

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

Chapter 1

Summarizing Categorical Data: Counts and Percents

Categorical data is data in which individuals are placed into groups or categories — for example gender, region, or type of movie. 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 (or count) 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 that you take a sample of 10 people and ask them all whether they own a cellphone. Each person falls into one of two categories: yes or no. The data are shown in the following table.

Person #	Cellphone	Person #	Cellphone
1	Y	6	Y
2	N	7	Y
3	Y	8	Y
4	N	9	N
5	Y	10	Y

- Summarize this data in a frequency table.
- What's an advantage of summarizing categorical data?

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

- The frequency table for this data is shown in the following table.
- 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 Cellphone?	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, 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.

- 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.
- Show the results in a frequency table.
 - Why is it important to include the total number at the bottom of a frequency table?

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.

4 Suppose that 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?

Relating with Percentages

Another way to summarize categorical data is to show the percentage of individuals who 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 $10 \div 50 = 0.20$ 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 cellphone data from the following table, make a relative frequency table and interpret the results.

Person #	Cellphone	Person #	Cellphone
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 cellphone data. Seventy percent of the people sampled reported owning cellphones, and 30 percent admitted to being technologically behind the times.

Own a Cellphone?	Relative Frequency
Y	70%
N	30%

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

- 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, A, A, A, B, A, A, A, A, B, B, A, A. Which brand do the shoppers prefer?
- Use a relative frequency table to determine the preferred brand.
 - In general, if you had to choose, which is easier to interpret: frequencies or relative frequencies? Explain.

- 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.

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).

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

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.” In this case, 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?



REMEMBER

With relative frequency tables, don't forget to check whether 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%

- What kind of table is this?
- Interpret the results. (Did the new cold medicine beat out the leading brand?)
- What important details are missing from this table?

- A.** Much like the cold medicines I always take, the table about “Nocold” does “Nogood.”

- 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 $100\% - (47\% + 18\%) = 35\%$ of the people say that the leading brand is better.
- If you put the two groups together, 65% 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.
- 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 whether the surveyors sampled 10 people, 100 people, or 1,000 people. This means that 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 10.)

9 Suppose that 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 Disney World: 216; New Orleans: 312; Las Vegas: 418; New York City: 359; and Washington, D.C.: 188.

- Explain why creating a traditional relative frequency table doesn’t make sense here.
- How can you summarize this data with percents in a way that makes sense?

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.

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.
- 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 $18,726 - 10,479 = 8,247$.
 - 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.
- The results are shown in the following table.
 - If you don't show the nonrespondents, 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 nonrespondents would respond the same way as the respondents.

Gone to the Zoo in the Last Year?	Frequency
Y	592
N	198
Nonrespondents	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 20 prefer Brand A, and nine shoppers out of the 20 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 $10,479 \div 18,726 = 55.96\%$. Because the total is 100%, the no percentage is $100\% - 55.96\% = 44.04\%$.

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 $59.2\% + 19.8\% = 79.0\%$ – the total percentage of people who responded in any way (yes or no) to the survey. (Note that 21% is the nonresponse rate.)

Gone to the Zoo in the Last Year?	Relative Frequency
Y	$592 \div 1,000 = 0.592 = 59.2\%$
N	19.8%
Nonrespondents	21.0%

- 8 One disadvantage of a relative frequency table is that 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. You can get around this problem by putting the total sample size somewhere at the top or bottom of your relative frequency table.



REMEMBER

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

9

Be careful about 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 1 (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 1 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
Disney World	$216 \div 1,000 = 21.6\%$	$100\% - 21.6\% = 78.4\%$
New Orleans	$312 \div 1,000 = 31.2\%$	68.8%
Las Vegas	$418 \div 1,000 = 41.8\%$	58.2%
New York City	$359 \div 1,000 = 35.9\%$	64.1%
Washington, D.C.	$188 \div 1,000 = 18.8\%$	81.2%



REMEMBER

Not all tables involving percents should sum to 1. Don't force tables to sum to 1 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 that 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, $80 = (100 * 0.80)$ people like ice cream. If the total is 50, you're looking at $40 = (50 * 0.80)$ positive answers. If the total is 5, you deal only with $4 = (5 * 0.80)$. This illustrates why relative frequency tables need to have the total sample size somewhere.



REMEMBER

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, because many are not.

