Lesson 6: Summarizing Bivariate Categorical Data

So far, we have been working with **univariate data** – data involving ONE variable.

Bivariate categorical data results from collecting data on two categorical variables. In this lesson, you will see examples involving categorical data collected from two survey questions.

Classwork

Superhero Powers

Superheroes have been popular characters in movies, television, books, and comics for many generations. Superman was one of the most popular series in the 1950's while Batman was a top rated series in the 1960's. Each of these characters was also popular in movies released from 1990 to 2013. Other notable characters portrayed in movies over the last several decades include Captain America, She-Ra, and the Fantastic Four. What is special about a superhero? Is there a special superhero power that makes these characters particularly popular?

High school students in the United States were invited to complete an online survey in 2010. Part of the survey included questions about superhero powers. *More than 1,000 students responded to this survey that included a question about a student's most favorite superhero power. 450 of the completed surveys were randomly selected to include in the study. A rather confusing breakdown of the data by gender was compiled from the 450 surveys:*

- 100 students indicated their favorite power was "to fly." 49 of those students were females.
- 131 students selected the power to "freeze time" as their favorite power. 71 of those students were males.
- 75 students selected "invisibility" as their favorite power. 48 of those students were females.
- 26 students indicated "super strength" as their favorite power. 25 of those students were males.
- And finally, 118 students indicated "telepathy" as their favorite power. 70 of those students were females.

The data in Example 1 prompted students in a mathematics class to pose the statistical question, "**Do high school males have different preferences for superhero powers than high school females?"** Answering this statistical question involves collecting data as well as anticipating variability in the data collected.

The data consist of two responses from each student completing a survey. The first response indicates a student's gender, and the second response indicates the student's favorite superpower. For example, data collected from one student was "male" and "to fly." The data are **bivariate categorical data**.

The first step in analyzing the statistical question posed by the students in their mathematics class is to organize this data in a two-way frequency table.

A Statistical Study Involving a Two-Way Frequency Table

A two-way frequency table that can be used to organize the categorical data is shown below

1. Complete the table below by determining a frequency count for each cell based on the summarized data (included below for convenience!

   450 of the completed surveys were randomly selected to include in the study.

   - 100 students indicated their favorite power was "to fly." 49 of those students were females.
   - 131 students selected the power to "freeze time" as their favorite power. 71 of those students were males.
   - 75 students selected "invisibility" as their favorite power. 48 of those students were females.
   - 26 students indicated "super strength" as their favorite power. 25 of those students were males.
   - And finally, 118 students indicated "telepathy" as their favorite power. 70 of those students were females.

	To Fly	Freeze time	Invisibility	Super Strength	Telepathy	Total
Females						
Males						
Total						

Extending the Frequency Table to a Relative Frequency Table

Determining the number of students in each cell presents the first step in organizing bivariate categorical data. Another way of analyzing the data in the table is to calculate the *relative frequency* for each cell. Relative frequencies relate each frequency count to the total number of observations. For each cell in this table, the **relative frequency** of a cell is found by dividing the frequency of that cell by the total number of responses.

2. Calculate the remaining relative frequencies in the table below. Write the value in the table as a decimal rounded to the nearest thousandth.

	To Fly	Freeze Time	Invisibility	Super Strength	Telepathy	Total
Females	$\frac{49}{450} \approx 0.109$					$\frac{228}{450} \approx 0.507$
Males			$\frac{27}{450} \approx 0.060$			
Total		$\frac{131}{450} \approx 0.291$			$\frac{118}{450} \approx 0.262$	

3.

4. What is the *joint relative frequency* for females and "invisibility"? Interpret the meaning of this value.

5. What is the *marginal relative frequency* (out of the total) for "freeze time"? Interpret the meaning of this value.

6. What is the difference in the joint relative frequencies for males and for females who selected "to fly" as their favorite superpower?

7. Is there a noticeable difference between the genders and their favorite superpowers?

Example 4: Interpreting Data

Interest in superheroes continues at Rufus King High School. The students who analyzed the data in the previous lesson decided to create a comic strip for the school website that involves a superhero. They thought the summaries developed from the data would be helpful in designing the comic strip.

Only one power will be given to the superhero. A debate arose as to what power the school's superhero would possess.

Scott initially indicated that the character created should have "super strength" as the special power. This suggestion was not well received by the other students. In particular, Jill argued, "Well, if you don't want to ignore more than half of the readers, then I suggest 'telepathy' is the better power for our character."

Scott acknowledged that "super strength" was probably not the best choice based on the data. "The data indicate that 'freeze time' is the most popular power for a super hero," continued Scott. Jill, however, still did not agree with Scott that this was a good choice. She argued that "telepathy" was a better choice.

8. How do the data support Scott's claim? Why do you think he selected *freeze time* as the special power for the comic strip superhero?

9. How do the data support Jill's claim? Why do you think she selected *telepathy* as the special power for the comic strip superhero?

10. Of the two special powers *freeze time* and *telepathy*, select one and justify why you think it is a better choice based on the data.

Conditional Relative Frequencies

A **conditional relative frequency** compares a frequency count to the marginal total (the total for a row or column) that represents the condition of interest. For example, the condition of interest in the first row is females. The row conditional relative frequency of females responding "Invisibility" as the favorite superpower is 48/228 or approximately 0.211. This conditional relative frequency indicates that approximately 21.1% of females prefer "Invisibility" as their favorite superpower. Similarly, 27/222, or approximately 0.122 or 12.2%, of males prefer "Invisibility" as their favorite superpower.

