CHAPTER 1

I tell a friend that my job is more fun than you'd think: What is statistics?

I have a friend who is a well-known DJ. He was once telling me about a gig he'd recently played in Milan—some kind of massive party for 20,000—when a neighbor of ours rolled up in a cab. A journalist, he was just returning from the Pacific Northwest where he'd been researching a story on wild mushrooms. The three of us stood chatting on the sidewalk for a while, and I thought I should mention that I was going to Cleveland to give a talk about bladder cancer.

I think they were impressed.

The DJ's unnaturally supportive wife once asked what it was that I did all day. What I said was: it's a lot of hard work and although it isn't as much of a thrill as DJ'ing a rave, it is more fun than you might think and is very satisfying. Best of all you get to meet a lot of other statisticians (actually, I didn't say that). This sounds like a bit of a non-answer, but as I'll explain, what I said made reference to *inference* and *estimation*, which is pretty much the A–Z of statistical analysis.

Estimation is about trying to work out how large or small something is. That "something" generally either can't be directly measured, or would take too much time and effort to do so. The two estimates in my answer were: "statistics is a lot of hard work" and "being a statistician is very satisfying." Ok, not very precise, but we might imagine that some psychologist had developed a questionnaire measuring mental effort, job satisfaction and, while we are at it, fun. To answer our question, "How much work is being a statistician?", what we'd ideally do is send the questionnaire to every single statistician in the world. But that would be kind of a pain, so we'd probably be better off sending the questionnaire to, say, 500 statisticians and hope that their answers were representative of statisticians in general. Let's say that our sample of 500 statisticians scored an average of 88% on the "mental effort at work" questionnaire: 88% is then our estimate of the average for all statisticians.

Inference is about drawing conclusions, and statisticians usually make inferences by testing hypotheses. My answer to the DJ's wife included two hypotheses: "doing statistics is not as much of a thrill as DJ'ing" and "statistics is more fun than you might think." To test the first hypothesis, we could give our "fun at work" questionnaire to 500 DJs and then compare their answers to those of the 500 statisticians (ok, it probably isn't worth doing this); to test the second hypothesis we just compare the statisticians' answers to some guess we'd made before we'd taken the data (e.g., statistics sounds like it would be about 0.3% fun).

This just leaves the issue of how we choose the statisticians to sample, which questionnaire we give them, whether we put a stamp on the return envelope or use "postage paid," how often we should chase up the ones that don't get back to us, how we enter data from all 500 questionnaires onto a computer, what you do if someone only fills in half of the questions, and so on and so forth. *Designing* a study is pretty complicated and statisticians know a lot about study design: it is said that R A Fisher, one of the great statisticians, used to clean out the rats' cages himself on the grounds that "if the rats are dirty, so will my data be." Indeed, all of the questions about the design of the questionnaire study (even whether you should put a stamp on the return envelope) have been written about by statisticians.

Given the choice, I guess most of us (me included) would rather be eating wild mushrooms, or spinning discs at a party, than running a complex statistical analysis. That is, until someone we love gets bladder cancer and we want to know what to do about it. Statistics sounds pretty cold—what is a number to me is somebody's living and breathing father, with stories to tell—but it has a very human goal: we want to live our lives better. To do that, we have to make good decisions and sometimes looking at numerical data in the right way can help us do so.

• Things to Remember •

1. Statistics involves estimation, inference and study design.
2. Estimation is about trying to work out how large or small something is.
3. Inference is about drawing conclusions, usually by conducting a statistical test of a hypothesis.
4. A hypothesis is a statement about the world that could be tested to see whether it is true or false.
5. Many studies produce numbers; as experts in numbers, statisticians often have a lot to say about how exactly a study should be designed.
6. Cleveland gets a bad rap, but the faculty dinner I had was actually pretty good.

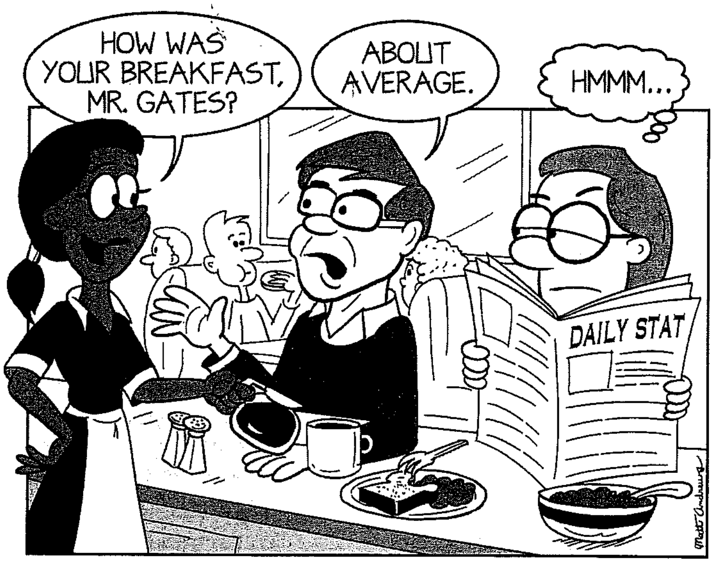
for Discussion

1. I defined a hypothesis as "a statement about the world that could be tested to see whether it is true or false." Are there some statements that can't be tested?
2. There are two sorts of estimates that statisticians make: how big or small something is and how big or small something is compared to something else. An example of the first sort of estimate is "the mean height of an American male is close to 5 ft 9 $\frac{1}{2}$ in." An example of the second sort of estimate is "men who smoke are 23 times more likely to develop lung cancer than men who have never smoked." Write down some examples of estimates of both sorts.
3. Most hypotheses can be rephrased in terms of estimates. I mixed up some estimates and hypotheses below. Match each estimate with the corresponding hypothesis and say which is the estimate and which the hypothesis.
   a. Crimes decreased 21% comparing the year before and the year after completion of a program to improve street lighting.
   b. Men and women do not differ in their voting behavior for presidential candidates.
   c. Obesity rates in California increased during the 1990's.
   d. Recurrence rates were 5% lower in women receiving chemotherapy after surgery compared to women receiving surgery alone.
   e. The proportion of women voting for Democratic presidential candidates is 5% higher than men.
   f. Improvements in street lighting decrease crime.
   g. Electric shocks (punishment) are more effective than sugar (reward) for improving learning in rats, as measured by time to complete a maze learning task.
   h. Mean time to complete a maze was 20 seconds shorter in rats exposed to shocks than those given sugar.
   i. Obesity rates in California almost doubled between 1990 and 2000, from 10% to nearly 20%.
   j. Chemotherapy plus surgery is no more effective than surgery for breast cancer.
4. Who said "there are lies, damned lies and statistics"?

