The Effect
An Introduction to Research Design and Causality

Nick Huntington-Klein

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

3
Describing Variables

3.1 Descriptions of Variables

This chapter will be all about how to describe a variable. That seems like an odd goal.[1] The opening to this book was all about setting up research questions and how empirical research can help us understand the world. And we jet right from that into describing variables? What gives?

It turns out that empirical research questions really come down entirely to describing the density distributions of statistical variables. That's, well, that's really all that quantitative empirical research is. Sorry.

Maybe that's the wrong approach for me to take. Perhaps I should say that all the interesting empirical research findings you've ever heard about—in physics, sociology, biology, medicine, economics, political science, and so on— can be all connected by a single thread. That thread is laid delicately on top of a mass of probability. The shape it takes as it lies is the density of a statistical variable, tying together all empirical knowledge, throughout the universe, forever.

[1] And what does it even mean? We'll get there.

DOI: 10.1201/9781003226055-3

Is that better? Am I at the top of *The New York Times* nonfiction bestseller list yet?

Look, in order to make any sense of data we have to know how to take some observations and describe them. The way we do that is by describing the types of variables we have and the distributions they take. Part of that description will be in the form of describing how different variables interact with each other. That will be Chapter 4. In this chapter, we'll be describing variables all on their own. It will be less interesting than Chapter 4, but, I am sorry to say, more important.

Variable. A set of observations of the same thing.

A variable, in the context of empirical research, is a bunch of observations of the same measurement—the monthly incomes of 433 South Africans, the number of business mergers in France in each year from 1984—2014, the psychological "neuroticism" score from interviews with 744 children, the color of 532 flowers, the top headline from 2,348 consecutive days of *The Washington Post*. Successfully *describing a variable* means being able to take those observations and clearly explain what was observed without making someone look through all 744 neuroticism scores themselves. Trickier than it sounds.

3.2 Types of Variables

THE FIRST STEP in figuring out how to describe a variable is figuring out what kind of variable it is.

While there are always exceptions, in general the most common kinds of variables you will encounter are:

Continuous Variables. Continuous variables are variables that could take any value (perhaps within some range). For example, the monthly income of a South African would be a continuous variable. It could be 20,000 ZAR,[2] or it could be 34,123.32 ZAR, or anything in between, or from 0 to infinity. There's no such thing as "the next highest value," since the variable changes, well, continuously. 20,000 ZAR isn't followed by 20,001 ZAR, because 20,000.5 ZAR is between them. And before you get there you have to go through 20,000.25 ZAR, and 20,000.10 ZAR, and so on.

[2] ZAR is the South African rand, the official currency in South Africa.

Count Variables. Count variables are those that, well, count something. Perhaps how many times something happened or how many of something there are. The number of business mergers in France in a given year is an example of a count variable. Count variables can't be negative, and they certainly can't take fractional values. They can be a little tougher to deal with

than continuous variables. Sometimes, if a count variable takes many different values, it acts a lot like a continuous variable and so researchers often treat them as continuous.

Ordinal Variables. Ordinal variables are variables where some values are "more" and others are "less," but there's not necessarily a rule as to how *much* more "more" is. A "neuroticism" score with the options "low levels of neuroticism," "medium levels of neuroticism," and "high levels of neuroticism" would be an example of an ordinal variable. High is higher than low, but how much higher? It's not clear. We don't even know if the difference between "low" and "medium" is the same as the difference between "medium" and "high." Another example of an ordinal variable that might make this clear is "final completeted level of schooling" with options like "elementary school," "middle school," "high school," and "college." Sure, completing high school means you got more schooling than people who completed middle school. But how much more? Is that… two more school? That's not really how it works. It's just "more." So that's an ordinal variable.

Categorical Variables. Categorical variables are variables recording which category an observation is in—simple enough! The color of a flower is an example of a categorical variable. Is the flower white, orange, or red? None of those options is "more" than the others; they're just different. Categorical variables are very common in social science research, where lots of things we're interested in, like religious affiliation, race, or geographic location, are better described as categories than as numbers.

