(Adobe Stock)

chapter 7

Correlation

So far in this text, all we have covered is characteristics of a single distribution of data. In other words, we have looked at only a single variable and how that variable appears when graphed with a histogram. We looked at many kinds of distributions that can come from a variable, and we have found sophisticated ways to describe the shapes of those distributions (e.g., kurtosis and skew).

We have also learned some characteristics of those distributions. The measures of central tendency give us a way of describing the value that best represents all of the data in the set. How different these measures are from each other informs us how skewed the distribution is. Furthermore, the standard deviation is a tool that describes how wide the distribution is around the mean. Using both of these tools, we can then calculate the meaningfulness of a single score within the dataset (by using *z* scores and percentiles). All of these tools are helpful to understand the characteristics of a single dataset.

Because these things merely *describe* the dataset, these tools are called **descriptive statistics**. The mean, median, mode, range, and standard deviation are all ways to describe the patterns of a single variable's appearance. Descriptive statistics are useful on their own, but they are also vital building blocks for most of the rest of what statistical methods can do. In this chapter, we will now step into new territory and use these tools to learn new things.

The rest of the text will focus on methods of **inferential statistics**. Whereas descriptive statistics are methods that let us describe a single dataset, *inferential* statistics are tools that allow us to make inferences about whether and how two variables (or more) are related to each other, if at all. For example, although it is neat to be able to say with descriptive statistics, "The average tastiness rating of my cheeseburgers was 7 out of 10," it may be much more useful information to say with inferential statistics, "My cheeseburgers are rated as tastier among high school students than college students."

In this simple example about the cheeseburgers, there are two variables: (a) the tastiness of the cheeseburgers, and (b) the class of the students who ate the cheeseburgers. Knowing only the characteristics of one variable (tastiness) is some good information, but if we can see how its variation is influenced by another variable (students' class), then maybe we can change one of them to influence the other. For example, knowing how these two variables are related, this restaurateur may wish to focus their advertising on high school students.

This is the potential of inferential statistics—uncovering clues about how one variable may influence another. If the researcher uses sound research methods with careful observations, and then uses the correct inferential statistical tests, the researcher may uncover evidence that one variable affects another. That knowledge is power, whether it is used to increase burger sales, or to find out what causes or cures depression.

7.1 Correlation Overview

The first inferential statistics method we will cover is called "correlation." Correlation is a useful tool in our statistical arsenal, but it is also often misunderstood or even abused. We will work carefully to understand the strengths and limitations of correlation so that we may use it properly.

The **correlation** test seeks to answer what one continuous variable does on average (increase or decrease) when another continuous variable increases or decreases. For example, a researcher may wish to know what happens to a person's blood pressure depending on how much fat they consume. This researcher wants to know whether the person's blood pressure rises or falls as their fat intake rises or falls. There are two continuous variables, and the researcher is interested in how one of them behaves in relation to the other.

The typical correlation is sometimes referred to as a "Pearson correlation," named after the statistician who developed it, Karl Pearson (who adapted it from Francis Galton).

The traditional correlation requires the following:

- Two continuous variables with at least interval properties*
- Each unit of measurement has a value for both variables

What the first point means is that it does not make sense to correlate a categorical variable with something that is continuous. Hopefully that is intuitive, because a categorical variable's "increase" does not make sense. Categorical variables can take on *different* values, but those values indicate only difference, and not order or equal intervals between values, so it makes little sense to try to understand how one continuous variable changes as a categorical variable's value changes categories. For full disclosure, there are some very sophisticated ways to examine this possibility with advanced statistical methods (check out something called "dummy coding"), but it is beyond the scope of this course, and beyond the utility of basic correlation.

The second point means that for each unit of measurement one has (like for each person), that unit needs to provide a value for both variables that interest the researcher.

For example, imagine Irving wants to know if there is a relationship between his hours of sleep from the night before and how much coffee he drinks the following day. Irving needs two observations that are paired together by the day to which they refer. In other words, N is not referring to people, because Irving is the only person in this study. N in this case refers to the number of days that he is collecting these observations. His coffee intake one morning would need to be paired with the hours of sleep from the preceding night; otherwise, he cannot actually assess whether they are related. He needs his data organized in such a way that the two variables go together. Another way of thinking about this is to say that it matters which row in which Irving puts the data. Each row of his data collection should be from one of the days upon which he recorded the datapoints. He cannot just take all of the values from one variable and toss them in randomly next to any of the values of the other variable, because then the correlation will not make any sense:

* There are some special versions of correlational tests that do not require both variables to be continuous. For example, a point-biserial correlation uses one continuous variable and the other variable is dichotomous (i.e., can take on only two values). There are some other special tests, but for the traditional correlation in this textbook, both variables must be continuous.

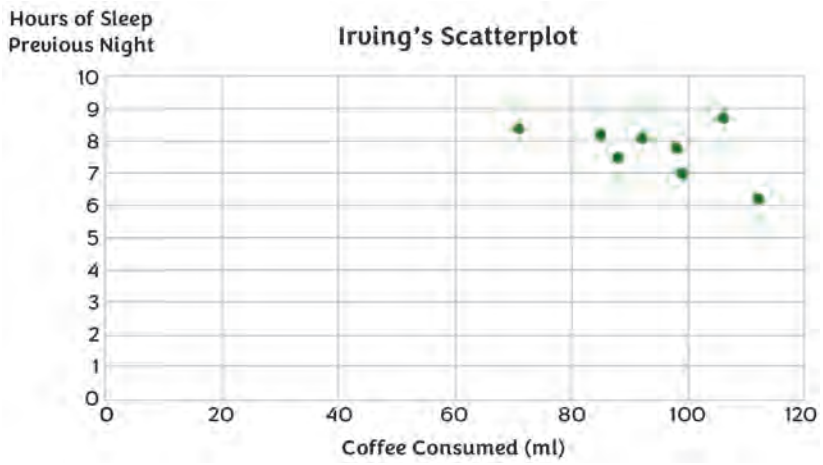
Table 7–1 Irving's Observations of One Variable Need to Pair With Observations of the Other Variable (in This Case, They Pair by the Day They Were Presumed to be Related)