NOTE: See page 154 for answer sets.

CHAPTER 2

So Bill Gates walks into a diner: On means and medians



A statistician's joke: So Bill Gates walks into a diner . . . and the average salary changes.

Ok, not very funny, I realize.

The point of the Bill Gates "joke" is to illustrate the difference between two different types of average, the *mean* and the *median*. Let's imagine that the salaries in the diner before Bill walked in were as follows:

Eric	$85,000
Jose	$50,000
Barrett	$45,000
Sandra	$40,000
Todd	$35,000
Michael	$30,000
Katie	$30,000

The mean is what we normally think of as the average: you add up all the salaries and divide the total by the number of people. If you add up all the salaries ($315,000) and divide by the number of people (7), you get $45,000. The median is best thought of as the "middle" number: line up all the salaries from lowest to highest; the median salary is the one halfway along. The middle of 7 is 4, so the median salary in our diner is that of Sandra, who has the fourth highest salary, $40,000.

So now Bill Gates walks in with an annual income of, say, $1 billion (most people would call this rich, statisticians call it an *outlier*). Bill's salary changes the mean salary to a little over $125m. As for the median, there is no middle of 8, so we go halfway between 4 and 5. The 4th highest salary is now $45,000 and the 5th is $40,000, giving a new median salary of $42,500. Most people would say that $42,500 was a fair reflection of the salaries in the diner and that $125m had nothing to do with anything. And so we end up with a neat little rule: if there are outliers in the data—which is exactly what happens whenever a major software entrepreneur feels like having a greasy breakfast—use a median.

Here is the key point: if you stumble across a "neat little rule" in statistics, be very careful. Wanting a "fair reflection" of the data is not the only thing we want to use a statistic for; the other is to plan and make decisions. So let's imagine that instead of a diner we had a hospital, and instead of salaries, we had costs of surgery; in place of Bill Gates, we have a patient who has a series of complications after surgery, leading to costs of $250,000.

Patient	Cost of surgery
1	$85,000
2	$50,000
3	$45,000
4	$40,000
5	$35,000
6	$30,000
7	$30,000
8	$250,000

This gives a mean of just over $70,000; the median is the same as the Bill Gates example, $42,500. Which number would be most important if you were, say, a hospital administrator? $42,500 may well be a "fair reflection" of the typical cost of a patient's care, but writing next year's budget assuming costs of $42,500 per patient will likely lead to a shortfall. Thinking about means and medians is also why I buy health insurance (the median yearly expenditure of Americans like me is far less than the premium but when I look at mean yearly expenditure, I reckon I get a pretty good deal) and wear a seat-belt (even though the median number of injuries per car trip is zero).

So next time you are in a diner, and are bored and depressed because Bill Gates hasn't shown up yet, here is a neat little rule to scrawl on your coffee-sodden napkin: sometimes there are no right and wrong statistics; it all depends on what you want to use them for.

• Things to Remember •

1. What most people call an "average" is what statisticians call a mean. To calculate a mean, think of your data as a list of numbers, add up all the numbers and then divide by the number of items on your list.
2. The median is the half way point of your list of numbers: half of the sample have values higher than the median and half have values lower than the median.
3. An outlier is when you have an observation that doesn't follow the pattern of the data.
4. When you have outliers, the median often gives a fairer reflection of the data than the mean.
5. Generally speaking, means are better than medians for planning and making decisions.

for Discussion

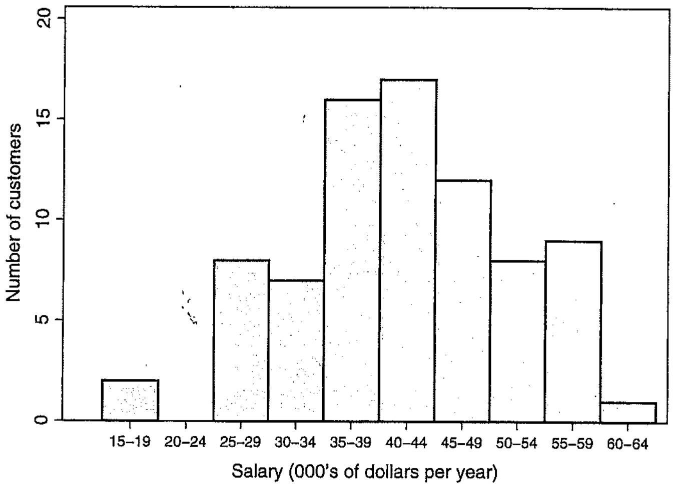
1. I said that "half of the sample have values higher than the median and half have values lower than the median." Is that always true?
2. Here is a die rolling game: you roll a die and if you get 1–5, I give you $20; if you roll a 6, you give me $1,000. Would you play? Explain your answer.

NOTE: See page 156 for answer sets.

CHAPTER 3

Bill Gates goes back to the diner: Standard deviation and interquartile range

Ok, I know, I can't really see the world's richest man going into a cheap diner in the first place, let alone going back. But I wanted to stick with the example I had, so let's imagine that Bill had enjoyed the good-natured joshing about the design flaws in Windows and didn't have much to do the next day. Also, let's imagine that the diner was much, much busier, with 80 people in total having passed through at some point that morning. Here is a *histogram* showing the salaries of customers.



The *x*-axis (going from left to right) shows different salary levels put into different groups (statisticians call these "bins," which has the unfortunate implication that the data are garbage). The *y*-axis (going up and down) shows the number of people within each salary level. For example, the histogram shows that 16 of the individuals who had eaten in the diner have salaries in the range $35,000–$39,000 per year.