A special version of categorical variables are *binary variables*, which are categorical variables that only take two values. Often, these values are "yes" and "no." That is, "Were you ever in the military?" Yes or no. "Was this animal given the medicine?" Yes or no. Binary variables are handy because they're a little easier to deal with than categorical variables, because they're useful in asking about the effects of treatments (Did you get the treatment? Yes or no) and also because categorical variables can be turned into a series of binary variables. Instead of our religious affiliation variable being categorical with options for "Christian," "Jewish," "Muslim," etc., we could have a bunch of binary variables—"Are you Christian?" Yes or no. "Are you Jewish?" Yes or no. Why would we want that? As you'll find out throughout this book, it just happens to be kind of convenient. Plus it allows for things like someone in your data being both Christian *and* Jewish.

Qualitative Variables Qualitative variables are a sort of catch-all category for everything else. They aren't numeric in nature, but also they're not categorical. The text of a *Washington Post* headline is an example of a qualitative variable. These can be very tricky to work with and describe, as these kinds of variables tend to contain a lot of detail that resists boiling-down-and-summarizing. Often, in order to summarize these variables, they get turned into one of the other variable types above first. For example, instead of trying to describe the *Washington Post* headlines as a whole, perhaps asking first "how many times is a president referred to in this headline?"—a count variable—and summarizing that instead.

3.3 The Distribution

ONCE WE HAVE AN IDEA OF WHAT KIND OF VARIABLE WE'RE DEALING WITH, the next step is to look at the *distribution* of that variable.

A variable's *distribution* is a description of *how often different values occur*. That's it! So, for example, the distribution of a coin flip is that it will be heads 50% of the time and tails 50% of the time. Or, the distribution of "the number of limbs a person has" is that it will be 4 most often, and each of 0, 1, 2, 3, and 5+ will occur less often.

Distribution. A description of the probability that each possible value of a variable will occur.

When it comes to categorical or ordinal variables, the variable's distribution can be described by simply giving the percentage of observations that are in each category or value. The full distribution can be shown in a frequency table or bar graph, which just shows the percentage of the sample or population that has each value.

Variable	N	Percent
Primary Degree Type Awarded	7424	
... Less than Two-Year Degree	3495	47.1%
... Two-Year Degree	1647	22.2%
... Four-Year or More	2282	30.7%

Data from College Scorecard

Table 3.1: Distribution of Kinds of Degrees US Colleges Award

Frequency table. A table that shows the proportion of the time that a variable takes a given value. A **bar graph** is a graphical representation of a frequency table.

These tables tell you all you need to know. From Table 3.1 we can see that, of the 7,424 colleges in our data,[3] 3,495 of them (47.1%) predominantly grant degrees that take less than two years to complete, 1,647 of them (22.2%) predominantly grant degrees that take two years to complete, and 2,282 of

[3] Generally in statistics, as well as in this book, "N" means "number of observations."

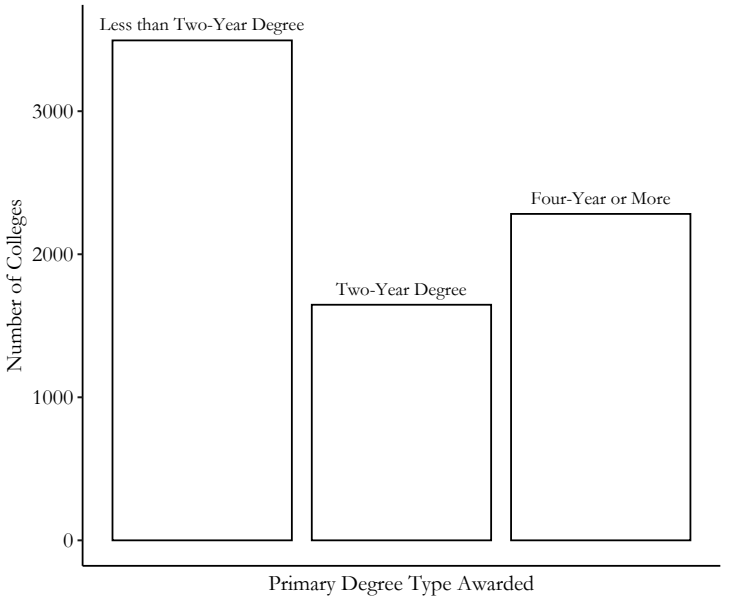

Figure 3.1: Distribution of Kinds of Degrees US Colleges Award

them (30.7%) predominantly grant degrees that take four years or more to complete or advanced degrees. Figure 3.1 shows the exact same information in graph format.

There are only so many possibilities to consider, and the table and graph each show you how often each of these possibilities comes up. Once you've done that, you've fully described the distribution of the variable. There's literally no more information in this variable to show you! If we wanted to show more detail (maybe *which majors* each college tends to specialize in) we'd need a different data source with a different variable.

