Alan Krueger's study of terrorists did not follow thousands of youth over multiple decades to observe which of them evolved into terrorists. It's just not possible. Nor can we create two identical nations—except that one is highly repressive and the other is not—and then compare the number of suicide bombers that emerge in each. Even when we can conduct large, controlled experiments on human beings, they are neither easy nor cheap. Researchers did a large-scale study on whether or not prayer reduces postsurgical complications, which was one of the questions raised earlier in this chapter. *That study cost $2.4 million.* (For the results, you'll have to wait until Chapter 13.)

Secretary of Defense Donald Rumsfeld famously said, "You go to war with the army you have—not the army you might want or wish to have at a later time." Whatever you may think of Rumsfeld (and the Iraq war that he was explaining), that aphorism applies to research, too. We conduct statistical analysis using the best data and methodologies and resources available. The approach is not like addition or long division, in which the correct technique yields the "right" answer and a computer is always more precise and less fallible than a human. Statistical analysis is more like good detective work (hence the commercial potential of *CSI: Regression Analysis*). Smart and honest people will often disagree about what the data are trying to tell us.

But who says that everyone using statistics is smart or honest? As mentioned, this book began as an homage to *How to Lie with Statistics*, which was first published in 1954 and has sold over a million copies. The reality is that you *can* lie with statistics. Or you can make inadvertent errors. In either case, the mathematical precision attached to statistical analysis can dress up some serious nonsense. This book will walk through many of the most common statistical errors and misrepresentations (so that you can recognize them, not put them to use).

So, to return to the title chapter, what is the point of learning statistics?

To summarize huge quantities of data.

To make better decisions.

To answer important social questions.

To recognize patterns that can refine how we do everything from selling diapers to catching criminals.

To catch cheaters and prosecute criminals.

To evaluate the effectiveness of policies, programs, drugs, medical procedures, and other innovations.

And to spot the scoundrels who use these very same powerful tools for nefarious ends.

If you can do all of that while looking great in a Hugo Boss suit or a short black skirt, then you might also be the next star of *CSI: Regression Analysis*.

* The Gini index is sometimes multiplied by 100 to make it a whole number. In that case, the United States would have a Gini Index of 45.

* The word "data" has historically been considered plural (e.g., "The data are very encouraging.") The singular is "datum," which would refer to a single data point, such as one person's response to a single question on a poll. Using the word "data" as a plural noun is a quick way to signal to anyone who does serious research that you are conversant with statistics. That said, many authorities on grammar and many publications, such as the *New York Times*, now accept that "data" can be singular or plural, as the passage that I've quoted from the *Times* demonstrates.

* This is a gross simplification of the fascinating and complex field of medical ethics.

CHAPTER 2

Descriptive Statistics

Who was the best baseball player of all time?

Let us ponder for a moment two seemingly unrelated questions: (1) What is happening to the economic health of America's middle class? and (2) Who was the greatest baseball player of all time?

The first question is profoundly important. It tends to be at the core of presidential campaigns and other social movements. The middle class is the heart of America, so the economic well-being of that group is a crucial indicator of the nation's overall economic health. The second question is trivial (in the literal sense of the word), but baseball enthusiasts can argue about it endlessly. What the two questions have in common is that they can be used to illustrate the strengths and limitations of descriptive statistics, which are the numbers and calculations we use to summarize raw data.

If I want to demonstrate that Derek Jeter is a great baseball player, I can sit you down and describe every at bat in every Major League game that he's played. That would be raw data, and it would take a while to digest, given that Jeter has played seventeen seasons with the New York Yankees and taken 9,868 at bats.

Or I can just tell you that at the end of the 2011 season Derek Jeter had a career batting average of .313. That is a descriptive statistic, or a "summary statistic."

The batting average is a gross simplification of Jeter's seventeen seasons. It is easy to understand, elegant in its simplicity—and limited in what it can tell us. Baseball experts have a bevy of descriptive statistics that they consider to be more valuable than the batting average. I called Steve Moyer, president of Baseball Info Solutions (a firm that provides a lot of the raw data for the *Moneyball* types), to ask him, (1) What are the most important statistics for evaluating baseball talent? and (2) Who was the greatest player of all time? I'll share his answer once we have more context.

Meanwhile, let's return to the less trivial subject, the economic health of the middle class. Ideally we would like to find the economic equivalent of a batting average, or something even better. We would like a simple but accurate measure of how the economic well-being of the typical American worker has been changing in recent years. Are the people we define as middle class getting richer, poorer, or just running in place? A reasonable answer—though by no means the "right" answer—would be to calculate the change in per capita income in the United States over the course of a generation, which is roughly thirty years. Per capita income is a simple average: total income divided by the size of the population. By that measure, average income in

the United States climbed from $7,787 in 1980 to $26,487 in 2010 (the latest year for which the government has data).[1] Voilà! Congratulations to us.

There is just one problem. My quick calculation is technically correct and yet totally wrong in terms of the question I set out to answer. To begin with, the figures above are not adjusted for inflation. (A per capita income of $7,787 in 1980 is equal to about $19,600 when converted to 2010 dollars.) That's a relatively quick fix. The bigger problem is that the average income in America is not equal to the income of the average American. Let's unpack that clever little phrase.

Per capita income merely takes all of the income earned in the country and divides by the number of people, which tells us absolutely nothing about who is earning how much of that income—in 1980 or in 2010. As the Occupy Wall Street folks would point out, explosive growth in the incomes of the top 1 percent can raise per capita income significantly without putting any more money in the pockets of the other 99 percent. In other words, average income can go up without helping the average American.

As with the baseball statistic query, I have sought outside expertise on how we ought to measure the health of the American middle class. I asked two prominent labor economists, including President Obama's top economic adviser, what descriptive statistics they would use to assess the economic well-being of a typical American. Yes, you will get that answer, too, once we've taken a quick tour of descriptive statistics to give it more meaning.

From baseball to income, the most basic task when working with data is to summarize a great deal of information. There are some 330 million residents in the United States. A spreadsheet with the name and income history of every American would contain all the information we could ever want about the economic health of the country—yet it would also be so unwieldy as to tell us nothing at all. The irony is that more data can often present less clarity. So we simplify. We perform calculations that reduce a complex array of data into a handful of numbers that describe those data, just as we might encapsulate a complex, multifaceted Olympic gymnastics performance with one number: 9.8.

The good news is that these descriptive statistics give us a manageable and meaningful summary of the underlying phenomenon. That's what this chapter is about. The bad news is that any simplification invites abuse. Descriptive statistics can be like online dating profiles: technically accurate and yet pretty darn misleading.

Suppose you are at work, idly surfing the Web when you stumble across a riveting day-by-day account of Kim Kardashian's failed seventy-two-day marriage to professional basketball player Kris Humphries. You have finished reading about day seven of the marriage when your boss shows up with two enormous files of data. One file has warranty claim information for each of the 57,334 laser printers that your firm sold last year. (For each printer sold, the file documents the number of quality problems that were reported during the warranty period.) The other file has the same information for each of the 994,773 laser printers that your chief competitor sold during the same stretch. Your boss wants to know how your firm's printers compare in terms of quality with the competition.

Fortunately the computer you've been using to read about the Kardashian marriage has a basics statistics package, but where do you begin? Your instincts are probably correct: The first descriptive task is often to find some measure of the "middle" of a set of data, or what statisticians might describe as its "central tendency." What is the typical quality experience for your printers compared with those of the competition? The most basic measure of the "middle" of a distribution is the mean, or average. In this case, we want to know the average number of quality problems per printer sold for your firm and for your competitor. You would simply tally the total number of quality problems reported for all printers during the warranty period and then divide by the total number of printers sold. (Remember, the same printer can have multiple problems while under warranty.) You would do that for each firm, creating an important descriptive statistic: the average number of quality problems per printer sold.

Suppose it turns out that your competitor's printers have an average of 2.8 quality-related problems per printer during the warranty period compared with your firm's average of 9.1 reported defects. That was easy. You've just taken information on a million printers sold by two different companies and distilled it to the essence of the problem: your printers break a lot. Clearly it's time to send a short e-mail to your boss quantifying this quality gap and then get back to day eight of Kim Kardashian's marriage.

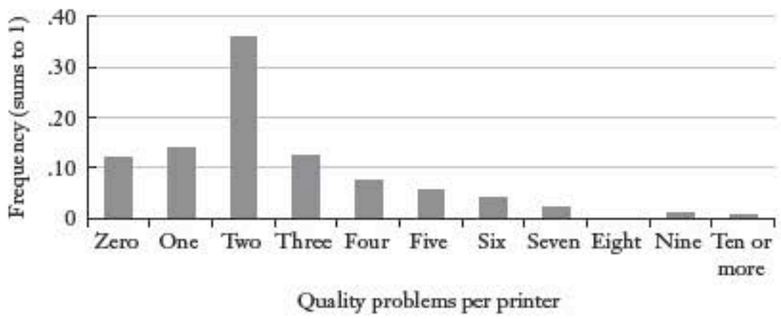
Or maybe not. I was deliberately vague earlier when I referred to the "middle" of a distribution. The mean, or average, turns out to have some problems in that regard, namely, that it is prone to distortion by "outliers," which are observations that lie farther from the center. To get your mind around this concept, imagine that ten guys are sitting on bar stools in a middle-class drinking establishment in Seattle; each of these guys earns $35,000 a year, which makes the mean annual income for the group $35,000. Bill Gates walks into the bar with a talking parrot perched on his shoulder. (The parrot has nothing to do with the example, but it kind of spices things up.) Let's assume for the sake of the example that Bill Gates has an annual income of $1 billion. When Bill sits down on the eleventh bar stool, the mean annual income for the bar patrons rises to about $91 million. Obviously none of the original ten drinkers is any richer (though it might be reasonable to expect Bill Gates to buy a round or two). If I were to describe the patrons of this bar as having an average annual income of $91 million, the statement would be both statistically correct and grossly misleading. This isn't a bar where multimillionaires hang out; it's a bar where a bunch of guys with relatively low incomes happen to be sitting next to Bill Gates and his talking parrot. The sensitivity of the mean to outliers is why we should not gauge the economic health of the American middle class by looking at per capita income. Because there has been explosive growth in incomes at the top end of the distribution—CEOs, hedge fund managers, and athletes like Derek Jeter—the average income in the United States could be heavily skewed by the megarich, making it look a lot like the bar stools with Bill Gates at the end.

For this reason, we have another statistic that also signals the "middle" of a distribution, albeit differently: the median. The median is the point that divides a distribution in half,

meaning that half of the observations lie above the median and half lie below. (If there is an even number of observations, the median is the midpoint between the two middle observations.) If we return to the bar stool example, the median annual income for the ten guys originally sitting in the bar is $35,000. When Bill Gates walks in with his parrot and perches on a stool, the median annual income for the eleven of them is still $35,000. If you literally envision lining up the bar patrons on stools in ascending order of their incomes, the income of the guy sitting on the sixth stool represents the median income for the group. If Warren Buffett comes in and sits down on the twelfth stool next to Bill Gates, the median still does not change.*

For distributions without serious outliers, the median and the mean will be similar. I've included a hypothetical summary of the quality data for the competitor's printers. In particular, I've laid out the data in what is known as a frequency distribution. The number of quality problems per printer is arrayed along the bottom; the height of each bar represents the percentages of printers sold with that number of quality problems. For example, 36 percent of the competitor's printers had two quality defects during the warranty period. Because the distribution includes all possible quality outcomes, including zero defects, the proportions must sum to 1 (or 100 percent).

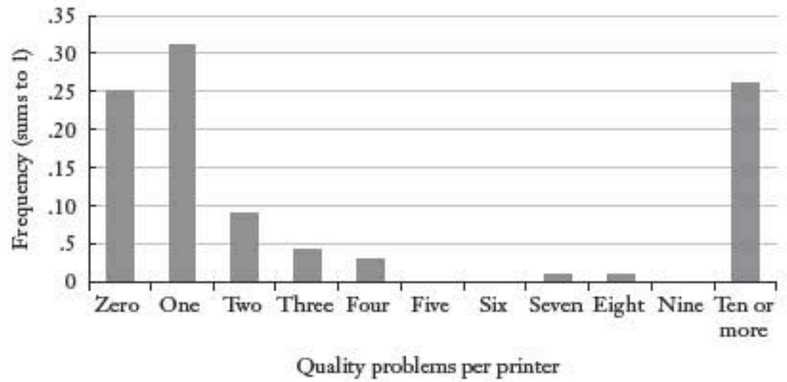
Frequency Distribution of Quality Complaints for Competitor's Printers



Because the distribution is nearly symmetrical, the mean and median are relatively close to one another. The distribution is slightly skewed to the right by the small number of printers with many reported quality defects. These outliers move the mean slightly rightward but have no impact on the median. Suppose that just before you dash off the quality report to your boss you decide to calculate the *median* number of quality problems for your firm's printers and the competition's. With a few keystrokes, you get the result. The median number of quality complaints for the competitor's printers is 2; the median number of quality complaints for your company's printers is 1.