Day	Hours of Sleep Previous Night	Coffee Intake in Morning (ml)
1	8.1	92
2	8.7	106
3	7.5	88
4	8.4	71
5	7.0	99
6	6.2	112
7	7.8	98
8	8.2	85

That is, in this case, the correlation test wants to look at the value of one of the variables (e.g., hours of sleep) on one day, see what the value is of the other variable (e.g., coffee intake) on that same day, and then it does that for each of the days for which there is information available. Finally, it explains whether and how closely the two values changed in a patterned manner.

It may be illustrative to see how such data are graphed, and then what that has to do with the correlation result itself. Data in a correlation test are best illustrated using something called a **scatterplot**. A scatterplot graphs two continuous variables. One of those variables is plotted on the *x* axis, and the other is plotted on the *y* axis. It uses dots (or something similar) to show where a single unit lies in relation to the rest of the units that were measured. To apply that to Irving, the scatterplot puts one of the variables—let us say hours of sleep the previous night—on the *y* axis, and then the other variable—let us say caffeine intake—on the *x* axis. It actually makes no difference which variable goes on which axis in this case, so it could easily go the other way around if we prefer. In this scatterplot, we see 8 dots because the number of times we had paired observations (*N*) of our two variables was 8. In this case, each of these data pairs refers to observations from a specific day that Irving recorded, but other data sets could have paired observations of a person, hamsters, bank accounts, and so on, depending on what sort of unit the research concerned. Each dot in this scatterplot represents a day that Irving observed, and each dot's location on our scatterplot. Each dot's location on our scatterplot is determined by its measurement of either variable, so that it looks like this:

Figure 7–1 A Scatterplot of the Data in Table 7–1

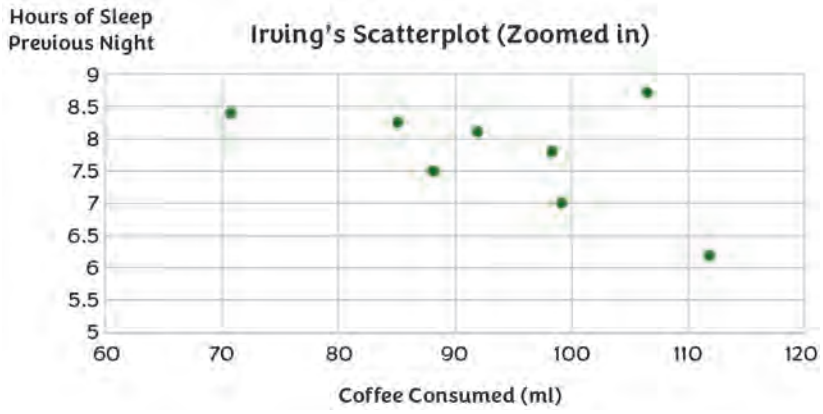


Take a moment now and find which dot represents Day 7, for example. On that day, Irving got 7.8 hours of sleep, so find which dot seems to line up with that value along the *y* axis. Also see that the dot lines up with the value of his coffee consumed for that same paired observation, which was 98 ml.

Just a quick note here before we move on—on this scatterplot each value of one variable must be paired with another value of a variable, or it will not actually go anywhere on the scatterplot. Say Irving got 6.9 hours of sleep one night, but then forgot to measure how much coffee he drank the next morning. In that case, he cannot use the observed value of 6.9 for that day, because he had no value for the paired variable. He would stare at the scatterplot wondering where to put the dot, because he would know where it should be on the *y* axis, but with no paired value along the *x* axis, he would have no idea where it should properly go. However, if Irving had drunk no coffee that morning, then that *would be* a value, because 0 is a value: it is information, whereas forgetting to record is no information.

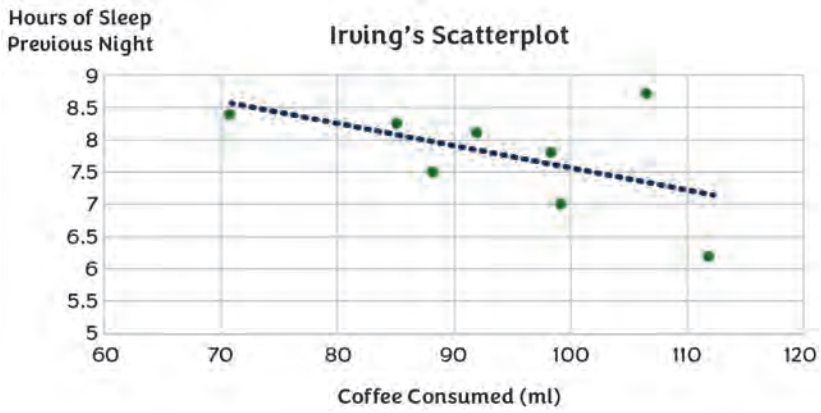
If we zoom in where most of the dots are, we can see that there seems to be a little bit of a pattern to them:

Figure 7–2 Scatterplot of the Data in Table 7–1, but Zoomed In



We can see that there is possibly a gradual slope to the dots, sort of from the upper left side to the lower right side. If we were to draw a line to show this slope, it would look like this:

Figure 7-3 There is, on Average, a Gradual Slope to These Datapoints



That slope of the line is important. We will cover even more of it in the next chapter on regression, but for now, let us just focus on the main point of that line. We can call it a "trendline" for now. Because it is slanted, that means that there is some pattern present in these two variables.

