Chapter 1

Summarizing Categorical Data: Counts and Percents

In This Chapter

- Making tables to summarize categorical data
- Highlighting the difference between frequencies and relative frequencies
- Interpreting and evaluating tables

Summarizing categorical data involves boiling down all the information into just a few numbers that tell its basic story. Because categorical data involves pieces of data that belong in categories, you have to look at how many individuals fall into each group and summarize the numbers appropriately. In this chapter, you practice making, interpreting, and evaluating frequency and relative frequency tables for categorical data.

Counting On the Frequency

One way to summarize categorical data is to simply count, or tally up, the number of individuals that fall into each category. The number of individuals in any given category is called the *frequency* for that category. If you list all the possible categories along with the frequency for each, you create a *frequency table.* The total of all the frequencies should equal the size of the sample (because you place each individual in one category).

See the following for an example of summarizing data by using a frequency table.

Q. Suppose you take a sample of 10 people and ask them all whether or not they own a cell phone. Each person falls into one of two categories: yes or no. The data is shown in the following table.

Person #	*Cell Phone*	*Person #*	*Cell Phone*
1	Y	6	Y
2	N	7	Y
3	Y	8	Y
4	N	9	N
5	Y	10	Y

a. Summarize this data in a frequency table.

b. What's an advantage of summarizing categorical data?

A. Data summaries boil down the data quickly and clearly.

a. The frequency table for this data is shown in the following table.

b. A data summary allows you to see patterns in the data, which aren't clear if you look only at the original data.

Own a Cell Phone?	*Frequency*
Y	7
N	3
Total	10

1. You survey 20 shoppers to see what type of soft drink they like best, Brand A or Brand B. The results are: A, A, B, B, B, B, B, B, A, A, A, B, A, A, A, A, B, B, A, A. Which brand do the shoppers prefer? Make a frequency table and explain your answer.

Solve It

2. A local city government asks voters to vote on a tax levy for the local school district. A total of 18,726 citizens vote on the issue. The yes count comes in at 10,479, and the rest of the voters said no.

a. Show the results in a frequency table.

b. Why is it important to include the total number at the bottom of a frequency table?

Solve It

3. A zoo asks 1,000 people whether they've been to the zoo in the last year. The surveyors count that 592 say yes, 198 say no, and 210 don't respond.

a. Show the results in a frequency table.

b. Explain why you need to include the people who don't respond.

Solve It

4. Suppose instead of showing the number in each group, you show just the percentage (called a *relative frequency*). What's one advantage a relative frequency table has over a frequency table?

Solve It

Relating with Percentages

Another way to summarize categorical data is to show the percentage of individuals that fall into each category, thereby creating a relative frequency. The *relative frequency* of a given category is the frequency (number of individuals in that category) divided by the total sample size, multiplied by 100 to get the percentage. For example, if you survey 50 people and 10 are in favor of a certain issue, the relative frequency of the "in-favor" category is 10 ÷ 50 = .20 times 100, which gives you 20 percent. If you list all the possible categories along with their relative frequencies, you create a *relative frequency table.* The total of all the relative frequencies should equal 100 percent (subject to possible round-off error).

See the following for an example of summarizing data by using a relative frequency table.

Q. Using the cell phone data from the following table, make a relative frequency table and interpret the results.

Person #	*Cell Phone*	*Person #*	*Cell Phone*
1	Y	6	Y
2	N	7	Y
3	Y	8	Y
4	N	9	N
5	Y	10	Y

A. The following table shows a relative frequency table for the cell phone data. Seventy percent of the people sampled reported owning cell phones, and thirty percent admitted to being technologically behind the times.

Own a Cell Phone?	*Relative Frequency*
Y	70%
N	30%

You get the 70 percent by taking 7 ÷ 10 * 100, and you calculate the 30 percent by taking 3 ÷ 10 * 100.

5. You survey 20 shoppers to see what type of soft drink they like best, Brand A or Brand B. The results are: A, A, B, B, B, B, B, B, A, A, A, B, A, A, A, A, B, B, A, A. Which brand do the shoppers prefer?

a. Use a relative frequency table to determine the preferred brand.

b. In general, if you had to choose, which is easier to interpret: frequencies or relative frequencies? Explain.