Continuous variables are a little trickier. We can't just do a frequency table for continuous variables since it's unlikely that more than one observation takes any specific value. Sure, one person's 24,201 ZAR income is very close to someone else's 24,202 ZAR. But they're not the same and so wouldn't take the same spot on a frequency table or bar chart.

For continuous variables, distributions are described not by the probability that the variable takes a given value, but by the probability that the variable takes a value *close* to that one.

One common way of expressing the distribution of a continuous variable is with a *histogram*. A histogram carves up the potential range of the data into bins, and shows the proportion of observations that fall into each bin. It's the exact same thing as the frequency table or graph we used for the categorical variable, except that the categories are ranges of the variable rather than the full list of values it could take.

Histogram. A graph showing the proportion of the time that a variable falls into a given range between two values.

For example, Figure 3.2 shows the distribution of the earnings of college graduates a few years after they graduate, with one observation per college per graduating class ("cohorts"). We can see that there are over 20,000 college cohorts whose graduates make between $20,000 and $40,000 per year. There are a smaller number—about 4,000—making on average $10,000 to $20,000. For a very tiny number of college cohorts, the cohort average is between $80,000 and $160,000. There are so few between $160,000 and $320,000 that you can't even really see them on the graph.

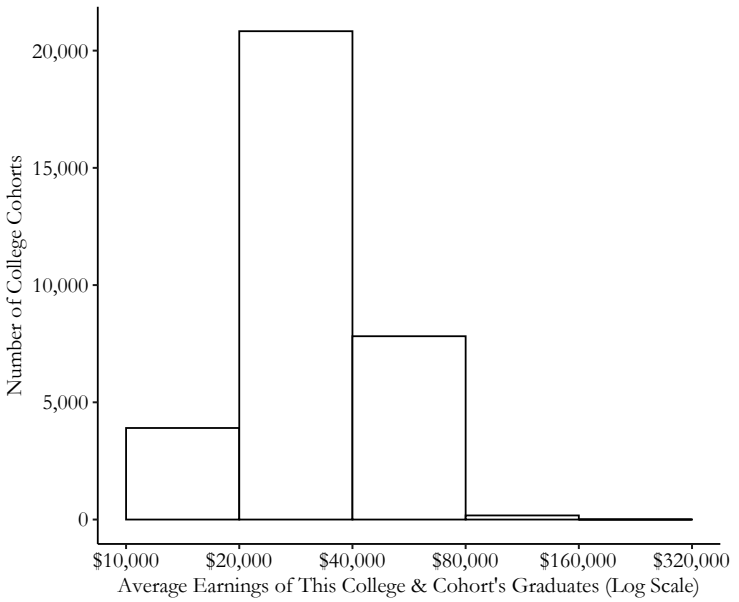

Figure 3.2: Distribution of Average Earnings across US College Cohorts

With a continuous variable we can go one step further than a histogram all the way to a *density*.[4] A density shows what would happen to a histogram if the bins got narrower and narrower, as you can see in Figure 3.3.[5]

When we have a density plot, we can describe the probability of being in a given range of the variable by seeing how large the area underneath the distribution is. For example, Figure 3.4 shows the distribution of earnings and has shaded the area between $40,000 and $50,000. That area, relative to the size of the area under the *entire* distribution curve, is the probability of being between $40,000 and $50,000. That particular shaded area makes up about 16% of the area underneath the curve, and so 16% of all cohorts have average earnings between $40,000 and $50,000.[6]

[4] Density distribution if you're graphing it, probability density if you're describing it.

[5] This only works because we have enough observations to fill in each bin. Otherwise it looks a lot choppier. Formally we need the bins to get narrower *and* the number of observations to get bigger.

[6] Calculus-heads will recognize all of this as taking integrals of the density curve. Did you know there's calculus hidden inside statistics? The things your professor won't tell you until it's too late to drop the class. We won't be doing calculus in this book, though.

