Associations in Categorical Data

Let's look for relationships between categorical variables.

3.1 English or Math

The table displays the course preference and dominant hand (left- or right-handed) for a sample of 300 people.

	prefers English	prefers math	total
left-handed	10	20	30
right-handed	90	180	270
total	100	200	300

For each of the calculations, describe the interpretation of the percentage in terms of the situation.

1. 10% (from $\frac{10}{100} = 0.1$)

2. 67% (from $\frac{180}{270} \approx 0.67$)

3. 30% (from $\frac{90}{300} = 0.3$)

3.2 Associations in Categorical Data

1. The two-way table displays data about 55 different locations. Scientists have a list of possible chemicals that may influence the health of the coral. They first look at how nitrate concentration might be related to coral health. The table displays the health of the coral (healthy or unhealthy) and the nitrate concentration (low or high).

	low nitrate concentration	high nitrate concentration	total
healthy	20	5	25
unhealthy	8	22	30
total	28	27	55

   a. Complete the two-way relative frequency table for the data in the two-way table in which the relative frequencies are calculated using the total for each column.

	low nitrate concentration	high nitrate concentration
healthy		
unhealthy		
total	100%	100%

   b. When there is a low nitrate concentration, which has a higher relative frequency, healthy or unhealthy coral?

   c. When there is a high nitrate concentration, is there a higher relative frequency of healthy or unhealthy coral?

d. Considering this data, is there a possible **association** between coral health and the level of nitrate concentration? Explain your reasoning.

e. The scientists next look at how silicon dioxide concentration might be related to coral health. The relative frequencies calculating using the total for each column are shown in the table. Considering this data, is there a possible association between coral health and the level of silicon dioxide concentration? Explain your reasoning.

	low silicon dioxide concentration	high silicon dioxide concentration
healthy	44%	46%
unhealthy	56%	54%
total	100%	100%

2. Jada surveyed 300 people from various age groups about their shoe preferences. The two-way table summarizes the results of the survey.

	prefers sneakers without laces	prefers sneakers with laces	prefers shoes that are not sneakers	total
4–10 years old	21	12	3	36
11–17 years old	21	48	39	108
18–24 years old	15	54	87	156
total	57	114	129	300

Jada concludes that there is a possible association between age and shoe preference. Is Jada's conclusion reasonable? Explain your reasoning.

3. The two-way table summarizes data on writing utensil preference and the dominant hand for a sample of 100 people.

	left-handed	right-handed	total
prefers pen	7	82	89
prefers pencil	6	5	11
total	13	87	100

Is there a possible association between the dominant hand and writing utensil preference? Explain your reasoning.

Are you ready for more?

The incomplete two-way table displays the results of a survey about the type of sports medicine treatment and recovery time for 33 student athletes who visited the athletic trainer.

	returned to playing in less than 2 days	**returned to playing in 2 or more days**
treated with ice	8	4
treated with heat		

1. What 2 values could you use to complete the two-way table to show that there is an association between returning to playing in less than 2 days and the treatment (ice or heat)? Explain your reasoning.

2. What 2 values could you use to complete the two-way table to show that there is no association between returning to playing in less than 2 days and the treatment (ice or heat)? Explain your reasoning.

3. Which 2 values were easier to choose, the 2 values showing an association, or the 2 values showing no association? Explain your reasoning.

3.3 Associating Your Own Variables

1. Work with your group to identify a pair of categorical variables you think might be associated and another pair you think would not be associated.

2. Imagine your group collected data for each pair of categorical variables. Create a two-way table that could represent each set of data. Invent some data with 100 total values to complete each table. Remember that one table shows a possible association, and the other table shows no association.

3. Explain or show why there appears to be an association for the first pair of variables and why there appears to be no association for the other pair of variables.

4. Prepare a display of your work to share.

Lesson 3 Summary

An **association** between two variables means that the two variables are statistically related to each other. For example, we might expect that ice cream sales would be higher on sunny days than on snowy days. If sales were higher on sunny days than on snowy days, then we would say that there is a possible association between ice cream sales and whether it is sunny or snowing. When dealing with categorical variables, row or column relative frequency tables are often used to look for associations in the data.

Here is a two-way table displaying ice cream cone sales and weather conditions for 41 days for a particular creamery.

	sunny day	snowy day	total
sold fewer than 50 cones	8	7	15
sold 50 cones or more	22	4	26
total	30	11	41

Noticing a pattern in the raw data can be difficult, especially when the row or column totals are not the same for different categories, so the data should be converted into a row or column relative frequency table to better compare the categories. For the creamery, notice that the number of days with low sales is about the same for the two weather types, which contradicts our intuition. In this case, it makes sense to look at the percentage of days that sold well under each weather condition separately. That is, consider the column relative frequencies.

	sunny day	snowy day
sold fewer than 50 cones	27%	64%
sold 50 cones or more	73%	36%
total	100%	100%

From the column relative frequency table, it is clear that most of the sunny days resulted in sales of at least 50 cones (73%), while most of the snowy days resulted in fewer than 50 cones sold (64%). Because these percentages are quite different, this suggests there is an association between the weather condition and the number of cone sales. A bakery might wonder if the weather conditions impact their muffin sales as well.

	sunny day	snowy day
sold fewer than 50 muffins	32%	35%
sold 50 muffins or more	68%	65%
total	100%	100%

For the bakery, it seems there is not an association between weather conditions and muffin sales since the percentage of days with low sales are very similar under the different weather conditions, and the percentages are also close on days when they sold many muffins.

Using row or column relative frequency tables helps organize data so that columns (or rows) can be easily compared between different categories for a variable. This comparison can be accomplished using a two-way table, but the differences in the number of data values in a given category must be accounted for.