Specifically, a correlation is summarized with two pieces of information:

1. The **direction** of the relationship
2. The **strength** of the relationship

7.2 Direction

We can see the first characteristic in the scatterplot. Because the line is sloped from the upper left side of the graph to the lower right side of the graph, that means that as the hours of Irving's sleep increased, the amount of coffee he consumed the next morning *decreased* on average.

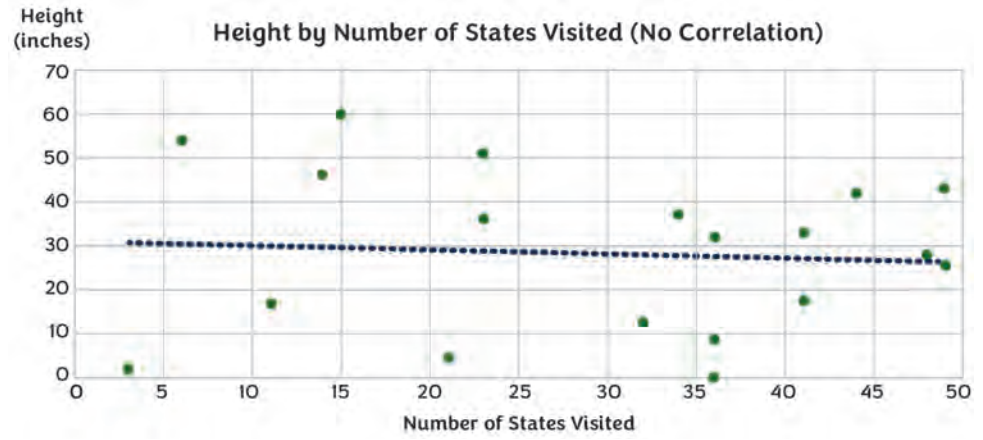
When the two variables go in opposite directions, they are "negatively correlated." That is, as one variable rises, the other one falls. It does not matter which one trends up and which one trends down: as long as they are not doing the same thing together, it is a negative correlation.

When the two variables go in the same direction, they are "positively correlated." That means that both variables rise and fall together, on average. For example, as a person moves more quickly, their heart rate increases, on average. As one increases, the other tends to also increase, and as one decreases, the other also tends to decrease. A positive correlation in a scatterplot has the opposite slope of the negative correlation. The "trendline" will slope from the lower left to the upper right of the scatterplot.

Figure 7-4 A Positive Correlation Trends From the Lower Left to the Upper Right Corners

When the trendline is horizontal, or very close to horizontal, that means that there is "no correlation," or at least no *discernible* correlation, between the two variables.

Figure 7-5 When the Pattern is Relatively Flat, There is No Correlation



In this chapter, we will compute a number that represents the correlation between two variables. It is called the correlation coefficient, and it is represented with r (it is sometimes called Pearson's r). When two variables are negatively correlated, r is a negative number (that is not very close to zero). When the two variables are positively correlated, r is a positive number (that is not very close to zero).

7.3 Strength

When we refer to the "strength" of the relationship between two variables, we basically mean how closely these two variables appear to move together. If they nearly always move predictably in relation to each other (positive or negative), then that would be a very strong relationship. However, if they do not tend to move together in a predictable way, then it is a weaker relationship. In other words, if I can be pretty sure how Irving's coffee intake will be

affected by his sleep the previous night, then there is a strong correlation between his sleep and coffee intake the next morning. But if it is more like a guess at how his coffee intake will change due to his sleep, then it is a weak correlation.

When we compute the *r* for a correlation, the size of the number will tell us the strength of the relationship. An *r* that is close to zero (whether it is positive or negative) is weaker, and an *r* that is farther away from zero and gets closer to ±1 is stronger. In fact, this is an important thing to note, so I will put it in some fancy, eye-catching text:

Math Check

The *r* cannot be less than −1 or more than +1. If an *r* goes beyond those limits, there must be a miscalculation.

That is, if we find $r = 1.28$, we miscalculated. If Danny finds that $r = -7.64$, Danny miscalculated. I can hear Danny saying, "But what if it's, like, a really crazy correlation?" The answer is still no. It is mathematically impossible for *r* to ever go beyond the limits of −1 to +1.

We will revisit this rule as we practice the math that goes along with the correlation, but remember that it is a fast way to know whether the math got off track somewhere in the calculations.

Assigning adjectives to the correlation values is a little tricky, because "strong" and "weak" are rather subjective terms. Because of this fact, I generally hesitate to give clear cutoff criteria for when to call a correlation "weak" or "strong," but for the purpose of assisting in interpretation of what we find, we may use the criteria below. However, please keep in mind that these cutoffs are somewhat arbitrary. Having a "weak" correlation does not necessarily mean that it is uninteresting, useless, or negligible.

Table 7-2 Some Adjectives That May be Used to Describe Correlations

Strength of Relationship	None	"Weak"	"Moderate"	"Strong"	"Perfect"
r =	+/− 0.0 to .29	+/− .30 to .39	+/− .40 to .49	+/− .50 to .99	+/− 1.0

So, if a correlation (*r*) is close to zero, that means there is no discernible pattern to the rising and falling of one variable based on the other (the scatterplot of height by # states visited shows this sort of pattern). However, if the correlation is equal to 1 (one), then that means there is a "perfect" relationship. In other words, with a perfect correlation between one variable and another, we can know the value of one of the variables as long as we know the value of the other. They are perfectly correlated.

