Lesson 14: Association Between Categorical Variables

Classwork

Example 1

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

Smartphone Use and Gender

	Use Smartphone	Do not Use Smartphone	Total
Male	30	10	40
Female	45	15	60
Total	75	25	100

Smartphone Use and Age

	Use Smartphone	Do not Use Smartphone	Total
Under 40 years of age	45	5	50
40 years of age or older	30	20	50
Total	75	25	100

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 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/student) were recorded for each participant.

Exercises 1–7

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

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

	Movie Preference				
	Action	Drama	Science Fiction	Comedy	Total
Student	120	60	30	90	300
Teacher	40	20	10	30	100
Total	160	80	40	120	400

   a. What proportion of participants who are teachers would prefer "action" movies?

   b. What proportion of participants who are teachers would prefer "drama" movies?

   c. What proportion of participants who are teachers would prefer "science fiction" movies?

   d. What proportion of participants who are teachers would prefer "comedy" movies?

The answers to Exercise 2 are called row relative frequencies. Notice that you divided each cell frequency in the teacher row by the row total for that row. Below is a blank relative frequency table.

Table of Row Relative Frequencies

	Movie Preference			
	Action	Drama	Science Fiction	Comedy
Student				
Teacher	a)	b)	c)	d)

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 would prefer "action" movies?
   b. What proportion of participants who are students would prefer "drama" movies?
   c. What proportion of participants who are students would prefer "science fiction" movies?
   d. What proportion of participants who are students would prefer "comedy" movies?

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

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

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/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 or her 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 of Row Relative Frequencies

	Movie Preference			
	Action	Drama	Science Fiction	Comedy
Student				
Teacher				

   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
- 78 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 summaries the data on movie preference and gender.

	Movie Preference				
	Action	Drama	Science Fiction	Comedy	Total
Student					
Teacher					
Total					

9. What proportion of the participants is female?

10. If there were 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 through 4 to review how to complete the table below.

	Movie Preference			
	Action	Drama	Science Fiction	Comedy
Student				
Teacher				

Suppose that you randomly pick 1 of the 400 participants. Use the table of row relative frequencies above 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 was 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. So what can be said when two variables are associated? Read the following sentences. Decide if the 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 moves.

    c. Being female causes one to prefer drama movies.

Lesson Summary

- Saying that two variables ARE NOT associated means that knowing the value of one variable provides no information about the value of the other variable.
- Saying that two variables ARE associated means that knowing the value of one variable provides information about the value of the other variable.
- To determine if two variables are associated, calculate row relative frequencies. If the row relative frequencies are about the same for all of the rows, it is reasonable to say that there is no association between the two variables that define the table.
- Another way to decide if there is an association between two categorical variables is to calculate column relative frequencies. If the column relative frequencies are about the same for all of the rows, it is reasonable to say that there is no association between the two variables that define the table.
- If the row relative frequencies are quite different for some of the rows, it is reasonable to say that there is an association between the two variables that define the table.

Problem Set

A sample of 200 middle school students was randomly selected from the middle schools in a large city. Answers to several survey questions were recorded for each student. The tables below summarize the results of the survey.

For each table, calculate the row relative frequencies for the female row and for the male row. Write the row relative frequencies beside the corresponding frequencies in each table below.

1. This table summarizes the results of the survey data for the two variables, gender and which sport the students prefer to play. Is there an association between gender and which sport the students prefer to play? Explain.

		Sport				
		Football	Basketball	Volleyball	Soccer	Total
Gender	Female	2	29	28	38	97
	Male	35	26	8	24	103
	Total	37	65	36	62	200

2. This table summarizes the results of the survey data for the two variables, gender and the students' T-shirt sizes. Is there an association between gender and T-Shirt size? Explain.

		School T-Shirt Sizes				
		Small	Medium	Large	X-Large	Total
Gender	Female	47	35	13	2	97
	Male	11	41	42	9	103
	Total	58	76	55	11	200

3. This table summarizes the results of the survey data for the two variables, gender and favorite type of music. Is there an association between gender and favorite type of music? Explain

		Favorite Type of Music				
		Pop	Hip Hop	Alternative	Country	Total
Gender	Female	35	28	11	23	97
	Male	37	30	13	23	103
	Total	72	58	24	46	200