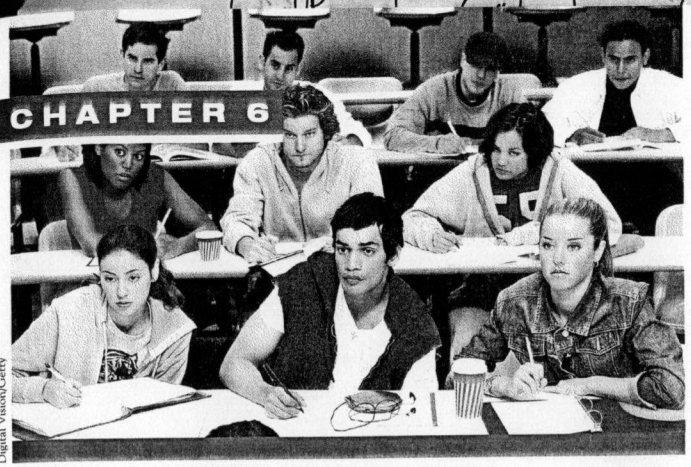
CHAPTER 6



Digital Vision/Getty

Two-Way Tables*

IN THIS CHAPTER WE COVER...

- Marginal distributions
- Conditional distributions
- Simpson's paradox

We have concentrated on relationships in which at least the response variable is quantitative. Now we will describe relationships between two categorical variables. Some variables—such as sex, race, and occupation—are categorical by nature. Other categorical variables are created by grouping values of a quantitative variable into classes. Published data often appear in grouped form to save space. To analyze categorical data, we use the *counts* or *percents* of individuals that fall into various categories.

EXAMPLE 6.1 I think I'll be rich by age 30

A sample survey of young adults (aged 19 to 25) asked, "What do you think are the chances you will have much more than a middle-class income at age 30?" Table 6.1 shows the responses, omitting a few people who refused to respond or who said they were already rich.[1] This is a **two-way table** because it describes two categorical variables: sex and opinion about becoming rich. Opinion is the **row variable** because each row in the table describes young adults who held one of the five opinions about their chances. Because the opinions have a natural order from "Almost no chance" to

two-way table
row variable

*This material is important in statistics, but it is needed later in this book only for Chapter 22. You may omit it if you do not plan to read Chapter 22 or delay reading it until you reach Chapter 22.

TABLE 6.1 Young adults by sex and chance of getting rich

	SEX		
OPINION	FEMALE	MALE	TOTAL
Almost no chance	96	98	194
Some chance but probably not	426	286	712
A 50-50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

column variable

"Almost certain," the rows are also in this order. Sex is the **column variable** because each column describes one sex. The entries in the table are the counts of individuals in each opinion-by-sex class. ■

Marginal distributions

How can we best grasp the information contained in Table 6.1? First, *look at the distribution of each variable separately*. The distribution of a categorical variable says how often each outcome occurred. The "Total" column at the right of the table contains the totals for each of the rows. These row totals give the distribution of opinions about becoming rich in the entire group of 4826 young adults: 194 felt that they had almost no chance, 712 thought they had just some chance, and so on.

If the row and column totals are missing, the first thing to do in studying a two-way table is to calculate them. The distributions of opinion alone and sex alone are called **marginal distributions** because they appear at the right and bottom margins of the two-way table.

marginal distribution

Percents are often more informative than counts. We can display the marginal distribution of opinions in percents by dividing each row total by the table total and converting to a percent.

EXAMPLE 6.2 Calculating a marginal distribution

The percent of these young adults who think they are almost certain to be rich by age 30 is

$$\frac{\text{almost certain total}}{\text{table total}} = \frac{1083}{4826} = 0.224 = 22.4\%$$

Do four more such calculations to obtain the marginal distribution of opinion in percents. Here is the complete distribution:

Response	Percent
Almost no chance	$\frac{194}{4826} = 4.0\%$
Some chance	$\frac{712}{4826} = 14.8\%$
A 50-50 chance	$\frac{1416}{4826} = 29.3\%$
A good chance	$\frac{1421}{4826} = 29.4\%$
Almost certain	$\frac{1083}{4826} = 22.4\%$

It seems that many young adults are optimistic about their future income. The total should be 100% because everyone holds one of the five opinions. In fact, the percents add to 99.9% because we rounded each one to the nearest tenth. This is **roundoff error.** ■

roundoff error

Each marginal distribution from a two-way table is a distribution for a single categorical variable. As we saw in Chapter 1, we can use a bar graph or a pie chart to display such a distribution. Figure 6.1 is a bar graph of the distribution of opinion among young adults.

