CHAPTER 2

Data

AP Stats Summer Assignment

-Read Chapters #2 & #3

-Do the circled problems at the end of each chapter

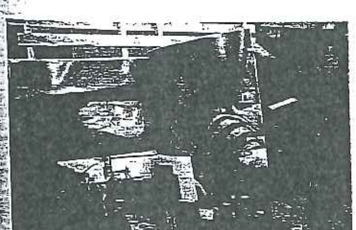
Mr. Steve Strayer

EMAIL: sstrayer@cps.k12.in.us

Many years ago, most stores in small towns knew their customers personally. If you walked into the hobby shop, the owner might tell you about a new bridge that had come in for your Lionel train set. The tailor knew your dad's size, and the hairdresser knew how your mom liked her hair. There are still some stores like that around today, but we're increasingly likely to shop at large stores, by phone, or on the Internet. Even so, when you phone an 800 number to buy new running shoes, customer service representatives may call you by your first name or ask about the socks you bought 6 weeks ago. Or the company may send an e-mail in October offering new head warmers for winter running. This company has millions of customers, and you called without identifying yourself. How did the sales rep know who you are, where you live, and what you had bought?

"Data is king at Amazon. Clickstream and purchase data are the crown jewels at Amazon. They help us build features to personalize the Web site experience."

—Ronny Kohavi, Director of Data Mining and Personalization, Amazon.com



The answer is data. Collecting data on their customers, transactions, and sales lets companies track their inventory and helps them predict what their customers prefer. These data can help them predict what their customers may buy in the future so they know how much of each item to stock. The store can use the data and what it learns from the data to improve customer service, mimicking the kind of personal attention a shopper had 50 years ago.

Amazon.com opened for business in July 1995, billing itself as "Earth's Biggest Bookstore." By 1997, Amazon had a catalog of more than 2.5 million book titles and had sold books to more than 1.5 million customers in 150 countries. In 2006, the company's revenue reached $10.7 billion. Amazon has expanded into selling a wide selection of merchandise, from $400,000 necklaces[1] to yak cheese from Tibet to the largest book in the world.

Amazon is constantly monitoring and evolving its Web site to serve its customers better and maximize sales performance. To decide which changes to make to the site, the company experiments, collecting data and analyzing what works best. When you visit the Amazon Web site, you may encounter a different look or different suggestions and offers. Amazon statisticians want to know whether you'll follow the links offered, purchase the items suggested, or even spend a

[1] Please get credit card approval before purchasing online.

longer time browsing the site. As Ronny Kohavi, director of Data Mining and Personalization, said, "Data trumps intuition. Instead of using our intuition, we experiment on the live site and let our customers tell us what works for them."

But What *Are* Data?

THE W'S:
WHO
WHAT
and in what units
WHEN
WHERE
WHY
HOW

We bet you thought you knew this instinctively. Think about it for a minute. What exactly *do* we mean by "data"?

Do data have to be numbers? The amount of your last purchase in dollars is numerical data, but some data record names or other labels. The names in Amazon.com's database are data, but not numerical.

Sometimes, data can have values that look like numerical values but are just numerals serving as labels. This can be confusing. For example, the ASIN (Amazon Standard Item Number) of a book, like 0321570448, may have a numerical value, but it's really just another name for *Stats: Modeling the World.*

Data values, no matter what kind, are useless without their context. Newspaper journalists know that the lead paragraph of a good story should establish the "Five W's": *Who, What, When, Where,* and (if possible) *Why.* Often we add *How* to the list as well. Answering these questions can provide the **context** for data values. The answers to the first two questions are essential. If you can't answer *Who* and *What,* you don't have **data,** and you don't have any useful information.

Data Tables

Here are some data Amazon might collect:

B0000010AA	10.99	Chris G.	902	15783947	15.98	Kansas	Illinois	Boston
Canada	Samuel P.	Orange County	N	B000068ZVQ	Bad Blood	Nashville	Katherine H.	N
Mammals	10783489	Ohio	N	Chicago	12837593	11.99	Massachusetts	16.99
312	Monique D.	10675489	413	B00000I5Y6	440	B000002BK9	Let Go	Y

Activity: What Is (Are) Data? Do you really know what's data and what's just numbers?

Try to guess what they represent. Why is that hard? Because these data have no *context.* If we don't know *Who* they're about or *What* they measure, these values are meaningless. We can make the meaning clear if we organize the values into a **data table** such as this one:

Purchase Order	Name	Ship to State/Country	Price	Area Code	Previous CD Purchase	Gift?	ASIN	Artist
10675489	Katharine H.	Ohio	10.99	440	Nashville	N	B00000I5Y6	Kansas
10783489	Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
12837593	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
15783947	Monique D.	Canada	11.99	902	Let Go	N	B0000010AA	Mammals

Now we can see that these are four purchase records, relating to CD orders from Amazon. The column titles tell *What* has been recorded. The rows tell us *Who.* But be careful. Look at all the variables to see *Who* the variables are about. Even if people are involved, they may not be the *Who* of the data. For example, the *Who* here are the purchase orders (not the people who made the purchases).

A common place to find the *Who* of the table is the leftmost column. The other W's might have to come from the company's database administrator.[2]

Who

In general, the rows of a data table correspond to individual **cases** about *Whom* (or about which—if they're not people) we record some characteristics. These cases go by different names, depending on the situation. Individuals who answer a survey are referred to as *respondents*. People on whom we experiment are *subjects* or (in an attempt to acknowledge the importance of their role in the experiment) *participants*, but animals, plants, Web sites, and other inanimate subjects are often just called *experimental units*. In a database, rows are called *records*—in this example, purchase records. Perhaps the most generic term is **cases.** In the Amazon table, the cases are the individual CD orders.

Sometimes people just refer to data values as *observations*, without being clear about the *Who*. Be sure you know the *Who* of the data, or you may not know what the data say.

Often, the cases are a **sample** of cases selected from some larger **population** that we'd like to understand. Amazon certainly cares about its customers, but also wants to know how to attract all those other Internet users who may never have made a purchase from Amazon's site. To be able to generalize from the sample of cases to the larger population, we'll want the sample to be *representative* of that population—a kind of snapshot image of the larger world.

A S ***Activity:* Consider the Context . . .** Can you tell who's *Who* and what's *What?* And *Why?* This activity offers real-world examples to help you practice identifying the context.

FOR EXAMPLE **Identifying the "Who"**

In March 2007, *Consumer Reports* published an evaluation of large-screen, high-definition television sets (HDTVs). The magazine purchased and tested 98 different models from a variety of manufacturers.

Question: Describe the population of interest, the sample, and the *Who* of this study.

The magazine is interested in the performance of all HDTVs currently being offered for sale. It tested a sample of 98 sets, the "Who" for these data. Each HDTV set represents all similar sets offered by that manufacturer.

What and Why

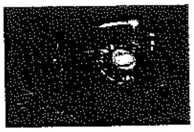
The characteristics recorded about each individual are called **variables.** These are usually shown as the columns of a data table, and they should have a name that identifies *What* has been measured. Variables may seem simple, but to really understand your variables, you must *Think* about what you want to know.

Although area codes are numbers, do we use them that way? Is 610 twice 305? Of course it is, but is that the question? Why would we want to know whether Allentown, PA (area code 610), is twice Key West, FL (305)? Variables play different roles, and you can't tell a variable's role just by looking at it.

Some variables just tell us what group or category each individual belongs to. Are you male or female? Pierced or not? . . . What kinds of things can we learn about variables like these? A natural start is to *count* how many cases belong in each category. (Are you listening to music while reading this? We could count

[2] In database management, this kind of information is called "metadata."

It is wise to be careful. The *What* and *Why* of area codes are not as simple as they may first seem. When area codes were first introduced, AT&T was still the source of all telephone equipment, and phones had dials.



To reduce wear and tear on the dials, the area codes with the lowest digits (for which the dial would have to spin least) were assigned to the most populous regions—those with the most phone numbers and thus the area codes most likely to be dialed. New York City was assigned 212, Chicago 312, and Los Angeles 213, but rural upstate New York was given 607, Joliet was 815, and San Diego 619. For that reason, at one time the numerical value of an area code could be used to guess something about the population of its region. Now that phones have push-buttons, area codes have finally become just categories.

the number of students in the class who were and the number who weren't.) We'll look for ways to compare and contrast the sizes of such categories.

Some variables have measurement **units.** Units tell how each value has been measured. But, more importantly, units such as yen, cubits, carats, angstroms, nanoseconds, miles per hour, or degrees Celsius tell us the *scale* of measurement. The units tell us how much of something we have or how far apart two values are. Without units, the values of a measured variable have no meaning. It does little good to be promised a raise of 5000 a year if you don't know whether it will be paid in euros, dollars, yen, or Estonian krooni.

By international agreement, the International System of Units links together all systems of weights and measures. There are seven base units from which all other physical units are derived:

• Distance	Meter
• Mass	Kilogram
• Time	Second
• Electric current	Ampere
• Temperature	°Kelvin
• Amount of substance	Mole
• Intensity of light	Candela

What kinds of things can we learn about measured variables? We can do a lot more than just counting categories. We can look for patterns and trends. (How much did you pay for your last movie ticket? What is the range of ticket prices available in your town? How has the price of a ticket changed over the past 20 years?)

When a variable names categories and answers questions about how cases fall into those categories, we call it a **categorical variable.**[3] When a measured variable with units answers questions about the quantity of what is measured, we call it a **quantitative variable.** These types can help us decide what to do with a variable, but they are really more about what we hope to learn from a variable than about the variable itself. It's the questions we ask a variable (the *Why* of our analysis) that shape how we think about it and how we treat it.