In reality, we do not see perfectly correlated variables unless they are measuring the same thing. For example, consider how one's age and the year of their birth are correlated.

Figure 7-6 In a Perfect Correlation, Positive or Negative, All Points Fall on a Line

If someone tells me what year they were born, then I can tell them how old they are now (depending a little on the month). If I find a person who is older, the year they were born decreases predictably, so that there is a perfect negative relationship between one's age and the year they were born.

Of course, as noted above, these have a perfect relationship because they are actually just two ways of measuring the same thing. In fact, if we find a correlation between two variables that is getting to be very close to 1, then we may want to consider the possibility that the two variables are the same thing, or at least are not different enough to treat them as separate.

This sort of thing could happen in psychology. Say a researcher carefully constructs a measure of confidence, and then asks 50 people to complete it. The researcher may think that it will be correlated with a scale of arrogance, because those phenomena are similar in many ways. So, the researcher may be pleased to see that the two scales are correlated at .70 (a strong, positive relationship). However, if the researcher finds a correlation of .96 (a nearly perfect relationship), that looks like the researcher was not really measuring two different things and should spend some more time working on their scale to find what distinguishes confidence and arrogance.

We will see how to calculate a correlation shortly, but before moving on it is important to note what we can and cannot conclude from a correlation. Correlation is frequently misunderstood and misused, and so we must be clear about its limitations.

7.4 Correlation Never Implies Causation

Students should memorize that phrase, tell all of their friends, make a cross-stitch that says that and put it above their bed so that it is the first thing they see in the morning and the last thing they see at night.

That phrase means that just because two things seem related, it does not mean that one of them made the other one occur. There are several reasons that correlation does not imply causation, which we will now explore.

Consider an example. Imagine that a researcher does a large study and concludes the following:

"People who eat more organic food tend to live longer."

In other words, the researcher found a positive correlation between organic food intake and longevity. Let us assume that it was even a good study of a large, randomly selected

sample, with careful observations. A tempting misinterpretation of this study is that eating organically grown (or raised) foods causes longer life. That is, organic foods must be more nutritious or in some other way better for the human body than non-organic foods, right? The answer is, "Perhaps, but there are other explanations for why this correlation exists that are not due to organic foods' inherent properties causing longer life." In fact, a positive correlation will appear if any of the following possibilities is true:

Possibility 1: Organic food really does cause longer life. It could be that there really is something about the process of organic farming and ranching that reduces cancer-causing substances to appear in foods, or that the organic foods contain significantly more nutrients in appropriate amounts that would support human health, compared to the alternatives.

Possibility 2: Living longer increases the opportunity to eat more organic foods. It makes sense that as a person lives longer, they have more contact with organic foods, especially as organic foods find their way onto our grocery shelves more easily as they get cheaper to produce and demand for them rises. So maybe living longer is what causes a person to eat more organic food.

As we can see, one reason that correlation does not imply causation is because we cannot (usually) infer *directionality*, or which variable led to the change in the other.

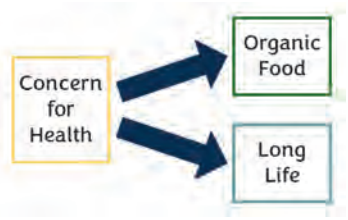
I slipped in that "usually" because sometimes it is not possible for one of the variables to have had an influence on the other. For example, if I find that longevity is correlated with how long one's funeral is, it would not really be possible for the funeral length to have had any effect on longevity, because chronologically it would present a problem (unless we acknowledge that time is an illusion; e.g., Rovelli, 2018).

Possibility 3: Buying organic food makes one feel better about oneself, which in turn makes them live longer. In this possibility, organic food by itself has no effect on longevity whatsoever, but it does affect how one feels about themselves, and then maybe it is that feeling about themselves that causes them to live longer.

Another reason that we cannot infer causation from correlation is because the correlation does not account for variables that come in between the two we measured.

Possibility 4: Being more concerned about one's health causes both longer life and an increase in organic food consumption. If a person chooses to eat organic food, they probably are also more concerned about their health and diet generally. It is unlikely that they

eat lots of foods that are high in saturated fats or loaded with sugar. So then maybe it is not the organic food per se that is leading to their living longer, but it is simply the fact that they are mindful of what they put in their bodies. In other words, maybe eating an apple is way healthier than a cookie, regardless of whether it was grown organically or conventionally.



We cannot infer causation from correlation, because the correlation does not account for a third variable that could cause the change in both of the variables that we measured.

And of course, there are many other possibilities besides only these four. The problem for researchers is that all they learn from a correlation is that one of these possibilities could be true, but the correlation does not shed light on which one is actually true. To figure that out, more sophisticated research designs and statistical tests are required.

7.5 The Equation

Now that we are all on board with how a correlation is used, it is time to learn how it is computed. We will now get to know the equation better to see what it does, and then we will practice the math associated with the correlation.

Just like with the standard deviation formula, there is a correlation formula that does the long-hand work of computing the correlation coefficient (the raw-score equation), and there is also a simplified version that arrives at the same answer but takes fewer steps (the computational equation). Let us first get to know the long-form version.

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\Sigma(x_i - \bar{x})^2][\Sigma(y_i - \bar{y})^2]}}$$

That looks like a lot, right? It is going to be fine, though. Try not to get lost in the size of the equation. We will go through each piece so that we know what it does, and then will solve the equation using the simplified version of the formula.

