



Lesson 14: Association Between Categorical Variables

Student Outcomes

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

Lesson Notes

In this lesson, students consider whether conclusions are reasonable based on a two-way table. Students think about what it means to have similar row relative frequencies for all rows in a table or to have similar column relative frequencies for all columns in a table. They also consider what it means to have row relative frequencies that are not similar for all rows in the table. Students study the meaning of association between two categorical variables. For example, students are asked to predict the favorite movies of a person whose gender is not known, and then they are asked if knowing that the person is female would change their prediction. This lesson provides a foundation for more detailed coverage of association in Algebra I.

This lesson is designed to have students work in groups of 2–3. Prior to class, prepare the list of students in each group, and arrange desks or tables to allow for group work.

Classwork

Example 1 (2–3 minutes)

Let students compare the two tables. Use the following questions to lead into a discussion about association. Some students may calculate row relative frequencies to justify their answers.

- What are the variables being recorded?
 - Smartphone use, gender, and age.*
- What can you conclude about the table “Smartphone Use and Gender”?
 - Answers will vary. Possible responses: 75% of those surveyed use smartphones. The percentage is the same for males and females, which is 75%.*
- What can you conclude about the table “Smartphone Use and Age”?
 - Answers will vary. Possible responses: 75% of those surveyed use smartphones. However, a larger percentage of those under 40 years old use a smartphone (90%) compared to the percentage of those 40 or older (60%).*
- If you knew that someone was 20 years old, would you expect that person to use a smartphone? Explain.
 - Yes. Possible response: One would expect a young person to use a smartphone based on the results in the table because 90% of people under 40 use smartphones.*

Scaffolding:

Some ELL students may need to learn the word *smartphone*. Consider providing a visual aid.

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 (2 minutes)

Read the beginning of Example 2 to the class. Ask students:

- What are the variables being recorded?
 - *Movie preference and teacher or student status.*

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 (12–15 minutes)

Have students work in small groups. Give groups 1–2 minutes to answer Exercise 1, and then confirm their answers as a class.

Students should read the results of the survey. Remind them that a row relative frequency is the cell frequency divided by the corresponding row total. Allow groups to answer Exercises 2–5, and then confirm answers as a class. Give groups adequate time to discuss Exercises 6 and 7, and then discuss as a class.

Exercises 1–7

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

Both variables are categorical.

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?

$$\frac{40}{100} = 0.40$$

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

$$\frac{20}{100} = 0.20$$

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

$$\frac{10}{100} = 0.10$$

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

$$\frac{30}{100} = 0.30$$

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	0.40	0.20	0.10	0.30
Teacher	(a) 0.40	(b) 0.20	(c) 0.10	(d) 0.30

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.

- What proportion of participants who are students prefers action movies?
- What proportion of participants who are students prefers drama movies?
- What proportion of participants who are students prefers science-fiction movies?
- What proportion of participants who are students prefers comedy movies?

See table above.

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.

No, because teachers are just as likely to prefer each movie type as students are, according to the row relative frequencies.

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

Answers will vary. No association means that knowing the value of one variable does not tell you 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/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?

It means that there is an association or tendency between the two variables.

Scaffolding:

For ELL students, the concept of *no association* may be difficult. Although, for students working in groups, consider explicitly modeling the thinking employed in Exercise 6.

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 Column Relative Frequencies

	Movie Preference			
	Action	Drama	Science-Fiction	Comedy
Student	0.75	0.75	0.75	0.75
Teacher	0.25	0.25	0.25	0.25

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

The column relative frequencies are equal for all four columns.

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

Because the column relative frequencies are the same for all four columns, we would conclude that there is no association between movie preference and status.

In this part of the lesson, students should understand that there is a mathematical way to determine if there is no association between two categorical variables. Students can look to see if the row relative frequencies are the same (or approximately the same) for each row in the table. Discuss the mathematical method for determining if there is no association between two categorical variables.

Example 3 (2 minutes)

Introduce the data in Example 3. Give students a moment to read the results. Take a quick movie-preference poll in class. Ask the following:

- Who likes action movies?
- Do you think movie preference is equal among males and females?
 - *Answers will vary. Encourage students to explain why they think the preferences might be equal or different.*

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 8–11 (8–10 minutes)

Let students work with their groups on Exercises 8–10, and then confirm answers as a class. Give students 2–3 minutes to complete Exercise 11.

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.

	Movie Preference				Total
	Action	Drama	Science-Fiction	Comedy	
Female	60	78	8	94	240
Male	100	2	32	26	160
Total	160	80	40	120	400

9. What proportion of the participants is female?

$$\frac{240}{400} = 0.60$$

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.