A S *Activity:* **Recognize variables measured in a variety of ways.** This activity shows examples of the many ways to measure data.

A S *Activities:* **Variables.** Several activities show you how to begin working with data in your statistics package.

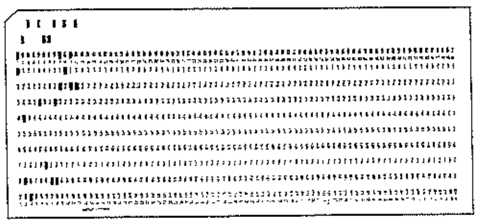
Some variables can answer questions only about categories. If the values of a variable are words rather than numbers, it's a good bet that it is categorical. But some variables can answer both kinds of questions. Amazon could ask for your *Age* in years. That seems quantitative, and would be if the company wanted to know the average age of those customers who visit their site after 3 a.m. But suppose Amazon wants to decide which CD to offer you in a special deal—one by Raffi, Blink-182, Carly Simon, or Mantovani—and needs to be sure to have adequate supplies on hand to meet the demand. Then thinking of your age in one of the categories—child, teen, adult, or senior—might be more useful. If it isn't clear whether a variable is categorical or quantitative, think about *Why* you are looking at it and what you want it to tell you.

A typical course evaluation survey asks, "How valuable do you think this course will be to you?": 1 = Worthless; 2 = Slightly; 3 = Middling; 4 = Reasonably; 5 = Invaluable. Is *Educational Value* categorical or quantitative? Once again, we'll look to the *Why.* A teacher might just count the number of students who gave each response for her course, treating *Educational Value* as a categorical variable. When she wants to see whether the course is improving, she might treat the responses as the *amount* of perceived value—in effect, treating the variable as quantitative. But what are the units? There is certainly an *order* of perceived worth: Higher numbers indicate higher perceived worth. A course that averages 4.5 seems more valuable than one that averages 2, but we should be careful about treating *Educational Value* as

[3] You may also see it called a *qualitative variable.*

purely quantitative. To treat it as quantitative, she'll have to imagine that it has "educational value units" or some similar arbitrary construction. Because there are no natural units, she should be cautious. Variables like this that report order without natural units are often called "ordinal" variables. But saying "that's an ordinal variable" doesn't get you off the hook. You must still look to the *Why* of your study to decide whether to treat it as categorical or quantitative.

One tradition that hangs on in some quarters is to name variables with cryptic abbreviations written in uppercase letters. This can be traced back to the 1960s, when the very first statistics computer programs were controlled with instructions punched on cards. The earliest punch card equipment used only uppercase letters, and the earliest statistics programs limited variable names to six or eight characters, so variables were called things like PRSRF3. Modern programs do not have such restrictive limits, so there is no reason for variable names that you wouldn't use in an ordinary sentence.



FOR EXAMPLE Identifying "What" and "Why" of HDTVs.

Recap: A *Consumer Reports* article about 98 HDTVs lists each set's manufacturer, cost, screen size, type (LCD, plasma, or rear projection), and overall performance score (0–100).

Question: Are these variables categorical or quantitative? Include units where appropriate, and describe the "Why" of this investigation.

The "what" of this article includes the following variables:

- manufacturer (categorical);
- cost (in dollars, quantitative);
- screen size (in inches, quantitative);
- type (categorical);
- performance score (quantitative).

The magazine hopes to help consumers pick a good HDTV set.

Counts Count

In Statistics, we often count things. When Amazon considers a special offer of free shipping to customers, it might first analyze how purchases are shipped. They'd probably start by counting the number of purchases shipped by ground transportation, by second-day air, and by overnight air. Counting is a natural way to summarize the categorical variable *Shipping Method*. So every time we see counts, does that mean the variable is categorical? Actually, no.

We also use counts to measure the amounts of things. How many songs are on your digital music player? How many classes are you taking this semester? To measure these quantities, we'd naturally count. The variables (*Songs, Classes*) would be quantitative, and we'd consider the units to be "number of . . ." or, generically, just "counts" for short.

So we use counts in two different ways. When we count the cases in each category of a categorical variable, the category labels are the *What* and the individuals counted are the *Who* of our data. The counts themselves are not the

> **A S** *Activity:* **Collect data in an experiment on yourself.** With the computer, you can experiment on yourself and then save the data. Go on to the subsequent related activities to check your understanding.

data, but are something we summarize about the data. Amazon counts the number of purchases in each category of the categorical variable *Shipping Method.* For this purpose (the *Why*), the *What* is shipping method and the *Who* is purchases.

Shipping Method	Number of Purchases
Ground	20,345
Second-day	7,890
Overnight	5,432

Other times our focus is on the amount of something, which we measure by counting. Amazon might record the number of teenage customers visiting thei site each month to track customer growth and forecast CD sales (the *Why*). Now the *What* is *Teens,* the *Who* is *Months,* and the units are *Number of Teenage Cus-tomers. Teen* was a category when we looked at the categorical variable *Age.* Bu now it is a quantitative variable in its own right whose amount is measured by counting the number of customers.

Month	Number of Teenage Customers
January	123,456
February	234,567
March	345,678
April	456,789
May	...
...	...

Identifying Identifiers

What's your student ID number? It is numerical, but is it a quantitative variable No, it doesn't have units. Is it categorical? Yes, but it is a special kind. Look at hov many categories there are and at how many individuals are in each. There are a many categories as individuals and only one individual in each category. Whil it's easy to count the totals for each category, it's not very interesting. Amazo wants to know who you are when you sign in again and doesn't want to confus you with some other customer. So it assigns you a unique identifier.

Identifier variables themselves don't tell us anything useful about the cate-gories because we know there is exactly one individual in each. However, they ar crucial in this age of large data sets. They make it possible to combine data fro different sources, to protect confidentiality, and to provide unique labels. Th variables *UPS Tracking Number, Social Security Number,* and Amazon's *ASIN* are a examples of identifier variables.

You'll want to recognize when a variable is playing the role of an identifier s you won't be tempted to analyze it. There's probably a list of unique ID numbe for students in a class (so they'll each get their own grade confidentially), but yo might worry about the professor who keeps track of the average of these numbe from class to class. Even though this year's average ID number happens to b higher than last's, it doesn't mean that the students are better.

Where, When, and How

A S *Self-Test:* **Review concepts about data.** Like the Just Checking sections of this textbook, but interactive. (Usually, we won't reference the *ActivStats* self-tests here, but look for one whenever you'd like to check your understanding or review material.)

We must know *Who, What,* and *Why* to analyze data. Without knowing these three, we don't have enough to start. Of course, we'd always like to know more. The more we know about the data, the more we'll understand about the world.

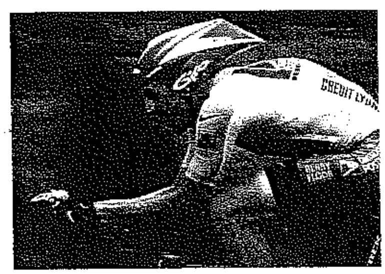
If possible, we'd like to know the **When** and **Where** of data as well. Values recorded in 1803 may mean something different than similar values recorded last year. Values measured in Tanzania may differ in meaning from similar measurements made in Mexico.

How the data are collected can make the difference between insight and nonsense. As we'll see later, data that come from a voluntary survey on the Internet are almost always worthless. One primary concern of Statistics, to be discussed in Part III, is the design of sound methods for collecting data.

Throughout this book, whenever we introduce data, we'll provide a margin note listing the W's (and H) of the data. It's a habit we recommend. The first step of any data analysis is to know why you are examining the data (what you want to know), whom each row of your data table refers to, and what the variables (the columns of the table) record. These are the *Why*, the *Who*, and the *What*. Identifying them is a key part of the *Think* step of any analysis. Make sure you know all three before you proceed to *Show* or *Tell* anything about the data.

JUST CHECKING

In the 2003 Tour de France, Lance Armstrong averaged 40.94 kilometers per hour (km/h) for the entire course, making it the fastest Tour de France in its 100-year history. In 2004, he made history again by winning the race for an unprecedented sixth time. In 2005, he became the only 7-time winner and once again set a new record for the fastest average speed. You can find data on all the Tour de France races on the DVD. Here are the first three and last ten lines of the data set. Keep in mind that the entire data set has nearly 100 entries.



1. List as many of the W's as you can for this data set.
2. Classify each variable as categorical or quantitative; if quantitative, identify the units.

Year	Winner	Country of origin	Total time (h/min/s)	Avg. speed (km/h)	Stages	Total distance ridden (km)	Starting riders	Finishing riders
1903	Maurice Garin	France	94.33.00	25.3	6	2428	60	21
1904	Henri Cornet	France	96.05.00	24.3	6	2388	88	23
1905	Louis Trousselier	France	112.18.09	27.3	11	2975	60	24
⋮								
1999	Lance Armstrong	USA	91.32.16	40.30	20	3687	180	141
2000	Lance Armstrong	USA	92.33.08	39.56	21	3662	180	128
2001	Lance Armstrong	USA	86.17.28	40.02	20	3453	189	144
2002	Lance Armstrong	USA	82.05.12	39.93	20	3278	189	153
2003	Lance Armstrong	USA	83.41.12	40.94	20	3427	189	147
2004	Lance Armstrong	USA	83.36.02	40.53	20	3391	188	147
2005	Lance Armstrong	USA	86.15.02	41.65	21	3608	189	155
2006	Óscar Periero	Spain	89.40.27	40.78	20	3657	176	139
2007	Alberto Contador	Spain	91.00.26	38.97	20	3547	189	141
2008	Carlos Sastre	Spain	87.52.52	40.50	21	3559	199	145

5. **The news.** Find a newspaper or magazine article in which some data are reported. For the data discussed in the article, answer the questions above. Include a copy of the article with your report.