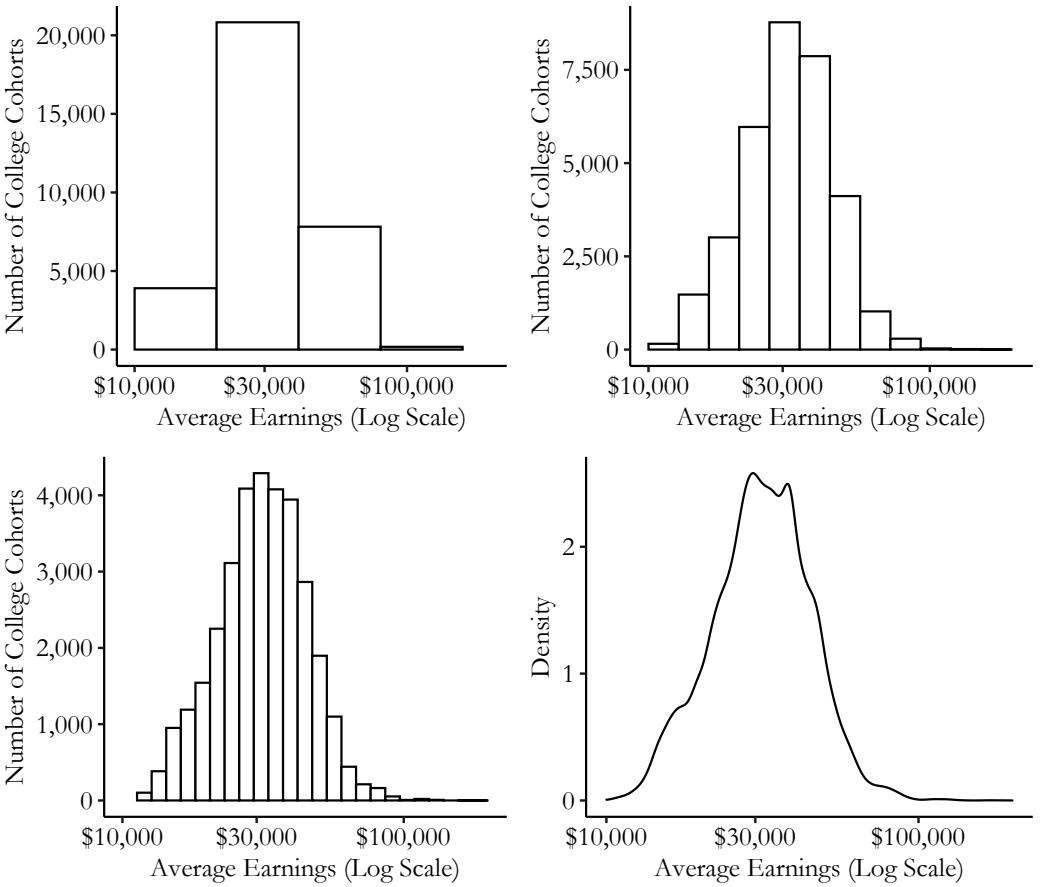

Figure 3.3: Distribution of Average Earnings across US College Cohorts

And that's it. Once you have the distribution of the variable, that's really all you can say about it.[7] After all, what do we have? For each possible value the variable *could* take, we know how likely that outcome, or at least an outcome like it, *is*. What else could you say about a variable?

Of course, in many cases these distributions are a little too detailed to show in full. Sure, for categorical variables with only a few categories we can easily show the full frequency table. But for any sort of continuous variable, even if we show someone the density plot, it's going to be difficult to take all that information in.

So, what can we do with the distribution to make it easier to understand? We pick a few key characteristics of it and tell you about them. In other words, we summarize it.

[7] Until you start to incorporate how it relates to *other* variables, as we'll talk about in Chapter 4.