Huh? Your firm's median number of quality complaints per printer is actually *lower* than your competitor's. Because the Kardashian marriage is getting monotonous, and because you are intrigued by this finding, you print a frequency distribution for your own quality problems.

Frequency Distribution of Quality Complaints at Your Company



What becomes clear is that your firm does not have a uniform quality problem; you have a "lemon" problem; a small number of printers have a huge number of quality complaints. These outliers inflate the mean but not the median. More important from a production standpoint, you do not need to retool the whole manufacturing process; you need only figure out where the egregiously low-quality printers are coming from and fix that.*

Neither the median nor the mean is hard to calculate; the key is determining which measure of the "middle" is more accurate in a particular situation (a phenomenon that is easily exploited). Meanwhile, the median has some useful relatives. As we've already discussed, the median divides a distribution in half. The distribution can be further divided into quarters, or quartiles. The first quartile consists of the bottom 25 percent of the observations; the second quartile consists of the next 25 percent of the observations; and so on. Or the distribution can be divided into deciles, each with 10 percent of the observations. (If your income is in the top decile of the American income distribution, you would be earning more than 90 percent of your fellow workers.) We can go even further and divide the distribution into hundredths, or percentiles. Each percentile represents 1 percent of the distribution, so that the 1st percentile represents the bottom 1 percent of the distribution and the 99th percentile represents the top 1 percent of the distribution.

The benefit of these kinds of descriptive statistics is that they describe where a particular observation lies compared with everyone else. If I tell you that your child scored in the 3rd percentile on a reading comprehension test, you should know immediately that the family should be logging more time at the library. You don't need to know anything about the test itself, or the number of questions that your child got correct. The percentile score provides a ranking of your child's score relative to that of all the other test takers. If the test was easy, then most test takers will have a high number of answers correct, but your child will have fewer correct than most of the others. If the test was extremely difficult, then all the test takers will have a low number of correct answers, but your child's score will be lower still.

Here is a good point to introduce some useful terminology. An "absolute" score, number, or figure has some intrinsic meaning. If I shoot 83 for eighteen holes of golf, that is an absolute

figure. I may do that on a day that is 58 degrees, which is also an absolute figure. Absolute figures can usually be interpreted without any context or additional information. When I tell you that I shot 83, you don't need to know what other golfers shot that day in order to evaluate my performance. (The exception might be if the conditions are particularly awful, or if the course is especially difficult or easy.) If I place ninth in the golf tournament, that is a relative statistic. A "relative" value or figure has meaning only in comparison to something else, or in some broader context, such as compared with the eight golfers who shot better than I did. Most standardized tests produce results that have meaning only as a relative statistic. If I tell you that a third grader in an Illinois elementary school scored 43 out of 60 on the mathematics portion of the Illinois State Achievement Test, that absolute score doesn't have much meaning. But when I convert it to a percentile—meaning that I put that raw score into a distribution with the math scores for all other Illinois third graders—then it acquires a great deal of meaning. If 43 correct answers falls into the 83rd percentile, then this student is doing better than most of his peers statewide. If he's in the 8th percentile, then he's really struggling. In this case, the percentile (the relative score) is more meaningful than the number of correct answers (the absolute score).

Another statistic that can help us describe what might otherwise be a jumble of numbers is the standard deviation, which is a measure of how dispersed the data are from their mean. In other words, how spread out are the observations? Suppose I collected data on the weights of 250 people on an airplane headed for Boston, and I also collected the weights of a sample of 250 qualifiers for the Boston Marathon. Now assume that the mean weight for both groups is roughly the same, say 155 pounds. Anyone who has been squeezed into a row on a crowded flight, fighting for the armrest, knows that many people on a typical commercial flight weigh more than 155 pounds. But you may recall from those same unpleasant, overcrowded flights that there were lots of crying babies and poorly behaved children, all of whom have enormous lung capacity but not much mass. When it comes to calculating the average weight on the flight, the heft of the 320-pound football players on either side of your middle seat is likely offset by the tiny screaming infant across the row and the six-year-old kicking the back of your seat from the row behind.

On the basis of the descriptive tools introduced so far, the weights of the airline passengers and the marathoners are nearly identical. *But they're not.* Yes, the weights of the two groups have roughly the same "middle," but the airline passengers have far more dispersion around that midpoint, meaning that their weights are spread farther from the midpoint. My eight-year-old son might point out that the marathon runners look like they all weigh the same amount, while the airline passengers have some tiny people and some bizarrely large people. The weights of the airline passengers are "more spread out," which is an important attribute when it comes to describing the weights of these two groups. The standard deviation is the descriptive statistic that allows us to assign a single number to this dispersion around the mean. The formulas for calculating the standard deviation and the variance (another common measure of dispersion from which the standard deviation is derived) are included in an appendix at the end of the chapter. For now, let's think about why the measuring of dispersion matters.

Suppose you walk into the doctor's office. You've been feeling fatigued ever since your promotion to head of North American printer quality. Your doctor draws blood, and a few days later her assistant leaves a message on your answering machine to inform you that your HCb2 count (a fictitious blood chemical) is 134. You rush to the Internet and discover that the mean HCb2 count for a person your age is 122 (and the median is about the same). Holy crap! If you're like me, you would finally draft a will. You'd write tearful letters to your parents, spouse, children, and close friends. You might take up skydiving or try to write a novel very fast. You would send your boss a hastily composed e-mail comparing him to a certain part of the human anatomy—IN ALL CAPS.

None of these things may be necessary (and the e-mail to your boss could turn out very badly). When you call the doctor's office back to arrange for your hospice care, the physician's assistant informs you that your count is within the normal range. But how could that be? "My count is 12 points higher than average!" you yell repeatedly into the receiver.

"The standard deviation for the HCb2 count is 18," the technician informs you curtly.

What the heck does that mean?

There is natural variation in the HCb2 count, as there is with most biological phenomena (e.g., height). While the mean count for the fake chemical might be 122, plenty of healthy people have counts that are higher or lower. The danger arises only when the HCb2 count gets excessively high or low. So how do we figure out what "excessively" means in this context? As we've already noted, the standard deviation is a measure of dispersion, meaning that it reflects how tightly the observations cluster around the mean. For many typical distributions of data, a high proportion of the observations lie within one standard deviation of the mean (meaning that they are in the range from one standard deviation below the mean to one standard deviation above the mean). To illustrate with a simple example, the mean height for American adult men is 5 feet 10 inches. The standard deviation is roughly 3 inches. A high proportion of adult men are between 5 feet 7 inches and 6 feet 1 inch.

Or, to put it slightly differently, any man in this height range would not be considered abnormally short or tall. Which brings us back to your troubling HCb2 results. Yes, your count is 12 above the mean, but that's less than one standard deviation, which is the blood chemical equivalent of being about 6 feet tall—not particularly unusual. Of course, far fewer observations lie two standard deviations from the mean, and fewer still lie three or four standard deviations away. (In the case of height, an American man who is three standard deviations above average in height would be 6 feet 7 inches or taller.)

Some distributions are more dispersed than others. Hence, the standard deviation of the weights of the 250 airline passengers will be higher than the standard deviation of the weights of the 250 marathon runners. A frequency distribution with the weights of the airline passengers would literally be fatter (more spread out) than a frequency distribution of the weights of the marathon runners. Once we know the mean and standard deviation for any collection of data, we have some serious intellectual traction. For example, suppose I tell you

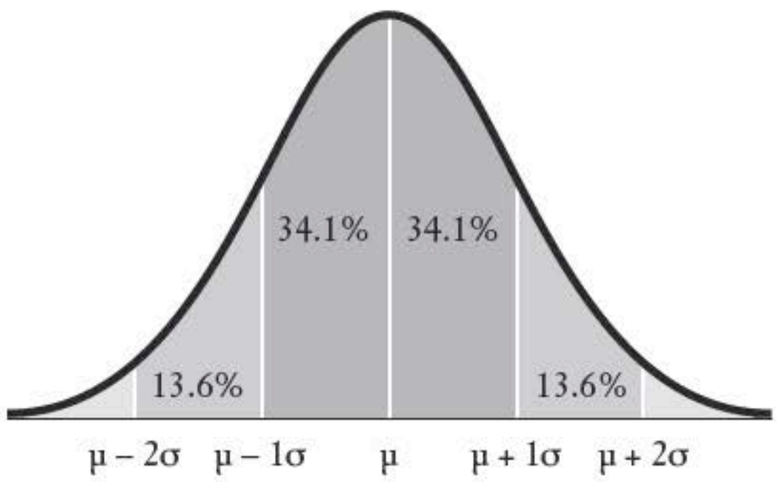
that the mean score on the SAT math test is 500 with a standard deviation of 100. As with height, the bulk of students taking the test will be within one standard deviation of the mean, or between 400 and 600. How many students do you think score 720 or higher? Probably not very many, since that is more than two standard deviations above the mean.

In fact, we can do even better than "not very many." This is a good time to introduce one of the most important, helpful, and common distributions in statistics: the normal distribution. Data that are distributed normally are symmetrical around their mean in a bell shape that will look familiar to you.

The normal distribution describes many common phenomena. Imagine a frequency distribution describing popcorn popping on a stove top. Some kernels start to pop early, maybe one or two pops per second; after ten or fifteen seconds, the kernels are exploding frenetically. Then gradually the number of kernels popping per second fades away at roughly the same rate at which the popping began. The heights of American men are distributed more or less normally, meaning that they are roughly symmetrical around the mean of 5 feet 10 inches. Each SAT test is specifically designed to produce a normal distribution of scores with mean 500 and standard deviation of 100. According to the *Wall Street Journal*, Americans even tend to park in a normal distribution at shopping malls; most cars park directly opposite the mall entrance—the "peak" of the normal curve—with "tails" of cars going off to the right and left of the entrance.

The beauty of the normal distribution—its Michael Jordan power, finesse, and elegance—comes from the fact that we know by definition exactly what proportion of the observations in a normal distribution lie within one standard deviation of the mean (68.2 percent), within two standard deviations of the mean (95.4 percent), within three standard deviations (99.7 percent), and so on. This may sound like trivia. In fact, it is the foundation on which much of statistics is built. We will come back to this point in much great depth later in the book.

The Normal Distribution



The mean is the middle line which is often represented by the Greek letter μ. The standard deviation is often represented by the Greek letter σ. Each band represents one standard deviation.

Descriptive statistics are often used to compare two figures or quantities. I'm one inch taller than my brother; today's temperature is nine degrees above the historical average for this date; and so on. Those comparisons make sense because most of us recognize the scale of the units involved. One inch does not amount to much when it comes to a person's height, so you can infer that my brother and I are roughly the same height. Conversely, nine degrees is a significant temperature deviation in just about any climate at any time of year, so nine degrees above average makes for a day that is much hotter than usual. But suppose that I told you that Granola Cereal A contains 31 milligrams more sodium than Granola Cereal B. Unless you know an awful lot about sodium (and the serving sizes for granola cereal), that statement is not going to be particularly informative. Or what if I told you that my cousin Al earned $53,000 less this year than last year? Should we be worried about Al? Or is he a hedge fund manager for whom $53,000 is a rounding error in his annual compensation?

In both the sodium and the income examples, we're missing context. The easiest way to give meaning to these relative comparisons is by using percentages. It *would* mean something if I told you that Granola Bar A has 50 percent more sodium than Granola Bar B, or that Uncle Al's income fell 47 percent last year. Measuring change as a percentage gives us some sense of scale.

You probably learned how to calculate percentages in fourth grade and will be tempted to skip the next few paragraphs. Fair enough. But first do one simple exercise for me. Assume that a department store is selling a dress for $100. The assistant manager marks down all merchandise by 25 percent. But then that assistant manager is fired for hanging out in a bar with Bill Gates,* and the new assistant manager raises all prices by 25 percent. What is the final price of the dress? If you said (or thought) $100, then you had better not skip any paragraphs.

