Looking at data: categorical data and relative frequencies (and superheros)

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 1950s while Batman was a top-rated series in the 1960s. Each of these characters was also popular in movies released from 1990 to present. 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?

Movie fans in the United States were invited to complete an online survey in 2010. Part of the survey included questions about superhero powers. More than $1,000$ people responded to this survey that included a question about a favorite superhero power. Researchers randomly selected 450 of the completed surveys. A rather confusing breakdown of the data by gender was compiled from the 450 surveys:

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

Exercises

Use the survey information given in Example 1 to answer the following questions.

1. How many more females than males indicated their favorite power is telepathy?

2. How many more males than females indicated their favorite power was to fly?

3. Write survey questions that you think might have been used to collect this data.

4. How do you think the 450 surveys used in the example might have been selected? You can assume that there were $1,000$ surveys to select from.

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG II-M3-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

A Statistical Study Involving a Two-Way Frequency Table

The data in the example may prompt us to pose the statistical question, "**Do males have different preferences for superhero powers than 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 person completing a survey. The first response indicates a person's gender, and the second response indicates the person's favorite superpower.

The first step in analyzing our statistical question is to organize this data in a two-way ***frequency table***.

A two-way frequency table that can be used to organize the categorical data is shown below. The letters below represent the frequency counts of the cells of the table.

	To Fly	Freeze time	Invisibility	Super Strength	Telepathy	Total
Females	(a)	(b)	(c)	(d)	(e)	(f)
Males	(g)	(h)	(i)	(j)	(k)	(l)
Total	(m)	(n)	(o)	(p)	(q)	(r)

- The shaded cells are called ***marginal frequencies***. They are located around the margins of the table and represent the totals of the rows or columns of the table.
- The non-shaded cells *within* the table are called ***joint frequencies***. Each joint cell is the frequency count of responses from the two categorical variables located by the intersection of a row and column.

Exercises

5. Complete the table below by determining a frequency count for each cell based on the summarized data.

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

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG II-M3-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

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.

Problem Set

1. Consider the following results from 100 randomly selected students regarding their after-school activities:
   - Of the 60 female students selected, 20 of them played intramural basketball, 10 played chess, and 10 were in the jazz band. The rest of them did not participate in the after-school program.
   - Of the male students, 10 did not participate in the after-school program, 20 played intramural basketball, 8 played in the jazz band, and the rest played chess.

   A two-way frequency table to summarize the survey data was started. Indicate what label is needed in the table cell identified with a ???.

	Intramural Basketball	Chess Club	Jazz Band	???	Total
Female					
Male					
Total					

2. Complete the above table for the 100 students who were surveyed.
3. The table shows the responses to the after-school activity question for males and females. Do you think there is a difference in the responses of males and females? Explain your answer.

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG II-M3-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Extending the Frequency Table to a Relative Frequency Table - Back to the Super Powers

Determining the number of people in each cell represents 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.

Two-Way Frequency Table for Super Power Preferences:

	To Fly	Freeze Time	Invisibility	Super Strength	Telepathy	Total
Females	**Error**	60	48	1	70	228
Males	51	71	27	25	48	222
Total	100	131	75	26	118	450

The relative frequency table would be found by dividing each of the above cell values by 450. For example, the relative frequency of females selecting to fly is $\frac{49}{450}$, or approximately 0.109, to the nearest thousandth. A few of the other relative frequencies to the nearest thousandth are shown in the following relative frequency table:

	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		$\frac{131}{450}\approx 0.291$			$\frac{118}{450}\approx 0.262$	

Exercises

1. Calculate the remaining relative frequencies in the table. You can leave the answers as fractions.
2. Which cells in this table would represent the joint relative frequencies?
3. Which cells in the relative frequency table would represent the marginal relative frequencies?

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG II-M3-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

4. What is the joint relative frequency for females who selected invisibility as their favorite superpower?
5. What is the marginal relative frequency 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?

Interpreting Data

If you were to make a new superhero movie and had to pick the single power for you superhero, what would it be?

Scott picked super strength as the special power (well, because it is his favorite special power).

Jill argued, that telepathy was the best choice that accommodates most of the potential viewers.

Scott acknowledged that super strength was probably not the best choice based on the data. But he did not agree with Jill that telepathy was the best choice. He argued that freeze time was clearly the most popular power among the surveyed movie fans.

Jill did not agree.

1. How do the data support Scott's claim? Why do you think he selected freeze time as the special power for the superhero?
2. How do the data support Jill's claim? Why do you think she selected telepathy as the special power for the superhero?
3. 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.

Summary

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***.

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG II-M3-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Problem Set

1. Consider the students and their after-school activities:

	Intramural Basketball	Chess Club	Jazz Band	Not Involved	Total
Males	20	Error	8	10	40
Females	20	10	10	20	60
Total	40	12	18	30	100

Calculate the relative frequencies for each of the cells to the nearest thousandth. Place the relative frequencies in the cells of the following table. (The first cell has been completed as an example.)