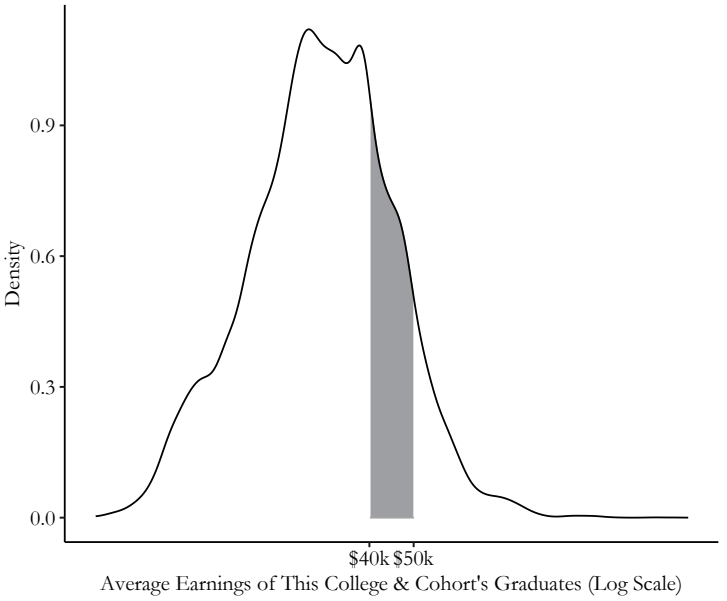

Figure 3.4: Shaded Distribution of Earnings across US College Cohorts

3.4 Summarizing the Distribution

Once we have the variable's distribution, we can turn our attention to *summarizing* that variable. The whole distribution might be a bit too much information for us to make any use of, especially for continuous variables. So our goal is to pick ways to take the *entire* distribution and produce a few numbers that describe that distribution pretty well.

Probably the most well-known example of a single number that tries to describe an entire distribution is the *mean*. The mean is what you get if you add up all the observations you have and then divide by the number of observations. So if you have 2, 5, 5, and 6, the mean is $(2 + 5 + 5 + 6)/4 = 18/4 = 4.5$.

A little more formally, what the mean does is it takes each value you might get, scales it by *how likely you are to get it*, and then adds it all up. And so on! What does the distribution look like for our data set of 2, 5, 5, and 6? Our frequency table is shown in Table 3.2.

Variable	N	Percent
Observed.Values	4	
... 2	1	25%
... 5	2	50%
... 6	1	25%

Table 3.2: Distribution of a Variable

Table 3.2 gives the distribution of our variable. Now we can calculate the mean. Again, this scales each value by how likely we are to get it. In 2, 5, 5, and 6, we get 5 half the time, so we count *half of five*. We get 2 a quarter of the time, so we count *a quarter of 2*.

Okay, so, since 2 only shows up 25% (1/4) of the time, we only count 1/4 of 2 to get .5. Next, 5 shows up 50% (1/2) of the time, so we count half of 5 and get 2.5. We see 6 shows up 25% (1/4) of the time as well, so we scale 6 by 1/4 and get 1.5. Add it all up to get our mean of $.5 + 2.5 + 1.5 = 4.5$.

So what the mean is *actually doing* is looking at the *distribution* of the variable and summarizing it, boiling it down to a single number. What is that number? The mean is supposed to represent a central tendency of the data—it's in the middle of what you might get. More specifically, it tries to produce a representative *value*. If the variable is "how many dollars this slot machine pays out" with a mean of $4.50, and it costs $4.50 to play, then if you played the slot machine a bunch of times you'd break even exactly.

SOMETIMES IT PAYS TO BE MORE DIRECT. We can certainly use the mean to describe a distribution, and in this book we will, many times. If the goal is to describe the distribution to someone, why bother doing a calculation of the mean when we could just tell people about the distribution itself?

The Xth percentile of a variable's distribution is the value for which X% of the observations are less. So for example, if you lined up 100 people by height, if the person in line with 5 people in front of them is 5 foot 4 inches tall, then 5 foot 4 inches tall is the 5th percentile.

Xth percentile. The value for which X percent of the variable's observations are smaller.

We can see percentiles on our distribution graphs. Figure 3.5 shows our distribution of college cohort earnings from before. We started shading in the left part of the distribution and kept going until we'd shaded in 5% of the area underneath the curve. The point on the x-axis where we stopped, $16,400, is the 5th percentile.

We can actually describe the entire distribution perfectly this way. What's the 1st percentile? Okay, now what's the 2nd? And so on.[8] Pretty soon we'll have mapped out the entire distribution by just shading in a little more each time. So percentiles are a fairly direct way of describing a distribution.

[8] Okay, fine, you can't actually get it *perfectly* unless you also do the 1.00001th percentile and the 1.00002nd and and the infinite percentiles between those two and so on. But you know what I mean.

There are a few percentiles that deserve special mention.