6. **The Internet.** Find an Internet source that reports on a study and describes the data. Print out the description and answer the questions above.

7. **Bicycle safety.** Ian Walker, a psychologist at the University of Bath, wondered whether drivers treat bicycle riders differently when they wear helmets. He rigged his bicycle with an ultrasonic sensor that could measure how close each car was that passed him. He then rode on alternating days with and without a helmet. Out of 2500 cars passing him, he found that when he wore his helmet, motorists passed 3.35 inches closer to him, on average, than when his head was bare. [*NY Times*, Dec. 10, 2006]

8. **Investments.** Some companies offer 401(k) retirement plans to employees, permitting them to shift part of their before-tax salaries into investments such as mutual funds. Employers typically match 50% of the employees' contribution up to about 6% of salary. One company, concerned with what it believed was a low employee participation rate in its 401(k) plan, sampled 30 other companies with similar plans and asked for their 401(k) participation rates.

9. **Honesty.** Coffee stations in offices often just ask users to leave money in a tray to pay for their coffee, but many people cheat. Researchers at Newcastle University alternately taped two posters over the coffee station. During one week, it was a picture of flowers; during the other, it was a pair of staring eyes. They found that the average contribution was significantly higher when the eyes poster was up than when the flowers were there. Apparently, the mere feeling of being watched—even by eyes that were not real—was enough to encourage people to behave more honestly. [*NY Times*, Dec. 10, 2006]

10. **Movies.** Some motion pictures are profitable and others are not. Understandably, the movie industry would like to know what makes a movie successful. Data from 120 first-run movies released in 2005 suggest that longer movies actually make *less* profit.

11. **Fitness.** Are physically fit people less likely to die of cancer? An article in the May 2002 issue of *Medicine and Science in Sports and Exercise* reported results of a study that followed 25,892 men aged 30 to 87 for 10 years. The most physically fit men had a 55% lower risk of death from cancer than the least fit group.

12. **Molten iron.** The Cleveland Casting Plant is a large, highly automated producer of gray and nodular iron automotive castings for Ford Motor Company. The company is interested in keeping the pouring temperature of the molten iron (in degrees Fahrenheit) close to the specified value of 2550 degrees. Cleveland Casting measured the pouring temperature for 10 randomly selected crankshafts.

START HERE !

(Exercises 13–26) For each description of data, identify the W's, name the variables, specify for each variable whether its use indicates that it should be treated as categorical or quantitative, and, for any quantitative variable, identify the units in which it was measured (or note that they were not provided).

13. **Weighing bears.** Because of the difficulty of weighing a bear in the woods, researchers caught and measured 54 bears, recording their weight, neck size, length, and sex. They hoped to find a way to estimate weight from the other, more easily determined quantities.

14. **Schools.** The State Education Department requires local school districts to keep these records on all students: age, race or ethnicity, days absent, current grade level, standardized test scores in reading and mathematics, and any disabilities or special educational needs.

15. **Arby's menu.** A listing posted by the Arby's restaurant chain gives, for each of the sandwiches it sells, the type of meat in the sandwich, the number of calories, and the serving size in ounces. The data might be used to assess the nutritional value of the different sandwiches.

16. **Age and party.** The Gallup Poll conducted a representative telephone survey of 1180 American voters during the first quarter of 2007. Among the reported results were the voter's region (Northeast, South, etc.), age, party affiliation, and whether or not the person had voted in the 2006 midterm congressional election.

17. **Babies.** Medical researchers at a large city hospital investigating the impact of prenatal care on newborn health collected data from 882 births during 1998–2000. They kept track of the mother's age, the number of weeks the pregnancy lasted, the type of birth (cesarean, induced, natural), the level of prenatal care the mother had (none, minimal, adequate), the birth weight and sex of the baby, and whether the baby exhibited health problems (none, minor, major).

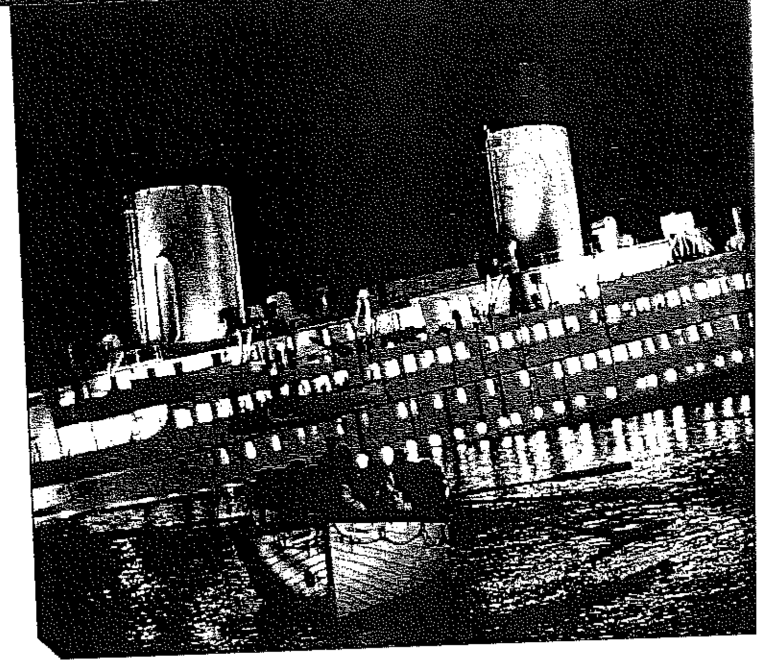
18. **Flowers.** In a study appearing in the journal *Science*, a research team reports that plants in southern England are flowering earlier in the spring. Records of the first flowering dates for 385 species over a period of 47 years show that flowering has advanced an average of 15 days per decade, an indication of climate warming, according to the authors.

19. **Herbal medicine.** Scientists at a major pharmaceutical firm conducted an experiment to study the effectiveness of an herbal compound to treat the common cold. They exposed each patient to a cold virus, then gave them either the herbal compound or a sugar solution known to have no effect on colds. Several days later they assessed each patient's condition, using a cold severity scale ranging from 0 to 5. They found no evidence of the benefits of the compound.

20. **Vineyards.** Business analysts hoping to provide information helpful to American grape growers compiled these data about vineyards: size (acres), number of years in existence, state, varieties of grapes grown, average case price, gross sales, and percent profit.

CHAPTER 3

Displaying and Describing Categorical Data



WHO	People on the *Titanic*
WHAT	Survival status, age, sex, ticket class
WHEN	April 14, 1912
WHERE	North Atlantic
HOW	A variety of sources and Internet sites
WHY	Historical interest

What happened on the *Titanic* at 11:40 on the night of April 14, 1912, is well known. Frederick Fleet's cry of "Iceberg, right ahead" and the three accompanying pulls of the crow's nest bell signaled the beginning of a nightmare that has become legend. By 2:15 a.m., the *Titanic*, thought by many to be unsinkable, had sunk, leaving more than 1500 passengers and crew members on board to meet their icy fate.

Here are some data about the passengers and crew aboard the *Titanic*. Each case (row) of the data table represents a person on board the ship. The variables are the person's *Survival* status (Dead or Alive), *Age* (Adult or Child), *Sex* (Male or Female), and ticket *Class* (First, Second, Third, or Crew).

The problem with a data table like this—and in fact with all data tables—is that you can't *see* what's going on. And seeing is just what we want to do. We need ways to show the data so that we can see patterns, relationships, trends, and exceptions.

A S *Video:* **The Incident** tells the story of the *Titanic*, and includes rare film footage.

Survival	Age	Sex	Class
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Alive	Adult	Female	First
Dead	Adult	Male	Third
Dead	Adult	Male	Crew

Table 3.1
Part of a data table showing four variables fo nine people aboard the *Titanic*.

The Three Rules of Data Analysis

FIGURE 3.1 *A Picture to Tell a Story*

Florence Nightingale (1820–1910), a founder of modern nursing, was also a pioneer in health management, statistics, and epidemiology. She was the first female member of the British Statistical Society and was granted honorary membership in the newly formed American Statistical Association.