The first thing to notice about the correlation equation is that it is a fraction. There is a numerator and a denominator. In a practical sense, that means that the correlation coefficient is a simplified version of a ratio of whatever is in the numerator to whatever is in the denominator. That is, it is looking at how much of the numerator there is compared to how much of the denominator there is. In the correlation, if there is the same amount in the numerator as in the denominator, then it will be a perfect correlation, $r = 1$. As we have covered, it is not mathematically possible to find more in the numerator than in the denominator for a correlation, so r cannot be more than 1.

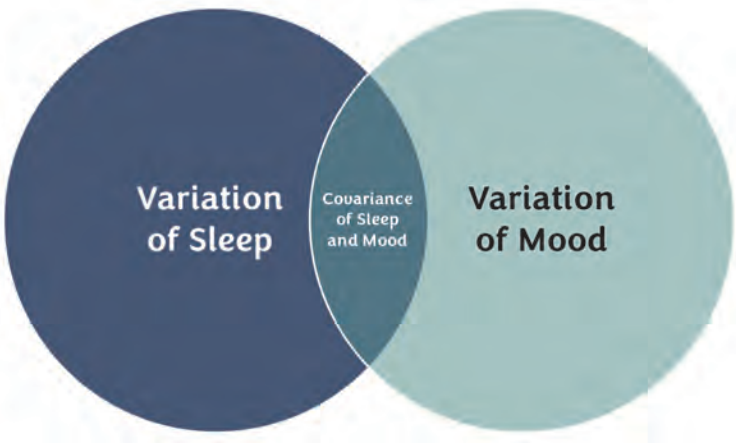
Now, let us explore what is in the numerator and the denominator of the equation. The numerator of this equation is $\Sigma(x_i - \bar{x})(y_i - \bar{y})$. In words, it tells us to subtract the mean of all of the scores of the x variable from each individual score of the x variable. We then do that same thing for the y variable scores, using the mean of the y variable scores. This step gives us two sets of deviation scores: one for the x scores and one for the y scores. Next, we would

pair each *x* deviation with its *y* deviation counterpart (the two observations that were paired) and multiply those deviation scores together for each unit for which we obtained paired observations. We add up the products we got from doing that step, and that is our numerator.

Even I got lost a little on that explanation, so I will now make it more concrete. Say Germaine has 30 best friends. She asks each friend their height and their weight. Germaine now has two paired variables: height and weight. They are both continuous data, and each unit (person) gave her a score on each variable, so a correlation is one way to analyze these observations. She first calculates the average height, and then the average weight for this group of best friends. Then, she would calculate a deviation score of each friend's height from their mean height, and also a deviation score of each friend's weight from the mean weight. Now that she has a deviation score of height and weight for each friend, she then multiplies those deviation scores together for each friend, and then adds up all of those multiplied deviation scores from all of the friends.

The numerator gives us what is called the **covariance** between the *x* and *y* variables. In other words, it is an index of how much both variables vary together. Let us dig into that a little deeper, using an example. Think about sleep and mood. Consider whether these two variables are related to each other in any way. On the one hand, sleep is affected by lots of things, like noise, temperature, the softness of one's mattress, what one ate before they lay down, and so on. Similarly, mood is affected by lots of things, including attitude, the weather, how others treat us, how well our thyroid is working, and so on. In some ways, sleep is affected by mood, and mood is affected by sleep. When we sleep poorly, that often affects our mood in some way. Additionally, when we are in a bad mood, that often has some influence on how well we sleep. In this way, mood and sleep may have some relationship to each other. Mood is not entirely in control of our sleep quality (or duration, or whatever), and our sleep does not entirely affect our mood, because plenty of other variables will influence both of these things. Still, there is some degree of influence between them, and so that is essentially what we are calling the "covariance." Here is a visualization:

Figure 7-7 Both Sleep and Mood Vary for Many Reasons. Where They Vary Together is Called "Covariance"



The circle on the left represents all the variation in sleep quality. The circle on the right represents all of the variation in mood. The overlapping sections of the circles represent the covariation of sleep and mood, or how they vary together. This overlapping section is what is sought out by the numerator of the correlation equation.

Now let us dissect the denominator of the equation: $\sqrt{[\Sigma(x_i - \bar{x})^2][\Sigma(y_i - \bar{y})^2]}$. This portion should look a little familiar, in that it asks us to create a deviation score for each value of the variable, square it, and then sum up those squares. We have done that before—that was a big step in the standard deviation equation. In fact, that is precisely what the denominator is in the correlation coefficient: the standard deviations of the two variables. The denominator is essentially just the combined standard deviations of the two variables for which we compute the correlation. The standard deviation, of course, is a way of measuring the (average) variability in a variable.

What this all means is that the correlation coefficient is the ratio of the covariance of the two variables to the average variation of those two variables. That is, if we were to put the correlation coefficient into words, it would look like this:

$$r = \frac{\text{How much the two variables change together}}{\text{How much the two variables change at all}}$$

So, the correlation coefficient gets bigger the more the two circles in our figure (like Figure 7–7) overlap. If they perfectly overlap, that means that the two variables always change together, and the total proportion of one circle overlapped by another circle is 100%, or $r = 1$. If they do not overlap at all, then they do not covary at all, and $r = 0$. Then, anything between 0 and ±1 is some degree of overlap.