The first is the *median*, or the 50th percentile. This is the person right in the middle—half the sample is taller than them,

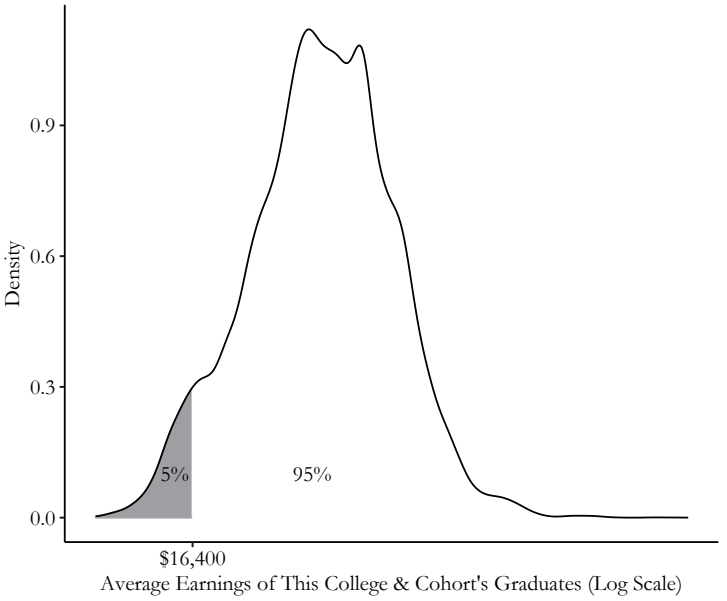

Figure 3.5: Distribution of Average Earnings across US College Cohorts

half the sample is shorter. Like the mean, the median is measuring a central tendency of the data. Instead of trying to produce a representative *value*, like the mean does, the median gives a representative *observation*.

For example, say you're looking at the wealth of 10,000 people, one of whom is Amazon founder Jeff Bezos. The mean says "hmm… sure, most people don't have much wealth, but once in a while you're Jeff Bezos and that makes up for it. The mean is very high." But the median says "Jeff Bezos isn't very representative of the rest of the people. He's going to count exactly the same as everyone else. The median is relatively low."[9]

For this reason, the median is generally used over the mean when you want to describe what a *typical observation* looks like, or when you have a variable that is highly skewed, with a few *really big* observations, like with wealth and Jeff Bezos. The mean wealth of that room might be $15,000,000, but that's almost all Jeff Bezos and doesn't really represent anyone else. As soon as he walks out of the room, the mean drops like a stone. The mean is very sensitive to Jeff! But the median might be closer to $90,000, a fairly typical net worth for an American family,[10] and it would stay pretty much exactly the same if Jeff left the room. The median is great for stuff like this![11]

The other two percentiles to focus on are the *minimum*, or the 0th percentile, and the *maximum*, or the 100th percentile. These are the lowest and highest values the variable takes. They're

[9] So why use the mean at all? It has a lot of nice properties too! For reasons too technical to go into here, it's usually much easier to work with mathematically and pops up in all sorts of statistical calculations beyond just describing a variable. Besides, sometimes you *do* want the representative value, not the representative observation. A place for everything.

[10] Yes, really. Start saving, kids.

[11] Sometimes the mean can work well with skewed data if it's transformed first. A common approach is to take the logarithm of the data, which reins in those really big observations. We'll talk about this more in the Theoretical Distributions section.

handy because they show you the kinds of values that the variable produces. The minimum and maximum height of a large group of people would tell you something about how tall or short humans can possibly be, for example.

Another nice thing about the minimum and maximum is that we can take the difference between them to get the *range*. The range is one way of seeing how much a variable *varies*. If the maximum and minimum are very far apart, as they would be for wealth with Jeff Bezos in the room, you know the variable can take a very wide range of values. If the maximum and minimum are close together, as they might be for "number of eyes you have," you know the range of values that a variable can take is fairly small.

Range. The difference between the maximum value of a variable and the minimum value.

