8.5.4 Summarizing Bivariate Categorical Data: Lesson 1

Classwork

Recall from your work in Grade 6 and Grade 8 that categorical data are data that are not numbers. 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.

**Exploratory Challenge 1: 2016 Summit Middle School Presidential Election Results.**

In 2016, there were four presidential candidates students were able to choose from: Dr. Jill Stein with the Green Party, Hillary Clinton with the Democratic Party, Donald Trump with the Republican Party, and Gary Johnson with the Libertarian Party. 339 students voted in our election.

* 23 students voted for Jill Stein. 9 of those were female and 2 of those were gender non-conforming.
* 249 students voted for Hillary Clinton. 118 of those students were males and 1 was gender non-conforming.
* 34 students voted for Donald Trump. 11 of those were femails and 0 non-conforming students.
* Finally, 33 students voted for Gary Johnson. 22 of those were males and 10 were females.

Exercises 1–4

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

1. How many more females than males indicated their preference for Hillary Clinton as President of the United States?
2. How many more males than females indicated their preference for Donald Trump as President?

**Exploratory Challenge 2: A Statistical Study Involving a Two-Way Frequency Table**

The data in Example 1 is fascinating to me. I have a statistical question, “Does gender make a difference regarding who middle schoolers prefer as president? 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 presidential candidate. For example, data collected from one student was *male* and *Dr. Jil Stein, Green Party*. 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 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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Dr. Jill Stein, Green Party** | **Hillary Clinton, Democratic Party** | **Donald Trump, Republican Party** | **Gary Johnson, Libertarian Party** | **Total** |
| **Females** | (a) | (b) | (c) | (d) | (e) |
| **Males** | (f) | (g) | (h) | (i) | (j) |
| **Non-Conforming** | (k) | (l) | (m) | (n) | (o) |
| **Total** | (p) | (q) | (r) | (s) | (t) |

* 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 5–12

1. Describe the data that would be counted in cell (a).
2. Describe the data that would be counted in cell (j).
3. Describe the data that would be counted in cell (l).
4. Describe the data that would be counted in cell (n).
5. Describe the data that would be counted in cell (r).
6. Cell (n) is the number of gender non-conforming students who selected *Gary Johnson* as their choice for President. Using the information given in Example 1, what is the value of this number?
7. Cell (b) is the number of females whose chose *Hillary Clinton*. Using the information given in Example 1, what is the value of this number?
8. Complete the table below by determining a frequency count for each cell based on the summarized data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Dr. Jill Stein, Green Party** | **Hillary Clinton, Democratic Party** | **Donald Trump, Republican Party** | **Gary Johnson, Libertarian Party** | **Total** |
| **Females** |  |  |  |  |  |
| **Males** |  |  |  |  |  |
| **Non-Conforming** |  |  |  |  |  |
| **Total** |  |  |  |  |  |

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.

Summarizing Bivariate Categorical Data with Relative Frequencies: Lesson 2

Classwork

This lesson expands on your work with two-way frequency tables from Lesson 9.

**Exploratory Challenge 1: 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.

Consider the two-way frequency table from #10.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Dr. Jill Stein, Green Party** | **Hillary Clinton, Democratic Party** | **Donald Trump, Republican Party** | **Gary Johnson, Libertarian Party** | **Total** |
| **Females** |  |  |  | 10 |  |
| **Males** |  |  |  |  | 175 |
| **Non-Conforming** | 2 |  |  |  |  |
| **Total** |  |  | 34 |  | 339 |

The relative frequency table would be found by dividing each of the above cell values by $339$. For example, the relative frequency of females selecting to voting for Gary Johnson is $\frac{10}{339}$, or approximately $0.029$, to the nearest thousandth. A few of the other relative frequencies to the nearest thousandth are shown in the following relative frequency table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Dr. Jill Stein, Green Party** | **Hillary Clinton, Democratic Party** | **Donald Trump, Republican Party** | **Gary Johnson, Libertarian Party** | **Total** |
| **Females** |  |  |  | 10/339 $≈$ 0.029 |  |
| **Males** |  |  |  |  | 175/339 $≈$ .515 |
| **Non-Conforming** | 2/339 $≈$ 0.059 |  |  |  |  |
| **Total** |  |  | 34/339 $≈$ 0.100 |  |  |

Exercises 1–7

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

Two-Way Frequency Table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Dr. Jill Stein, Green Party** | **Hillary Clinton, Democratic Party** | **Donald Trump, Republican Party** | **Gary Johnson, Libertarian Party** | **Total** |
| **Females** |  |  |  |  |  |
| **Males** |  |  |  |  |  |
| **Non-Conforming** |  |  |  |  |  |
| **Total** |  |  |  |  |  |

1. Based on previous work with frequency tables, which cells in this table would represent the joint relative frequencies?
2. Which cells in the relative frequency table would represent the marginal relative frequencies*?*
3. What is the joint relative frequency for females who selected Donald Trump for President?
4. What is the marginal relative frequency for the Green Party? Interpret the meaning of this value.
5. What is the difference in the joint relative frequencies for males and for females who voted for Gary Johnson for President?
6. Is there a noticeable difference between the genders and their choices for who should be our next President?