In working with two-way tables, you must calculate lots of percents. Here's a tip to help decide what fraction gives the percent you want. Ask, "What group represents the total of which I want a percent?" The count for that group is the denominator of the fraction that leads to the percent. In Example 6.2, we want a percent "of young adults," so the count of young adults (the table total) is the denominator.

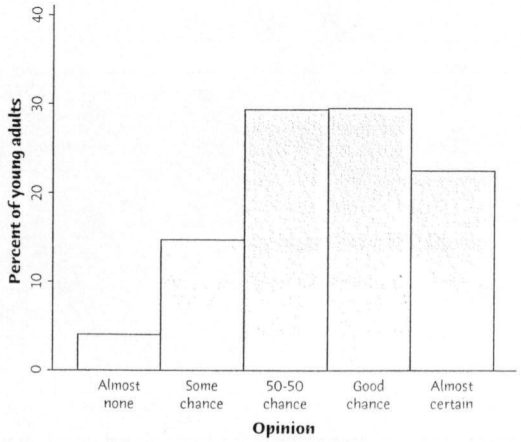


FIGURE 6.1
A bar graph of the distribution of opinions of young adults about becoming rich by age 30. This is one of the marginal distributions for Table 6.1.

APPLY YOUR KNOWLEDGE

6.1 Attitudes toward recycled products. Recycling is supposed to save resources. Some people think recycled products are lower in quality than other products, a fact that makes recycling less practical. Here are data on attitudes toward coffee filters made of recycled paper among people who had bought these filters and people who had not:[2]

	Think the quality of the recycled product is		
	Higher	The same	Lower
Buyers	20	7	9
Nonbuyers	29	25	43

(a) How many people does this table describe? How many of these were buyers of coffee filters made of recycled paper?

(b) Give the marginal distribution of opinion about the quality of recycled filters. What percent of consumers think the quality of the recycled product is the same or higher than the quality of other filters?

Digital Vision/Getty

6.2 Undergraduates' ages. Here is a two-way table of Census Bureau data describing the age and sex of all American undergraduate college students. The table entries are counts in thousands of students.[3]

Age group	Female	Male
15 to 17 years	116	61
18 to 24 years	5470	4691
25 to 34 years	1319	824
35 years or older	1075	616

(a) How many college undergraduates are there?

(b) Find the marginal distribution of age group. What percent of undergraduates are in the traditional 18 to 24 college age group?

Conditional distributions

Table 6.1 contains much more information than the two marginal distributions of opinion alone and sex alone. *Marginal distributions tell us nothing about the relationship between two variables.* To describe a relationship between two categorical variables, we must calculate some well-chosen percents from the counts given in the body of the table.

Let's say that we want to compare the opinions of women and men. To do this, compare percents for women alone with percents for men alone. To study the opinions of women, we look only at the "Female" column in Table 6.1. To find the percent *of young women* who think they are almost certain to be rich by age 30, divide the count of such women by the total number of women (the column total):

$$\frac{\text{women who are almost certain}}{\text{column total}} = \frac{486}{2367} = 0.205 = 20.5\%$$

Doing this for all five entries in the "Female" column gives the *conditional distribution* of opinion among women. We use the term "conditional" because this distribution describes only young adults who satisfy the condition that they are female.

MARGINAL AND CONDITIONAL DISTRIBUTIONS

The **marginal distribution** of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among all individuals described by the table.

A **conditional distribution** of a variable is the distribution of values of that variable among only individuals who have a given value of the other variable. There is a separate conditional distribution for each value of the other variable.

EXAMPLE 6.3 Comparing women and men

STATE: How do young men and young women differ in their responses to the question "What do you think are the chances you will have much more than a middle-class income at age 30?"

PLAN: Make a two-way table of response by sex. Find the two conditional distributions of response for men alone and for women alone. Compare these two distributions.