Solve It

6. A local city government asked voters in the last election to vote on a tax levy for the local school district. A record 18,726 voted on the issue. The Yes count came in at 10,479, and the rest of the voters checked the No box. Show the results in a relative frequency table.

Solve It

7. A zoo surveys 1,000 people to find out whether they've been to the zoo in the last year. The surveyors count that 592 say yes, 198 say no, and 210 don't respond. Make a relative frequency table and use it to find the *response rate* (percentage of people who respond to the survey).

Solve It

8. Name one disadvantage that comes with creating a relative frequency table compared to using a frequency table.

Solve It

Interpreting Counts and Percents with Caution

Not all summaries of categorical data are fair and accurate. Knowing what to look for can help you keep your eyes open for misleading and incomplete information.

Instructors often ask you to "interpret the results." Your instructor wants you to use the statistics available to talk about how they relate to the given situation. In other words, what do the results mean to the person who collects the data?

With relative frequency tables, don't forget to check if all categories sum to 1 or 100 percent (subject to round-off error), and remember to look for some indicator as to total sample size.

See the following for an example of critiquing a data summary.

Q. You watch a commercial where the manufacturer of a new cold medicine ("Nocold") compares it to the leading brand. The results are shown in the following table.

How Nocold Compares	*Percentage*
Much better	47%
At least as good	18%

a. What kind of table is this?

b. Interpret the results. (Did the new cold medicine beat out the leading brand?)

c. What important details are missing from this table?

A. Much like the cold medicines I always take, the table about "Nocold" does "Nogood."

a. This table is an incomplete relative frequency table. The remaining category is "not as good" for the Nocold brand, and the advertiser doesn't show it. But you can do the math and see that 100 percent – (47 percent + 18 percent) = 35 percent of the people say that the leading brand is better.

b. If you put the two groups together, 65 percent of the patients say that Nocold is at least as good as the leading brand, and almost half of the patients say Nocold is much better.

c. What's missing? The remaining percentage (to keep all possible results in perspective). But more importantly, the total sample size is missing. You don't know if the surveyors sampled 10 people, 100 people, or 1,000 people. This means the precision of the results is unknown. (Precision means how consistent the results will be from sample to sample; it's related to sample size, as you see in Chapter 8.)

9. Suppose you ask 1,000 people to identify from a list of five vacation spots which ones they've already visited. The frequencies you receive are Disneyworld: 216; New Orleans: 312; Las Vegas: 418; New York City: 359; and Washington, D.C.: 188.

a. Explain why creating a traditional relative frequency table doesn't make sense here.

b. How can you summarize this data with percents in a way that makes sense?

Solve It

10. If you have only a frequency table, can you find the corresponding relative frequency table? Conversely, if you have only a relative frequency table, can you find the corresponding frequency table? Explain.

Solve It

Answers to Problems in Summarizing Categorical Data

1 Eleven shoppers prefer Brand A, and nine shoppers prefer Brand B. The frequency table is shown in the following table. Brand A got more votes, but the results are pretty close.

Brand Preferred	*Frequency*
A	11
B	9
Total	20

2 Frequencies are fine for summarizing data as long as you keep the total number in perspective.

a. The results are shown in the following table. Because the total is 18,726, and the Yes count is 10,479, the No count is the difference between the two, which is 18,726 – 19,479 = 8,247.

b. The total is important because it helps keep the frequencies in perspective when you compare them to each other.

Vote	*Frequency*
Y	10,479
N	8,247
Total	18,726

3 This problem shows the importance of reporting not only the results of participants who respond, but also what percentage of the total actually respond.

a. The results are shown in the following table.

b. If you don't show the non-respondents, the total doesn't add up to 1,000 (the number surveyed). An alternative way to show the data is to base it on only the respondents, but the results would be biased. You can't definitively say that the non-respondents would respond the same way as the respondents.