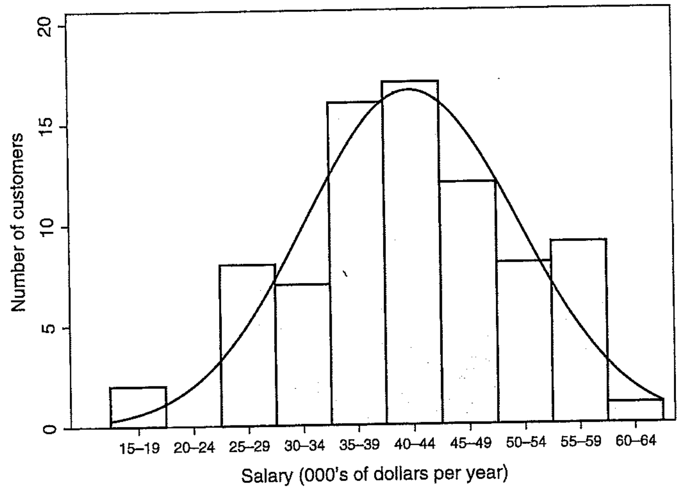
You can't work this out from this histogram, but I used the raw data to calculate that the mean salary was $42,360. So we took all those data points and turned them into just a single number. Now, there is an old joke that goes something like: a statistician had his head in the oven and his feet in the fridge. When he was asked how he felt, he said, "On average, pretty good." From this we learn two things: (a) statisticians tell bad jokes (am I repeating myself here?) and (b) a single number often doesn't describe a data set that well. Accordingly, it is generally a good idea to report not just a mean or median—what statisticians call the *central tendency* of the data—but some measure of how much the data vary—what statisticians call a *measure of spread* or a *measure of dispersion.*

One common measure of spread describing how much the study data vary is the *standard deviation.* The standard deviation is calculated from the data using a formula that I won't go into here (roughly speaking, you calculate the difference between each value and the mean, square it, take the mean of all the squares, and then take the square root). The thing to remember is that if the standard deviation for our data set was small, it would mean that everyone has pretty close to the same salary; if the standard deviation was large, it would mean that the salaries of the people in the diner vary widely.

To work out just how much variation we have, we can use some simple rules of thumb. The most well known is that "95% of observations are within about two standard deviations of the mean." This is the same as saying that only 5% of customers have salaries more than two standard deviations from the mean. From the raw data I used to create the bar chart, I worked out a standard deviation of $9,616. We can now work out that 5% of salaries are expected to be either higher than $61,592 (mean $42,360 + standard deviation $9,616 × 2 = $61,592) or lower than $23,128 (mean $42,360 – standard deviation $9,616 × 2 = $23,128). As it happens, one customer has a salary above $61,592 and two are below $23,128 (you can see his from the histogram); this is 3 out of 80, or 3.75%, which is reasonably close to 5%.

It is also true (although this doesn't seem to get mentioned much for some reason) that about two-thirds of observations are within one standard deviation of the mean, and that about half of observations are within two-thirds of a standard deviation from the mean. You can test out these rules of thumb using the histogram: for example, it does look as though about two-thirds of the customers have salaries between about $33,000 (i.e., $42,360 − $9,616) and $52,000 (i.e., $42,360 + $9,616).

That is, of course, until Bill Gates walks into the diner. Now we have a mean salary of about $12 million and a standard deviation of, let's see, $100 million. Clearly it is no longer the case that two-thirds of observations are within a standard deviation of the mean because no one in the diner (other than Bill) has a salary anywhere near $112 million and you can't have a salary of negative $88 million (although this guy I know from college is certainly trying). So the general rules of thumb don't work when the data are skewed away from the bell-shaped curve statisticians call the *normal* distribution (see *Chutes and Ladders and serum hemoglobin levels: Thoughts on the normal distribution*). Here is the normal curve on our data set without Bill Gates. As you can see it isn't a bad fit, which is why our rules of thumb work ok.



When Bill Gates walks in, however, we have data that we'd describe as "skewed": they don't seem to fit the normal distribution at all. What should you do about standard deviations if you don't have a good fit? Remember that the first time Bill Gates went into the diner, we said that you could use a median instead of a mean as the average. The measure of spread you use with medians is not a standard deviation, but what is called the *interquartile range*. The median is "halfway" along the data; comparably, the *quartiles* are a quarter and three-quarters of the way along. Using the data set in the histograms, I found the median income to be $41,900 and the interquartile range to be $36,000 to $49,300. These three numbers allow you to see a bunch of things immediately. For example:

- 50% of the customers have salaries of more than $41,900 and 50% have salaries of less than $41,900
- 25% of the customers have salaries of more than $49,300
- 25% of customers have salaries of less than $36,000

- 50% of customers have salaries between $36,000 and $49,300
- 25% of customers have salaries between $36,000 and $41,900
- 25% of customers have salaries between $41,900 and $49,300

Bill Gates causes the mean salary and standard deviation to go haywire, but the median and interquartile range remain pretty constant (e.g., the upper quartile goes from $49,300 to $49,500). This is one reason why we said, last time, that if data are very skewed, you often get a fairer reflection of the data if you use medians (and therefore the interquartile range) than means (along with standard deviations).

Here is another reason to use the median and interquartile range. I analyze the results of cancer studies and one of the first things I report in a scientific paper is the general characteristics of the patients in the study: how old were they? What is the ratio of male to female? How many had early stage cancer and how many had advanced disease? Let's imagine that the age of the patients in a study followed very closely to the normal distribution. I could report a mean and standard deviation, knowing that any reader could then work out whatever they wanted about the distribution of ages. But the point is, they are not going to. You can hardly see a busy cancer doctor thinking, "Ok, a mean of 64.3 and a standard deviation of 9.8; half the patients are within two-thirds of a standard deviation of the mean, that is, 64.3 + 9.8 × 0.667, which is—wait a minute, where's my calculator?" You can just glance at the median and interquartile range and get a good, quick idea about the sort of data that you are dealing with.

In other words, the median and interquartile range are very useful for describing a data set. And this is exactly what we want them to do: everything I have been talking about here—means, medians, standard deviations, interquartile ranges—are known as *descriptive statistics*.

• Things to Remember •

1. Means and medians are useful for describing a data set. Means and medians are types of average, or central tendency.
2. You generally want to know not only the average of a data set, but how much the data vary around that average: a measure of spread.
3. The measure of spread normally reported with a mean is the standard deviation.
4. The measure of spread normally reported with a median is the interquartile range.
5. If data follow something close to a normal distribution, the mean and standard deviation can be used to work out all sorts of things about your data, but you have to do some calculations.
6. The median and interquartile range give quick information about data without the need for any calculation.
7. The median and interquartile range are also useful for describing data that are skewed.
8. Statistics used to describe a data set, means and medians, standard deviations and interquartile ranges, are known as descriptive statistics.