The final price of the dress is actually $93.75. This is not merely a fun parlor trick that will win you applause and adulation at cocktail parties. Percentages are useful—but also potentially confusing or even deceptive. The formula for calculating a percentage difference (or change) is the following: (new figure – original figure)/original figure. The numerator (the part on the top of the fraction) gives us the size of the change in absolute terms; the denominator (the bottom of the fraction) is what puts this change in context by comparing it with our starting point. At first, this seems straightforward, as when the assistant store manager cuts the price of the $100 dress by 25 percent. Twenty-five percent of the original $100 price is $25; that's the discount, which takes the price down to $75. You can plug the numbers into the formula above and do some simple manipulation to get to the same place: ($100 – $75)/$100 = .25, or 25 percent.

The dress is selling for $75 when the new assistant manager demands that the price be raised 25 percent. That's where many of the people reading this paragraph probably made a mistake. The 25 percent markup is calculated as a percentage of the dress's new reduced price, which is $75. The increase will be .25($75), or $18.75, which is how the final price ends up at

$93.75 (and not $100). The point is that a percentage change always gives the value of some figure *relative to something else*. Therefore, we had better understand what that something else is.

I once invested some money in a company that my college roommate started. Since it was a private venture, there were no requirements as to what information had to be provided to shareholders. A number of years went by without any information on the fate of my investment; my former roommate was fairly tight-lipped on the subject. Finally, I received a letter in the mail informing me that the firm's profits were 46 percent higher than the year before. There was no information on the size of those profits in absolute terms, meaning that I still had absolutely no idea how my investment was performing. Suppose that last year the firm earned 27 cents—essentially nothing. This year the firm earned 39 cents—also essentially nothing. Yet the company's profits grew from 27 cents to 39 cents, which is technically a 46 percent increase. Obviously the shareholder letter would have been more of a downer if it pointed out that the firm's cumulative profits over two years were less than the cost of a cup of Starbucks coffee.

To be fair to my roommate, he eventually sold the company for hundreds of millions of dollars, earning me a 100 percent return on my investment. (Since you have no idea how much I invested, you also have no idea how much money I made—which reinforces my point here very nicely!)

Let me make one additional distinction. Percentage change must not be confused with a change in percentage points. Rates are often expressed in percentages. The sales tax rate in Illinois is 6.75 percent. I pay my agent 15 percent of my book royalties. These rates are levied against some quantity, such as income in the case of the income tax rate. Obviously the rates can go up or down; less intuitively, the *changes* in the rates can be described in vastly dissimilar ways. The best example of this was a recent change in the Illinois personal income tax, which was raised from 3 percent to 5 percent. There are two ways to express this tax change, both of which are technically accurate. The Democrats, who engineered this tax increase, pointed out (correctly) that the state income tax *rate* was increased by *2 percentage points* (from 3 percent to 5 percent). The Republicans pointed out (also correctly) that the state income tax had been raised by *67 percent*. [This is a handy test of the formula from a few paragraphs back: $(5 - 3)/3 = 2/3$, which rounds up to 67 percent.]

The Democrats focused on the absolute change in the tax rate; Republicans focused on the percentage change in the tax burden. As noted, both descriptions are technically correct, though I would argue that the Republican description more accurately conveys the impact of the tax change, since what I'm going to have to pay to the government—the amount that I care about, as opposed to the way it is calculated—really has gone up by 67 percent.

Many phenomena defy perfect description with a single statistic. Suppose quarterback Aaron Rodgers throws for 365 yards but no touchdowns. Meanwhile, Peyton Manning throws for a meager 127 yards but three touchdowns. Manning generated more points, but presumably Rodgers set up touchdowns by marching his team down the field and keeping the other team's offense off the field. Who played better? In Chapter 1, I discussed the NFL passer rating, which is the league's reasonable attempt to deal with this statistical challenge. The passer rating is an example of an index, which is a descriptive statistic made up of other descriptive statistics. Once these different measures of performance are consolidated into a single number, that statistic can be used to make comparisons, such as ranking quarterbacks on a particular day, or even over a whole career. If baseball had a similar index, then the question of the best player ever would be solved. Or would it?

The advantage of any index is that it consolidates lots of complex information into a single number. We can then rank things that otherwise defy simple comparison—anything from quarterbacks to colleges to beauty pageant contestants. In the Miss America pageant, the overall winner is a combination of five separate competitions: personal interview, swimsuit, evening wear, talent, and onstage question. (Miss Congeniality is voted on separately by the participants themselves.)

Alas, the disadvantage of any index is that it consolidates lots of complex information into a single number. There are countless ways to do that; each has the potential to produce a different outcome. Malcolm Gladwell makes this point brilliantly in a *New Yorker* piece critiquing our compelling need to rank things.[2] (He comes down particularly hard on the college rankings.) Gladwell offers the example of *Car and Driver*'s ranking of three sports cars: the Porsche Cayman, the Chevrolet Corvette, and the Lotus Evora. Using a formula that includes twenty-one different variables, *Car and Driver* ranked the Porsche number one. But Gladwell points out that "exterior styling" counts for only 4 percent of the total score in the *Car and Driver* formula, which seems ridiculously low for a sports car. If styling is given more weight in the overall ranking (25 percent), then the Lotus comes out on top.

But wait. Gladwell also points out that the sticker price of the car gets relatively little weight in the *Car and Driver* formula. If value is weighted more heavily (so that the ranking is based equally on price, exterior styling, and vehicle characteristics), the Chevy Corvette is ranked number one.

Any index is highly sensitive to the descriptive statistics that are cobbled together to build it, and to the weight given to each of those components. As a result, indices range from useful but imperfect tools to complete charades. An example of the former is the United Nations Human Development Index, or HDI. The HDI was created as a measure of economic well-being that is broader than income alone. The HDI uses income as one of its components but also includes measures of life expectancy and educational attainment. The United States ranks eleventh in the world in terms of per capita economic output (behind several oil-rich nations like Qatar, Brunei, and Kuwait) but fourth in the world in human development.[3] It's true that the HDI rankings would change slightly if the component parts of the index were reconfigured, but no reasonable change is going to make Zimbabwe zoom up the rankings past Norway. The HDI provides a handy and reasonably accurate snapshot of living standards around the globe.

Descriptive statistics give us insight into phenomena that we care about. In that spirit, we can

return to the questions posed at the beginning of the chapter. Who is the best baseball player of all time? More important for the purposes of this chapter, what descriptive statistics would be most helpful in answering that question? According to Steve Moyer, president of Baseball Info Solutions, the three most valuable statistics (other than age) for evaluating any player who is not a pitcher would be the following:

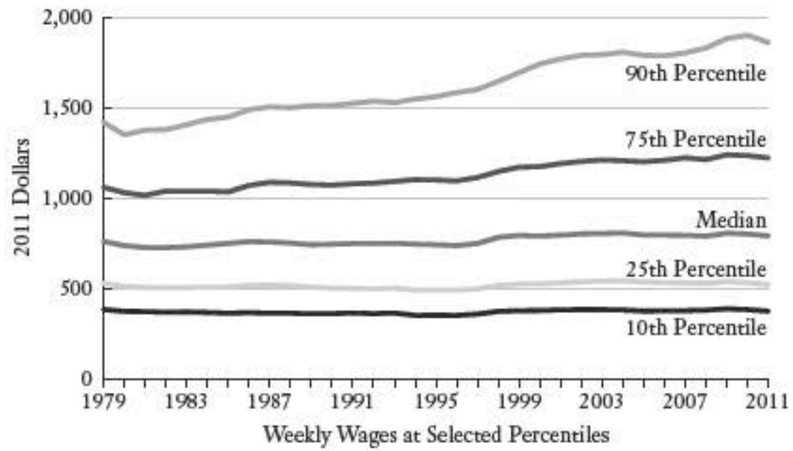
1. On-base percentage (OBP), sometimes called the on-base average (OBA): Measures the proportion of the time that a player reaches base successfully, including walks (which are not counted in the batting average).
2. Slugging percentage (SLG): Measures power hitting by calculating the total bases reached per at bat. A single counts as 1, a double is 2, a triple is 3, and a home run is 4. Thus, a batter who hit a single and a triple in five at bats would have a slugging percentage of (1 + 3)/5, or .800.
3. At bats (AB): Puts the above in context. Any mope can have impressive statistics for a game or two. A superstar compiles impressive "numbers" over thousands of plate appearances.

In Moyer's view (without hesitation, I might add), the best baseball player of all time was Babe Ruth because of his unique ability to hit and to pitch. Babe Ruth still holds the Major League career record for slugging percentage at .690.[4]

What about the economic health of the American middle class? Again, I deferred to the experts. I e-mailed Jeff Grogger (a colleague of mine at the University of Chicago) and Alan Krueger (the same Princeton economist who studied terrorists and is now serving as chair of President Obama's Council of Economic Advisers). Both gave variations on the same basic answer. To assess the economic health of America's "middle class," we should examine changes in the median wage (adjusted for inflation) over the last several decades. They also recommended examining changes to wages at the 25th and 75th percentiles (which can reasonably be interpreted as the upper and lower bounds for the middle class).

One more distinction is in order. When assessing economic health, we can examine income or wages. They are not the same thing. A wage is what we are paid for some fixed amount of labor, such as an hourly or weekly wage. Income is the sum of all payments from different sources. If workers take a second job or work more hours, their income can go up without a change in the wage. (For that matter, income can go up even if the wage is falling, provided a worker logs enough hours on the job.) However, if individuals have to work more in order to earn more, it's hard to evaluate the overall effect on their well-being. The wage is a less ambiguous measure of how Americans are being compensated for the work they do; the higher the wage, the more workers take home for every hour on the job.

Having said all that, here is a graph of American wages over the past three decades. I've also added the 90th percentile to illustrate changes in the wages for middle-class workers compared over this time frame to those workers at the top of the distribution.



Source: "Changes in the Distribution of Workers' Hourly Wages between 1979 and 2009," Congressional Budget Office, February 16, 2011. The data for the chart can be found at http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/120xx/doc12051/02-16-wagedispersion.pdf.

A variety of conclusions can be drawn from these data. They do not present a single "right" answer with regard to the economic fortunes of the middle class. They do tell us that the typical worker, an American worker earning the median wage, has been "running in place" for nearly thirty years. Workers at the 90th percentile have done much, much better. Descriptive statistics help to frame the issue. What we do about it, if anything, is an ideological and political question.

APPENDIX TO CHAPTER 2

Data for the printer defects graphics

	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten or more
Frequency of competitor's defects	12	14	36	13	8	6	5	3	0	2	1
	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten or more
Frequency of your defects	25	31	9	4	3	0	0	1	1	0	26

Formula for variance and standard deviation

Variance and standard deviation are the most common statistical mechanisms for measuring and describing the dispersion of a distribution. The variance, which is often represented by the

symbol σ^2, is calculated by determining how far the observations within a distribution lie from the mean. However, the twist is that the difference between each observation and the mean is squared; the sum of those squared terms is then divided by the number of observations.

Specifically:

For any set of *n* observations $x_1, x_2, x_3 \ldots x_n$ with mean μ,

$$\text{Variance} = \sigma^2 = [(x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + \ldots (x_n - \mu)^2]/n$$

Because the difference between each term and the mean is squared, the formula for calculating variance puts particular weight on observations that lie far from the mean, or outliers, as the following table of student heights illustrates.

Group 1	Height (μ = 70 inches)	Distance from the mean = Absolute value of $(x_n - \mu)$*	$(x_n - \mu)^2$	Group 2	Height (μ = 70 inches)	Distance from the mean = Absolute value of $(x_n - \mu)$*	$(x_n - \mu)^2$
Nick	74	4	16	Sahar	65	5	25
Elana	66	4	16	Maggie	68	2	4
Dinah	68	2	4	Faisal	69	1	1
Rebecca	69	1	1	Ted	70	0	0
Ben	73	3	9	Jeff	71	1	1
Charu	70	0	0	Narciso	75	5	25
		Total = 14	Total = 46			Total = 14	Total = 56
			Variance = 46/6 = 7.7				Variance = 56/6 = 9.3
			Standard deviation = $\sqrt{7.7} = 2.8$				Standard deviation = $\sqrt{9.3} = 3$

* Absolute value is the distance between two figures, regardless of direction, so that it is always positive. In this case, it represents the number of inches between the height of the individual and the mean.

Both groups of students have a mean height of 70 inches. The heights of students in both groups also differ from the mean by the same number of total inches: 14. By that measure of dispersion, the two distributions are identical. However, the variance for Group 2 is higher because of the weight given in the variance formula to values that lie particularly far from the mean—Sahar and Narciso in this case.