SOLVE: Table 6.1 is the two-way table we need. Look first at just the "Female" column to find the conditional distribution for women, then at just the "Male" column to find the conditional distribution for men. Here are the calculations and the two conditional distributions:

Response	Female	Male
Almost no chance	$\frac{96}{2367} = 4.1\%$	$\frac{98}{2459} = 4.0\%$
Some chance	$\frac{426}{2367} = 18.0\%$	$\frac{286}{2459} = 11.6\%$
A 50-50 chance	$\frac{696}{2367} = 29.4\%$	$\frac{720}{2459} = 29.3\%$
A good chance	$\frac{663}{2367} = 28.0\%$	$\frac{758}{2459} = 30.8\%$
Almost certain	$\frac{486}{2367} = 20.5\%$	$\frac{597}{2459} = 24.3\%$

Each set of percents adds to 100% because everyone holds one of the five opinions.

CONCLUDE: Men are somewhat more optimistic about their future income than are women. Men are less likely to say that they have "some chance but probably not" and more likely to say that they have "a good chance" or are "almost certain" to have much more than a middle-class income by age 30. ■

Smiling faces

Women smile more than men. The same data that produce this fact allow us to link smiling to other variables in two-way tables. For example, add as the second variable whether or not the person thinks they are being observed. If yes, that's when women smile more. If no, there's no difference between women and men. Next, take the second variable to be the person's social role (for example, is he or she the boss in an office?). Within each role, there is very little difference in smiling between women and men.

FIGURE 6.2
Minitab output for the two-way table of young adults by sex and chance of getting rich, along with each entry as a percent of its column total. The "Female" and "Male" columns give the conditional distributions of responses for women and men, and the "All" column shows the marginal distribution of responses for all these young adults.

Session

	Female	Male	All
A: Almost no chance	96	98	194
	4.06	3.99	4.02
B: Some chance but probably not	426	286	712
	18.00	11.63	14.75
C: A 50-50 chance	696	720	1416
	29.40	29.28	29.34
D: A good chance	663	758	1421
	28.01	30.83	29.44
E: Almost certain	486	597	1083
	20.53	24.28	22.44
All	2367	2459	4826
	100.00	100.00	100.00

Cell Contents: Count
% of Column

Software will do these calculations for you. Most programs allow you to choose which conditional distributions you want to compare. The output in Figure 6.2 presents the two conditional distributions of opinion, for women and for men, and also the marginal distribution of opinion for all of the young adults. The distributions agree (up to roundoff) with the results in Examples 6.2 and 6.3.

Remember that there are two sets of conditional distributions for any two-way table. Example 6.3 looked at the conditional distributions of opinion for the two sexes. We could also examine the five conditional distributions of sex, one for each of the five opinions, by looking separately at the five rows in Table 6.1. Because the variable "sex" has only two categories, comparing the five conditional distributions amounts to comparing the percents of women among young adults who hold each opinion. Figure 6.3 makes this comparison in a bar graph. The bar heights do *not* add to 100%, because each bar represents a different group of people.

No single graph (such as a scatterplot) portrays the form of the relationship between categorical variables. No single numerical measure (such as the correlation) summarizes the strength of the association. Bar graphs are flexible enough to be helpful, but you must think about what comparisons you want to display. For numerical measures, we rely on well-chosen percents. You must decide which percents you need. Here is a hint: *if there is an explanatory-response relationship, compare the conditional distributions of the response variable for the separate values of the explanatory variable.* If you think that sex influences young adults' opinions about their chances of getting rich by age 30, compare the conditional distributions of opinion for women and for men, as in Example 6.3.

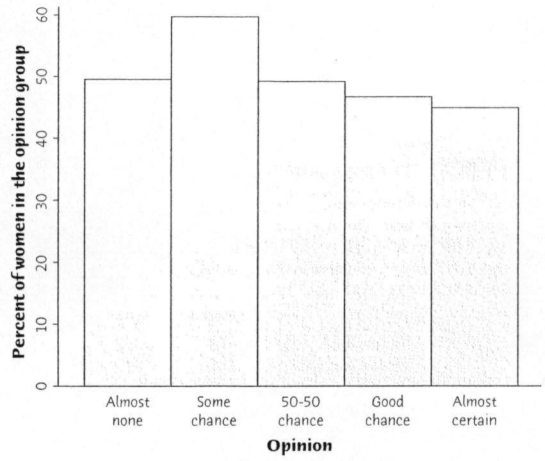


FIGURE 6.3
Bar graph comparing the percents of females among those who hold each opinion about their chances of getting rich by age 30.