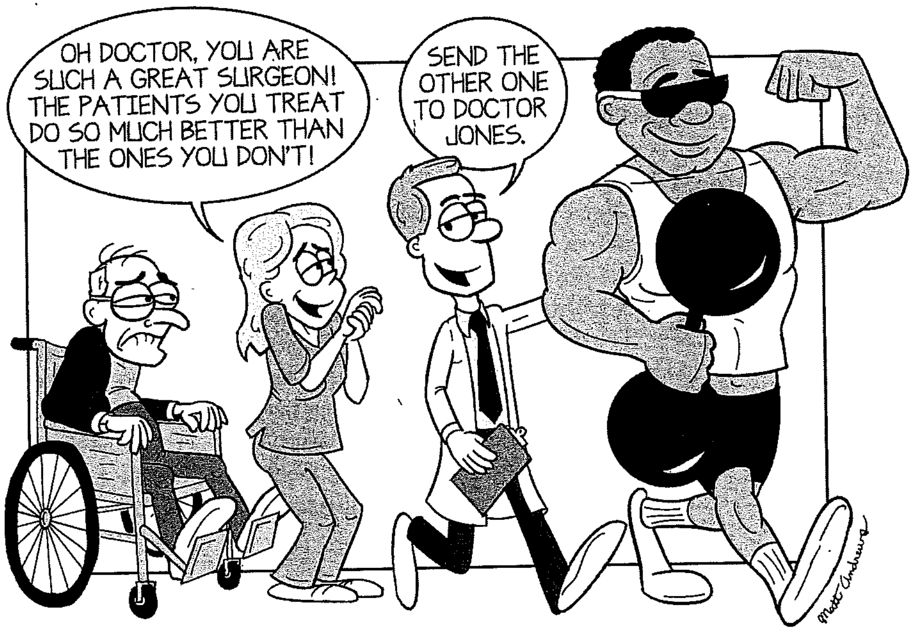
for Discussion

1. The upper and lower quartile are sometimes described as the 75th and 25th *centile* (or *percentile*). Explain this.
2. When talking about the interquartile range I said things like 25% of the customers in the diner have salaries of "$49,300 or more." When talking about standard deviations, I said 5% had salaries above $61,592 or below $23,128. Why did I sometimes say "*x* or more" and sometimes "higher than *x*?"
3. Is it really true that 95% of observations are within two standard deviations of the mean, even for a perfectly normal distribution?

NOTE: See page 157 for answer sets.

CHAPTER 4

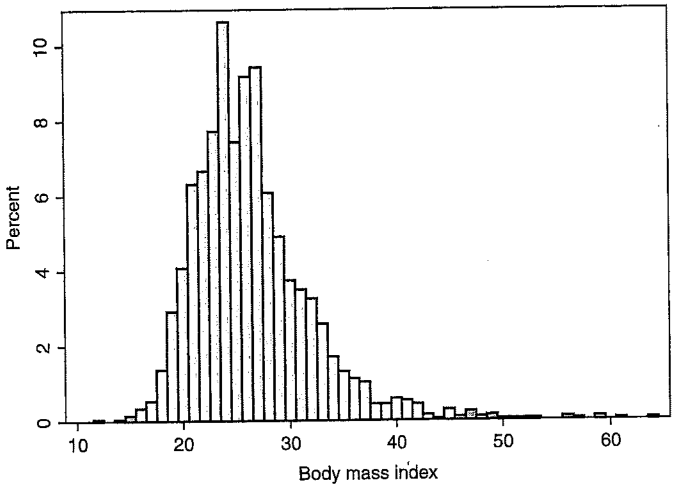
A skewed shot, a biased referee


OH DOCTOR, YOU ARE SUCH A GREAT SURGEON! THE PATIENTS YOU TREAT DO SO MUCH BETTER THAN THE ONES YOU DON'T!
SEND THE OTHER ONE TO DOCTOR JONES.
Matt Andrews

Like all great moments in my life, I remember it as if it were yesterday. England was playing Spain in the European soccer championships, a Spanish player mis hit a shot, which then fell to a team mate, who slotted it past the English keeper for a goal. But then the referee disallowed the goal for offside (relief!), even though the replay showed that the goal ought to have counted.

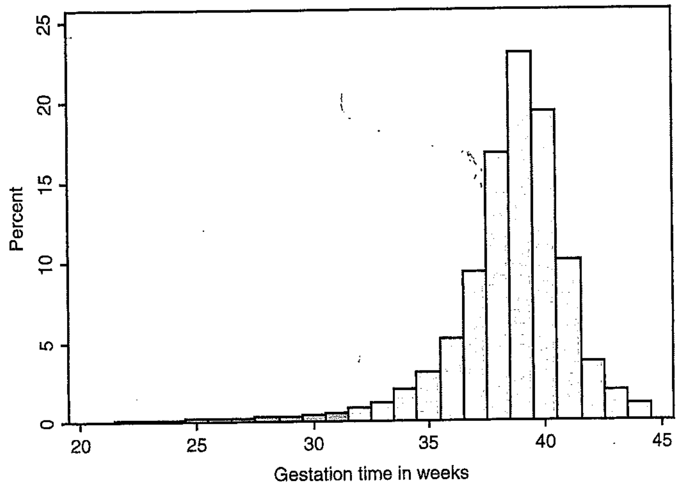
I couldn't believe it: *there had been a bad decision on the soccer field and, as an England fan, I wasn't going to suffer as a result.* The British press the next day were full of praise for "our brave lads" (England had squeaked through on penalties); the Spanish press were outraged at what had obviously been a biased referee. In my view, if the Spanish player hadn't skewed his shot in the first place, we wouldn't even be talking about the referee. But then again, you'd hardly call me unbiased about it.

Skew means "off to one side." Statistics can also get "off to one side," for one of two reasons. First, sometimes that is just how the data are: you have more observations on one side than the other. The following graph is from a study of US adults and gives the body mass index (which is weight in kilograms divided by the square of height in meters):



This data set is skewed to the right. There are some Americans who are perhaps a little underweight, but the Americans who are on the heavy side are often very heavy (20—25 is considered "normal"). These data are right-skewed because observations above the median tend to be further from the median than observations below the median. As a result, the mean is higher than the median (26.5 vs. 25.7).

Left-skewed data is when the mean is less than the median. Here is an interesting example of left-skewed data—duration of pregnancy:



There is a long tail off to the left of premature infants, but not a tail off to the right of infants born much later than their due date. This is because a very long pregnancy can be dangerous, and doctors don't let any woman carry a baby for more than two weeks longer than normal (although some women appear to have slipped through the net). As a result, the mean is lower than the median.