Variance is rarely used as a descriptive statistic on its own. Instead, the variance is most useful as a step toward calculating the standard deviation of a distribution, which is a more intuitive tool as a descriptive statistic.

The standard deviation for a set of observations is the square root of the variance:

For any set of *n* observations $x_1, x_2, x_3 \ldots x_n$ with mean μ,
standard deviation = σ = square root of this whole quantity =

$$\sqrt{[(x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + \ldots (x_n - \mu)^2]/n}$$

* With twelve bar patrons, the median would be the midpoint between the income of the guy on the sixth stool and the income of the guy on the seventh stool. Since they both make $35,000, the median is $35,000. If one made $35,000 and the other made $36,000, the median for the whole group would be $35,500.

* Manufacturing update: It turns out that nearly all of the defective printers were being manufactured at a plant in Kentucky where workers had stripped parts off the assembly line in order to build a bourbon distillery. Both the perpetually drunk employees and the random missing pieces on the assembly line appear to have compromised the quality of the printers being produced there.

* Remarkably, this person was one of the ten people with annual incomes of $35,000 who were sitting on bar stools when Bill Gates walked in with his parrot. Go figure!

Specifically:

For any set of *n* observations $x_1, x_2, x_3 \ldots x_n$ with mean μ,
Variance = $\sigma^2 = [(x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + \ldots (x_n - \mu)^2]/n$

Because the difference between each term and the mean is squared, the formula for calculating variance puts particular weight on observations that lie far from the mean, or outliers, as the following table of student heights illustrates.

Group 1	Height (μ = 70 inches)	Distance from the mean = Absolute value of $(x_n - \mu)$*	$(x_n - \mu)^2$	Group 2	Height (μ = 70 inches)	Distance from the mean = Absolute value of $(x_n - \mu)$*	$(x_n - \mu)^2$
Nick	74	4	16	Sahar	65	5	25
Elana	66	4	16	Maggie	68	2	4
Dinah	68	2	4	Faisal	69	1	1
Rebecca	69	1	1	Ted	70	0	0
Ben	73	3	9	Jeff	71	1	1
Charu	70	0	0	Narciso	75	5	25
		Total = 14	Total = 46			Total = 14	Total = 56
			Variance = 46/6 = 7.7				Variance = 56/6 = 9.3
			Standard deviation = $\sqrt{7.7}$ = 2.8				Standard deviation = $\sqrt{9.3}$ = 3

* Absolute value is the distance between two figures, regardless of direction, so that it is always positive. In this case, it represents the number of inches between the height of the individual and the mean.

Both groups of students have a mean height of 70 inches. The heights of students in both groups also differ from the mean by the same number of total inches: 14. By that measure of dispersion, the two distributions are identical. However, the variance for Group 2 is higher because of the weight given in the variance formula to values that lie particularly far from the mean—Sahar and Narciso in this case.

Variance is rarely used as a descriptive statistic on its own. Instead, the variance is most useful as a step toward calculating the standard deviation of a distribution, which is a more intuitive tool as a descriptive statistic.

The standard deviation for a set of observations is the square root of the variance:

For any set of *n* observations $x_1, x_2, x_3 \ldots x_n$ with mean μ,
standard deviation = σ = square root of this whole quantity =

$$\sqrt{[(x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + \ldots (x_n - \mu)^2]/n}$$

* With twelve bar patrons, the median would be the midpoint between the income of the guy on the sixth stool and the income of the guy on the seventh stool. Since they both make $35,000, the median is $35,000. If one made $35,000 and the other made $36,000, the median for the whole group would be $35,500.

* Manufacturing update: It turns out that nearly all of the defective printers were being manufactured at a plant in Kentucky where workers had stripped parts off the assembly line in order to build a bourbon distillery. Both the perpetually drunk employees and the random missing pieces on the assembly line appear to have compromised the quality of the printers being produced there.

* Remarkably, this person was one of the ten people with annual incomes of $35,000 who were sitting on bar stools when Bill Gates walked in with his parrot. Go figure!

CHAPTER 3

Deceptive Description

"He's got a great personality!" and other true but grossly misleading statements

To anyone who has ever contemplated dating, the phrase "he's got a great personality" usually sets off alarm bells, not because the description is necessarily wrong, but for what it may *not* reveal, such as the fact that the guy has a prison record or that his divorce is "not entirely final." We don't doubt that this guy has a great personality; we are wary that a true statement, the great personality, is being used to mask or obscure other information in a way that is seriously misleading (assuming that most of us would prefer not to date ex-felons who are still married). The statement is not a lie per se, meaning that it wouldn't get you convicted of perjury, but it still could be so inaccurate as to be untruthful.

And so it is with statistics. Although the field of statistics is rooted in mathematics, and mathematics is exact, the use of statistics to describe complex phenomena is not exact. That leaves plenty of room for shading the truth. Mark Twain famously remarked that there are three kinds of lies: lies, damned lies, and statistics.* As the last chapter explained, most phenomena that we care about can be described in multiple ways. Once there are multiple ways of describing the same thing (e.g., "he's got a great personality" *or* "he was convicted of securities fraud"), the descriptive statistics that we choose to use (or not to use) will have a profound impact on the impression that we leave. Someone with nefarious motives can use perfectly good facts and figures to support entirely disputable or illegitimate conclusions.

We ought to begin with the crucial distinction between "precision" and "accuracy." These words are not interchangeable. Precision reflects the exactitude with which we can express something. In a description of the length of your commute, "41.6 miles" is more precise than "about 40 miles," which is more precise than "a long f——ing way." If you ask me how far it is to the nearest gas station, and I tell you that it's 1.265 miles to the east, that's a precise answer. Here is the problem: That answer may be entirely inaccurate if the gas station happens to be in the other direction. On the other hand, if I tell you, "Drive ten minutes or so until you see a hot dog stand. The gas station will be a couple hundred yards after that on the right. If you pass the Hooters, you've gone too far," my answer is less precise than "1.265 miles to the east" but significantly better because I am sending you in the direction of the gas station. Accuracy is a measure of whether a figure is broadly consistent with the truth—hence the danger of confusing precision with accuracy. If an answer is accurate, then more precision is usually better. But no amount of precision can make up for inaccuracy.

In fact, precision can mask inaccuracy by giving us a false sense of certainty, either inadvertently or quite deliberately. Joseph McCarthy, the Red-baiting senator from Wisconsin, reached the apogee of his reckless charges in 1950 when he alleged not only that the U.S. State Department was infiltrated with communists, but that he had a list of their names. During a speech in Wheeling, West Virginia, McCarthy waved in the air a piece of paper and declared, "I have here in my hand a list of 205—a list of names that were made known to the Secretary of State as being members of the Communist Party and who nevertheless are still working and shaping policy in the State Department."[1] It turns out that the paper had

no names on it at all, but the specificity of the charge gave it credibility, despite the fact that it was a bald-faced lie.

I learned the important distinction between precision and accuracy in a less malicious context. For Christmas one year my wife bought me a golf range finder to calculate distances on the course from my golf ball to the hole. The device works with some kind of laser; I stand next to my ball in the fairway (or rough) and point the range finder at the flag on the green, at which point the device calculates the exact distance that I'm supposed to hit the ball. This is an improvement upon the standard yardage markers, which give distances only to the center of the green (and are therefore accurate but less precise). With my Christmas-gift range finder I was able to know that I was 147.2 yards from the hole. I expected the precision of this nifty technology to improve my golf game. Instead, it got appreciably worse.

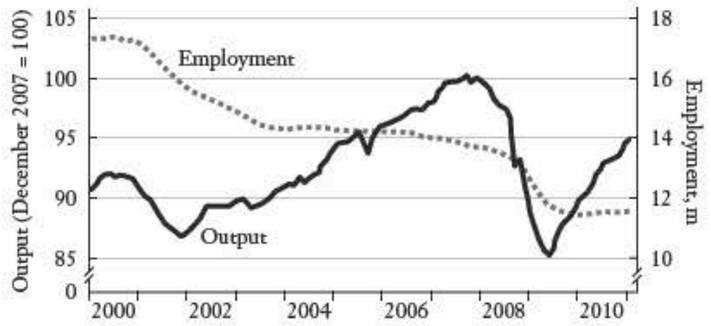
There were two problems. First, I used the stupid device for three months before I realized that it was set to meters rather than to yards; every seemingly precise calculation (147.2) was wrong. Second, I would sometimes inadvertently aim the laser beam at the trees behind the green, rather than at the flag marking the hole, so that my "perfect" shot would go exactly the distance it was supposed to go—right over the green into the forest. The lesson for me, which applies to all statistical analysis, is that even the most precise measurements or calculations should be checked against common sense.

To take an example with more serious implications, many of the Wall Street risk management models prior to the 2008 financial crisis were quite precise. The concept of "value at risk" allowed firms to quantify with precision the amount of the firm's capital that could be lost under different scenarios. The problem was that the supersophisticated models were the equivalent of setting my range finder to meters rather than to yards. The math was complex and arcane. The answers it produced were reassuringly precise. But the assumptions about what might happen to global markets that were embedded in the models were just plain wrong, making the conclusions wholly inaccurate in ways that destabilized not only Wall Street but the entire global economy.

Even the most precise and accurate descriptive statistics can suffer from a more fundamental problem: a lack of clarity over what exactly we are trying to define, describe, or explain. Statistical arguments have much in common with bad marriages; the disputants often talk past one another. Consider an important economic question: How healthy is American manufacturing? One often hears that American manufacturing jobs are being lost in huge numbers to China, India, and other low-wage countries. One also hears that high-tech manufacturing still thrives in the United States and that America remains one of the world's top exporters of manufactured goods. Which is it? This would appear to be a case in which sound analysis of good data could reconcile these competing narratives. Is U.S. manufacturing profitable and globally competitive, or is it shrinking in the face of intense foreign competition?

Both. The British news magazine the *Economist* reconciled the two seemingly contradictory views of American manufacturing with the following graph.

"The Rustbelt Recovery," March 10, 2011



The seeming contradiction lies in how one defines the "health" of U.S. manufacturing. In terms of output—the total value of goods produced and sold—the U.S. manufacturing sector grew steadily in the 2000s, took a big hit during the Great Recession, and has since bounced back robustly. This is consistent with data from the CIA's *World Factbook* showing that the United States is the third-largest manufacturing exporter in the world, behind China and Germany. The United States remains a manufacturing powerhouse.

But the graph in the *Economist* has a second line, which is manufacturing *employment*. The number of manufacturing jobs in the United States has fallen steadily; roughly six million manufacturing jobs were lost in the last decade. Together, these two stories—rising manufacturing output and falling employment—tell the complete story. Manufacturing in the United States has grown steadily more productive, meaning that factories are producing more output with fewer workers. This is good from a global competitiveness standpoint, for it makes American products more competitive with manufactured goods from low-wage countries. (One way to compete with a firm that can pay workers $2 an hour is to create a manufacturing process so efficient that one worker earning $40 can do twenty times as much.) *But there are a lot fewer manufacturing jobs*, which is terrible news for the displaced workers who depended on those wages.

Since this is a book about statistics and not manufacturing, let's go back to the main point, which is that the "health" of U.S. manufacturing—something seemingly easy to quantify—depends on how one chooses to define health: output or employment? In this case (and many others), the most complete story comes from including both figures, as the *Economist* wisely chose to do in its graph.

Even when we agree on a single measure of success, say, student test scores, there is plenty of statistical wiggle room. See if you can reconcile the following hypothetical statements, both of which could be true:

Politician A (the challenger): "Our schools are getting worse! Sixty percent of our schools had lower test scores this year than last year."

Politician B (the incumbent): "Our schools are getting better! Eighty percent of our students had higher test scores this year than last year."

Here's a hint: The schools do not all necessarily have the same number of students. If you take another look at the seemingly contradictory statements, what you'll see is that one politician is using schools as his *unit of analysis* ("Sixty percent of our schools . . ."), and the other is using students as the unit of analysis ("Eighty percent of our students . . ."). The unit of analysis is the entity being compared or described by the statistics—school performance by one of them and student performance by the other. It's entirely possible for most of the students to be improving and most of the schools to be getting worse—if the students showing improvement happen to be in very big schools. To make this example more intuitive, let's do the same exercise by using American states:

Politician A (a populist): "Our economy is in the crapper! Thirty states had falling incomes last year."

Politician B (more of an elitist): "Our economy is showing appreciable gains: Seventy percent of

Americans had rising incomes last year."