Now, as I noted above, rather than go through the math of the raw score formula for the correlation, let us use a mathematically equivalent formula that arrives at the same answer as the raw score formula, but uses fewer steps (we did this sort of thing with the standard deviations). It does not matter which formula we use, as we will arrive at the same answer, but this simplified formula makes it easier to arrive at the answer. We will call this the "computational formula for the correlation."

$$r = \frac{N\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{[N\Sigma x_i^2 - (\Sigma x_i)^2][N\Sigma y_i^2 - (\Sigma y_i)^2]}}$$

Take a careful look at the formula's elements, and notice that we are familiar with nearly everything in it already. We will start with the numerator:

N, when referring to a correlation, is the total number of paired observations of the variables included in the equation. That is, it represents how many observations of variable x that are also paired with an observation of the variable y.

As we can see, N is right next to $\Sigma x_i y_i$. Because capital sigma means to add up whatever is just to the right of it, we can first do the stuff next to it. The x_i and y_i right next to each other indicates that we multiply them together. What this really means is to multiply each value (hence the "i" subscript) of the x variable with its paired y value. Once we have multiplied each x with its paired y, the sigma tells us to add up those products.* Then, because the N is right next to sigma, that means that we multiply that sum of the products by the total number of pairings there were. That completes $N\Sigma x_i y_i$.

Still in the numerator is $(\Sigma x_i)(\Sigma y_i)$. Notice the parentheses. Those tell us that we need to add up all of the values of x we observed in the dataset (Σx_i), and then also add up all of the values of y we observed in the dataset (Σy_i). Then, because these two things are right next to each other, we multiply one sum by the other sum. We then subtract this piece from what we calculated in the $N\Sigma x_i y$ part. That concludes the calculation of the numerator, but we will practice it with real numbers in a moment.

* Even though the order of operations says that addition should come after multiplication, the sigma implies that there are parentheses around itself. If the sigma were before the N, then the multiplication would come before the addition.

Now, for the denominator. Notice that the left side of the denominator refers to all of the *x* values, whereas the right side of the denominator refers to all of the *y* values. It is the same steps, just first with the *x* values, then the *y* values.

Let us focus on the parentheses first. Again, the equation asks for Σx_i on the left side of the denominator and Σy_i on the right side. These are the same values as we used in the numerator, just before we multiplied them together. They are just the sums of the *x* values or *y* values, respectively. In either case, we are asked to square the sum of the values of the *x* variable on the left side: $(\Sigma x_i)^2$. We then do the same with the values of the *y* variable on the right side of the equation: $(\Sigma y_i)^2$.

We will subtract those portions from the $N\Sigma x_i^2$ or $N\Sigma y_i^2$. As we see, there are no parentheses, so we do the exponents first. That means we square each value of *x* we observed and square each value of *y* we observed. We then add up all the squared *x*s on the left, and multiply them by *N*, the total number of paired observations. Do the same with the *y*s on the other side of the equation.

Do not forget to subtract the squared sums from the sums of squares, then multiply the results together, and then take the square root. That will give us the denominator. Let us try this out with some real numbers so that we can put it all together.

Here is a scenario:

Baxter likes to go fishing on the weekends, but he is not sure which bait is most effective. He decides to test it out. Each weekend, he goes fishing at the same place from exactly 6:00 p.m. to exactly 8:00 p.m. Each weekend, he tries a new kind of bait and records the price of the bait. He then records how many fish he caught on that same night using that bait.

Here are his data after 12 weekends:

Table 7–3 Baxter Went Fishing for 12 Weekends and Kept Track of His Bait Costs and How Many Fish He Caught on Those Weekends

Weekend	Bait Cost in $ (x)	Fish Caught (y)
1	7.23	3
2	3.45	5
3	8.12	4
4	2.20	6
5	6.50	1
6	6.43	0
7	9.20	2
8	4.90	4
9	2.70	2
10	3.25	6
11	7.75	3
12	5.10	2

Notice here that each observation of the *x* variable must be matched with an observation of the *y* variable. In this case, they are matched according to the corresponding weekend on which both occurred. Notice also that both variables are continuous and have at least interval properties. Again, that is necessary for a standard correlation.

The correlation equation asks us to do several things with these data, so to keep it all organized, we will just add some sections to our table. One of the first and simplest things we can do is sum up the number of paired observations. Then we will sum up the total of our *x* values and then the total of our *y* values, so let us add a row at the bottom for those totals (I have color-coded these new elements just so it is easier to keep track of them):

Table 7-4 Baxter's Data With a Row for the Totals

	Weekend	Bait Cost in $ (*x*)	Fish Caught (*y*)
	1	7.23	3
	2	3.45	5
	3	8.12	4
	4	2.20	6
	5	6.50	1
	6	6.43	0
	7	9.20	2
	8	4.90	4
	9	2.70	2
	10	3.25	6
	11	7.75	3
	12	5.10	2
Σ	12	66.83	38

No sweat. Now we have $N = 12$, $\Sigma x_i = 66.83$, and $\Sigma y_i = 38$. In other words, Baxter went fishing for 12 weekends and recorded data on both variables for each of those 12 weekends, giving us a total of 12 paired observations (N). He spent a total of $66.83 on bait (whoa, must be nice to have that kind of cheddar lying around), and he caught a total of 38 fish in that time (imagine the smell!).