Another meaning of skew is "skewed away from the truth." Here is a famous example of an opinion poll that got things quite spectacularly wrong. When Franklin D. Roosevelt ran for reelection in 1936, the *Literary Digest* conducted an opinion poll to predict the result on election day. There were two problems. First, they selected much of their sample from the telephone book. Only a relatively small number of wealthier Americans had telephones during the Great Depression and richer folks tended not to like FDR and his "New Deal" policies. The second problem was that the poll had a very low response rate, with only about 20–25% of those polled returning their postal ballot. It seems likely that those voters who didn't like FDR were especially motivated to give the *Literary Digest* a piece of their mind. You'd probably want to say something like "the *Literary Digest* used a skewed sample of voters." As it happens, statisticians tend to reserve the word "skew" to describe data that is off to one side or another. In fact, you can work out the "skewness" of a set of data using a formula just as you can work out the mean and standard deviation. To describe an error in experimental methods or statistical analysis that leads to an incorrect estimate, statisticians use the word *bias*.

The *Literary Digest* study was biased because those responding to the survey were not a representative sample of American voters. A medical study might also be biased by this sort of *selection* bias. For example, a study might examine the survival rates of heart attack patients, comparing those undergoing a heart operation with those not treated surgically. This study would be biased because some patients are too sick to go through an operation; only the healthier patients go for surgery. So you would expect survival rates to be better in surgical patients even if the surgery didn't help at all.

You can also get bias even if you select your sample carefully and fairly. For example, if you are asking people questions as part of your study, the way that you ask them can introduce bias. An obvious example is "push polling," where political campaigns conduct phony opinion polls: "If you found out that Brown, candidate for governor, had fathered four children out of marriage and had paid off a judge to escape a bribery charge, would that make you more or less likely to vote for him?" My favorite case of biased questioning came in a study suggesting that rates of adultery were much lower than previously thought, only 2–3% rather than 15–20%. It turned out that the researchers had interviewed married couples sitting together in their own home. This is hardly likely to encourage frank answers to personal questions about something which almost everyone feels is wrong and tries to hide.

There are numerous other types of bias, each with their own name (if you are interested, a colleague and I have studied what is known as *verification bias*). But it is probably not worth remembering them all and what they mean, especially as statisticians themselves disagree on what to call things (when explaining our research at a conference, someone said, "Oh, you mean detection bias."). What you need to remember is that skew is out there as part of the world (the Spanish player did skew his shot on goal); bias is what we can sometimes introduce when we study the world but, like a referee, it is something we should try to avoid.

• Things to Remember •

1. Skewness describes the distribution of data.
2. Data are skewed if there are more observations below the mean than above, or vice versa.
3. The greater the proportion of observations above or below the mean, the more skewness you have.
4. Bias describes a problem with the design, conduct or analysis of a study.
5. A study is biased if the methods or statistical analyses cause an estimate to be too high or too low.
6. My joy was short lived. England lost the next time they played—a game they should have won.

SEE ALSO: *If the normal distribution is so normal, how come my data never are?*

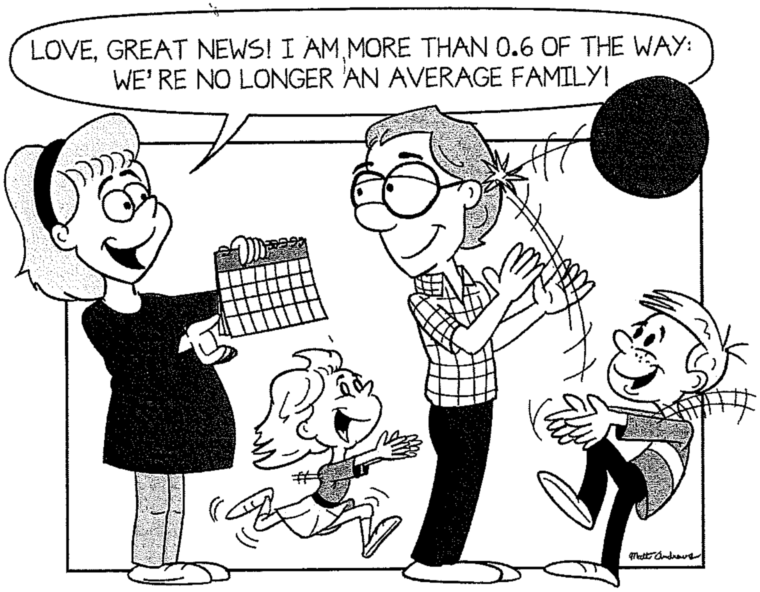
for Discussion

1. How would you avoid selection bias in the surgery study?
2. Imagine that you were conducting a study on cheating at college. Like the adultery research, this involves questions about bad behavior. How would you encourage truthful answers?

NOTE: See page 158 for answer sets.

CHAPTER 5

You can't have 2.6 children: On different types of data


LOVE, GREAT NEWS! I AM MORE THAN 0.6 OF THE WAY: WE'RE NO LONGER AN AVERAGE FAMILY!
Matt Andrews

My mother has a statistical insight

Growing up, my mother used to tell me, "Statisticians have got it wrong: you can't have 2.6 children." Now before any amateur psychologists out there chime in to tell me that I obviously became a statistician to annoy my mother, let me say that: (a) she is completely correct and (b) she has hit upon something quite profound. My mother's comment came after a survey finding that the "average" British woman had 2.6 children. This "average" was clearly what statisticians call a mean. To get a mean, you add everything up and divide by the number of observations, so you'd get a mean of 2.6 if there were, say, 1 million women and a total of 2.6 million children.

The other type of average most often used by statisticians is the median, the number higher than half the observations. My guess would be that the median number of children per woman in the survey was 2, that is, 50% of women have 0, 1 or 2 children, and 50% have 2, 3 or 4 (or some number even less conducive to a peaceful Sunday morning).

From my mother's point of view, the big difference between the mean and the median is that one is an artificial abstraction—a mathematical calculation—and the other relates to something you can actually go out and see. Generally speaking, someone in a data set will have a value at the median (a family of 2 kids); that often isn't the case for the mean (you can't have 2.6 children). The same is true of the measures of spread generally reported alongside means (standard deviation) and medians (interquartile range).

Why not stick to the median?

So why use "artificial" numbers like the mean or standard deviation and risk giving my mother something to complain about? The quick answer is that you can use the artificial numbers to answer a whole host of questions; if you want to use the real data, you have to look it up each time. Let's imagine that we had a data set consisting of heights from a sample of 10-year-old boys. A well known rule of thumb is that "95% of observations are within two standard deviations of the mean." But you can also calculate that about two-thirds of observations are within one standard deviation of the mean, that 90% of observations are greater than the mean minus 1.28 standard deviations and, just to show that you can do pretty much whatever you want, it is also true that 86.4% of observations are less than the mean plus 1.1 standard deviations. So, give me a mean and a standard deviation and I can immediately answer questions such as: What height is exceeded by only 5% of boys? What proportion of boys are more than 5 feet tall? What are the heights between which 50% of the boys' heights can be found? There is no need to go back to the data set and look anything up.