What I would infer from those statements is that the biggest states have the healthiest economies: New York, California, Texas, Illinois, and so on. The thirty states with falling average incomes are likely to be much smaller: Vermont, North Dakota, Rhode Island, and so on. Given the disparity in the size of the states, it's entirely possible that the majority of states are doing worse while the majority of Americans are doing better. The key lesson is to pay attention to the unit of analysis. Who or what is being described, and is that different from the "who" or "what" being described by someone else?

Although the examples above are hypothetical, here is a crucial statistical question that is not: Is globalization making income inequality around the planet better or worse? By one interpretation, globalization has merely exacerbated existing income inequalities; richer countries in 1980 (as measured by GDP per capita) tended to grow faster between 1980 and 2000 than poorer countries.[2] The rich countries just got richer, suggesting that trade, outsourcing, foreign investment, and the other components of "globalization" are merely tools for the developed world to extend its economic hegemony. Down with globalization! Down with globalization!

But hold on a moment. The same data can (and should) be interpreted entirely differently if one changes the unit of analysis. We don't care about poor countries; *we care about poor people*. And a high proportion of the world's poor people happen to live in China and India. Both countries are huge (with a population over a billion); each was relatively poor in 1980. Not only have China and India grown rapidly over the past several decades, but they have done so in large part because of their increased economic integration with the rest of the world. They are "rapid globalizers," as the *Economist* has described them. Given that our goal is to ameliorate human misery, it makes no sense to give China (population 1.3 billion) the same weight as Mauritius (population 1.3 million) when examining the effects of globalization on the poor.

The unit of analysis should be people, not countries. What really happened between 1980 and 2000 is a lot like my fake school example above. The bulk of the world's poor happened to live in two giant countries that grew extremely fast as they became more integrated into the global economy. The proper analysis yields an entirely different conclusion about the benefits of globalization for the world's poor. As the *Economist* points out, "If you consider people, not countries, global inequality is falling rapidly."

The telecommunications companies AT&T and Verizon have recently engaged in an advertising battle that exploits this kind of ambiguity about what is being described. Both companies provide cellular phone service. One of the primary concerns of most cell phone users is the quality of the service in places where they are likely to make or receive phone calls. Thus, a logical point of comparison between the two firms is the size and quality of their networks. While consumers just want decent cell phone service in lots of places, both AT&T and Verizon have come up with different metrics for measuring the somewhat amorphous demand for "decent cell phone service in lots of places." Verizon launched an aggressive advertising campaign touting the geographic coverage of its network; you may remember the maps of the United States that showed the large percentage of the country covered by the Verizon network compared with the relatively paltry geographic coverage of the AT&T network. The unit of analysis chosen by Verizon is geographic area covered—because the company has more of it.

AT&T countered by launching a campaign that changed the unit of analysis. Its billboards advertised that "AT&T covers 97 percent of Americans." Note the use of the word "Americans" rather than "America." AT&T focused on the fact that most people don't live in rural Montana or the Arizona desert. Since the population is not evenly distributed across the physical geography of the United States, the key to good cell service (the campaign argued implicitly) is having a network in place where callers actually live and work, not necessarily where they go camping. As someone who spends a fair bit of time in rural New Hampshire, however, my sympathies are with Verizon on this one.

Our old friends the mean and the median can also be used for nefarious ends. As you should recall from the last chapter, both the median and the mean are measures of the "middle" of a distribution, or its "central tendency." The mean is a simple average: the sum of the observations divided by the number of observations. (The mean of 3, 4, 5, 6, and 102 is 24.) The median is the midpoint of the distribution; half of the observations lie above the median and half lie below. (The median of 3, 4, 5, 6, and 102 is 5.) Now, the clever reader will see that there is a sizable difference between 24 and 5. If, for some reason, I would like to describe this group of numbers in a way that makes it look big, I will focus on the mean. If I want to make it look smaller, I will cite the median.

Now let's look at how this plays out in real life. Consider the George W. Bush tax cuts, which were touted by the Bush administration as something good for most American families. While pushing the plan, the administration pointed out that 92 million Americans would receive an average tax reduction of over $1,000 ($1,083 to be precise). But was that summary of the tax cut accurate? According to the *New York Times*, "The data don't lie, but some of them are mum."

Would 92 million Americans be getting a tax cut? Yes.

Would most of those people be getting a tax cut of around $1,000? No. The median tax cut was less than $100.

A relatively small number of extremely wealthy individuals were eligible for very large tax cuts; these big numbers skew the mean, making the average tax cut look bigger than what most Americans would likely receive. The median is not sensitive to outliers, and, in this case, is probably a more accurate description of how the tax cuts affected the typical household.

Of course, the median can also do its share of dissembling *because it is not sensitive to outliers*. Suppose that you have a potentially fatal illness. The good news is that a new drug has been developed that might be effective. The drawback is that it's extremely expensive and has many unpleasant side effects. "But does it work?" you ask. The doctor informs you that the new drug increases the median life expectancy among patients with your disease by two weeks. That is hardly encouraging news; the drug may not be worth the cost and unpleasantness. Your insurance company refuses to pay for the treatment; it has a pretty good case on the basis of the median life expectancy figures.

Yet the median may be a horribly misleading statistic in this case. Suppose that many patients do not respond to the new treatment but that some large number of patients, say 30 or 40 percent, are cured entirely. This success would not show up in the median (though the mean life expectancy of those taking the drug would look very impressive). In this case, the outliers—those who take the drug and live for a long time—would be highly relevant to your decision. And it is not merely a hypothetical case. Evolutionary biologist Stephen Jay Gould was diagnosed with a form of cancer that had a median survival time of eight months; he died of a different and unrelated kind of cancer twenty years later.[3] Gould subsequently wrote a famous article called "The Median Isn't the Message," in which he argued that his scientific knowledge of statistics saved him from the erroneous conclusion that he would necessarily be dead in eight months. The definition of the median tells us that half the patients will live at least eight months—and possibly much, much longer than that. The mortality distribution is "right-skewed," which is more than a technicality if you happen to have the disease.[4]

In this example, the defining characteristic of the median—that it does not weight observations on the basis of *how far* they lie from the midpoint, only on whether they lie above or below—turns out to be its weakness. In contrast, the mean *is* affected by dispersion. From the standpoint of accuracy, the median

versus mean question revolves around whether the outliers in a distribution distort what is being described or are instead an important part of the message. (Once again, judgment trumps math.) Of course, nothing says that you must choose the median or the mean. Any comprehensive statistical analysis would likely present both. When just the median or the mean appears, it may be for the sake of brevity—or it may be because someone is seeking to "persuade" with statistics.

Those of a certain age may remember the following exchange (as I recollect it) between the characters played by Chevy Chase and Ted Knight in the movie *Caddyshack*. The two men meet in the locker room after both have just come off the golf course:

> TED KNIGHT: What did you shoot?
> CHEVY CHASE: Oh, I don't keep score.
> TED KNIGHT: Then how do you compare yourself to other golfers?
> CHEVY CHASE: By height.

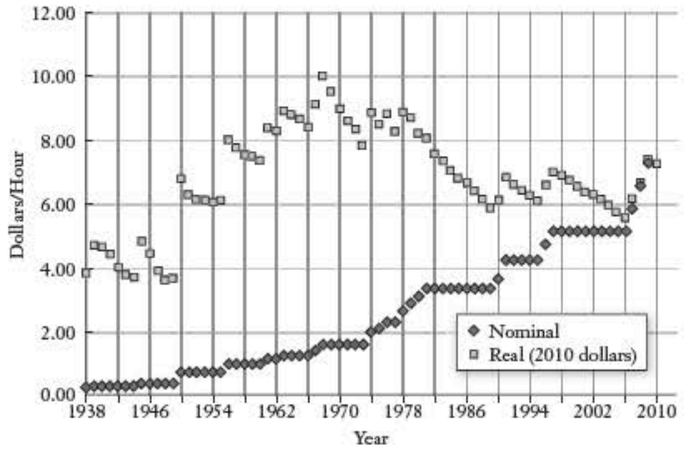
I'm not going to try to explain why this is funny. I will say that a great many statistical shenanigans arise from "apples and oranges" comparisons. Suppose you are trying to compare the price of a hotel room in London with the price of a hotel room in Paris. You send your six-year-old to the computer to do some Internet research, since she is much faster and better at it than you are. Your child reports back that hotel rooms in Paris are more expensive, around 180 a night; a comparable room in London is 150 a night.

You would likely explain to your child the difference between pounds and euros, and then send her back to the computer to find the exchange rate between the two currencies so that you could make a meaningful comparison. (This example is loosely rooted in truth; after I paid 100 rupees for a pot of tea in India, my daughter wanted to know why everything in India was so expensive.) Obviously the numbers on currency from different countries mean nothing until we convert them into comparable units. What is the exchange rate between the pound and the euro, or, in the case of India, between the dollar and the rupee?

This seems like a painfully obvious lesson—yet one that is routinely ignored, particularly by politicians and Hollywood studios. These folks clearly recognize the difference between euros and pounds; instead, they overlook a more subtle example of apples and oranges: inflation. A dollar today is not the same as a dollar sixty years ago; it buys much less. Because of inflation, something that cost $1 in 1950 would cost $9.37 in 2011. As a result, any monetary comparison between 1950 and 2011 without adjusting for changes in the value of the dollar would be less accurate than comparing figures in euros and pounds—*since the euro and the pound are closer to each other in value than a 1950 dollar is to a 2011 dollar*.

This is such an important phenomenon that economists have terms to denote whether figures have been adjusted for inflation or not. *Nominal* figures are not adjusted for inflation. A comparison of the nominal cost of a government program in 1970 to the nominal cost of the same program in 2011 merely compares the size of the checks that the Treasury wrote in those two years—without any recognition that a dollar in 1970 bought more stuff than a dollar in 2011. If we spent $10 million on a program in 1970 to provide war veterans with housing assistance and $40 million on the same program in 2011, *the federal commitment to that program has actually gone down*. Yes, spending has gone up in nominal terms, but that does not reflect the changing value of the dollars being spent. One 1970 dollar is equal to $5.83 in 2011; the government would need to spend $58.3 million on veterans' housing benefits in 2011 to provide support comparable to the $10 million it was spending in 1970.

Real figures, on the other hand, are adjusted for inflation. The most commonly accepted methodology is to convert all of the figures into a single unit, such as 2011 dollars, to make an "apples and apples" comparison. Many websites, including that of the U.S. Bureau of Labor Statistics, have simple inflation calculators that will compare the value of a dollar at different points in time.* For a real (yes, a pun) example of how statistics can look different when adjusted for inflation, check out the following graph of the U.S. federal minimum wage, which plots both the nominal value of the minimum wage and its real purchasing power in 2010 dollars.



Source: http://oregonstate.edu/instruct/anth484/minwage.html.

The federal minimum wage—the number posted on the bulletin board in some remote corner of your office—is set by Congress. This wage, currently $7.25, is a nominal figure. Your boss does not have to ensure that $7.25 buys as much as it did two years ago; he just has to make sure that you get a minimum of $7.25 for every hour of work that you do. It's all about the number on the check, not what that number can buy.

Yet inflation erodes the purchasing power of the minimum wage over time (and every other nominal wage, which is why unions typically negotiate "cost of living adjustments"). If prices rise faster than Congress raises the minimum wage, the real value of that minimum hourly payment will fall. Supporters of a minimum wage should care about the real value of that wage, since the whole point of the law is to guarantee low-wage workers some minimum level of consumption for an hour of work, not to give them a check with a big number on it that buys less than it used to. (If that were the case, then we could just pay low-wage workers in rupees.)

Hollywood studios may be the most egregiously oblivious to the distortions caused by inflation when comparing figures at different points in time—and deliberately so. What were the top five highest-grossing films (domestic) of all time as of 2011?[5]

1. *Avatar* (2009)
2. *Titanic* (1997)
3. *The Dark Knight* (2008)
4. *Star Wars Episode IV* (1977)
5. *Shrek 2* (2004)

Now you may feel that list looks a little suspect. These were successful films—but *Shrek 2*? Was that

really a greater commercial success than *Gone with the Wind*? *The Godfather*? *Jaws*? No, no, and no. Hollywood likes to make each blockbuster look bigger and more successful than the last. One way to do that would be to quote box office receipts in Indian rupees, which would inspire headlines such as the following: "Harry Potter Breaks Box Office Record with Weekend Receipts of 1.3 Trillion!" But even the most dim-witted moviegoers would be suspicious of figures that are large only because they are quoted in a currency with relatively little purchasing power. Instead, Hollywood studios (and the journalists who report on them) merely use nominal figures, which makes recent movies look successful largely because ticket prices are higher now than they were ten, twenty, or fifty years ago. (When *Gone with the Wind* came out in 1939, a ticket cost somewhere in the range of $.50.) The most accurate way to compare commercial success over time would be to adjust ticket receipts for inflation. Earning $100 million in 1939 is a lot more impressive than earning $500 million in 2011. So what are the top grossing films in the U.S. of all time, *adjusted for inflation*?[6]

1. *Gone with the Wind* (1939)
2. *Star Wars Episode IV* (1977)
3. *The Sound of Music* (1965)
4. *E.T.* (1982)
5. *The Ten Commandments* (1956)

In real terms, *Avatar* falls to number 14; *Shrek 2* falls all the way to 31st.