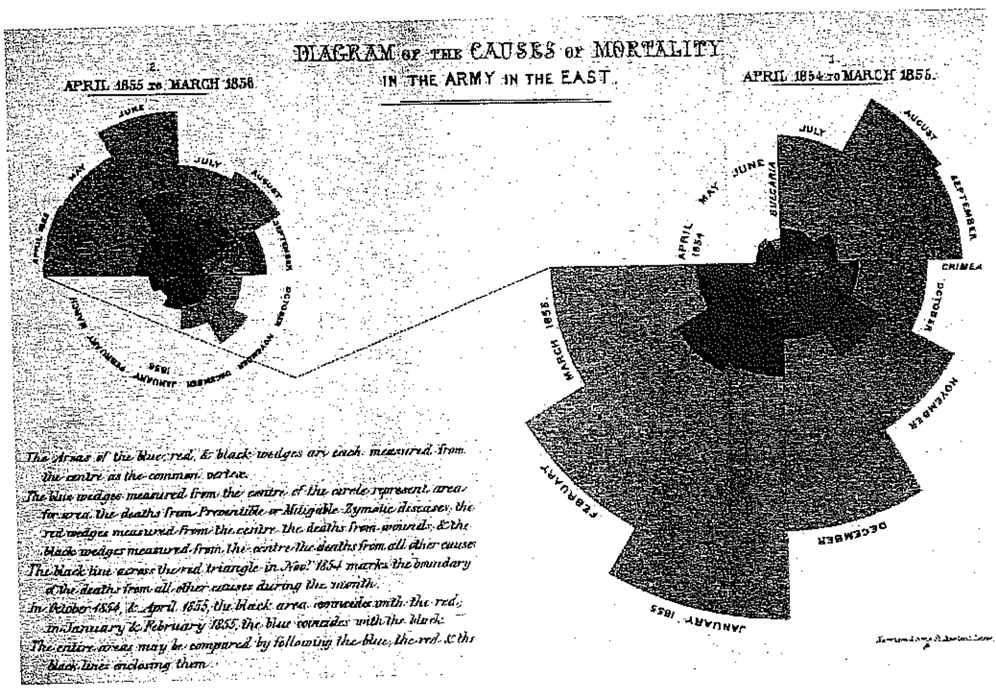
To argue forcefully for better hospital conditions for soldiers, she and her colleague, Dr. William Farr, invented this display, which showed that in the Crimean War, far more soldiers died of illness and infection than of battle wounds. Her campaign succeeded in improving hospital conditions and nursing for soldiers.

Florence Nightingale went on to apply statistical methods to a variety of important health issues and published more than 200 books, reports, and pamphlets during her long and illustrious career.

So, what should we do with data like these? There are three things you should always do first with data:

1. **Make a picture.** A display of your data will reveal things you are not likely to see in a table of numbers and will help you to *Think* clearly about the patterns and relationships that may be hiding in your data.
2. **Make a picture.** A well-designed display will *Show* the important features and patterns in your data. A picture will also show you the things you did not expect to see: the extraordinary (possibly wrong) data values or unexpected patterns.
3. **Make a picture.** The best way to *Tell* others about your data is with a well-chosen picture.

These are the three rules of data analysis. There are pictures of data throughout the book, and new kinds keep showing up. These days, technology makes drawing pictures of data easy, so there is no reason not to follow the three rules.



Frequency Tables: Making Piles

> **A S** *Activity:* **Make and examine a table of counts.** Even data on something as simple as hair color can reveal surprises when you organize it in a data table.

To make a picture of data, the first thing we have to do is to make piles. Making piles is the beginning of understanding about data. We pile together things that seem to go together, so we can see how the cases distribute across different categories. For categorical data, piling is easy. We just count the number of cases corresponding to each category and pile them up.

One way to put all 2201 people on the *Titanic* into piles is by ticket *Class*, counting up how many had each kind of ticket. We can organize these counts into a **frequency table,** which records the totals and the category names.

Even when we have thousands of cases, a variable like ticket *Class*, with only a few categories, has a frequency table that's easy to read. A frequency table with dozens or hundreds of categories would be much harder to read. We use the names of the categories to label each row in the frequency table. For ticket *Class*, these are "First," "Second," "Third," and "Crew."

Class	Count
First	325
Second	285
Third	706
Crew	885

Table 3.2

A frequency table of the *Titanic* passengers.

Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

Table 3.3
A relative frequency table for the same data.

Counts are useful, but sometimes we want to know the fraction or **proportion** of the data in each category, so we divide the counts by the total number of cases. Usually we multiply by 100 to express these proportions as **percentages**. A **relative frequency table** displays the *percentages*, rather than the counts, of the values in each category. Both types of tables show how the cases are distributed across the categories. In this way, they describe the **distribution** of a categorical variable because they name the possible categories and tell how frequently each occurs.

The Area Principle

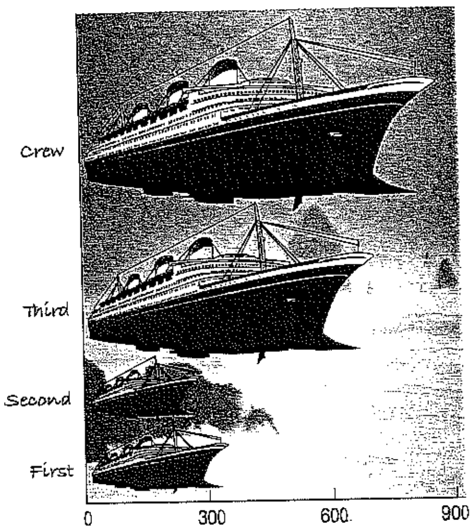


FIGURE 3.2
How many people were in each class on the Titanic? From this display, it looks as though the service must have been great, since most aboard were crew members. Although the length of each ship here corresponds to the correct number, the impression is all wrong. In fact, only about 40% were crew.

Now that we have the frequency table, we're ready to follow the three rules of data analysis and make a picture of the data. But a bad picture can distort our understanding rather than help it. Here's a graph of the *Titanic* data. What impression do you get about who was aboard the ship?

It sure looks like most of the people on the *Titanic* were crew members, with a few passengers along for the ride. That doesn't seem right. What's wrong? The lengths of the ships *do* match the totals in the table. (You can check the scale at the bottom.) However, experience and psychological tests show that our eyes tend to be more impressed by the *area* than by other aspects of each ship image. So, even though the *length* of each ship matches up with one of the totals, it's the associated *area* in the image that we notice. Since there were about 3 times as many crew as second-class passengers, the ship depicting the number of crew is about 3 times longer than the ship depicting second-class passengers, but it occupies about 9 times the area. As you can see from the frequency table (Table 3.2), that just isn't a correct impression.

The best data displays observe a fundamental principle of graphing data called the **area principle**. The area principle says that the area occupied by a part of the graph should correspond to the magnitude of the value it represents. Violations of the area principle are a common way to lie (or, since most mistakes are unintentional, we should say err) with Statistics.

Bar Charts

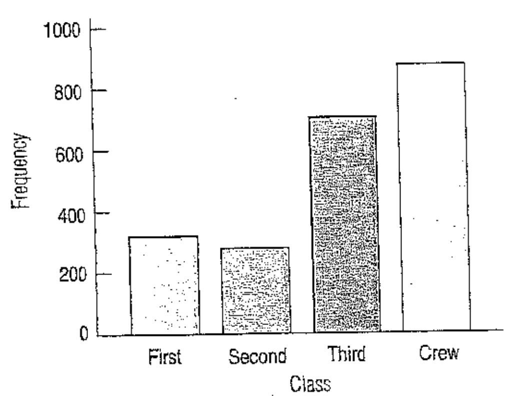
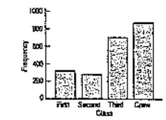
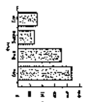


FIGURE 3.3 ***People on the* Titanic *by Ticket Class***
With the area principle satisfied, we can see the true distribution more clearly.

Here's a chart that obeys the area principle. It's not as visually entertaining as the ships, but it does give an *accurate* visual impression of the distribution. The height of each bar shows the count for its category. The bars are the same width, so their heights determine their areas, and the areas are proportional to the counts in each class. Now it's easy to see that the majority of people on board were *not* crew, as the ships picture led us to believe. We can also see that there were about 3 times as many crew as second-class passengers. And there were more than twice as many third-class passengers as either first- or second-class passengers, something you may have missed in the frequency table. Bar charts make these kinds of comparisons easy and natural.

A **bar chart** displays the distribution of a categorical variable showing the counts for each category next to each other for easy comparison. Bar charts should have small spaces between the bars to indicate that these are freestanding bars that could be rearranged into any order. The bars are lined up along a common base.

Usually they stick up like this  but sometimes they run sideways like this 

If we really want to draw attention to the relative *proportion* of passengers falling into each of these classes, we could replace the counts with percentages and use a **relative frequency bar chart.**

A S ***Activity:*** **Bar Charts.** Watch bar charts grow from data; then use your statistics package to create some bar charts for yourself.

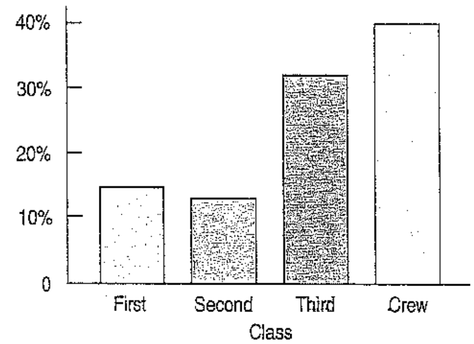


FIGURE 3.4

The relative frequency bar chart looks the same as the bar chart (Figure 3.3) but shows the proportion of people in each category rather than the counts.

For some reason, some computer programs give the name "bar chart" to any graph that uses bars. And others use different names according to whether the bars are horizontal or vertical. Don't be misled. "Bar chart" is the term for a *display of counts of a categorical variable* with bars.

Pie Charts

Another common display that shows how a whole group breaks into several categories is a pie chart. **Pie charts** show the whole group of cases as a circle. They slice the circle into pieces whose sizes are proportional to the fraction of the whole in each category.

Pie charts give a quick impression of how a whole group is partitioned into smaller groups. Because we're used to cutting up pies into 2, 4, or 8 pieces, pie charts are good for seeing relative frequencies near 1/2, 1/4, or 1/8. For example, you may be able to tell that the pink slice, representing the second-class passengers, is very close to 1/8 of the total. It's harder to see that there were about twice as many third-class as first-class passengers. Which category had the most passengers? Were there more crew or more third-class passengers? Comparisons such as these are easier in a bar chart.