We can plop those values right into our correlation equation so far:

$$r = \frac{N\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{[N\Sigma x_i^2 - (\Sigma x_i)^2][N\Sigma y_i^2 - (\Sigma y_i)^2]}}$$

$$r = \frac{12\Sigma x_i y_i - (66.83)(38)}{\sqrt{[12\Sigma x_i^2 - 66.83^2][12\Sigma y_i^2 - 38^2]}}$$

That information already fills out big portions of the equation! We still need to find the sum of squares for each variable (Σx_i^2 and Σy_i^2) as well as the summed products of the paired values ($\Sigma x_i y_i$). Let us add some new columns for those steps:

Table 7-5 Baxter's Data With Columns Added for the Squared Values and the Products of *x* and *y*

	Weekend	Bait Cost in $ (x)	x^2	Fish Caught (y)	y^2	xy
	1	7.23	52.27	3	9	21.69
	2	3.45	11.90	5	25	17.25
	3	8.12	65.93	4	16	32.48
	4	2.20	4.84	6	36	13.20
	5	6.50	42.25	1	1	6.50
	6	6.43	41.34	0	0	0.00
	7	9.20	84.64	2	4	18.40
	8	4.90	24.01	4	16	19.60
	9	2.70	7.29	2	4	5.40
	10	3.25	10.56	6	36	19.50
	11	7.75	60.06	3	9	23.25
	12	5.10	26.01	2	4	10.20
Σ	12	66.83	431.12	38	160	187.47

Hang on just a moment here. Look back at the total of the squared columns. Look first at the total of the squared *x* values (x_i^2). It is 431.12. Some students may be tempted to just square the 66.83 value we obtained after summing them, but that does not equal 431.12. $66.83^2 = 4{,}466.249$, so do not try to take that wrong shortcut. This is the whole reason why we have to square each value first, and then sum the squares in that separate column. Obtaining those separate values is what allows us to skip the steps in the raw score equation.

We have the rest of the values that we need to put into our equation:

$$r = \frac{12\Sigma x_i y_i - (66.83)(38)}{\sqrt{[12\Sigma x_i^2 - 66.83^2][12\Sigma y_i^2 - 38^2]}}$$

$$r = \frac{12(187.47) - (66.83)(38)}{\sqrt{[12(431.12) - 66.83^2][12(160) - 38^2]}}$$

From this point, all we have to do is complete the equation step-by-step, following the rules of the order of operations. That means we will first do anything in parentheses. However, the equation does not actually have anything needing to be done within parentheses. The parentheses now are just there to keep numbers separated that need to be multiplied. In that case, we can first begin with the exponents (from now on, the color coding does not correspond to the table, but is just to keep track of items we are computing in each step):

$$r = \frac{12(187.47) - (66.83)(38)}{\sqrt{[12(431.12) - 66.83^2][12(160) - 38^2]}}$$

$$r = \frac{12(187.47) - (66.83)(38)}{\sqrt{[12(431.12) - 4{,}466.25][12(160) - 1{,}444]}}$$

Next, we can move onto the multiplication:

$$r = \frac{12(187.47) - (66.83)(38)}{\sqrt{[12(431.12) - 4{,}466.25][12(160) - 1{,}444]}}$$

$$r = \frac{2{,}249.64 - 2{,}539.54}{\sqrt{[5{,}173.44 - 4{,}466.25][1{,}920 - 1{,}444]}}$$

Now we have steps within parentheses (or brackets, in this case) that we can complete, so that is what we will do next. Of course, in the numerator, we can also complete that subtraction without interrupting our order:

$$r = \frac{2{,}249.64 - 2{,}539.54}{\sqrt{[5{,}173.44 - 4{,}466.25][1{,}920 - 1{,}444]}}$$

$$r = \frac{-289.9}{\sqrt{(707.19)(476)}}$$

Pause here for just a second. One quick clue about whether we made a mistake in our calculations is these numbers in the denominator (707.19 and 476). Both of these numbers *absolutely must be positive*. It is impossible to get a negative number for either of these, so if we find a negative number in the denominator at this point, we made a mistake somewhere. One hint at this fact is that if one of them is negative, then we would be asked to take the square root of a negative number, which, if we attempt, causes the universe to implode (or something like that—I have never actually tried it).

Math Check

The denominator of the correlation equation absolutely must be positive. If it is negative, there must be a miscalculation.

However, just because both numbers there are positive, that is not an assurance that our math is correct. Many mistakes can still result in positive numbers there. Still, we can use this quick math check to help ensure that we are on track.

Another thing to notice in this result is that the numerator is a negative number. That is totally possible because we can have *r* values that are anywhere between −1 and +1. We can tell already at this point that whatever our *r* value is, it is negative. That is half of the battle in interpreting a correlation, but we still need to know the *strength* of the correlation...

In any case, the next step is of course to multiply these two numbers in the denominator:

$$r = \frac{-289.9}{\sqrt{(707.19)(476)}} = \frac{-289.9}{\sqrt{336{,}622.44}}$$

We can now take the square root of the denominator. Do not forget this step—it is a very common mistake for students to just leave out the square root operator as they write out the equation over and over again, but this will ultimately lead to the wrong answer, so do not skip this step!

$$r = \frac{-289.9}{\sqrt{336{,}622.44}} = \frac{-289.9}{580.25}$$

Let us pause now and think again about what we have so far. In the numerator, we have −289.9. That number represents the covariance between our two variables. In other words, that is about how much the cost of the bait and the number of fish caught change together. The number so far does not really mean anything that is easy to interpret by itself, because it is not in dollars or in number of fish, but some weird hybrid of both of those. Still, we can see something interesting already about how different the numerator is from the denominator.

The denominator is 580.25. That number represents just how much variation there was in either variable by themselves. Again, the number is not easily interpretable because it is not in dollars or number of fish caught, but what is noteworthy is just how large the numerator is in relation to the denominator. It is very close to half of the size, which means that there was a pretty good amount of covariance between these variables. In fact, we will continue the interpretation after we have finished the equation's final step:

$$r = \frac{-289.9}{580.25} = -.50$$

We now have our correlation coefficient, $r = -.50$. Think back to what we covered about r on the previous pages. First, it cannot be lower than −1 or above +1. Ours is within those limits, so that is a good sign that we did not miscalculate. Now, we need to think about the two things that the correlation coefficient communicates: (a) the direction and (b) the strength of the relationship.