*If there were no association between gender and movie preference, then I would expect **MORE** females than males to prefer action movies just because there are more females in the sample. However, if there were an association between gender and movie preference, then I would expect either fewer females than males who prefer action movies or delete considerably more females than males who prefer action movies.*

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.

	Movie Preference			
	Action	Drama	Science-Fiction	Comedy
Female	0.25	0.325	0.033	0.392
Male	0.625	0.0125	0.2	0.1625

Exercises 12–15 (12–15 minutes)

Read the next instructions. Make sure that students understand that 1 of the 400 participants is randomly selected. Allow groups about 5 minutes to discuss and answer Exercises 12 and 13.

Then, discuss as a class what association means. Allow students 3 minutes to answer Exercise 14.

Allow 5 minutes for groups to discuss whether the statements in Exercise 15 are correct. Call on groups to share their answers.

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.

The participant likely prefers action movies because the largest proportion of participants preferred action movies.

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?

No, a female participant is more likely to prefer comedy since it has the greatest row relative frequency in the female row.

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.

Yes. The row relative frequencies are not the same (not even close) in each row in the table.

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.

Correct

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

Correct

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

Incorrect—association does not imply a cause-and-effect relationship.

Closing (3 minutes)

Read through the Lesson Summary with students.

If time allows, have students refer back to Example 1 and calculate row relative frequencies for each table to determine if there is evidence of association between variables.

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 columns, 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.

Exit Ticket (5 minutes)

Name _____

Date _____

Lesson 14: Association Between Categorical Variables

Exit Ticket

A random sample of 100 eighth-grade students is asked to record two variables, whether they have a television in their bedroom and if they passed or failed their last math test. The results of the survey are summarized below.

- 55 students have a television in their bedroom.
- 35 students do not have a television in their bedroom and passed their last math test.
- 25 students have a television and failed their last math test.
- 35 students failed their last math test.

1. Complete the two-way table.

	Pass	Fail	Total
Television in Bedroom			
No Television in Bedroom			
Total			

2. Calculate the row relative frequencies and enter the values in the table above. Round to the nearest thousandth.
3. Is there evidence of association between the variables? If so, does this imply there is a cause-and-effect relationship? Explain.

Exit Ticket Sample Solutions

A random sample of 100 eighth grade students is asked to record two variables, whether they have a television in their bedroom and if they passed or failed their last math test. The results of the survey are summarized below.

- 55 students have a television in their bedroom.
- 35 students do not have a television in their bedroom and passed their last math test.
- 25 students have a television and failed their last math test.
- 35 students failed their last math test.

1. Complete the two-way table.

	Pass	Fail	Total
Television in Bedroom	30 ≈ 0.545	25 ≈ 0.455	55 1.000
No Television in Bedroom	35 ≈ 0.778	10 ≈ 0.222	45 1.000
Total	65 ≈ 0.650	35 ≈ 0.350	100 1.000

2. Calculate the row relative frequencies and enter the values in the table above. Round to the nearest thousandth.

Row relative frequencies are displayed in the table above.

3. Is there evidence of association between the variables? If so, does this imply there is a cause-and-effect relationship? Explain.

Yes, there is evidence of association between the variables because the relative frequencies are different among the rows. However, this does not necessarily imply a cause-and-effect relationship. The fact that a student has a television in their room does not cause the student to fail a test. Rather, it may be that the student is spending more time watching television or playing video games instead of studying.

Problem Set Sample Solutions

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				Total
		Football	Basketball	Volleyball	Soccer	
Gender	Female	2 ≈ 0.021	29 ≈ 0.299	28 ≈ 0.289	38 ≈ 0.392	97
	Male	35 ≈ 0.340	36 ≈ 0.350	8 ≈ 0.078	24 ≈ 0.233	103
	Total	37	65	36	62	200

Yes, there appears to be an association between gender and sports preference. The row relative frequencies are not the same for the male and the female rows, as shown in the table above.

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				Total
		Small	Medium	Large	X-Large	
Gender	Female	47 ≈ 0.484	35 ≈ 0.361	13 ≈ 0.134	2 ≈ 0.021	97
	Male	11 ≈ 0.107	41 ≈ 0.398	42 ≈ 0.408	9 ≈ 0.087	103
	Total	58	76	55	11	200

Yes, there appears to be an association between gender and T-shirt size. The row relative frequencies are not the same for the male and the female rows, as shown in the table above.

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				Total
		Pop	Hip Hop	Alternative	Country	
Gender	Female	35 ≈ 0.361	28 ≈ 0.289	11 ≈ 0.113	23 ≈ 0.237	97
	Male	37 ≈ 0.359	30 ≈ 0.291	13 ≈ 0.126	23 ≈ 0.223	103
	Total	72	58	24	46	200

No, there may not be an association between gender and favorite type of music. The row relative frequencies are about the same for the male and female rows, as shown in the table above.