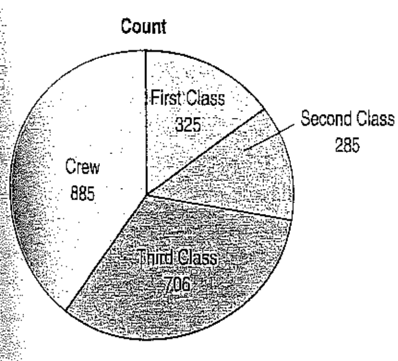


FIGURE 3.5 ***Number of Titanic passengers in each class***

Think before you draw. Our first rule of data analysis is *Make a picture*. But what kind of picture? We don't have a lot of options—yet. There's more to Statistics than pie charts and bar charts, and knowing when to use each type of graph is a critical first step in data analysis. That decision depends in part on what type of data we have.

It's important to check that the data are appropriate for whatever method of analysis you choose. Before you make a bar chart or a pie chart, always check the

Categorical Data Condition: The data are counts or percentages of individuals in categories.

If you want to make a relative frequency bar chart or a pie chart, you'll need to also make sure that the categories don't overlap so that no individual is counted twice. If the categories do overlap, you can still make a bar chart, but the percentages won't add up to 100%. For the *Titanic* data, either kind of display is appropriate because the categories don't overlap.

Throughout this course, you'll see that doing Statistics right means selecting the proper methods. That means you have to *Think* about the situation at hand. An important first step, then, is to check that the type of analysis you plan is appropriate. The Categorical Data Condition is just the first of many such checks.

Contingency Tables: Children and First-Class Ticket Holders First?

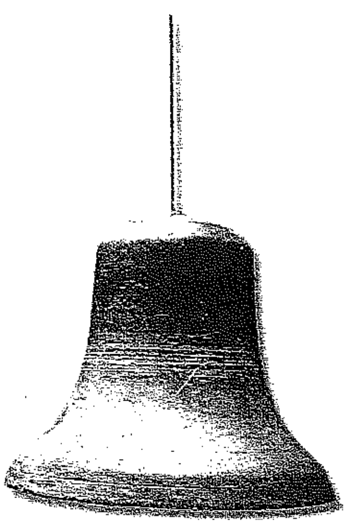
A S *Activity:* **Children at Risk.** This activity looks at the fates of children aboard the *Titanic*; the subsequent activity shows how to make such tables on a computer.

We know how many tickets of each class were sold on the *Titanic*, and we know that only about 32% of all those aboard the *Titanic* survived. After looking a the distribution of each variable by itself, it's natural and more interesting tc ask how they relate. Was there a relationship between the kind of ticket a passenger held and the passenger's chances of making it into the lifeboat? To answer this question, we need to look at the two categorical variables *Class* anc *Survival* together.

To look at two categorical variables together, we often arrange the counts ir a two-way table. Here is a two-way table of those aboard the *Titanic*, classifiec according to the class of ticket and whether the ticket holder survived or didn't Because the table shows how the individuals are distributed along each variable, contingent on the value of the other variable, such a table is called a **contingency table.**

Contingency table of ticket *Class* and *Survival*. The bottom line of "Totals" is the same as the previous frequency table.
Table 3.4

		Class				
		First	Second	Third	Crew	**Total**
Survival	**Alive**	203	118	178	212	**711**
	Dead	122	167	528	673	**1490**
	Total	**325**	**285**	**706**	**885**	**2201**



A bell-shaped artifact from the Titanic.

The margins of the table, both on the right and at the bottom, give totals. The bottom line of the table is just the frequency distribution of ticket *Class*. The righ column of the table is the frequency distribution of the variable *Survival*. When presented like this, in the margins of a contingency table, the frequency distribution of one of the variables is called its **marginal distribution.**

Each **cell** of the table gives the count for a combination of values of the twc variables. If you look down the column for second-class passengers to the firs cell, you can see that 118 second-class passengers survived. Looking at the third-class passengers, you can see that more third-class passengers (178) survived Were second-class passengers more likely to survive? Questions like this are easier to address by using percentages. The 118 survivors in second class were 41.4% of the total 285 second-class passengers, while the 178 surviving third-class passengers were only 25.2% of that class's total.

We know that 118 second-class passengers survived. We could display this number as a percentage—but as a percentage of what? The total number of passengers? (118 is 5.4% of the total: 2201.) The number of second-class passengers'

(118 is 41.4% of the 285 second-class passengers.) The number of survivors? (118 is 16.6% of the 711 survivors.) All of these are possibilities, and all are potentially useful or interesting. You'll probably wind up calculating (or letting your technology calculate) lots of percentages. Most statistics programs offer a choice of total percent, row percent, or column percent for contingency tables. Unfortunately, they often put them all together with several numbers in each cell of the table. The resulting table holds lots of information, but it can be hard to understand:

Another contingency table of ticket Class. This time we see not only the counts for each combination of *Class* and *Survival* (in bold) but the percentages these counts represent. For each count, there are three choices for the percentage: by row, by column, and by table total. There's probably too much information here for this table to be useful.
Table 3.5

Survival		Class: First	Second	Third	Crew	Total
Alive	Count	**203**	**118**	**178**	**212**	**711**
	% of Row	28.6%	16.6%	25.0%	29.8%	100%
	% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
	% of Table	9.2%	5.4%	8.1%	9.6%	32.3%
Dead	Count	**122**	**167**	**528**	**673**	**1490**
	% of Row	8.2%	11.2%	35.4%	45.2%	100%
	% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
	% of Table	5.6%	7.6%	24.0%	30.6%	67.7%
Total	Count	**325**	**285**	**706**	**885**	**2201**
	%of Row	14.8%	12.9%	32.1%	40.2%	100%
	% of Column	100%	100%	100%	100%	100%
	% of Table	14.8%	12.9%	32.1%	40.2%	100%

To simplify the table, let's first pull out the percent of table values:

A contingency table of *Class* by *Survival* with only the table percentages
Table 3.6

Survival	Class: First	Second	Third	Crew	Total
Alive	9.2%	5.4%	8.1%	9.6%	32.3%
Dead	5.6%	7.6%	24.0%	30.6%	67.7%
Total	14.8%	12.9%	32.1%	40.2%	100%

These percentages tell us what percent of *all* passengers belong to each combination of column and row category. For example, we see that although 8.1% of the people aboard the *Titanic* were surviving third-class ticket holders, only 5.4% were surviving second-class ticket holders. Is this fact useful? Comparing these percentages, you might think that the chances of surviving were better in third class than in second. But be careful. There were many more third-class than second-class passengers on the *Titanic*, so there were more third-class survivors. That group is a larger percentage of the passengers, but is that really what we want to know?

Percent of what? The English language can be tricky when we talk about percentages. If you're asked "What percent *of the survivors* were in second class?" it's pretty clear that we're interested only in survivors. It's as if we're restricting the *Who* in the question to the survivors, so we should look at the number of second-class passengers among all the survivors—in other words, the row percent.

But if you're asked "What percent were second-class passengers who survived?" you have a different question. Be careful; here, the *Who* is everyone on board, so 2201 should be the denominator, and the answer is the table percent.

And if you're asked "What percent of the second-class passengers survived?" you have a third question. Now the *Who* is the second-class passengers, so the denominator is the 285 second-class passengers, and the answer is the column percent.

Always be sure to ask "percent of what?" That will help you to know the *Who* and whether we want *row, column,* or *table* percentages.

FOR EXAMPLE Finding marginal distributions

In January 2007, a Gallup poll asked 1008 Americans age 18 and over whether they planned to watch the upcoming Super Bowl. The pollster also asked those who planned to watch whether they were looking forward more to seeing the football game or the commercials. The results are summarized in the table:

Response	Sex: Male	Female	Total
Game	279	200	479
Commercials	81	156	237
Won't watch	132	160	292
Total	492	516	1008

Question: What's the marginal distribution of the responses?

To determine the percentages for the three responses, divide the count for each response by the total number of people polled:

$$\frac{479}{1008} = 47.5\% \quad \frac{237}{1008} = 23.5\% \quad \frac{292}{1008} = 29.0\%$$

According to the poll, 47.5% of American adults were looking forward to watching the Super Bowl game, 23.5% were looking forward to watching the commercials, and 29% didn't plan to watch at all.

Conditional Distributions

The more interesting questions are *contingent.* We'd like to know, for examp what percentage of *second-class passengers* survived and how that compares w the survival rate for third-class passengers.

It's more interesting to ask whether the chance of surviving the *Titanic* si ing *depended* on ticket class. We can look at this question in two ways. First, could ask how the distribution of ticket *Class* changes between survivors and n survivors. To do that, we look at the *row percentages:*

The conditional distribution of ticket *Class* conditioned on each value of *Survival: Alive* and *Dead.*
Table 3.7

Survival	Class: First	Second	Third	Crew	Total
Alive	203 28.6%	118 16.6%	178 25.0%	212 29.8%	711 100%
Dead	122 8.2%	167 11.2%	528 35.4%	673 45.2%	1490 100%

By focusing on each row separately, we see the distribution of class under *condition* of surviving or not. The sum of the percentages in each row is 100%, we divide that up by ticket class. In effect, we temporarily restrict the *Who* fir survivors and make a pie chart for them. Then we refocus the *Who* on the non vivors and make their pie chart. These pie charts show the distribution of ti classes *for each row* of the table: survivors and nonsurvivors. The distribution create this way are called **conditional distributions,** because they show the distr tion of one variable for just those cases that satisfy a condition on another variab

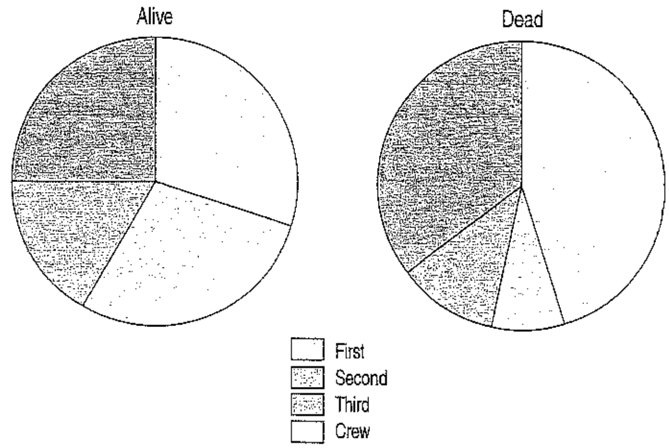


FIGURE 3.6
Pie charts of the conditional distributions of ticket Class for the survivors and nonsurvivors, separately. Do the distributions appear to be the same? We're primarily concerned with percentages here, so pie charts are a reasonable choice.