I can also do some hypothesis testing. If I have the mean and standard deviations of boys raised as vegans, and similar data from a control group from the general population, I can work out whether avoiding animal products affects growth in boys.

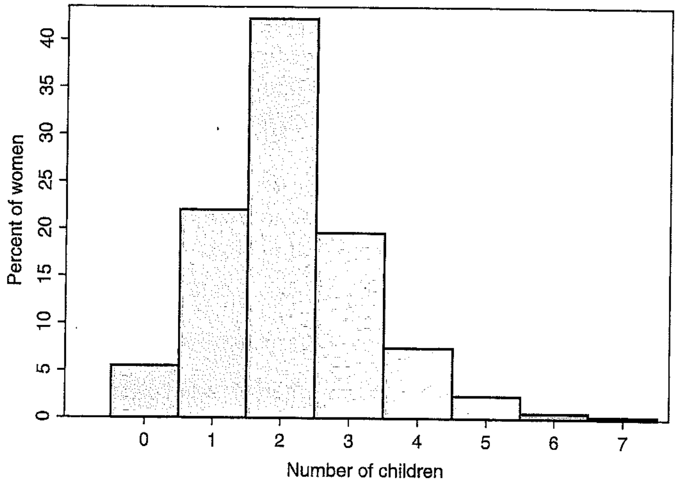
How to use 2.6

As you know, mothers are usually right, and my mother was certainly right in this particular case—citing an average of 2.6 children per woman is a bit of a silly statistic. We use statistics to do one of two things: *estimate* or *infer.* Take, for instance, the data on the height of 10-year-old boys. Estimates for the average height (e.g., a mean) and how it varies (e.g., a standard deviation) are great for description. We could look at these statistics and get a great idea about the usual heights of 10-year-old boys and just how common it is to be more than, say, 6 inches taller or smaller than average. We might then use our statistics to do something useful, like design a playground, or choose a range of sizes for a line of rain coats.

Inference—hypothesis testing—also has a practical application. Say we conduct a statistical hypothesis test to compare the heights of boys raised on vegan and regular diets, and conclude that the vegan diet does indeed appear to inhibit growth. We might then consider advising parents about the implications of raising their children on restrictive diets.

It is not clear how we could use "an average of 2.6 children" for testing a hypothesis. For example, imagine that we had a data set of women, giving the number of children they had given birth to and where they lived. We can't use the mean number of children to test a hypothesis such as "women in rural areas have more children than those living in cities." This is because statistical tests that use means assume that data can take a lot of different values, what statisticians call a *continuous* (or *quantitative*) variable. Height is a good example of a continuous variable: a 10-year old boy can be 4 ft 2 in., or 5 ft 2 in. or 4 ft 7 $^1/_2$ or pretty much anywhere in between. Family size is more like what is called a *categorical* variable, because almost all families (in industrialized nations, at least) have a family size in one of a limited number of categories: zero, one, two, three or four children, and so on. Statisticians don't like to use the sort of statistical tests designed for continuous variables on categorical variables like family size.

Moreover, "an average of 2.6 children" doesn't even give us a clear idea of how many children women typically bear. It sounds as though most women have either 2 or 3 children, but that might not be it at all; it might be that most women have either 1 or 2 children, and a few women have lots and lots. This is an example of the general rule that a single number can rarely be used to describe data and more specifically, that measures of spread should be reported alongside estimates. The problem is that the measure of spread associated with the mean is the standard deviation, and this means very little if data are skewed (see *Bill Gates goes back to the diner: Standard deviation and interquartile range)*. For example, here is a histogram with some example data for the number of children per woman:



I couldn't actually find a data set with a mean of 2.6 (I think the survey my mother was referring to was conducted in 1968 or something). This data set has a mean of 2.1 and a standard deviation of 1.1. Applying the rule that 95% of the data are within 2 standard deviations of the mean, you get that 95% of families have between −0.1 and 4.3 children—having a negative number of children is even more ridiculous than having 2.6. So, what statistics should we use in place of our mean of 2.1? The median (2) and interquartile range (1 to 3) aren't that helpful. For example, you know that 25% of families have 1 or fewer children, but (a) I would have kind of guessed that and (b) this would be true if either 1% or 24% of families were childless.

If not means or medians, then we are left with the histogram. And that's fine; it tells you pretty much everything you need to know. Or you could have a table like this, from which you could deduce not only that, for example, 7.4% of women have 4 children, but that 96.7% have 4 or fewer and 3.3% have more than 4.

Number of children in the family	Percentage	Cumulative percentage
0	5.5	5.5
1	22.0	27.5
2	42.2	69.7
3	19.6	89.3
4	7.4	96.7
5	2.4	99.1
6	0.6	99.7
7	0.3	100

True, we have lost the sound bite, with neither the table nor the graph as punchy as "an average of 2.6 children." And neither is likely to annoy my mother, which I see as a considerable downside. But here, finally, is something you could actually use.

• Things to Remember •

1. Statisticians sometimes calculate numbers from a data set, the mean and the standard deviation being good examples.
2. These numbers often take values that don't occur on the data set.
3. Statisticians sometimes choose particular numbers from the data as being illustrative, such as the median and quartiles.
4. Means, medians, standard deviations and interquartile ranges are used to describe variables that can take a large number of different values. These are known as continuous or quantitative variables.
5. Variables are sometimes best described in terms of categories: sometimes means and medians aren't used and the statistician just gives the number and percentage in each category.