11. Use the frequency counts from the table in Example 1 to calculate the missing row conditional relative frequencies. Round the answers to the nearest thousandth.

	To Fly	Freeze Time	Invisibility	Super Strength	Telepathy	Total
Females			$\frac{48}{228} \approx 0.211$			
Males	$\frac{51}{222} \approx 0.230$					$\frac{222}{222} = 1.000$
Total						

12. Suppose that a student is selected at random from those who completed the survey. What do you think is the gender of the student selected? What would you predict for this student's response to the superpower question?

13. Suppose that a student is selected at random from those who completed the survey. If the selected student is male, what do you think was his response to the selection of a favorite superpower? Explain your answer.

14. Suppose that a student is selected at random from those who completed the survey. If the selected student is female, what do you think was her response to the selection of a favorite superpower? Explain your answer.

Conditional Relative Frequencies from the other Perspective

15. Use the frequency counts from the table in Example 1 to calculate the missing COLUMN conditional relative frequencies. Round the answers to the nearest thousandth.

	To Fly	Freeze Time	Invisibility	Super Strength	Telepathy	Totals
Females						
Males						
Totals						

16. What different types of information does this table provide in comparison to the table you created in question 2 and question 11?

Possible Association Based on Conditional Relative Frequencies

Two categorical variables are ***associated*** if the row conditional relative frequencies (or column relative frequencies) are significantly different for the rows (or columns) of the table. For example, if the selection of superpower selected for females is significantly different than the selection of superpowers for males, then gender and superpower favorites are associated. This difference indicates that knowing the gender of a person in the sample indicates something about their superpower preference.

The evidence of an association is strongest when the conditional relative frequencies are quite different. If the conditional relative frequencies are nearly equal for all categories, then there is probably not an association between variables.

Examine the conditional relative frequencies in the two-way table of conditional relative frequencies you created in Exercise 1. Note that for each superpower, the conditional relative frequencies are different for females and males.

17. For what superpowers would you say that the conditional relative frequencies for females and males are very different?

18. For what superpowers are the conditional relative frequencies nearly equal for males and females?

19. Suppose a student is selected at random from the students who completed the survey. If you had to predict which superpower this student selected, would it be helpful to know the student's gender? Explain your answer.

20. Is there evidence of an association between gender and superpower selected? Explain why or why not.

21. What superpower would you recommend the students at Rufus King High School select for their superhero character? Justify your choice.

Lesson Summary:

- Categorical data are data that take on values that are categories rather than numbers. Examples include male or female for the categorical variable of gender, or the five superpower categories for the categorical variable of superpower qualities.
- A two-way frequency table is used to summarize bivariate categorical data.
- The number in a two-way frequency table at the intersection of a row and column of the response to two categorical variables represents a joint frequency.
- The total number of responses for each value of a categorical variable in the table represents the marginal frequency for that value.
- A relative frequency compares a frequency count to the total number of observations. It can be written as a decimal or percent. A two-way table summarizing the relative frequencies of each cell is called a relative frequency table.
- The marginal cells in a two-way relative frequency table are called the marginal relative frequencies, while the joint cells are called the joint relative frequencies.
- A conditional relative frequency compares a frequency count to the marginal total that represents the *condition* of interest.
- The differences in conditional relative frequencies are used to assess whether or not there is an association between two categorical variables.
- The greater the differences in the conditional relative frequencies, the stronger the evidence that an association exits.
- An observed association between two variables does not necessarily mean that there is a cause-and-effect relationship between the two variables.

PROBLEM SET: Bivariate Categorical Data – Association vs. Cause and Affect

Students were given the opportunity to prepare for a college placement test in mathematics by taking a review course. The students could be placed into one of two classes as a result of the placement test:

Math 50: A remedial course reviewing much of the mathematics learned in high school.

Math 100: An introductory college mathematics class extending concepts from high school.

Math 200: A more advanced mathematics class that goes much further and quicker through new concepts and ideas.

Not all students took advantage of the review course. The following results were obtained from a random sample of students who took the placement test:

	Placed in Math 200	Placed in Math 100	Placed in Math 50	Total
Took Review Course	40	13	7	60
Did not take Review Course	10	15	15	40
Total	50	28	22	100

1. Construction a relative frequency table based on this data.

	Placed in Math 200	Placed in Math 100	Placed in Math 50	Total
Took Review Course				
Did not take Review Course				
Total				

2. Construct a ROW conditional relative frequency table of the above data.

	Placed in Math 200	Placed in Math 100	Placed in Math 50	Total
Took review course				
Did not take review course				

3. Construct a COLUMN conditional relative frequency table of the data.

	Placed in Math 200	Placed in Math 100	Placed in Math 50
Took review course			
Did not take review course			
Total			

4. What is the percent of all students that took the review course and placed into Math 100? Which table was most useful to find this?

5. What is the percent of students taking the review course that got into Math 200? Which table was most useful to find this?

6. What is the percent of students placing into Math 50 that did not take the review course? Which table was most useful to find this?

7. The college wishes to know if there is an association between taking the review class and the math class a student is placed in. Which table is most helpful and does that table show that there is or is not an association? Would you conclude that there is an association?

8. If you DID conclude there was an association, then does this mean that taking the review course CAUSED a student to be placed into a higher-level mathematics class? Could you think of another explanation?