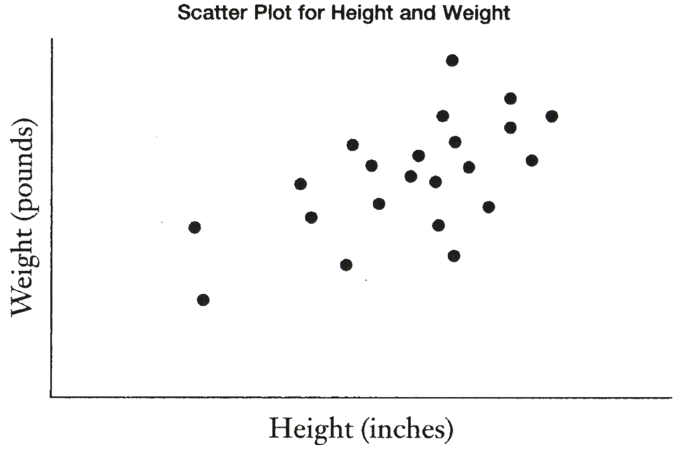
21

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.

22

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.

Excerpts from Wheelan, C. (2013). *Naked Statistics*. New York, NY. W. W. Norton and Company.

23

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.

24

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.

Excerpts from Wheelan, C. (2013). *Naked Statistics*. New York, NY. W. W. Norton and Company.