	Intramural Basketball	Chess Club	Jazz Band	Not Involved	Total
Males	$\frac{20}{100} = 0.200$				
Females					
Total					

2. If a student were randomly selected from the students at the school, do you think the student selected would be a male or a female?
3. If a student were selected at random from school, do you think this student would be involved in an after-school program? Explain your answer.Why might someone question whether or not the students who completed the survey were randomly selected? If the students completing the survey were randomly selected, what do the marginal relative frequencies possibly tell you about the school? Explain your answer.
4. Why might females think they are more involved in after-school activities than males? Explain your answer.

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG II-M3-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Conditional Relative Frequencies

A ***conditional relative frequency*** compares a frequency count to the marginal total 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 $\frac{48}{228}$, or approximately 0.211. This conditional relative frequency indicates that approximately 21.1% of females prefer invisibility as their favorite superpower. Similarly, $\frac{27}{222}$, or approximately 0.122 or 12.2%, of males prefer invisibility as their favorite superpower.

Exercises

1. Use the frequency counts from the table in previous sections to calculate the missing row of 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						

2. Suppose that a person is selected at random from those who completed the survey.
   a. What do you think is the gender of the person selected? What would you predict for this person's response to the superpower question?
   b. If the selected person is male, what do you think was his response to the selection of a favorite superpower? Explain your answer.
   c. If the selected person is female, what do you think was her response to the selection of a favorite superpower? Explain your answer.
3. What superpower was selected by approximately one-third of the females? What superpower was selected by approximately one-third of the males? How did you determine each answer from the conditional relative frequency table?

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG II-M3-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Possible Association Based on Conditional Relative Frequencies

Two categorical variables are associated if the row conditional relative frequencies (or column relative frequencies) are different for the rows (or columns) of the table. For example, if the selection of superpowers selected for females is 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.

Exercises

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

1. For what superpowers would you say that the conditional relative frequencies for females and males are very different?
2. For what superpowers are the conditional relative frequencies nearly equal for males and females?
3. Suppose a student is selected at random from the students who completed the survey. Would knowing the student's gender be helpful in predicting which superpower this student selected? Explain your answer.
4. Is there evidence of an association between gender and a favorite superpower? Explain why or why not.
5. What superpower would you recommend for the superhero character in the movie? Justify your choice.

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG II-M3-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Association and Cause-and-Effect

Students applying to college were given the opportunity to prepare for a college placement test in mathematics by taking a review course. Not all students took advantage of this opportunity. 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

Exercises

1. 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				
Total				

2. Based on the conditional relative frequencies, is there evidence of an association between whether a student takes the review course and the math course in which the student was placed? Explain your answer.
3. Looking at the conditional relative frequencies, the proportion of students who placed into Math 200 is much higher for those who took the review course than for those who did not. One possible explanation is that taking the review course caused improvement in placement test scores. What is another possible explanation?

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG II-M3-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Now consider the following statistical study:

Fifty people were selected at random. Each of these people was classified according to sugar consumption (high or low) and exercise level (high or low). The resulting data are summarized in the following frequency table.

		Exercise Level		
		High	**Low**	**Total**
Sugar Consumption	**High**	14	18	32
	Low	14	4	18
	Total	28	22	50

4. Calculate the row conditional relative frequencies, and display them in a row conditional relative frequency table.

		Exercise Level		
		High	**Low**	**Total**
Sugar Consumption	**High**			
	Low			
	Total			

5. Is there evidence of an association between sugar consumption category and exercise level? Support your answer using conditional relative frequencies.

6. Is it reasonable to conclude that high sugar consumption is the cause of the observed differences in the conditional relative frequencies? What other explanations could explain a difference in the conditional relative frequencies? Explain your answer.

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG II-M3-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Summary

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 exists.

An observed association between two variables does not necessarily mean that there is a cause-and-effect relationship between the two variables.

Problem Set

Consider again the summary of data from the 100 randomly selected students regarding their after-school activities and gender.

	Intramural Basketball	Chess Club	Jazz Band	Not Involved	Total
Females	20	10	10	20	60
Males	20	Error	8	10	40
Total	40	12	18	30	100

1. Construct a row conditional relative frequency table for this data. Decimal values are given to the nearest thousandth.

	Intramural Basketball	Chess Club	Jazz Band	Not Involved	Total
Females					60
Males					40
Total					

2. For what after-school activities do you think the row conditional relative frequencies for females and males are very different? What might explain why males or females select different activities?

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG II-M3-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

3. If John, a male student, completed the after-school survey, what would you predict was his response? Explain your answer.
4. If Beth, a female student, completed the after-school survey, what would you predict was her response? Explain your answer.
5. Notice that 20 female students participate in intramural basketball and that 20 male students participate in intramural basketball. Is it accurate to say that females and males are equally involved in intramural basketball? Explain your answer.
6. Do you think there is an association between gender and choice of after-school program? Explain.

Relative frequencies table from page 4.

	To Fly	Freeze Time	Invisibility	Super Strength	Telepathy	Total
Females	0.109 10.9%	0.133 13.3%	0.107 10.7%	0.002 0.2%	0.156 15.6%	0.507 50.7%
Males	0.113 11.3%	0.158 15.8%	0.060 6.0%	0.056 5.6%	0.107 10.7%	0.493 49.3%
Total	0.222 22.2%	0.291 29.1%	0.167 16.7%	0.058 5.8%	0.262 26.2%	1.00 100%

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG II-M3-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.