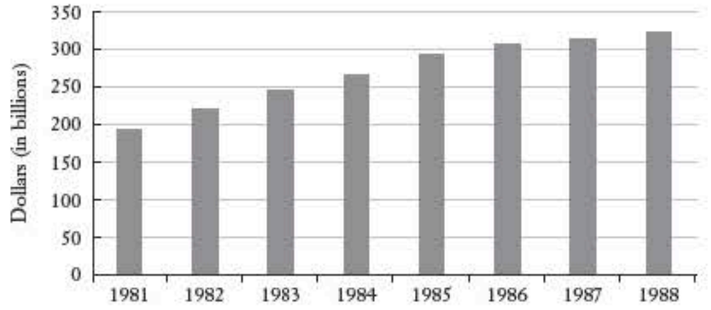
Even comparing apples and apples leaves plenty of room for shenanigans. As discussed in the last chapter, one important role of statistics is to describe changes in quantities over time. Are taxes going up? How many cheeseburgers are we selling compared with last year? By how much have we reduced the arsenic in our drinking water? We often use percentages to express these changes because they give us a sense of scale and context. We understand what it means to reduce the amount of arsenic in the drinking water by 22 percent, whereas few of us would know whether reducing arsenic by one microgram (the absolute reduction) would be a significant change or not. Percentages don't lie—but they can exaggerate. One way to make growth look explosive is to use percentage change to describe some change relative to a very low starting point. I live in Cook County, Illinois. I was shocked one day to learn that the portion of my taxes supporting the Suburban Cook County Tuberculosis Sanitarium District was slated to rise by 527 percent! However, I called off my massive antitax rally (which was really still in the planning phase) when I learned that this change would cost me less than a good turkey sandwich. The Tuberculosis Sanitarium District deals with roughly a hundred cases a year; it is not a large or expensive organization. The *Chicago Sun-Times* pointed out that for the typical homeowner, the tax bill would go from $1.15 to $6.[7] Researchers will sometimes qualify a growth figure by pointing out that it is "from a low base," meaning that any increase is going to look large by comparison.

Obviously the flip side is true. A small percentage of an enormous sum can be a big number. Suppose the secretary of defense reports that defense spending will grow only 4 percent this year. Great news! Not really, given that the Defense Department budget is nearly $700 billion. Four percent of $700 billion is $28 billion, which can buy a lot of turkey sandwiches. In fact, that seemingly paltry 4 percent increase in the defense budget is *more than the entire NASA budget and about the same as the budgets of the Labor and Treasury Departments combined.*

In a similar vein, your kindhearted boss might point out that as a matter of fairness, every employee will be getting the same raise this year, 10 percent. What a magnanimous gesture—except that if your boss makes $1 million and you make $50,000, his raise will be $100,000 and yours will be $5,000. The statement "everyone will get the same 10 percent raise this year" just sounds so much better than "my raise will be twenty times bigger than yours." Both are true in this case.

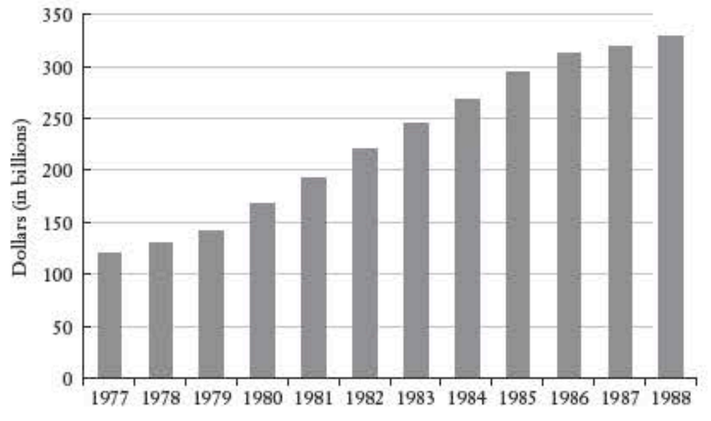
Any comparison of a quantity changing over time must have a start point and an end point. One can sometimes manipulate those points in ways that affect the message. I once had a professor who liked to speak about his "Republican slides" and his "Democratic slides." He was referring to data on defense spending, and what he meant was that he could organize the same data in different ways in order to please either Democratic or Republican audiences. For his Republican audiences, he would offer the following slide with data on increases in defense spending under Ronald Reagan. Clearly Reagan helped restore our commitment to defense and security, which in turn helped to win the Cold War. No one can look at these numbers and not appreciate the steely determination of Ronald Reagan to face down the Soviets.

Defense Spending in Billions, 1981–1988



For the Democrats, my former professor merely used the same (nominal) data, but a longer time frame. For this group, he pointed out that Jimmy Carter deserves credit for beginning the defense buildup. As the following "Democratic" slide shows, the defense spending increases from 1977 to 1980 show the same basic trend as the increases during the Reagan presidency. Thank goodness that Jimmy Carter—a graduate of Annapolis and a former naval officer—began the process of making America strong again!

Defense Spending in Billions, 1977–1988



Source: http://www.usgovernmentspending.com/spend.php?span=usgs302&year=1988&view=1&expand=30&expandC=&units=b&fy=fy12&local=s&state=US&pie=#usgs302.

While the main point of statistics is to present a meaningful picture of things we care about, in many cases we also hope to act on these numbers. NFL teams want a simple measure of quarterback quality so that they can find and draft talented players out of college. Firms measure the performance of their employees so that they can promote those who are valuable and fire those who are not. There is a common business aphorism: "You can't manage what you can't measure." True. *But you had better be darn sure that what you are measuring is really what you are trying to manage.*

Consider school quality. This is a crucial thing to measure, since we would like to reward and emulate "good" schools while sanctioning or fixing "bad" schools. (And within each school, we have the similar challenge of measuring teacher quality, for the same basic reason.) The most common measure of quality for both schools and teachers is test scores. If students are achieving impressive scores on a well-conceived standardized test, then presumably the teacher and school are doing a fine job. Conversely, bad test scores are a clear signal that lots of people should be fired, sooner rather than later. These statistics can take us a long way toward fixing our public education system, right?

Wrong. Any evaluation of teachers or schools that is based solely on test scores will present a dangerously inaccurate picture. Students who walk through the front door of different schools have vastly different backgrounds and abilities. We know, for example, that the education and income of a student's parents have a significant impact on achievement, regardless of what school he or she attends. The statistic that we're missing in this case happens to be the only one that matters for our purposes: How much of a student's performance, good or bad, can be attributed to what happens inside the school (or inside a particular classroom)?

Students who live in affluent, highly educated communities are going to test well from the moment their parents drop them off at school on the first day of kindergarten. The flip side is also true. There are schools with extremely disadvantaged populations in which teachers may be doing a remarkable job but the student test scores will still be low—albeit not nearly as low as they would have been if the teachers had not been doing a good job. What we need is some measure of "value-added" at the school level, or even at the classroom level. We don't want to know the absolute level of student achievement; we want to know how much that student achievement has been affected by the educational factors we are trying to evaluate.

At first glance, this seems an easy task, as we can simply give students a pretest and a posttest. If we know student test scores when they enter a particular school or classroom, then we can measure their performance at the end and attribute the difference to whatever happened in that school or classroom.

Alas, wrong again. Students with different abilities or backgrounds may also learn *at different rates*. Some students will grasp the material faster than others for reasons that have nothing to do with the quality of the teaching. So if students in Affluent School A and Poor School B both start algebra at the same time and level, the explanation for the fact that students at Affluent School A test better in algebra a year later may be that the teachers are better, or it may be that the students were capable of learning faster—or both. Researchers are working to develop statistical techniques that measure instructional quality in ways that account appropriately for different student backgrounds and abilities. In the meantime, our attempts to identify the "best" schools can be ridiculously misleading.

Every fall, several Chicago newspapers and magazines publish a ranking of the "best" high schools in the region, usually on the basis of state test score data. Here is the part that is laugh-out-loud funny from a statistical standpoint: Several of the high schools consistently at the top of the rankings are selective enrollment schools, meaning that students must apply to get in, and only a small proportion of those students are accepted. One of the most important admissions criteria is standardized test scores. So let's summarize: (1) these schools are being recognized as "excellent" for having students with high test scores; (2) to get into such a school, one must have high test scores. This is the logical equivalent of giving an award to the basketball team for doing such an excellent job of producing tall students.

Even if you have a solid indicator of what you are trying to measure and manage, the challenges are not over. The good news is that "managing by statistics" can change the underlying behavior of the person or institution being managed for the better. If you can measure the proportion of defective products coming off an assembly line, and if those defects are a function of things happening at the plant, then some kind of bonus for workers that is tied to a reduction in defective products would presumably change behavior in the right kinds of ways. Each of us responds to incentives (even if it is just praise or a better parking spot). Statistics measure the outcomes that matter; incentives give us a reason to improve those outcomes.

Or, in some cases, just to make the statistics look better. That's the bad news.

If school administrators are evaluated—and perhaps even compensated—on the basis of the high school graduation rate for students in a particular school district, they will focus their efforts on boosting the number of students who graduate. Of course, they may also devote some effort to improving the graduation rate, which is not necessarily the same thing. For example, students who leave school before graduation can be classified as "moving away" rather than dropping out. This is not merely a hypothetical example; it is a charge that was leveled against former secretary of education Rod Paige during his tenure as the Houston school superintendent. Paige was hired by President George W. Bush to be U.S. secretary of education because of his remarkable success in Houston in reducing the dropout rate and boosting test scores.

If you're keeping track of the little business aphorisms I keep tossing your way, here is another one: "It's never a good day when *60 Minutes* shows up at your door." Dan Rather and the *60 Minutes II* crew made a trip to Houston and found that the manipulation of statistics was far more impressive than the educational improvement.[8] High schools routinely classified students who quit high school as transferring to another school, returning to their native country, or leaving to pursue a General Equivalency Diploma (GED)—none of which count as dropping out in the official statistics. Houston reported a citywide dropout rate of 1.5 percent in the year that was examined; *60 Minutes* calculated that the true dropout rate was between 25 and 50 percent.

The statistical chicanery with test scores was every bit as impressive. One way to improve test scores (in Houston or anywhere else) is to improve the quality of education so that students learn more and test better. This is a good thing. Another (less virtuous) way to improve test scores is to prevent the worst students from taking the test. If the scores of the lowest-performing students are eliminated, the average test score for the school or district will go up, even if all the rest of the students show no improvement at all. In Texas, the statewide achievement test is given in tenth grade. There was evidence that Houston schools were trying to keep the weakest students from reaching tenth grade. In one particularly egregious example, a student spent three years in ninth grade and then was promoted straight to eleventh grade—a deviously clever way of keeping a weak student from taking a tenth-grade benchmark exam without forcing him to drop out (which would have showed up on a different statistic).

It's not clear that Rod Paige was complicit in this statistical trickery during his tenure as Houston superintendent; however, he did implement a rigorous accountability program that gave cash bonuses to principals who met their dropout and test score goals and that fired or demoted principals who failed to meet their targets. Principals definitely responded to the incentives; that's the larger lesson. But you had better be darn certain that the folks being evaluated can't make themselves look better (statistically) in ways that are not consistent with the goal at hand.

The state of New York learned this the hard way. The state introduced "scorecards" that evaluate the mortality rates for the patients of cardiologists performing coronary angioplasty, a common treatment for

heart disease.[9] This seems like a perfectly reasonable and helpful use of descriptive statistics. The proportion of a cardiologist's patients who die in surgery is an important thing to know, and it makes sense for the government to collect and promulgate such data since individual consumers would not otherwise have access to it. So is this a good policy? Yes, other than the fact that it probably ended up killing people.