APPLY YOUR KNOWLEDGE

6.3 Attitudes toward recycled products. Exercise 6.1 gives data on the opinions of people who have and have not bought coffee filters made from recycled paper. To see the relationship between opinion and experience with the product, find the conditional distributions of opinion (the response variable) for buyers and nonbuyers. What do you conclude?

6.4 Undergraduates' ages. Exercise 6.2 gives Census Bureau data describing the age and sex of all American college undergraduates. We suspect that the percent of women is higher among older students than in the traditional 18 to 24 college age group. Do the data support this suspicion? Follow the four-step process as illustrated in Example 6.3.

6.5 Marginal distributions aren't the whole story. Here are the row and column totals for a two-way table with two rows and two columns:

a	b	50
c	d	50
60	40	100

Make up *two different* sets of counts a, b, c, and d for the body of the table that give these same totals. This shows that the relationship between two variables cannot be obtained from the two individual distributions of the variables.

Simpson's paradox

As is the case with quantitative variables, the effects of lurking variables can change or even reverse relationships between two categorical variables. Here is an example that demonstrates the surprises that can await the unsuspecting user of data.

Ashley/Cooper/PICIMPACT/CORBIS

EXAMPLE 6.3 Do medical helicopters save lives?

Accident victims are sometimes taken by helicopter from the accident scene to a hospital. Helicopters save time. Do they also save lives? Let's compare the percents of accident victims who die with helicopter evacuation and with the usual transport to a hospital by road. Here are hypothetical data that illustrate a practical difficulty:[4]

	Helicopter	Road
Victim died	64	260
Victim survived	136	840
Total	200	1100

We see that 32% (64 out of 200) of helicopter patients died, but only 24% (260 out of 1100) of the others did. That seems discouraging.

The explanation is that the helicopter is sent mostly to serious accidents, so that the victims transported by helicopter are more often seriously injured. They are more likely to die with or without helicopter evacuation. Here are the same data broken down by the seriousness of the accident:

Serious Accidents		
	Helicopter	Road
Died	48	60
Survived	52	40
Total	100	100

Less Serious Accidents		
	Helicopter	Road
Died	16	200
Survived	84	800
Total	100	1000

Inspect these tables to convince yourself that they describe the same 1300 accident victims as the original two-way table. For example, 200 (100 + 100) were moved by helicopter, and 64 (48 + 16) of these died.

Among victims of serious accidents, the helicopter saves 52% (52 out of 100) compared with 40% for road transport. If we look only at less serious accidents, 84% of those transported by helicopter survive, versus 80% of those transported by road. Both groups of victims have a higher survival rate when evacuated by helicopter. •

How can it happen that the helicopter does better for both groups of victims but worse when all victims are lumped together? Examining the data makes the explanation clear. Half the helicopter transport patients are from serious accidents, compared with only 100 of the 1100 road transport patients. So the helicopter carries patients who are more likely to die. The seriousness of the accident was a lurking variable that, until we uncovered it, hid the true relationship between survival and mode of transport to a hospital. Example 6.4 illustrates *Simpson's paradox*.

SIMPSON'S PARADOX

An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson's paradox.**

The lurking variable in Simpson's paradox is categorical. That is, it breaks the individuals into groups, as when accident victims are classified as injured in a "serious accident" or a "less serious accident." Simpson's paradox is just an extreme form of the fact that observed associations can be misleading when there are lurking variables.

APPLY YOUR KNOWLEDGE

6.6 Airline flight delays. Here are the numbers of flights on time and delayed for two airlines at five airports in one month. Overall on-time percents for each airline are often reported in the news. The airport that flights serve is a lurking variable that can make such reports misleading.[5]

	Alaska Airlines		America West	
	On time	Delayed	On time	Delayed
Los Angeles	497	62	694	117
Phoenix	221	12	4840	415
San Diego	212	20	383	65
San Francisco	503	102	320	129
Seattle	1841	305	201	61

(a) What percent of all Alaska Airlines flights were delayed? What percent of all America West flights were delayed? These are the numbers usually reported.