FOR EXAMPLE Finding conditional distributions

Recap: The table shows results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn't plan to watch.

Question: How do the conditional distributions of interest in the commercials differ for men and women?

Response	Male	Female	Total
Game	279	200	479
Commercials	81	156	237
Won't watch	132	160	292
Total	492	516	1008

(Sex: Male, Female)

Look at the group of people who responded "Commercials" and determine what percent of them were male and female:

$$\frac{81}{237} = 34.2\% \quad \frac{156}{237} = 65.8\%$$

Women make up a sizable majority of the adult Americans who look forward to seeing Super Bowl commercials more than the game itself. Nearly 66% of people who voiced a preference for the commercials were women, and only 34% were men.

But we can also turn the question around. We can look at the distribution of *Survival* for each category of ticket *Class*. To do this, we look at the *column percentages*. Those show us whether the chance of surviving was roughly the same *for each of the four classes*. Now the percentages in each column add to 100%, because we've restricted the *Who*, in turn, to each of the four ticket classes:

A contingency table of *Class* by *Survival* with only counts and column percentages. Each column represents the conditional distribution of *Survival* for a given category of ticket *Class*.

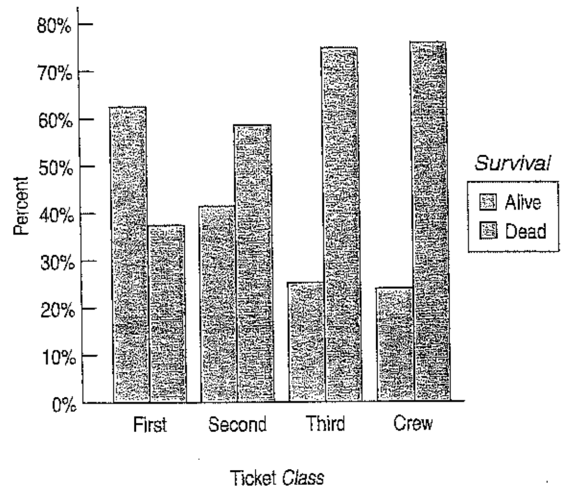
Table 3.8

Survival		First	Second	Third	Crew	Total
Alive	Count	203	118	178	212	711
	% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
Dead	Count	122	167	528	673	1490
	% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
Total	Count	325	285	706	885	2201
		100%	100%	100%	100%	100%

(Class)

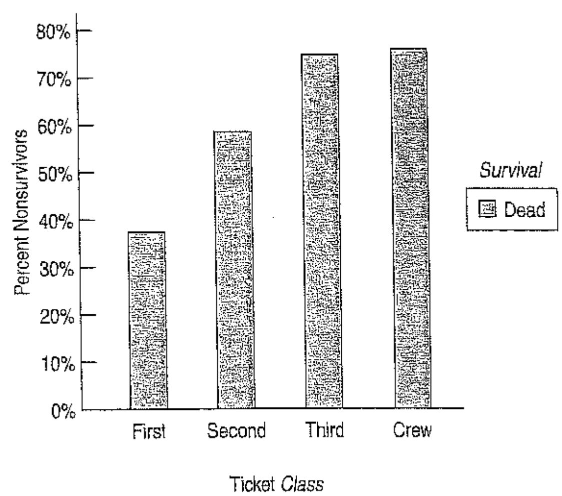
Looking at how the percentages change across each row, it sure looks like ticket class mattered in whether a passenger survived. To make it more vivid, we could show the distribution of *Survival* for each ticket class in a display. Here's a side-by-side bar chart showing percentages of surviving and not for each category:

FIGURE 3.7
Side-by-side bar chart *showing the conditional distribution of Survival for each category of ticket Class. The corresponding pie charts would have only two categories in each of four pies, so bar charts seem the better alternative.*



These bar charts are simple because, for the variable *Survival*, we have only two alternatives: Alive and Dead. When we have only two categories, we really need to know only the percentage of one of them. Knowing the percentage that survived tells us the percentage that died. We can use this fact to simplify the display even more by dropping one category. Here are the percentages of dying *across the classes* displayed in one chart:

FIGURE 3.8
Bar chart showing just nonsurvivor percentages for each value of ticket Class. Because we have only two values, the second bar doesn't add any information. Compare this chart to the side-by-side bar chart shown earlier.



TI-nspire

Conditional distributions and association. Explore the *Titanic* data to see which passengers were most likely to survive.

Now it's easy to compare the risks. Among first-class passengers, 37.5% perished, compared to 58.6% for second-class ticket holders, 74.8% for those in third class, and 76.0% for crew members.

If the risk had been about the same across the ticket classes, we would have said that survival was *independent* of class. But it's not. The differences we see among these conditional distributions suggest that survival may have depended on ticket class. You may find it useful to consider conditioning on each variable in a contingency table in order to explore the dependence between them.

It is interesting to know that *Class* and *Survival* are associated. That's an important part of the *Titanic* story. And we know how important this is because the margins show us the actual numbers of people involved.

Variables can be associated in many ways and to different degrees. The best way to tell whether two variables are associated is to ask whether they are *not*.[1] In a contingency table, when the distribution of *one* variable is the same for all categories of another, we say that the variables are **independent**. That tells us there's no association between these variables. We'll see a way to check for independence formally later in the book. For now, we'll just compare the distributions.

FOR EXAMPLE Looking for associations between variables

Recap: The table shows results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn't plan to watch.

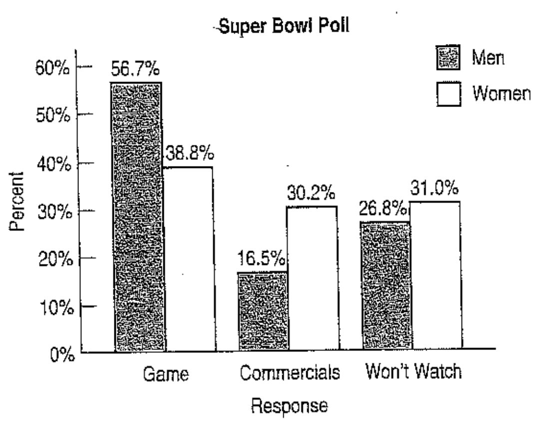
Question: Does it seem that there's an association between interest in Super Bowl TV coverage and a person's sex?

Response	Sex: Male	Female	Total
Game	279	200	479
Commercials	81	156	237
Won't watch	132	160	292
Total	492	516	1008

First find the distribution of the three responses for the men (the column percentages):

$$\frac{279}{492} = 56.7\% \quad \frac{81}{492} = 16.5\% \quad \frac{132}{492} = 26.8\%$$

Then do the same for the women who were polled, and display the two distributions with a side-by-side bar chart:



Based on this poll it appears that women were only slightly less interested than men in watching the Super Bowl telecast: 31% of the women said they didn't plan to watch, compared to just under 27% of men. Among those who planned to watch, however, there appears to be an association between the viewer's sex and what the viewer is most looking forward to. While more women are interested in the game (39%) than the commercials (30%), the margin among men is much wider: 57% of men said they were looking forward to seeing the game, compared to only 16.5% who cited the commercials.

[1] This kind of "backwards" reasoning shows up surprisingly often in science—and in Statistics. We'll see it again.

JUST CHECKING

A Statistics class reports the following data on Sex and Eye Color for students in the class:

Sex	Blue	Brown	Green/Hazel/Other	Total
Males	6	20	6	32
Females	4	16	12	32
Total	10	36	18	64

(Columns: Eye Color)

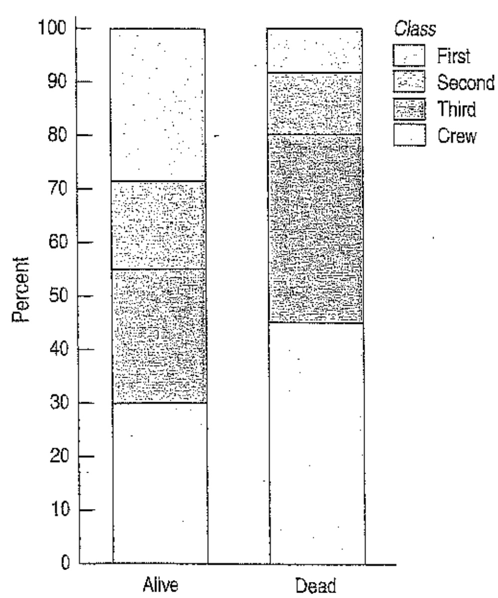
1. What percent of females are brown-eyed?
2. What percent of brown-eyed students are female?
3. What percent of students are brown-eyed females?
4. What's the distribution of Eye Color?
5. What's the conditional distribution of Eye Color for the males?
6. Compare the percent who are female among the blue-eyed students to the percent of all students who are female.
7. Does it seem that Eye Color and Sex are independent? Explain.

