized in the two-way table below.

<table>
<thead>
<tr>
<th>Actual Gender</th>
<th>Predicted Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>432</td>
</tr>
<tr>
<td>Male</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>390</td>
</tr>
</tbody>
</table>

7. Write a sentence explaining the meaning of the frequency 130 in this table.

Use the table provided above to calculate the following relative frequencies.

8. What is the proportion of babies who were predicted to be male but were actually female?
   \[ \frac{48}{1000} = 0.048 = 4.8\% \]

9. What is the proportion of incorrect ultrasound gender predictions?
   \[ \frac{178}{1000} = 0.178 = 17.8\% \]

10. For babies predicted to be female, what proportion of the predictions were correct?
    \[ \frac{432}{48} = 0.769 = 76.9\% \]

11. For babies predicted to be male, what proportion of the predictions were correct?
    \[ \frac{390}{438} = 0.890 = 89\% \]
Example 2

Suppose a sample of 400 participants (teachers and students) was randomly selected from the middle schools and high schools in a large city. These participants responded to the following question:

Which type of movie do you prefer to watch?

1. Action (The Avengers, Man of Steel, etc.)
2. Drama (42 (The Jackie Robinson Story), The Great Gatsby, etc.)
3. Science Fiction (Star Trek into Darkness, World War Z, etc.)
4. Comedy (Monsters University, Despicable Me 2, etc.)

Movie preference and status (teacher or student) were recorded for each participant.

Exercises 1–7

1. Two variables were recorded. Are these variables categorical or numerical?

   Movie Preference
   Status (teacher/student) > Categorical

2. The results of the survey are summarized in the table below.

<table>
<thead>
<tr>
<th></th>
<th>Action</th>
<th>Drama</th>
<th>Science Fiction</th>
<th>Comedy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>120</td>
<td>60</td>
<td>30</td>
<td>90</td>
<td>300</td>
</tr>
<tr>
<td>Teacher</td>
<td>40</td>
<td>20</td>
<td>10</td>
<td>30</td>
<td>100</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>160</td>
<td>80</td>
<td>40</td>
<td>120</td>
<td>400</td>
</tr>
</tbody>
</table>

a. What proportion of participants who are teachers prefer action movies?
   \[
   \frac{40}{100} = 0.40 = 40\% 
   \]

b. What proportion of participants who are teachers prefer drama movies?
   \[
   \frac{20}{100} = 0.20 = 20\% 
   \]

c. What proportion of participants who are teachers prefer science fiction movies?
   \[
   \frac{10}{100} = 0.10 = 10\% 
   \]

d. What proportion of participants who are teachers prefer comedy movies?
   \[
   \frac{30}{100} = 0.30 = 30\% 
   \]
The answers to Exercise 2 are called row relative frequencies. Notice that you divided each cell frequency in the Teacher row by the total for that row. Below is a blank relative frequency table.

**Table of Row Relative Frequencies**

<table>
<thead>
<tr>
<th>Movie Preference</th>
<th>Action</th>
<th>Drama</th>
<th>Science Fiction</th>
<th>Comedy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Student</strong></td>
<td>0.40 = 40%</td>
<td>0.20 = 20%</td>
<td>0.10 = 10%</td>
<td>0.30 = 30%</td>
</tr>
<tr>
<td><strong>Teacher</strong></td>
<td>(a) 0.40 = 40%</td>
<td>(b) 0.20 = 20%</td>
<td>(c) 0.10 = 10%</td>
<td>(d) 0.30 = 30%</td>
</tr>
</tbody>
</table>

Write your answers from Exercise 2 in the indicated cells in the table above.

3. Find the row relative frequencies for the Student row. Write your answers in the table above.
   a. What proportion of participants who are students prefer action movies?
   b. What proportion of participants who are students prefer drama movies?
   c. What proportion of participants who are students prefer science fiction movies?
   d. What proportion of participants who are students prefer comedy movies?

4. Is a participant’s status (i.e., teacher or student) related to what type of movie he would prefer to watch? Why or why not? Discuss this with your group.