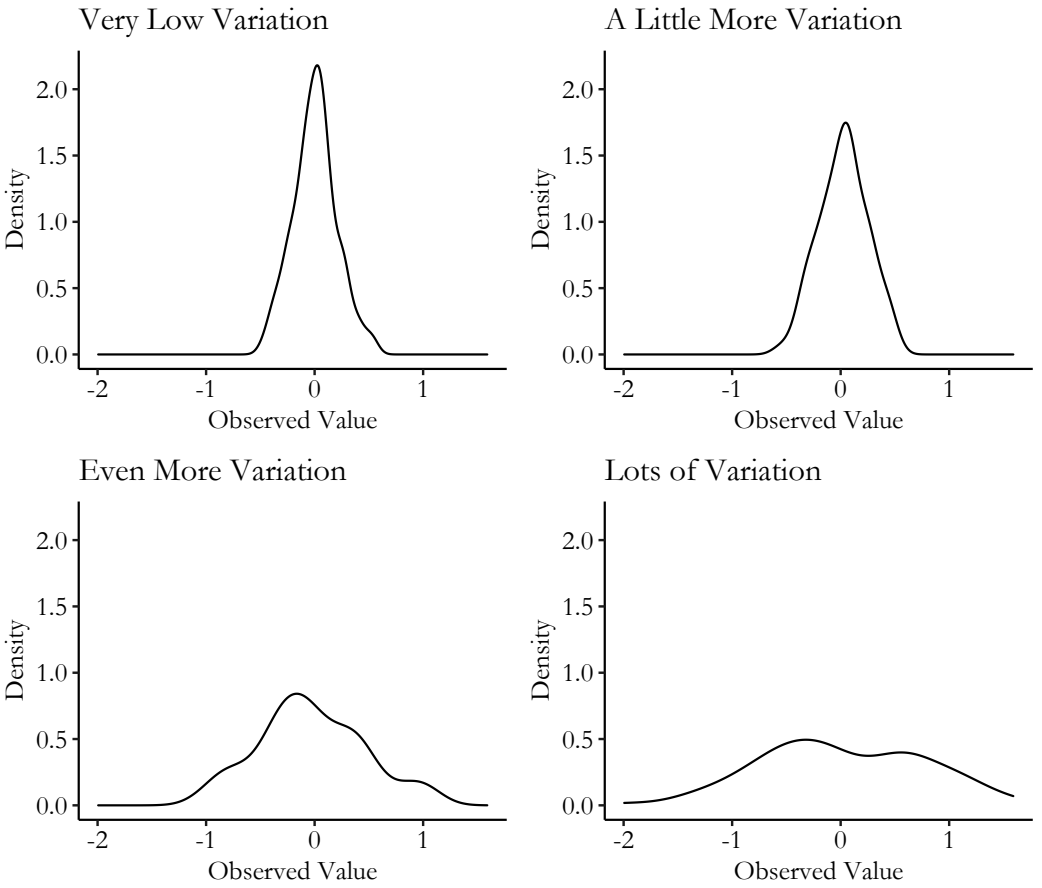

Figure 3.6: Four Variables with Different Levels of Variation

SOME VARIABLES VARY A LITTLE, OTHERS VARY A LOT. For example, take "the number of children a person has." For many people, this is zero. The mean, for people in their thirties

perhaps, is somewhere around 2. Some people have lots and lots of children, though. A few rare women have ten or more. There are men in the world with dozens of children. The number of children someone has can vary quite a bit.

Compare that to "the number of eyes a person has." For a small number of people, this might be 0, or 1, or maybe even 3. But the vast majority of people have two eyes. The number of eyes a person has varies fairly little.

These two variables—number of children and number of eyes—have similar means and medians, but they are clearly very different kinds of variables. We need ways to describe *variation* in addition to central tendencies or percentiles.

Variation. How a variable changes in value from observation to observation.

The way that variation shows up in a distribution graph is in how *wide* the distribution is. If the distribution is tall and skinny, then all of the observations are scrunched in very close to the mean. Low variation. If it's flat and wide, then there are a lot of observations in those fat "tails" on the left and right that are far away from the mean. High variation! See Figure 3.6 as an example. The distributions with more area "piled in the middle" have little variation—not a lot of area far from that middle point! The distributions with less "piled in the middle" and more in the "tails" on either side have plenty of observations far away from the middle.

There are quite a few ways to describe variation. Some of them, like the mean, focus on *values*, and others, like the median and percentiles, focus on *observations*.

Variation around the mean. The difference between the value of an observation and the mean.

VARIANCE IS A MEASURE OF VARIATION that focuses on values and is derived from the mean. To calculate the variance in a sample of observations of our data, we:

1. Find the mean. If our data is 2, 5, 5, 6, we get $(2 + 5 + 5 + 6)/4 = 18/4 = 4.5$.

2. Subtract the mean from each value. This turns our 2, 5, 5, 6 into -2.5, .5, .5, 1.5. This is our *variation around the mean*.

3. Square each of these values.[12] So now we have 6.25, .25, .25, and 2.25.

4. Add them up! We have $6.25 + .25 + .25 + 2.25 = 9$.

5. Divide by the number of observations minus 1.[13] So our sample variance is $9/(4-1) = 3$.

The bigger the variance is, the more variation there is in that variable. How does this work? Well, notice in steps 4 and 5 previously that we're sort of taking a mean. But the thing we're taking the mean *of* is squared variation around the actual mean. So any observations that are far from the mean get squared—making them even bigger and count for more in our mean! In this way we get a sense of how far from the mean our data is, on average.