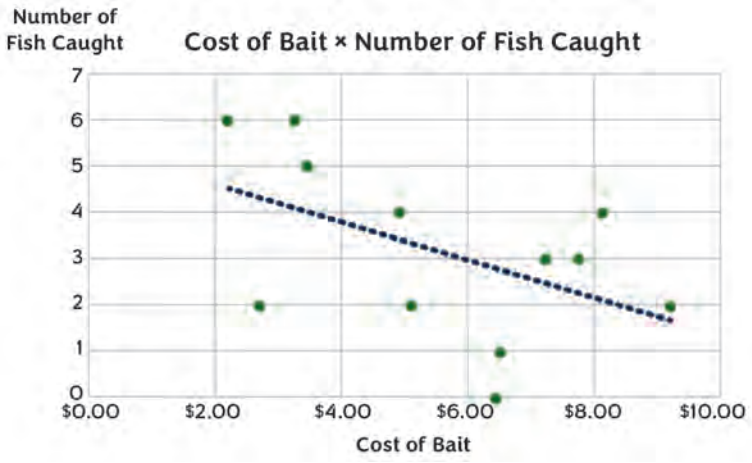
The direction of the relationship is clear to see from the minus sign preceding the ".50." Because of that minus sign, we can state that this is a *negative* correlation. In other words, as one of our variables increased in its value, the other one tended to decline in its value, on average.

The strength of the relationship has to do with the value of the number. This one is ".50," and so we can refer back to Table 7–2 to see that it just barely falls within the range of a "strong" relationship. That basically means that these two variables change together relatively well.

All together, we can say that there is a strong, negative correlation between the amount of money the bait cost and the number of fish Baxter caught. To be clear, that means that as Baxter spent *more* money on bait, he caught *fewer* fish with that bait. That is, as one of the variables increased, the other decreased (the definition of a negative correlation).

Perhaps a scatterplot would help to illustrate the relationship:

Figure 7-8 Scatterplot of Baxter's Data, Including the Trendline Showing the Negative Relationship



Again, each green dot in the scatterplot represents one of the pairings of bait with fish caught. We can think of each green dot as the day that Baxter bought bait and also went

fishing (assuming he did these things on the same day). The blue line shows the average trend of the data. It is a line that is as close to each of the green dots as possible (we will see how to make one in the next chapter). This line is sloping from the upper-left of the graph towards the lower-right, which is another definition of a negative relationship.

To drive home the earlier points about correlation not equating causation, Baxter might make the mistake of saying to himself, "Well tarnation! Looks like the more I spent on the ding dang bait, the fewer fish I caught! That means that this expensive bait ain't worth spit!" Surely, Baxter may be correct—maybe the more expensive bait is actually the reason (and the only reason) that he caught fewer fish, on average, as he spent more on the bait. However, it could also be that he happened to buy expensive bait on days when there were more anglers around catching the fish, and so there were simply fewer fish to be caught. It could also be that the fish were more or less hungry as the season changed (12 weeks is a long time). Baxter did not take into account the time of day that he caught the fish—maybe one type of bait was better for catching fish in the morning versus the evening, for example. Baxter did not tell us whether he used the same fishing line, weights, rod, and so on with each type of bait. Maybe catching more fish from the week before led Baxter to subconsciously feel more comfortable spending extra money on bait the next week (so that he assumed the wrong pairing of the variables). And of course, there are dozens of other legitimate criticisms of the way Baxter went about answering his question. Again, we do not assume a causal relationship if a correlation is the only information we have.

In the next chapter, we will build more upon the information we gleaned from the correlation and put it to more practical use. In the meantime, let us make sure we understand the concepts of a correlation by answering the questions below. Additionally, try out the practice problems in **Appendix F**.

Practice

Here are a few questions we can answer to see if we can correctly interpret a correlation coefficient:

Task – Describe these correlations in terms of both strength and direction.

1. $r = .78$*
2. $r = -.12$†
3. $r = -.92$‡
4. $r = .38$§
5. $r = 3.65$¶

Task – Explain whether the following are positive or negative correlations, or if it is impossible to tell.

6. Fritz is a dachshund whose barking tends to increase when his energy level decreases. What is the correlation between Fritz's barking and energy level?**
7. As Suzanne spends less money on movies in a given month, her happiness also decreases. What is the correlation between Suzanne's spending on movies and happiness?††
8. Dr. Blasenshirm discovers that injecting mice with a hormone that affects their ability to smell changes their mating behavior.‡‡

---

* Strong, positive correlation.
† Negative correlation that is so weak (close to zero), that we could say there is no correlation at all.
‡ Very strong (nearly perfect) negative correlation.
§ Weak, positive correlation.
¶ This one is not a possible correlation, because it falls outside the range of +1 or −1. Looks like somebody miscalculated...
** This is a negative correlation because as one of the variables increases, the other decreases on average.
†† This is positive correlation because the variables decrease (and increase) together, on average.
‡‡ This one is impossible to tell, because the statement gives no indication about the *direction* of the change.

Summary

- Correlation is a statistical method that summarizes how one continuous variable tends to change with the increase or decrease of another continuous variable.
- When reporting a correlation, we must report not only the strength of the relationship (from 0 to 1), but also the direction of that relationship (− or +).
- Correlation never implies a causal relationship, because there are many circumstances that may produce a correlation when two variables are not actually related. Interpret correlation with caution.