

**Association Between Categorical Variables****Student Objective**

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

**Classwork****Example 1**

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?

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

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

**Exercises 1–6**

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

What is the relative frequency of participants who are teachers who would prefer:

- “action” movies?
- “drama” movies?
- “science fiction” movies?
- “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.

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.

- Find the row relative frequencies for the “student” row. Write your answers in the table above.
- Is a participant’s status (teacher or student) related to what type of movie he or she would prefer to watch? Why or why not?

- What does it mean when we say that there is “no association” between two variables? 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, there is “no association” between two variables. 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?

6. 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				
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?

### 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.

**Homework Homework Homework Homework Homework**

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 of the three (3) tables**, calculate the row relative frequencies for the 7<sup>th</sup> Grade row and for the 8<sup>th</sup> Grade row to the nearest hundredths. Write the row relative frequencies beside the corresponding frequencies in each table below.

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

		Sport				Total
		Football	Basketball	Volleyball	Soccer	
Grade	7 <sup>th</sup> Grade	2	29	28	38	97
	8 <sup>th</sup> Grade	35	36	8	24	103
Total		37	65	36	62	200

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

		School T-Shirt Sizes				Total
		Small	Medium	Large	X-Large	
Grade	7 <sup>th</sup> Grade	47	35	13	2	97
	8 <sup>th</sup> Grade	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, grade and favorite type of music. Is there an association between grade and favorite type of music? Explain

		Favorite Type of Music				Total
		Pop	Hip Hop	Alternative	Country	
Grade	7 <sup>th</sup> Grade	35	28	11	23	97
	8 <sup>th</sup> Grade	37	30	13	23	103
Total		72	58	24	46	200