Gone to the Zoo in the Last Year?	*Frequency*
Y	592
N	198
Non-respondents	210
Total	1,000

4 Showing the percents rather than counts means making a relative frequency table rather than a frequency table. One advantage of a relative frequency table is that everything sums to 100 percent, making it easier to interpret the results, especially if you have a large number of categories.

5 Relative frequencies do just what they say: They help you relate the results to each other (by finding percentages).

a. Eleven shoppers out of the twenty prefer Brand A, and nine shoppers out of the twenty prefer Brand B. The relative frequency table is shown in the following table. Brand A got more votes, but the results are pretty close, with 55 percent of the shoppers preferring Brand A and 45 percent preferring Brand B.

Brand Preferred	*Relative Frequency*
A	55%
B	45%

b. You often have an easier time interpreting percents, because when you need to interpret counts, you have to put them in perspective in terms of "out of how many?"

6 The results are shown in the following table. The Yes percentage is 10,479 ÷ 18,726 = 55.96 percent. Because the total is 100 percent, the No percentage is 100 percent – 55.96 percent = 44.04 percent.

Vote	*Relative Frequency*
Y	55.96%
N	44.04%

7 You can see the relative frequency table that follows this answer. Knowing the response rate is critical for interpreting the results of a survey. The higher the response rate the better. The response rate is 59.2 percent + 19.8 percent = 79.0 percent – the total percentage of people who responded in any way (yes or no) to the survey. (Note that 21 percent is the non-response rate.)

Gone to the Zoo in the Last Year?	*Relative Frequency*
Y	592 ÷ 1,000 = .592 = 59.2%
N	19.8%
Non-respondents	21.0%

8 One disadvantage of a relative frequency table is if you see only the percents, you don't know how many people participated in the study; therefore, you don't know how precise the results are. Remember the commercial about four out of five dentists surveyed? Maybe the company only asked five dentists! You can get around this problem by putting the total sample size somewhere at the top or bottom of your relative frequency table.

When making a relative frequency table, include the total sample size somewhere on the table.

9 Be careful how you interpret tables where an individual can be in more than one category at the same time.

a. The frequencies don't sum to 1,000, because people have the option to choose multiple locations or none at all, so each person doesn't end up in exactly one group. If you take the grand total of all the frequencies (1,493) and divide each frequency by 1,493 to get a relative frequency, the relative frequencies sum to one (or 100 percent). But what does that mean? It makes it hard to interpret these percents because they don't account for the total number of people.

b. One way you can summarize this data is by showing the percentage of people who have been at each location separately (compared to the percentage who haven't been there before). These percents add up to one for each location. The following table shows the results summarized with this method. ***Note:*** The table isn't a relative frequency table; however, it uses relative frequencies.

Location	*% Who Have Been There*	*% Who Haven't Been There*
Disneyworld	216 ÷ 1,000 = 21.6%	100% – 21.6% = 78.4%
New Orleans	312 ÷ 1,000 = 31.2 %	68.8%
Las Vegas	418 ÷ 1,000 = 41.8%	58.2%
New York City	359 ÷ 1,000 = 35.9%	64.1%
Washington, D.C.	188 ÷ 1,000 = 18.8%	81.2%

Not all tables involving percents should sum to one. Don't force tables to sum to one when they shouldn't; do make sure you understand whether each individual can fall under more than one category. In those cases, a typical relative frequency table isn't appropriate.

10 You can always sum all the frequencies to get a total and then find each relative frequency by taking the frequency divided by the total. However, if you have only the percents, you can't go back and find the original counts unless you know the total number of individuals. Suppose you know that 80 percent of the people in a survey like ice cream. How many people in the survey like ice cream? If the total number of respondents is 100, 80 (100 * .80) people like ice cream. If the total is 50, you're looking at 40 (50 * .80) positive answers. If the total is five, you deal only with four (5 * .80). This illustrates why relative frequency tables need to have the total sample size somewhere.

Watch for total sample sizes when given a relative frequency table. Don't be misled by percentages alone, thinking they're always based on large sample sizes. Many are not.