SEE ALSO: *Numbers that mean something: Linking math and science*

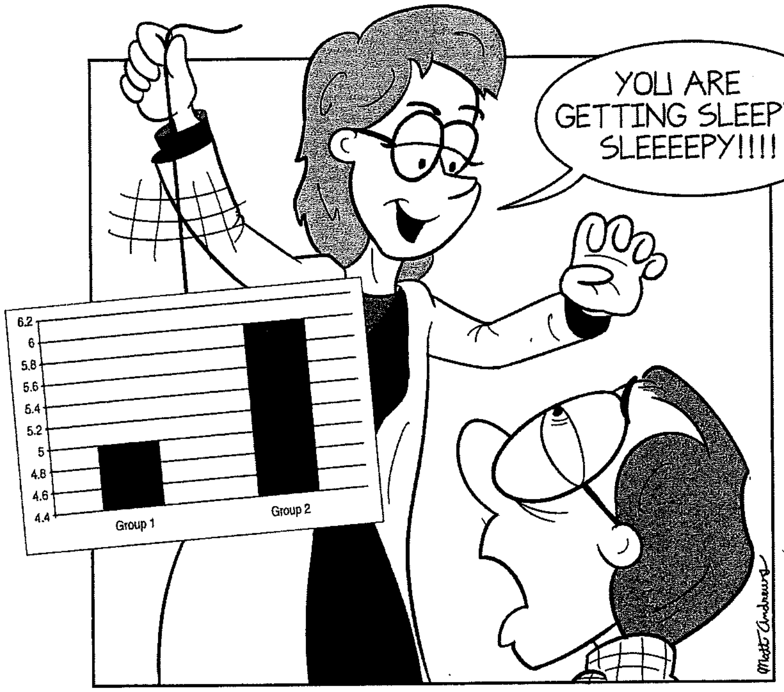
for Discussion

1. Does the median (or, say, upper quartile) always take a number that is part of the data set?
2. I described a continuous variable as one that can take "a lot of different values." How many different values is "a lot?"
3. Here are some variables. Which of these are continuous and which are categorical?
   - **a.** Height
   - **b.** Gender
   - **c.** Years of education
   - **d.** Pain score
   - **e.** Depression
   - **f.** Income
   - **g.** Race
   - **h.** Unemployment rate
4. Saying that an "average of 2.6 children is a silly statistic" allowed me to make some nice teaching points about different types of data. But as it happens, the "average" number of children that a woman bears over the course of her lifetime is actually pretty useful. How do you think that this statistic is used?

NOTE: See page 159 for answer sets.

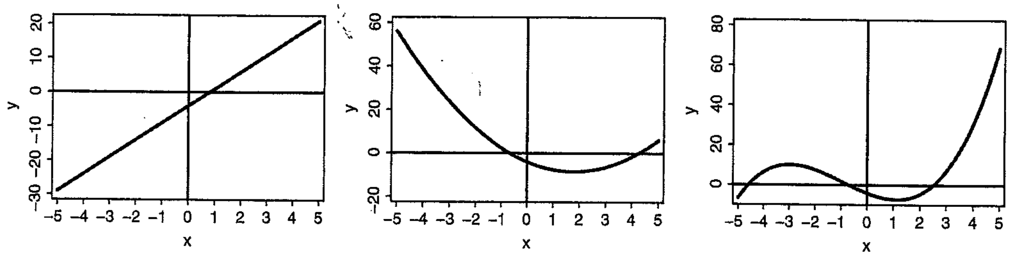
CHAPTER 6

Why your high school math teacher was right: How to draw a graph

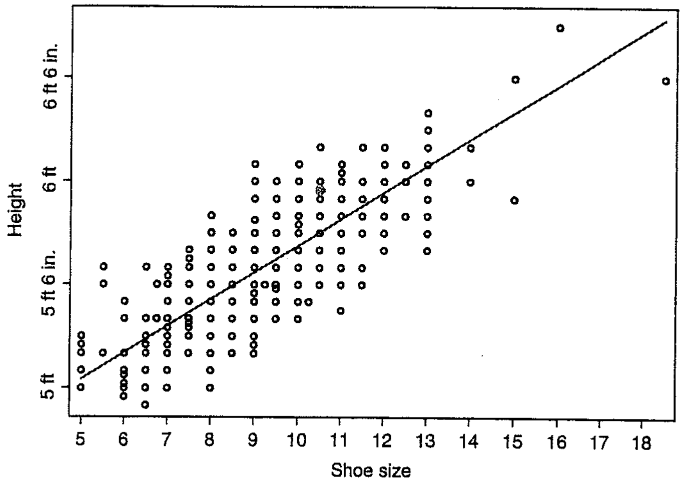


Another difference between normal people and statisticians: a normal person might say, "High school . . . those were the happiest days of my life." Statisticians tend to be more particular in stating that it was "high school math" that made them happiest. As it happens, much of the math I did in high school is far more advanced than what I do in my day to day work. (I don't need calculus on the average Tuesday morning.) But one thing I learned really stuck with me—how to draw a graph.

What I was taught about graphs was that you have an x-axis and a y-axis and, to draw a line, you state the value of y in terms of x, for example, $y = 5x - 4$. I was also taught that you can also put in other powers of x (e.g., $y = 1.4x^2 - 5x - 4$). Doing so allows you to have a curve, rather than a straight line (this is called a "non-linear" relationship). The number of times this curve can change direction is related to the number of different x terms that you have. From left to right, here are graphs for $y = 5x - 4, y = 1.4x^2 - 5x - 4, y = 0.5x^3 + 1.4x^2 - 5x + 4$.

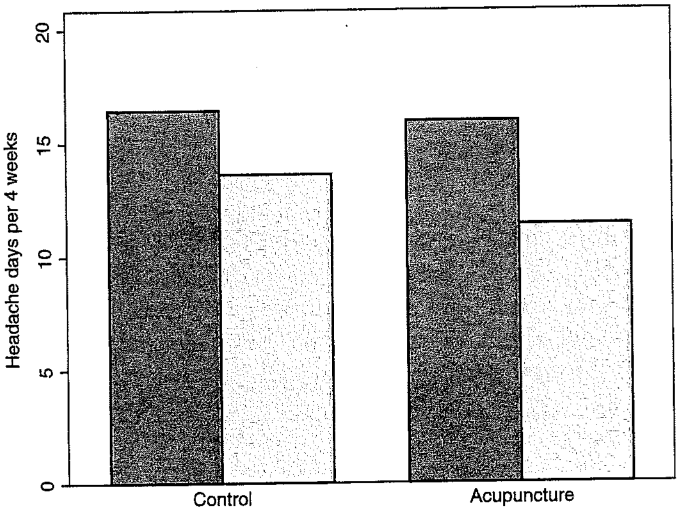


What I also did in school was to have the x- and y-axis represent something, like shoe size and height, and put a mark for each observation at the appropriate coordinate: for someone 5 ft $11^1/_2$ in. with a shoe size of $10^1/_2$ (which may or not be me, I couldn't possibly say), I'd draw an imaginary line up from the x-axis at $10^1/_2$ and an imaginary line across from the y-axis at 5 ft $11^1/_2$ in. and then put a dot where the two lines meet. I would then draw a line through the graph so that the line comes closest to the dots representing each person's height and shoe size.