Segmented Bar Charts

We could display the *Titanic* information by dividing up bars rather than circles. The resulting **segmented bar chart** treats each bar as the "whole" and divides it proportionally into segments corresponding to the percentage in each group. We can clearly see that the distributions of ticket *Class* are different, indicating again that survival was not independent of ticket *Class*.

FIGURE 3.9 ***A segmented bar chart for Class by Survival***

Notice that although the totals for survivors and nonsurvivors are quite different, the bars are the same height because we have converted the numbers to percentages. Compare this display with the side-by-side pie charts of the same data in Figure 3.6.



STEP-BY-STEP EXAMPLE Examining Contingency Tables

Medical researchers followed 6272 Swedish men for 30 years to see if there was any association between the amount of fish in their diet and prostate cancer ("Fatty Fish Consumption and Risk of Prostate Cancer," *Lancet*, June 2001). Their results are summarized in this table:

We asked for a picture of a man eating fish. This is what we got.

Fish Consumption	Prostate Cancer: No	Prostate Cancer: Yes
Never/seldom	110	14
Small part of diet	2420	201
Moderate part	2769	209
Large part	507	42

Table 3.9

Question: Is there an association between fish consumption and prostate cancer?

THINK

Plan Be sure to state what the problem is about.

I want to know if there is an association between fish consumption and prostate cancer.

Variables Identify the variables and report the W's.

The individuals are 6272 Swedish men followed by medical researchers for 30 years. The variables record their fish consumption and whether or not they were diagnosed with prostate cancer.

Be sure to check the appropriate condition.

✓ **Categorical Data Condition:** I have counts for both fish consumption and cancer diagnosis. The categories of diet do not overlap, and the diagnoses do not overlap. It's okay to draw pie charts or bar charts.

SHOW

Mechanics It's a good idea to check the marginal distributions first before looking at the two variables together.

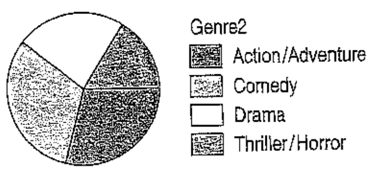
Fish Consumption	Prostate Cancer: No	Prostate Cancer: Yes	Total
Never/seldom	110	14	124 (2.0%)
Small part of diet	2420	201	2621 (41.8%)
Moderate part	2769	209	2978 (47.5%)
Large part	507	42	549 (8.8%)
Total	5806 (92.6%)	466 (7.4%)	6272 (100%)

Two categories of the diet are quite small, with only 2.0% Never/Seldom eating fish and 8.8% in the "Large part" category. Overall, 7.4% of the men in this study had prostate cancer.

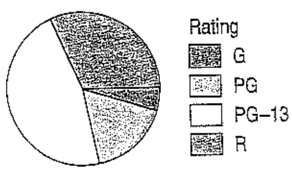
3. **Tables in the news.** Find a frequency table of categorical data from a newspaper, a magazine, or the Internet.
   a) Is it clearly labeled?
   b) Does it display percentages or counts?
   c) Does the accompanying article tell the W's of the variable?
   d) Do you think the article correctly interprets the data? Explain.

4. **Tables in the news II.** Find a contingency table of categorical data from a newspaper, a magazine, or the Internet.
   a) Is it clearly labeled?
   b) Does it display percentages or counts?
   c) Does the accompanying article tell the W's of the variables?
   d) Do you think the article correctly interprets the data? Explain.

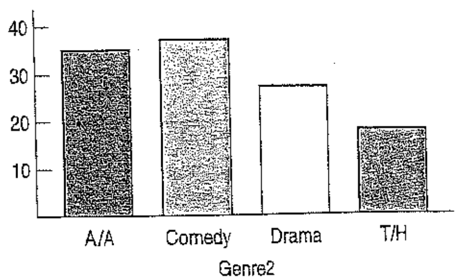
5. **Movie genres.** The pie chart summarizes the genres of 120 first-run movies released in 2005.
   a) Is this an appropriate display for the genres? Why/why not?
   b) Which genre was least common?



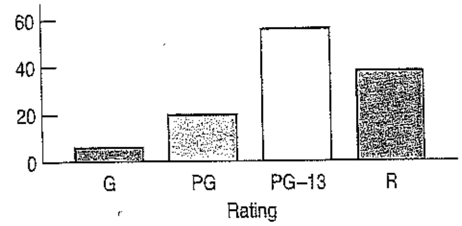
6. **Movie ratings.** The pie chart shows the ratings assigned to 120 first-run movies released in 2005.
   a) Is this an appropriate display for these data? Explain.
   b) Which was the most common rating?



7. **Genres again.** Here is a bar chart summarizing the 2005 movie genres, as seen in the pie chart in Exercise 5.
   a) Which genre was most common?
   b) Is it easier to see that in the pie chart or the bar chart? Explain.



8. **Ratings again.** Here is a bar chart summarizing the 2005 movie ratings, as seen in the pie chart in Exercise 6.
   a) Which was the least common rating?
   b) An editorial claimed that there's been a growth in PG-13 rated films that, according to the writer, "have too much sex and violence," at the expense of G-rated films that offer "good, clean fun." The writer offered the bar chart below as evidence to support his claim. Does the bar chart support his claim? Explain.



9. **Magnet schools.** An article in the Winter 2003 issue of *Chance* magazine reported on the Houston Independent School District's magnet schools programs. Of the 1755 qualified applicants, 931 were accepted, 298 were wait-listed, and 526 were turned away for lack of space. Find the relative frequency distribution of the decisions made, and write a sentence describing it.

10. **Magnet schools again.** The *Chance* article about the Houston magnet schools program described in Exercise 9 also indicated that 517 applicants were black or Hispanic, 292 Asian, and 946 white. Summarize the relative frequency distribution of ethnicity with a sentence or two (in the proper context, of course).

11. **Causes of death 2004.** The Centers for Disease Control and Prevention (www.cdc.gov) lists causes of death in the United States during 2004:

Cause of Death	Percent
Heart disease	27.2
Cancer	23.1
Circulatory diseases and stroke	6.3
Respiratory diseases	5.1
Accidents	4.7

   a) Is it reasonable to conclude that heart or respiratory diseases were the cause of approximately 33% of U.S. deaths in 2003?
   b) What percent of deaths were from causes not listed here?
   c) Create an appropriate display for these data.

12. **Plane crashes.** An investigation compiled informatior about recent nonmilitary plane crashes (www.planecrashinfo.com). The causes, to the extent that they could be determined, are summarized in the table.

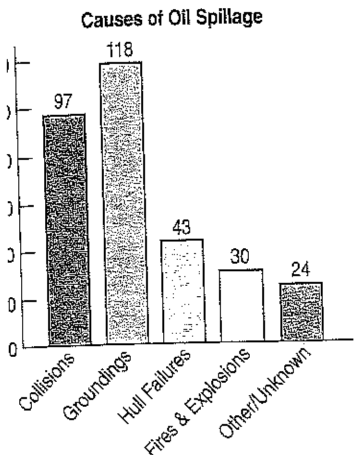
Cause	Percent
Pilot error	40
Other human error	5
Weather	6
Mechanical failure	14
Sabotage	6

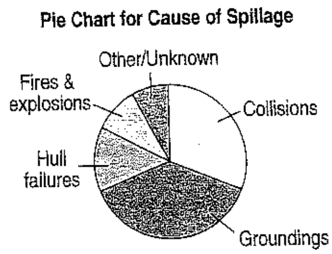
   a) Is it reasonable to conclude that the weather or mechanical failures caused only about 20% of recent plane crashes?
   b) In what percent of crashes were the causes not determined?
   c) Create an appropriate display for these data.

Oil spills 2006. Data from the International Tanker Owners Pollution Federation Limited (www.itopf.com) give the cause of spillage for 312 large oil tanker accidents from 1974–2006. Here are displays.

a) Write a brief report interpreting what the displays show.

b) Is a pie chart an appropriate display for these data? Why or why not?





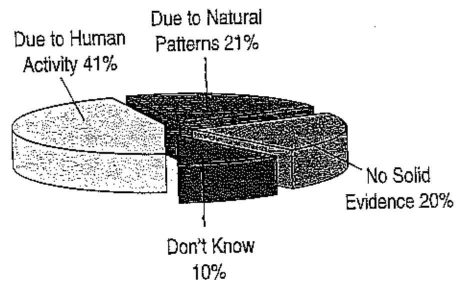
14. Winter Olympics 2006. Twenty-six countries won medals in the 2006 Winter Olympics. The table lists them, along with the total number of medals each won:

Country	Medals	Country	Medals
Germany	29	Finland	9
United States	25	Czech Republic	4
Canada	24	Estonia	3
Austria	23	Croatia	3
Russia	22	Australia	2
Norway	19	Poland	2
Sweden	14	Ukraine	2
Switzerland	14	Japan	1
South Korea	11	Belarus	1
Italy	11	Bulgaria	1
China	11	Great Britain	1
France	9	Slovakia	1
Netherlands	9	Latvia	1

a) Try to make a display of these data. What problems do you encounter?