(b) Now find the percent of delayed flights for Alaska Airlines at each of the five airports. Do the same for America West.

(c) America West did worse at *every one* of the five airports, yet did better overall. That sounds impossible. Explain carefully, referring to the data, how this can happen. (The weather in Phoenix and Seattle lies behind this example of Simpson's paradox.)

6.7 Which hospital is safer? To help consumers make informed decisions about health care, the government releases data about patient outcomes in hospitals. You want to compare Hospital A and Hospital B, which serve your community. The table presents data on all patients undergoing surgery in a recent time period. The data include the condition of the patient ("good" or "poor") before the surgery. "Survived" means that the patient lived at least 6 weeks following surgery.

	Good Condition			Poor Condition	
	Hospital A	Hospital B		Hospital A	Hospital B
Died	6	8	Died	57	8
Survived	594	592	Survived	1443	192
Total	600	600	Total	1500	200

(a) Compare percents to show that Hospital A has a higher survival rate for both groups of patients.

(b) Combine the data into a single two-way table of outcome ("survived" or "died") by hospital (A or B). The local paper reports just these overall survival rates. Which hospital has the higher rate?

(c) Explain from the data, in language that a reporter can understand, how Hospital B can do better overall even though Hospital A does better for both groups of patients.

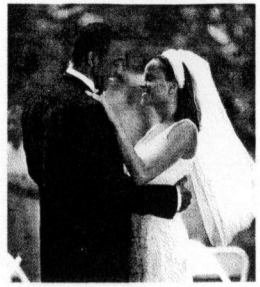
CHAPTER 6 SUMMARY

- A **two-way table** of counts organizes data about two categorical variables. Values of the **row variable** label the rows that run across the table, and values of the **column variable** label the columns that run down the table. Two-way tables are often used to summarize large amounts of information by grouping outcomes into categories.
- The **row totals** and **column totals** in a two-way table give the **marginal distributions** of the two individual variables. It is clearer to present these distributions as percents of the table total. Marginal distributions tell us nothing about the relationship between the variables.
- There are two sets of **conditional distributions** for a two-way table: the distributions of the row variable for each fixed value of the column variable, and the distributions of the column variable for each fixed value of the row variable. Comparing one set of conditional distributions is one way to describe the association between the row and the column variables.
- To find the **conditional distribution** of the row variable for one specific value of the column variable, look only at that one column in the table. Find each entry in the column as a percent of the column total.
- **Bar graphs** are a flexible means of presenting categorical data. There is no single best way to describe an association between two categorical variables.
- A comparison between two variables that holds for each individual value of a third variable can be changed or even reversed when the data for all values of the third variable are combined. This is **Simpson's paradox.** Simpson's paradox is an example of the effect of lurking variables on an observed association.

CHECK YOUR SKILLS

The National Longitudinal Study of Adolescent Health interviewed several thousand teens (grades 7 to 12). One question asked was "What do you think are the chances you will be married in the next 10 years?" Here is a two-way table of the responses by sex:[6]

Opinion	Female	Male
Almost no chance	119	103
Some chance but probably not	150	171
A 50-50 chance	447	512
A good chance	735	710
Almost certain	1174	756



Comstock Images/Age fotostock

Exercises 6.8 to 6.16 are based on this table.

6.8 How many individuals are described by this table?

(a) 2625 (b) 4877 (c) Need more information

6.9 How many females were among the respondents?

(a) 2625 (b) 4877 (c) Need more information

6.10 The percent of females among the respondents was

(a) about 46%. (b) about 54%. (c) about 86%.

6.11 Your percent from the previous exercise is part of

(a) the marginal distribution of sex.

(b) the marginal distribution of opinion about marriage.

(c) the conditional distribution of sex among adolescents with a given opinion.

6.12 What percent of females thought that they were almost certain to be married in the next 10 years?

(a) about 40% (b) about 45% (c) about 61%

6.13 Your percent from the previous exercise is part of

(a) the marginal distribution of opinion about marriage.

(b) the conditional distribution of sex among those who thought they were almost certain to be married.

(c) the conditional distribution of opinion about marriage among women.