   **No, teachers are just as likely to prefer each type of movie as students are.**

5. What does it mean when we say that there is no association between two variables? Discuss this with your group.

   **No association means that knowing the value of one variable does not tell anything about the value of the other variable.**

6. Notice that the row relative frequencies for each movie type are the same for both the Teacher and Student rows. When this happens, we say that the two variables, movie preference and status (student or teacher), are not associated. Another way of thinking about this is to say that knowing if a participant is a teacher (or a student) provides no information about his movie preference.

   What does it mean if row relative frequencies are not the same for all rows of a two-way table?
7. You can also evaluate whether two variables are associated by looking at column relative frequencies instead of row relative frequencies. A column relative frequency is a cell frequency divided by the corresponding column total. For example, the column relative frequency for the Student/Action cell is \( \frac{120}{160} = 0.75 \).

a. Calculate the other column relative frequencies, and write them in the table below.

<table>
<thead>
<tr>
<th>Movie Preference</th>
<th>Action</th>
<th>Drama</th>
<th>Science Fiction</th>
<th>Comedy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>( \frac{120}{160} = 0.75 ) = 75%</td>
<td>( \frac{40}{160} = 0.25 = 25% )</td>
<td>( \frac{30}{160} = 0.1875 = 18.75% )</td>
<td>( \frac{30}{160} = 0.1875 = 18.75% )</td>
</tr>
<tr>
<td>Teacher</td>
<td>( \frac{40}{160} = 0.25 = 25% )</td>
<td>( \frac{20}{160} = 0.125 = 12.5% )</td>
<td>( \frac{40}{160} = 0.25 = 25% )</td>
<td>( \frac{40}{160} = 0.25 = 25% )</td>
</tr>
</tbody>
</table>

b. What do you notice about the column relative frequencies for the four columns?

c. What would you conclude about association based on the column relative frequencies?
Example 3

In the survey described in Example 2, gender for each of the 400 participants was also recorded. Some results of the survey are given below:

- 160 participants preferred action movies.
- 80 participants preferred drama movies.
- 40 participants preferred science fiction movies.
- 240 participants were females.
- 70 female participants preferred drama movies.
- 32 male participants preferred science fiction movies.
- 60 female participants preferred action movies.

Exercises 8–15

Use the results from Example 3 to answer the following questions. Be sure to discuss these questions with your group members.

8. Complete the two-way frequency table that summarizes the data on movie preference and gender.

<table>
<thead>
<tr>
<th>Movie Preference</th>
<th>Action</th>
<th>Drama</th>
<th>Science Fiction</th>
<th>Comedy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>160</td>
<td>70</td>
<td>80</td>
<td>90</td>
<td>240</td>
</tr>
<tr>
<td>Male</td>
<td>100</td>
<td>20</td>
<td>40</td>
<td>50</td>
<td>160</td>
</tr>
<tr>
<td>Total</td>
<td>160</td>
<td>90</td>
<td>120</td>
<td>140</td>
<td>480</td>
</tr>
</tbody>
</table>

9. What proportion of the participants are female?

\[
\frac{240}{480} = 0.5 = 50\%
\]

10. If there was no association between gender and movie preference, should you expect more females than males or fewer females than males to prefer action movies? Explain.

11. Make a table of row relative frequencies of each movie type for the Male row and the Female row. Refer to Exercises 2–4 to review how to complete the table below.

<table>
<thead>
<tr>
<th>Movie Preference</th>
<th>Action</th>
<th>Drama</th>
<th>Science Fiction</th>
<th>Comedy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Suppose that you randomly pick 1 of the 400 participants. Use the table of row relative frequencies on the previous page to answer the following questions.

12. If you had to predict what type of movie this person chose, what would you predict? Explain why you made this choice.

13. If you know that the randomly selected participant is female, would you predict that her favorite type of movie is action? If not, what would you predict, and why?

14. If knowing the value of one of the variables provides information about the value of the other variable, then there is an association between the two variables. Is there an association between the variables gender and movie preference? Explain.

15. What can be said when two variables are associated? Read the following sentences. Decide if each sentence is a correct statement based upon the survey data. If it is not correct, explain why not.

a. More females than males participated in the survey.

b. Males tend to prefer action and science fiction movies.

c. Being female causes one to prefer drama movies.
Homework

Lesson #14 Problem Set