Cardiologists obviously care about their "scorecard." However, the easiest way for a surgeon to improve his mortality rate is *not* by killing fewer people; presumably most doctors are already trying very hard to keep their patients alive. The easiest way for a doctor to improve his mortality rate is by refusing to operate on the sickest patients. According to a survey conducted by the School of Medicine and Dentistry at the University of Rochester, the scorecard, which ostensibly serves patients, can also work to their detriment: 83 percent of the cardiologists surveyed said that, because of the public mortality statistics, some patients who might benefit from angioplasty might not receive the procedure; 79 percent of the doctors said that some of their personal medical decisions had been influenced by the knowledge that mortality data are collected and made public. The sad paradox of this seemingly helpful descriptive statistic is that cardiologists responded rationally by withholding care from the patients who needed it most.

A statistical index has all the potential pitfalls of any descriptive statistic—plus the distortions introduced by combining multiple indicators into a single number. By definition, any index is going to be sensitive to how it is constructed; it will be affected both by what measures go into the index and by how each of those measures is weighted. For example, why does the NFL passer rating not include any measure of third down completions? And for the Human Development Index, how should a country's literacy rate be weighted in the index relative to per capita income? In the end, the important question is whether the simplicity and ease of use introduced by collapsing many indicators into a single number outweighs the inherent inaccuracy of the process. Sometimes that answer may be no, which brings us back (as promised) to the *U.S. News & World Report* (*USNWR*) college rankings.

The *USNWR* rankings use sixteen indicators to score and rank America's colleges, universities, and professional schools. In 2010, for example, the ranking of national universities and liberal arts colleges used "student selectivity" as 15 percent of the index; student selectivity is in turn calculated on the basis of a school's acceptance rate, the proportion of the entering students who were in the top 10 percent of their high school class, and the average SAT and ACT scores of entering students. The benefit of the *USNWR* rankings is that they provide lots of information about thousands of schools in a simple and accessible way. Even the critics concede that much of the information collected on America's colleges and universities is valuable. Prospective students should know an institution's graduation rate and the average class size.

Of course, providing meaningful information is an enterprise entirely different from that of collapsing all of that information into a single ranking that purports to be authoritative. To critics, the rankings are sloppily constructed, misleading, and detrimental to the long-term interests of students. "One concern is simply about its being a list that claims to rank institutions in numerical order, which is a level of precision that those data just don't support," says Michael McPherson, the former president of Macalester College in Minnesota.[10] Why should alumni giving count for 5 percent of a school's score? And if it's important, why does it not count for ten percent?

According to *U.S. News & World Report*, "Each indicator is assigned a weight (expressed as a percentage) based on our judgments about which measures of quality matter most."[11] Judgment is one thing; arbitrariness is another. The most heavily weighted variable in the ranking of national universities and colleges is "academic reputation." This reputation is determined on the basis of a "peer assessment survey" filled out by administrators at other colleges and universities and from a survey of high school guidance counselors. In his general critique of rankings, Malcolm Gladwell offers a scathing (though humorous) indictment of the peer assessment methodology. He cites a questionnaire sent out by a former chief justice of the Michigan Supreme Court to roughly one hundred lawyers asking them to rank ten law schools in order of quality. Penn State's was one of the law schools on the list; the lawyers ranked it near the middle. *At the time, Penn State did not have a law school.*[12]

For all the data collected by *USNWR*, it's not obvious that the rankings measure what prospective students ought to care about: How much learning is going on at any given institution? Football fans may quibble about the composition of the passer index, but no one can deny that its component parts—completions, yardage, touchdowns, and interceptions—are an important part of a quarterback's overall performance. That is not necessarily the case with the *USNWR* criteria, most of which focus on inputs (e.g., what kind of students are admitted, how much faculty are paid, the percentage of faculty who are full-time) rather than educational outputs. Two notable exceptions are the freshman retention rate and the graduation rate, but even those indicators do not measure learning. As Michael McPherson points out, "We don't really learn anything from U.S. News about whether the education they got during those four years actually improved their talents or enriched their knowledge."

All of this would still be a harmless exercise, but for the fact that it appears to encourage behavior that is not necessarily good for students or higher education. For example, one statistic used to calculate the rankings is financial resources per student; the problem is that there is no corresponding measure of how well that money is being spent. An institution that spends less money to better effect (and therefore can charge lower tuition) is punished in the ranking process. Colleges and universities also have an incentive to encourage large numbers of students to apply, including those with no realistic hope of getting in, because it makes the school appear more selective. This is a waste of resources for the schools soliciting bogus applications and for students who end up applying with no meaningful chance of being accepted.

Since we are about to move on to a chapter on probability, I will bet that the *U.S. News & World Report* rankings are not going away anytime soon. As Leon Botstein, president of Bard College, has pointed out, "People love easy answers. What is the best place? Number 1."[13]

The overall lesson of this chapter is that statistical malfeasance has very little to do with bad math. If anything, impressive calculations can obscure nefarious motives. The fact that you've calculated the mean correctly will not alter the fact that the median is a more accurate indicator. Judgment and integrity turn out to be surprisingly important. A detailed knowledge of statistics does not deter wrongdoing any more than a detailed knowledge of the law averts criminal behavior. With both statistics and crime, the bad guys often know exactly what they're doing!

* Twain attributed this phrase to British prime minister Benjamin Disraeli, but there is no record of Disraeli's ever saying or writing it.
* Available at http://www.bls.gov/data/inflation_calculator.htm.

CHAPTER 4

Correlation

How does Netflix know what movies I like?

Netflix insists that I'll like the film *Bhutto*, a documentary that offers an "in-depth and at times incendiary look at the life and tragic death of former Pakistani prime minister Benazir Bhutto." I probably will like the film *Bhutto*. (I've added it to my queue.) The Netflix recommendations that I've watched in the past have been terrific. And when a film is recommended that I've already seen, it's typically one I've really enjoyed.

How does Netflix do that? Is there some massive team of interns at corporate headquarters who have used a combination of Google and interviews with my family and friends to determine that I might like a documentary about a former Pakistani prime minister? Of course not. Netflix has merely mastered some very sophisticated statistics. *Netflix doesn't know me*. But it does know what films I've liked in the past (because I've rated them). Using that information, along with ratings from other customers and a powerful computer, Netflix can make shockingly accurate predictions about my tastes.

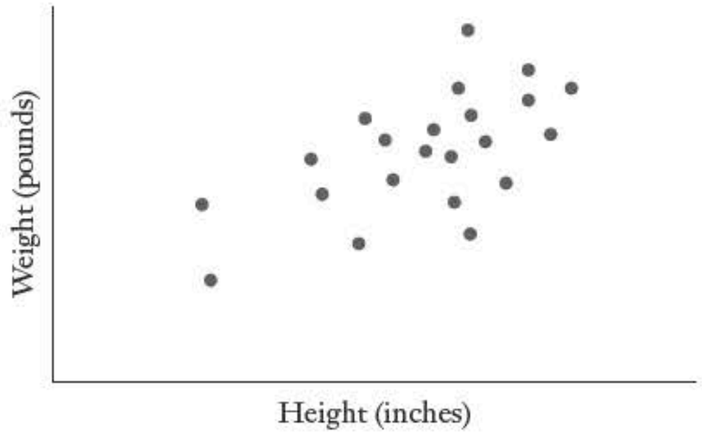
I'll come back to the specific Netflix algorithm for making these picks; for now, the important point is that it's all based on correlation. Netflix recommends movies that are similar to other films that I've liked; it also recommends films that have been highly rated by other customers whose ratings are similar to mine. *Bhutto* was recommended because of my five-star ratings for two other documentaries, *Enron: The Smartest Guys in the Room* and *Fog of War*.

Correlation measures the degree to which two phenomena are related to one another. For example, there is a correlation between summer temperatures and ice cream sales. When one goes up, so does the other. Two variables are positively correlated if a change in one is associated with a change in the other in the same direction, such as the relationship between height and weight. Taller people weigh more (on average); shorter people weigh less. A correlation is negative if a positive change in one variable is associated with a negative change in the other, such as the relationship between exercise and weight.

The tricky thing about these kinds of associations is that not every observation fits the pattern. Sometimes short people weigh more than tall people. Sometimes people who don't exercise are skinnier than people who exercise all the time. Still, there is a meaningful relationship between height and weight, and between exercise and weight.

If we were to do a scatter plot of the heights and weights of a random sample of American adults, we would expect to see something like the following:

Scatter Plot for Height and Weight



If we were to create a scatter plot of the association between exercise (as measured by minutes of intensive exercise per week) and weight, we would expect a negative correlation, with those who exercise more tending to weigh less. But a pattern consisting of dots scattered across the page is a somewhat unwieldy tool. (If Netflix tried to make film recommendations for me by plotting the ratings for thousands of films by millions of customers, the results would bury the headquarters in scatter plots.) Instead, the power of correlation as a statistical tool is that we can encapsulate an association between two variables in a single descriptive statistic: the correlation coefficient.

The correlation coefficient has two fabulously attractive characteristics. First, for math reasons that have been relegated to the appendix, it is a single number ranging from −1 to 1. A correlation of 1, often described as perfect correlation, means that every change in one variable is associated with an equivalent change in the other variable in the same direction.

A correlation of −1, or perfect negative correlation, means that every change in one variable is associated with an equivalent change in the other variable in the opposite direction.

The closer the correlation is to 1 or −1, the stronger the association. A correlation of 0 (or close to it) means that the variables have no meaningful association with one another, such as the relationship between shoe size and SAT scores.

The second attractive feature of the correlation coefficient is that it has no units attached to it. We can calculate the correlation between height and weight—even though height is measured in inches and weight is measured in pounds. We can even calculate the correlation between the number of televisions high school students have in their homes and their SAT scores, which I assure you will be positive. (More on that relationship in a moment.) The correlation coefficient does a seemingly miraculous thing: It collapses a complex mess of data measured in different units (like our scatter plots of height and weight) into a single, elegant descriptive statistic.

How?

As usual, I've put the most common formula for calculating the correlation coefficient in the appendix at the end of the chapter. This is not a statistic that you are going to be calculating by hand. (After you've entered the data, a basic software package like Microsoft Excel will calculate the correlation between two variables.) Still, the intuition is not that difficult. The formula for calculating the correlation coefficient does the following:

1. Calculates the mean and standard deviation for both variables. If we stick with the height and weight

example, we would then know the mean height for people in the sample, the mean weight for people in the sample, and the standard deviation for both height and weight.

2. Converts all the data so that each observation is represented by its distance (in standard deviations) from the mean. Stick with me; it's not that complicated. Suppose that the mean height in the sample is 66 inches (with a standard deviation of 5 inches) and that the mean weight is 177 pounds (with a standard deviation of 10 pounds). Now suppose that you are 72 inches tall and weigh 168 pounds. We can also say that you your height is 1.2 standard deviations above the mean in height [(72 – 66)/5)] and .9 standard deviations below the mean in weight, or –0.9 for purposes of the formula [(168 – 177)/10]. *Yes, it's unusual for someone to be above the mean in height and below the mean in weight, but since you've paid good money for this book, I figured I should at least make you tall and thin.* Notice that your height and weight, formerly in inches and pounds, have been reduced to 1.2 and –0.9. This is what makes the units go away.
3. Here I'll wave my hands and let the computer do the work. The formula then calculates the relationship between height and weight across all the individuals in the sample as measured by standard units. When individuals in the sample are tall, say, 1.5 or 2 standard deviations above the mean, what do their weights tend to be *as measured in standard deviations from the mean for weight*? And when individuals are near to the mean in terms of height, what are their weights as measured in standard units?

If the distance from the mean for one variable tends to be broadly consistent with distance from the mean for the other variable (e.g., people who are far from the mean for height in either direction tend also to be far from the mean in the same direction for weight), then we would expect a strong positive correlation.

If distance from the mean for one variable tends to correspond to a similar distance from the mean for the second variable *in the other direction* (e.g., people who are far above the mean in terms of exercise tend to be far below the mean in terms of weight), then we would expect a strong negative correlation.

If two variables do not tend to deviate from the mean in any meaningful pattern (e.g., shoe size and exercise) then we would expect little or no correlation.

You suffered mightily in that section; we'll get back to film rentals soon. Before we return to Netflix, however, let's reflect on another aspect of life where correlation matters: the SAT. Yes, that SAT. The SAT Reasoning Test, formerly known as the Scholastic Aptitude Test, is a standardized exam made up of three sections: math, reading, and writing. You probably took the SAT, or will soon. You probably did not reflect deeply on *why* you had to take the SAT. The purpose of the test is to measure academic ability and predict college performance. Of course, one might reasonably ask (particularly those who don't like standardized tests): Isn't that what high school is for? Why is a four-hour test so important when college admissions officers have access to *four years* of high school grades?