6.14 What percent of those who thought they were almost certain to be married were female?

(a) about 40% (b) about 45% (c) about 61%

6.15 Your percent from the previous exercise is part of

(a) the marginal distribution of opinion about marriage.

(b) the conditional distribution of sex among those who thought they were almost certain to be married.

(c) the conditional distribution of opinion about marriage among women.

6.16 A bar graph showing the conditional distribution of opinion among female respondents would have

(a) 2 bars. (b) 5 bars. (c) 10 bars.

6.17 A college looks at the grade point average (GPA) of its full-time and part-time students. Grades in science courses are generally lower than grades in other courses. There are few science majors among part-time students but many science majors among full-time students. The college finds that full-time students who are science majors have higher GPA than part-time students who are science majors. Full-time students who are not science majors also have higher GPA than part-time students who are not science majors. Yet part-time students as a group have higher GPA than full-time students. This finding is

(a) not possible: if both science and other majors who are full-time have higher GPA than those who are part-time, then all full-time students together must have higher GPA than all part-time students together.

(b) an example of Simpson's paradox: full-time students do better in both kinds of courses but worse overall because they take more science courses.

(c) due to comparing two conditional distributions that should not be compared.

CHAPTER 6 EXERCISES

6.18 Graduate school for men and women. The College of Liberal Arts at a large university looks at its graduate students classified by their sex and field of study. Here are the data:[7]

	Female	Male
English	136	89
Foreign languages	61	25
History	35	55
Philosophy	10	54
Political science	29	35

Find the two conditional distributions of field of study, one for women and one for men. Based on your calculations, describe the differences between women and men with a graph and in words.

6.19 Helping cocaine addicts. Will giving cocaine addicts an antidepressant drug help them break their addiction? An experiment assigned 24 chronic cocaine users to take the antidepressant drug desipramine, another 24 to take lithium, and another 24 to take a placebo. (Lithium is a standard drug to treat cocaine addiction. A placebo is a dummy pill, used so that the effect of being in the study but not taking any drug can be seen.) After three years, 14 of the 24 subjects in the desipramine group had remained free of cocaine, along with 6 of the 24 in the lithium group and 4 of the 24 in the placebo group.[8]

(a) Make up a two-way table of "Treatment received" by whether or not the subject remained free of cocaine.

(b) Compare the effectiveness of the three treatments in preventing use of cocaine by former addicts. Use percents and draw a bar graph. What do you conclude?

Marital status and job level. *We sometimes hear that getting married is good for your career. Table 6.2 presents data from one of the studies behind this generalization. To avoid gender effects, the investigators looked only at men. The data describe the marital status and the job level of all 8235 male managers and professionals employed by a large manufacturing firm.*[9] *The firm assigns each position a grade that reflects the value of that particular job to the company. The authors of the study grouped the many job grades into quarters. Grade 1 contains jobs in the lowest quarter of the job grades, and Grade 4 contains those in the highest quarter. Exercises 6.20 to 6.24 are based on these data.*

TABLE 6.2 Marital status and job level

	MARITAL STATUS				
JOB GRADE	SINGLE	MARRIED	DIVORCED	WIDOWED	TOTAL
1	58	874	15	8	955
2	222	3927	70	20	4239
3	50	2396	34	10	2490
4	7	533	7	4	551
Total	337	7730	126	42	8235

6.20 Marginal distributions. Give (in percents) the two marginal distributions, for marital status and for job grade. Do each of your two sets of percents add to exactly 100%? If not, why not?

6.21 Percents. What percent of single men hold Grade 1 jobs? What percent of Grade 1 jobs are held by single men?

6.22 Conditional distribution. Give (in percents) the conditional distribution of job grade among single men. Should your percents add to 100% (up to roundoff error)?

6.23 Marital status and job grade. One way to see the relationship is to look at who holds Grade 1 jobs.

(a) There are 874 married men with Grade 1 jobs, and only 58 single men with such jobs. Explain why these counts by themselves don't describe the relationship between marital status and job grade.

(b) Find the percent of men in each marital status group who have Grade 1 jobs. Then find the percent in each marital group who have Grade 4 jobs. What do these percents say about the relationship?

6.24 Association is not causation. The data in Table 6.2 show that single men are more likely to hold lower-grade jobs than are married men. We should not conclude that