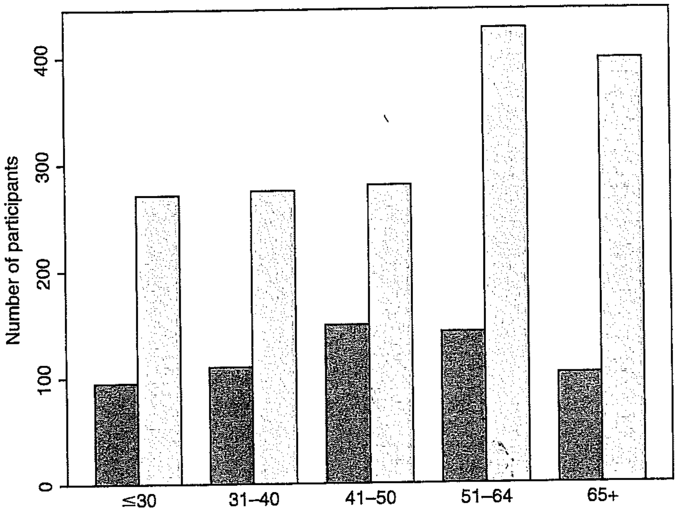


It turns out that this is just about all you need to know for most graphs. This begs the question of why good graphs are so few and far between in the scientific literature.

Here is a graph that is fairly typical of what you see in reports and papers:



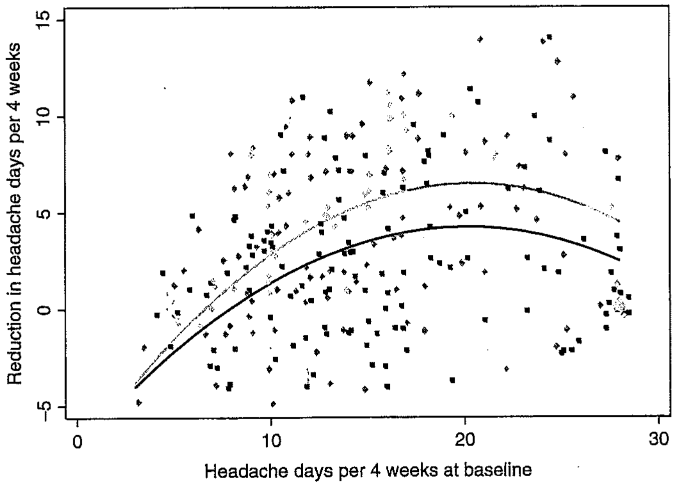
This shows the results of a clinical trial of acupuncture for headache. The bars show the number of days with headache per month before treatment (dark gray) and after treatment (light gray). Or how about this, from a survey on the state lottery:



This shows the number of survey respondents who played (dark gray bar) or did not play (light gray bar) the lottery in the previous 12 months, separately for different ages.

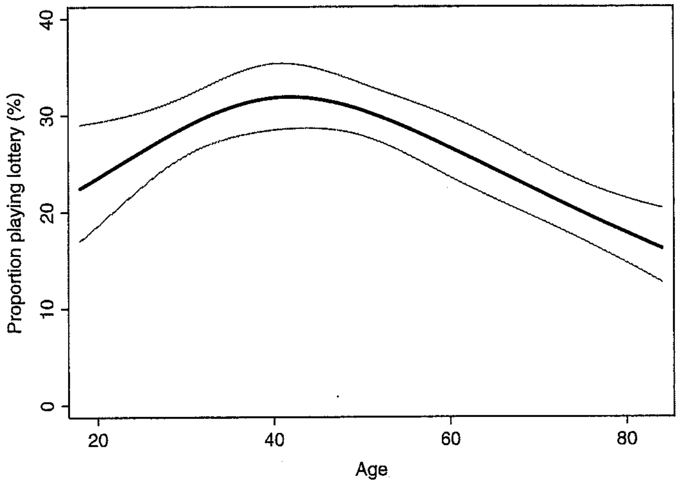
Now, for those of you still awake, a short review of why those graphs are so bad: (a) they are boring to look at, (b) they don't give an immediate visual impression of the results and (c) they don't provide information that anyone could actually use.

So, back to high school: let's present the results on an x and y graph. Here are the results of the acupuncture trial:



The x-axis shows patients' level of headache at the start of the trial; the y axis shows whether they got better (a reduction of "+5" means that the patient had 5 fewer days with a headache at the end of the trial than they had at the beginning). The gray diamonds show the results of each patient in the acupuncture group; the black squares show the results of the patients in the control group. In general, it looks as though, irrespective of baseline headache, the gray dots are higher up the graph. This suggests a greater reduction in headache in the acupuncture group. I have also shown a gray line drawn to come closest to all the gray diamonds and a black line drawn to come closest to the black squares. These lines give the expected outcome of treatment: if a patient went to a doctor and said, "I have a headache about 20 days a month." The doctor could say, "A year from now you would expect to have a reduction of about 3 days a month. If you have acupuncture treatment, you should expect to have a reduction of 5 days a month." Another nice thing about this graph is that it shows the actual results of each patient in the trial. It's what graphs are meant to do—show the data.

Here are the results from the lottery survey:



The black line shows the proportion of survey participants playing the lottery in terms of age. The thin grey lines show what is known as the 95% confidence interval, which reflect a plausible range for the relationship between age and lottery playing (see *How to avoid a rainy wedding: Variation and confidence*). You can instantly see that young and old play the lottery less than those of middle age, with a peak around age 42. The nice thing about this graph is that it gives an immediate visual impression of the data. The graph might, for example, be of interest to a psychologist trying to understand gambling addiction.

A final point: have a look at the far left of the acupuncture graph. It looks as though acupuncture actually makes things worse for patients who didn't have many headaches at baseline. But look again and you'll see that this is based on only a small number of patients. This is a reminder that we have to be careful about applying average statistical results to patients at the extremes. Then again, at least you have something to apply other than a boring gray bar.

• Things to Remember •

1. Often in research, we want to understand one thing (such as playing the lottery) in terms of another (such as age).
2. The thing we want to understand can be called *y*.
3. The thing we use to understand *y* is called *x*. So *y* might be lottery playing and *x* might be age.
4. Drawing a graph of *y* and *x*, just like you did in high school, is a good way of understanding the relationship between them.
5. You can mark a point at the *x* and *y* of each observation (this is called a *scatterplot*).
6. You can also draw a line or curve closest to each point.
7. You can use different colors for your points and lines to indicate different categories (such as headache separately for patients who did and did not receive acupuncture).

for Discussion

1. Can you always draw a line?

NOTE: See page 162 for answer sets.