**Summit’s Presidential Election Survey Results are located at** [**https://goo.gl/tm9VKw**](https://goo.gl/tm9VKw)

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.
* 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*.

Conditional Relative Frequencies and Association: Lesson 3

Classwork

When examining data, sometimes we need to reorganize the information in another way to help us gain insight into the situation. One way to do this is examine each row of the table can help determine whether or not there is an association.

**Exploratory Challenge 1: Conditional Relative Frequencies**

Recall the two-way table from the previous lesson.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gender | Dr. Jill Stein, Green Party | Hillary Clinton, Democratic Party | Donald Trump, Republican Party | Gary Johnson, Libertarian Party | Total |
| Female | 9 | 130 | 11 | 10 | 160 |
| Male | 12 | 118 | 23 | 22 | 175 |
| Non-conforming | 2 | 1 | 0 | 1 | 4 |
| Total | 23 | 249 | 34 | 33 | 339 |

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 Jill Stein for President is 9/160, or, 0.05625. This conditional relative frequency indicates that approximately $5.6\%$ of females prefer Jill Stein as our next President. Similarly, 12/175, or approximately 0.0686 or $6.9\%$, of males prefer Jill Stein for President.

Exercises 1–5

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gender | Dr. Jill Stein, Green Party | Hillary Clinton, Democratic Party | Donald Trump, Republican Party | Gary Johnson, Libertarian Party | Total |
| Female | 9/160 = 0.056 |  |  |  | 160/160 = 1 |
| Male | 12/175 = 0.069 |  |  |  |  |
| Non-conforming |  |  |  |  |  |
| Total |  |  |  |  |  |

1. 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?

1. 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.
2. 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.
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?

**Exploratory Challenge 2: 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 presidential candidates selected for females is different than the selection of presidential candidates for males, then gender and presidential candidates are associated. This difference indicates that knowing the gender of a person in the sample indicates something about their political 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 6–10**

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

1. For what presidential candidates would you say that the conditional relative frequencies for females and males are very different?
2. For what presidential candidates 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 candidate this student voted for? Explain your answer.
4. Is there evidence of an association between gender and a favorite person for President? Explain why or why not.

**Exploratory Challenge 3: Association and Cause-and-Effect**

As part of the Summit student body vote, we also asked students to look at one of the proposals on the Ballot. Namely, should tax sugary drinks in Boulder? Here are the responses based on Gender and Grade Level. The differences in the total number of responses is due to whether or not a student voted in that category.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tax on sugar drinks: Yes | Tax on sugar drinks: No | Total |
| Male | 76 | 98 | 174 |
| Female | 97 | 63 | 160 |
| Non-conforming | 1 | 3 | 4 |
| Total | 174 | 164 | 338 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Grade Level-6th | Grade Level-7th | Grade Level-8th | Total |
| Yes on Sugar Tax | 64 | 61 | 46 | 171 |
| No on Sugar Tax | 48 | 54 | 62 | 164 |
| Total | 112 | 115 | 108 | 335 |

Exercises 11–16

1. Construct a row conditional relative frequency table of the above data. Be careful to make certain that you use conditional relative frequencies with the information you want to compare.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tax on sugar drinks: Yes | Tax on sugar drinks: No | Total |
| Male |  |  |  |
| Female |  |  |  |
| Non-conforming |  |  |  |
| Total |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Grade Level-6th | Grade Level-7th | Grade Level-8th | Total |
| Yes on Sugar Tax |  |  |  |  |
| No on Sugar Tax |  |  |  |  |
| Total |  |  |  |  |

1. Based on the conditional relative frequencies, is there evidence of an association between gender and whether or not a student supported the sugar tax? Explain your answer.
2. Based on the conditional relative frequencies, is there evidence of an association between grade level and whether or not a student supported the sugar tax?
3. Go to Summit’s Presidential Election Survey Results (<https://goo.gl/tm9VKw>) and pick a relationship to study we have not looked at yet together in class. Here are some options:

-Ethnicity versus Presidential preference
-Ethnicity versus Sugar Tax

What relationship did you choose to study?
4. Create a conditional relative frequency table for your information. You may choose to do this in Google Sheets or in Excel. I suggest copying and pasting the information from my spreadsheet to yours. Summarize your findings below.
5. Is there evidence of an association between your two categories? Support your answer using conditional relative frequencies.
6. Is there any other possible explanation for the differences you see?

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