The answer to those questions is lurking back in Chapters 1 and 2. High school grades are an imperfect descriptive statistic. A student who gets mediocre grades while taking a tough schedule of math and science classes may have more academic ability and potential than a student at the same school with better grades in less challenging classes. Obviously there are even larger potential discrepancies across schools. According to the College Board, which produces and administers the SAT, the test was created to "democratize access to college for all students." Fair enough. The SAT offers a standardized measure of ability that can be compared easily across all students applying to college. *But is it a good measure of ability?* If we want a metric that can be compared easily across students, we could also have all high school seniors run the 100 yard dash, which is cheaper and easier than administering the SAT. The problem, of course, is that performance in the 100 yard dash is uncorrelated with college performance. It's easy to get the data; they just won't tell us anything meaningful.

So how well does the SAT fare in this regard? Sadly for future generations of high school students, the SAT does a reasonably good job of predicting first-year college grades. The College Board publishes the relevant correlations. On a scale of 0 (no correlation at all) to 1 (perfect correlation), the correlation between high school grade point average and first-year college grade point average is .56. (To put that in perspective, the correlation between height and weight for adult men in the United States is about .4.) The correlation between the SAT composite score (critical reading, math, and writing) and first-year college GPA is also .56.[1] That would seem to argue for ditching the SAT, as the test does not seem to do any better at predicting college performance than high school grades. In fact, the best predictor of all is a combination of SAT scores and high school GPA, which has a correlation of .64 with first-year college grades. Sorry about that.

One crucial point in this general discussion is that correlation does not imply causation; a positive or negative association between two variables does not necessarily mean that a change in one of the variables is causing the change in the other. For example, I alluded earlier to a likely positive correlation between a student's SAT scores and the number of televisions that his family owns. This does not mean that overeager parents can boost their children's test scores by buying an extra five televisions for the house. Nor does it likely mean that watching lots of television is good for academic achievement.

The most logical explanation for such a correlation would be that highly educated parents can afford a lot of televisions and tend to have children who test better than average. Both the televisions and the test scores are likely caused by a third variable, which is parental education. I can't prove the correlation between TVs in the home and SAT scores. (The College Board does not provide such data.) However, I can prove that students in wealthy families have higher mean SAT scores than students in less wealthy families. According to the College Board, students with a family income over $200,000 have a mean SAT math score of 586, compared with a mean SAT math score of 460 for students with a family income of $20,000 or less.[2] Meanwhile, it's also likely that families with incomes over $200,000 have more televisions in their (multiple) homes than families with incomes of $20,000 or less.

I began writing this chapter many days ago. Since then, I've had a chance to watch the documentary film *Bhutto*. Wow! This is a remarkable film about a remarkable family. The original footage, stretching all the way from the partition of India and Pakistan in 1947 to the assassination of Benazir Bhutto in 2007, is extraordinary. Bhutto's voice is woven effectively throughout the film in the form of speeches and interviews. Anyway, I gave the film five stars, which is pretty much what Netflix predicted.

At the most basic level, Netflix is exploiting the concept of correlation. First, I rate a set of films. Netflix compares my ratings with those of other customers to identify those whose ratings are highly correlated with mine. Those customers tend to like the films that I like. Once that is established, Netflix can recommend films that like-minded customers have rated highly but that I have not yet seen.

That's the "big picture." The actual methodology is much more complex. In fact, Netflix launched a contest in 2006 in which members of the public were invited to design a mechanism that improved on existing Netflix recommendations by at least 10 percent (meaning that the system was 10 percent more accurate in predicting how a customer would rate a film after seeing it). The winner would get $1,000,000.

Every individual or team that registered for the contest was given "training data" consisting of more than 100 million ratings of 18,000 films by 480,000 Netflix customers. A separate set of 2.8 million ratings was "withheld," meaning that Netflix knew how the customers rated these films but the contest

participants did not. The competitors were judged on how well their algorithms predicted the actual customer reviews for these withheld films. Over three years, thousands of teams from over 180 countries submitted proposals. There were two requirements for entry. First, the winner had to license the algorithm to Netflix. And second, the winner had to "describe to the world how you did it and why it works."[3]

In 2009 Netflix announced a winner: a seven-person team made up of statisticians and computer scientists from the United States, Austria, Canada, and Israel. Alas, I cannot describe the winning system, even in an appendix. The paper explaining the system is ninety-two pages long.* I'm impressed by the quality of the Netflix recommendations. Still, the system is just a super fancy variation on what people have been doing since the dawn of film: find someone with similar tastes and ask for a recommendation. You tend to like what I like, and to dislike what I dislike, so what did you think of the new George Clooney film?

That is the essence of correlation.

APPENDIX TO CHAPTER 4

To calculate the correlation coefficient between two sets of numbers, you would perform the following steps, each of which is illustrated by use of the data on heights and weights for 15 hypothetical students in the table below.

1. Convert the height of each student to standard units: (height – mean)/standard deviation.
2. Convert the weight of each student to standard units: (weight – mean)/standard deviation.
3. Calculate the product for each student of (weight in standard units) × (height in standard units). You should see that this number will be largest in absolute value when a student's height and weight are both relatively far from the mean.
4. The correlation coefficient is the sum of the products calculated above divided by the number of observations (15 in this case). The correlation between height and weight for this group of students is .83. Given that the correlation coefficient can range from –1 to 1, this is a relatively high degree of positive correlation, as we would expect with height and weight.

A	B	C	D	E	F
Student	Height	Weight	Height in standard units	Weight in standard units	(Weight in standard units) × (Height in standard units)
Nick	74	193	1.21	0.99	1.19
Elana	66	133	–0.63	–0.67	0.42
Dinah	68	155	–0.17	–0.06	0.01
Rebecca	69	147	0.06	–0.29	–0.02
Ben	73	175	0.98	0.49	0.48
Charu	70	128	0.29	–0.81	–0.24
Sahar	60	100	–2.00	–1.59	3.18
Maggie	63	128	–1.32	–0.81	1.07
Faisal	67	170	–0.40	0.35	–0.14
Ted	70	182	0.29	0.68	0.20
Narciso	70	178	0.29	0.57	0.17
Katrina	70	118	0.29	–1.09	–0.32
CJ	75	227	1.44	1.93	2.77
Sophia	62	115	–1.54	–1.17	1.81
Will	74	211	1.21	1.49	1.80
Mean	68.73	157.33			Total = 12.39
Standard Deviation	4.36	36.12		Correlation coefficient = Total/n = 12.39/15 = 0.83	

The formula for calculating the correlation coefficient requires a little detour with regard to notation. The figure $\sum$, known as the summation sign, is a handy character in statistics. It represents the summation of the quantity that comes after it. For example, if there is a set of observations x_1, x_2, x_3, and x_4, then $\sum (x_i)$ tells us that we should sum the four observations: $x_1 + x_2 + x_3 + x_4$. Thus, $\sum (x_i) = x_1 + x_2 + x_3 + x_4$. Our formula for the mean of a set of i observations could be represented as the following: mean = $\sum (x_i)/n$.

We can make the formula even more adaptable by writing $\sum_{i=1}^{n} (x_i)$, which sums the quantity $x_1 + x_2 + x_3 + \ldots x_n$, or, in other words, all the terms beginning with x_1 (because i = 1) up to x_n (because i = n). Our formula for the mean of a set of n observations could be represented as the following:

$$\text{mean} = \sum_{i=1}^{n} (x_i)/n$$

Given that general notation, the formula for calculating the correlation coefficient, r, for two variables x and y is the following:

$$r = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sigma_x} \frac{(y_i - \bar{y})}{\sigma_y}$$

where

n = the number of observations;
$\bar{x}$ is the mean for variable *x*;
$\bar{y}$ is the mean for variable *y*;
σ_x is the standard deviation for variable x;
σ_y is the standard deviation for variable y.

The formula for calculating the correlation coefficient requires a little detour with regard to notation. The figure $\sum$, known as the summation sign, is a handy character in statistics. It represents the summation of the quantity that comes after it. For example, if there is a set of observations x_1, x_2, x_3, and x_4, then $\sum (x_i)$ tells us that we should sum the four observations: $x_1 + x_2 + x_3 + x_4$. Thus, $\sum (x_i) = x_1 + x_2 + x_3 + x_4$. Our formula for the mean of a set of i observations could be represented as the following: mean = $\sum (x_i)/n$.

We can make the formula even more adaptable by writing $\sum_{i=1}^{n} (x_i)$, which sums the quantity $x_1 + x_2 + x_3 + \ldots x_n$, or, in other words, all the terms beginning with x_1 (because i = 1) up to x_n (because i = n). Our formula for the mean of a set of n observations could be represented as the following:

$$\text{mean} = \sum_{i=1}^{n} (x_i)/_n$$

Given that general notation, the formula for calculating the correlation coefficient, r, for two variables x and y is the following:

$$r = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sigma_x} \frac{(y_i - \bar{y})}{\sigma_y}$$

where

n = the number of observations;
$\bar{x}$ is the mean for variable *x*;
$\bar{y}$ is the mean for variable *y*;
σ_x is the standard deviation for variable x;
σ_y is the standard deviation for variable y.

Any statistical software program with statistical tools can also calculate the correlation coefficient between two variables. In the student height and weight example, using Microsoft Excel yields the same correlation between height and weight for the fifteen students as the hand calculation in the chart above: 0.83.

* You can read it at http://www.netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf.

CHAPTER 5

Basic Probability

Don't buy the extended warranty on your $99 printer

In 1981, the Joseph Schlitz Brewing Company spent $1.7 million for what appeared to be a shockingly bold and risky marketing campaign for its flagging brand, Schlitz. At halftime of the Super Bowl, in front of 100 million people around the world, the company broadcast a live taste test pitting Schlitz Beer against a key competitor, Michelob.[1] Bolder yet, the company did not pick random beer drinkers to evaluate the two beers; *it picked 100 Michelob drinkers*. This was the culmination of a campaign that had run throughout the NFL playoffs.[2] There were five live television taste tests in all, each of which had 100 consumers of a competing brand (Budweiser, Miller, or Michelob) conduct a blind taste test between their supposed favorite beer and Schlitz. Each of the beer taste-offs was promoted aggressively, just like the playoff game during which it would be held (e.g., "Watch Schlitz v. Bud, Live during the AFC Playoffs").

The marketing message was clear: Even beer drinkers who think they like another brand will prefer Schlitz in a blind taste test. For the Super Bowl spot, Schlitz even hired a former NFL referee to oversee the test. Given the risky nature of conducting blind taste tests in front of huge audiences on live TV, one can assume that Schlitz produced a spectacularly delicious beer, right?

Not necessarily. Schlitz needed only a mediocre beer and a solid grasp of statistics to know that this ploy—a term I do not use lightly, even when it comes to beer advertising—would almost certainly work out in its favor. Most beers in the Schlitz category taste about the same; ironically, that is exactly the fact that this advertising campaign exploited. Assume that the typical beer drinker off the street cannot tell Schlitz from Budweiser from Michelob from Miller. In that case, a blind taste test between any two of the beers is essentially a coin flip. On average, half the taste testers will pick Schlitz, and half will pick the beer it is "challenging." This fact alone would probably *not* make a particularly effective advertising campaign. ("You can't tell the difference, so you might as well drink Schlitz.") And Schlitz absolutely, positively would not want to do this test among its own loyal customers; roughly half of these Schlitz drinkers would pick the competing beer. It looks bad when the beer drinkers supposedly most committed to your brand choose a competitor in a blind taste test—which is exactly what Schlitz was trying to do to its competitors.

Schlitz did something cleverer. The genius of the campaign was conducting the taste test exclusively among beer drinkers who stated that they preferred a competing beer. If the blind taste test is really just a coin flip, then roughly half of the Budweiser or Miller or Michelob drinkers will end up picking Schlitz. That makes Schlitz look really good. *Half of all Bud drinkers like Schlitz better!*

And it looks particularly good at halftime of the Super Bowl with a former NFL referee (in uniform) conducting the taste test. Still, it's live television. Even if the statisticians at Schlitz had determined with loads of previous private trials that the typical Michelob drinker will pick Schlitz 50 percent of the time, what if the 100 Michelob drinkers taking the test at halftime of the Super Bowl turn out to be quirky? Yes, the blind taste test is the equivalent of a coin toss, but what if most of the tasters chose Michelob *just by chance*? After all, if we lined up the same 100 guys and asked them to flip a coin, it's entirely possible that they would flip 85 or 90 tails. That kind of bad luck in the taste test would be a disaster for the Schlitz