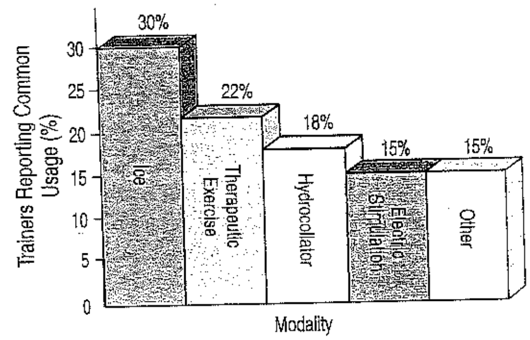
b) Can you find a way to organize the data so that the graph is more successful?

15. Global Warming. The Pew Research Center for the People and the Press (http://people-press.org) has asked a representative sample of U.S. adults about global warming, repeating the question over time. In January 2007, the responses reflected an increased belief that global warming is real and due to human activity. Here's a display of the percentages of respondents choosing each of the major alternatives offered:



List the errors in this display.

16. Modalities. A survey of athletic trainers (Scott F. Nadler, Michael Prybicien, Gerard A. Malanga, and Dan Sicher. "Complications from Therapeutic Modalities: Results of a National Survey of Athletic Trainers." *Archives of Physical Medical Rehabilitation* 84 [June 2003]) asked what modalities (treatment methods such as ice, whirlpool, ultrasound, or exercise) they commonly use to treat injuries. Respondents were each asked to list three modalities. The article included the following figure reporting the modalities used:



a) What problems do you see with the graph?

b) Consider the percentages for the named modalities. Do you see anything odd about them?

17. Teen smokers. The organization Monitoring the Future (www.monitoringthefuture.org) asked 2048 eighth graders who said they smoked cigarettes what brands they preferred. The table below shows brand preferences for two regions of the country. Write a few sentences describing the similarities and differences in brand preferences among eighth graders in the two regions listed.

Brand preference	South	West
Marlboro	58.4%	58.0%
Newport	22.5%	10.1%
Camel	3.3%	9.5%
Other (over 20 brands)	9.1%	9.5%
No usual brand	6.7%	12.9%

18. Handguns. In an effort to reduce the number of gun-related homicides, some cities have run buyback programs in which the police offer cash (often $50) to anyone who turns in an operating handgun. *Chance* magazine looked at results from a four-year period in Milwaukee. The table on the next page shows what types of guns were turned in and what types were used in homicides during a four-year period. Write a few sentences comparing the two distributions.

Caliber of gun	Buyback	Homicide
Small (.22, .25, .32)	76.4%	20.3%
Medium (.357, .38, 9 mm)	19.3%	54.7%
Large (.40, .44, .45)	2.1%	10.8%
Other	2.2%	14.2%

19. **Movies by Genre and Rating.** Here's a table that classifies movies released in 2005 by genre and MPAA rating:

	G	PG	PG-13	R	Total
Action/Adventure	66.7	25	30.4	23.7	**29.2**
Comedy	33.3	60.0	35.7	10.5	**31.7**
Drama	0	15.0	14.3	44.7	**23.3**
Thriller/Horror	0	0	19.6	21.1	**15.8**
Total	**100%**	**100%**	**100%**	**100%**	**100%**

a) The table gives column percents. How could you tell that from the table itself?
b) What percentage of these movies were comedies?
c) What percentage of the PG-rated movies were comedies?
d) Which of the following can you learn from this table? Give the answer if you can find it from the table.
   i) The percentage of PG-13 movies that were comedies
   ii) The percentage of dramas that were R-rated
   iii) The percentage of dramas that were G-rated
   iv) The percentage of 2005 movies that were PG-rated comedies

20. **The Last Picture Show.** Here's another table showing information about 120 movies released in 2005. This table gives percentages of the table total:

	G	PG	PG-13	R	Total
Action/Adventure	3.33%	4.17	14.2	7.50	**29.2**
Comedy	1.67	10	16.7	3.33	**31.7**
Drama	0	2.50	6.67	14.2	**23.3**
Thriller/Horror	0	0	9.17	6.67	**15.8**
Total	**5**	**16.7**	**46.7**	**31.7**	**100%**

a) How can you tell that this table holds table percentages (rather than row or column percentages)?
b) What was the most common genre/rating combination in 2005 movies?
c) How many of these movies were PG-rated comedies?
d) How many were G-rated?
e) An editorial about the movies noted, "More than three-quarters of the movies made today can be seen only by patrons 13 years old or older." Does this table support that assertion? Explain.

21. **Seniors.** Prior to graduation, a high school class was surveyed about its plans. The following table displays the results for white and minority students (the "Minority" group included African-American, Asian, Hispanic, and Native American students):

Seniors

Plans	White	Minority
4-year college	198	44
2-year college	36	6
Military	4	1
Employment	14	3
Other	16	3

a) What percent of the seniors are white?
b) What percent of the seniors are planning to attend a 2-year college?
c) What percent of the seniors are white and planning to attend a 2-year college?
d) What percent of the white seniors are planning to attend a 2-year college?
e) What percent of the seniors planning to attend a 2-year college are white?

22. **Politics.** Students in an Intro Stats course were asked to describe their politics as "Liberal," "Moderate," or "Conservative." Here are the results:

Politics

Sex	L	M	C	Total
Female	35	36	6	77
Male	50	44	21	115
Total	85	80	27	192

a) What percent of the class is male?
b) What percent of the class considers themselves to be "Conservative"?
c) What percent of the males in the class consider themselves to be "Conservative"?
d) What percent of all students in the class are males who consider themselves to be "Conservative"?

23. **More about seniors.** Look again at the table of post-graduation plans for the senior class in Exercise 21.
a) Find the conditional distributions (percentages) of plans for the white students.
b) Find the conditional distributions (percentages) of plans for the minority students.
c) Create a graph comparing the plans of white and minority students.
d) Do you see any important differences in the post-graduation plans of white and minority students? Write a brief summary of what these data show, including comparisons of conditional distributions.

24. **Politics revisited.** Look again at the table of political views for the Intro Stats students in Exercise 22.
a) Find the conditional distributions (percentages) of political views for the females.
b) Find the conditional distributions (percentages) of political views for the males.
c) Make a graphical display that compares the two distributions.
d) Do the variables *Politics* and *Sex* appear to be independent? Explain.

5. **Magnet schools revisited.** The *Chance* magazine article described in Exercise 9 further examined the impact of an applicant's ethnicity on the likelihood of admission to the Houston Independent School District's magnet schools programs. Those data are summarized in the table below:

Admission Decision

Ethnicity	Accepted	Wait-listed	Turned away	Total
Black/Hispanic	485	0	32	517
Asian	110	49	133	292
White	336	251	359	946
Total	931	300	524	1755

a) What percent of all applicants were Asian?
b) What percent of the students accepted were Asian?
c) What percent of Asians were accepted?
d) What percent of all students were accepted?

26. **More politics.** Look once more at the table summarizing the political views of Intro Stats students in Exercise 22.
   a) Produce a graphical display comparing the conditional distributions of males and females among the three categories of politics.
   b) Comment briefly on what you see from the display in a.

27. **Back to school.** Examine the table about ethnicity and acceptance for the Houston Independent School District's magnet schools program, shown in Exercise 25. Does it appear that the admissions decisions are made independent of the applicant's ethnicity? Explain.

28. **Cars.** A survey of autos parked in student and staff lots at a large university classified the brands by country of origin, as seen in the table.

Driver

Origin	Student	Staff
American	107	105
European	33	12
Asian	55	47

a) What percent of all the cars surveyed were foreign?
b) What percent of the American cars were owned by students?
c) What percent of the students owned American cars?
d) What is the marginal distribution of origin?
e) What are the conditional distributions of origin by driver classification?
f) Do you think that the origin of the car is independent of the type of driver? Explain.

29. **Weather forecasts.** Just how accurate are the weather forecasts we hear every day? The following table compares the daily forecast with a city's actual weather for a year:

Actual Weather

Forecast	Rain	No rain
Rain	27	63
No rain	7	268

a) On what percent of days did it actually rain?
b) On what percent of days was rain predicted?
c) What percent of the time was the forecast correct?
d) Do you see evidence of an association between the type of weather and the ability of forecasters to make an accurate prediction? Write a brief explanation, including an appropriate graph.

30. **Twins.** In 2000, the *Journal of the American Medical Association (JAMA)* published a study that examined pregnancies that resulted in the birth of twins. Births were classified as preterm with intervention (induced labor or cesarean), preterm without procedures, or term/post-term. Researchers also classified the pregnancies by the level of prenatal medical care the mother received (inadequate, adequate, or intensive). The data, from the years 1995–1997, are summarized in the table below. Figures are in thousands of births. (*JAMA* 284 [2000]:335–341)

TWIN BIRTHS 1995–1997 (IN THOUSANDS)

Level of Prenatal Care	Preterm (induced or cesarean)	Preterm (without procedures)	Term or post-term	Total
Intensive	18	15	28	61
Adequate	46	43	65	154
Inadequate	12	13	38	63
Total	76	71	131	278

a) What percent of these mothers received inadequate medical care during their pregnancies?
b) What percent of all twin births were preterm?
c) Among the mothers who received inadequate medical care, what percent of the twin births were preterm?
d) Create an appropriate graph comparing the outcomes of these pregnancies by the level of medical care the mother received.
e) Write a few sentences describing the association between these two variables.

31. **Blood pressure.** A company held a blood pressure screening clinic for its employees. The results are summarized in the table below by age group and blood pressure level:

Age

Blood Pressure	Under 30	30–49	Over 50
Low	27	37	31
Normal	48	91	93
High	23	51	73