One downside of the variance is that it's a little hard to interpret, since it is in "squared units." For example, the variance of the college cohort earnings variable is 153,287,962. 153,287,962... dollars squared? I'm not entirely sure what to make of that. So we often convert the variance into the *standard deviation* by taking the square root of it to get us back to our original units. The standard deviation of the college cohort earnings variable is $\sqrt{153,287,962} = 12,380.95$. And that 12,380.95 we can think of in dollars, like the original variable!

If we know that the mean is \$33,348.62, and we see a particular college cohort with average earnings of \$38,000, we know that that cohort is $(38,000 - 33,348.62)/12,380.95 = .376$, or *37.6% of a standard deviation above the mean.* This lets us know not just how much money that cohort earned (\$38,000), and how far above the mean they are (\$38,000 − \$33,348.62 = \$4,651.38), but how unusual that is relative to the amount of variation we typically see (37.6% of a standard deviation).

Figuring out how much variation one standard deviation is can be kind of tricky, and largely just takes practice and intuition. But a graph can help. Figure 3.7 shows how far to the left and right you have to go to find a one standard-deviation distance. It just so happens that 32.7% of the sample is under the curve between the "Mean − 1 SD" line and the "Mean" line, and another 35.5% of the sample is between the "Mean" line and the "Mean + 1 SD" line. In this case, more than 60% of the sample is closer than a single standard deviation.

[12] Why square? Why not something else? We could do something else! The kind of thing I'm calculating here for the variance is called a *moment* of the distribution. The mean is the first moment, and the variance is the second moment (this step squares), telling us about variation. The third moment (this step cubes) tells us about how lopsided the distribution is to one side or the other. The fourth moment (to the fourth power) tells us how much of the distribution is out in the "tails" (in the left and right edges). The third and fourth moments, after being "standardized," are called skewness and kurtosis.

[13] Why minus 1? Well, we estimated that mean already, and in different samples we might get slightly different calculations for the mean, in ways that are related to the variance. That introduces a little bias into the calculation. Specifically, if we divide by the number of observations n, we'll be off by $(n-1)/n$. So we divide by $1/(n-1)$ instead of $1/n$, in effect multiplying by $n/(n-1)$ and getting rid of that bias!

So how weird is being one standard deviation away from the mean? Well, roughly a third of people are between you and the average. Make of that what you will.

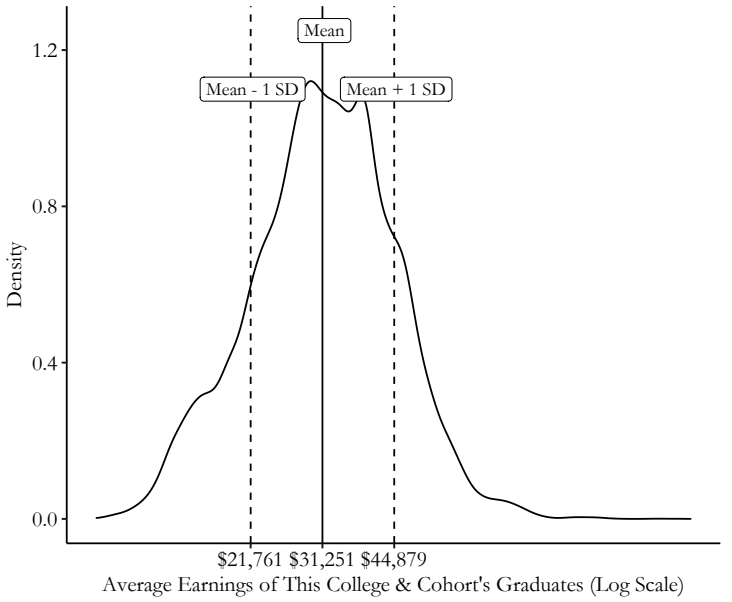

Figure 3.7: Distribution of Average Earnings across US College Cohorts

We can also compare percentiles to see how much a variable varies.

This is actually quite a straightforward process. All we have to do is pick a percentile above the median, and a percentile below the median, and see how different they are. That's it!

We've already discussed the range, which gives the distance between the biggest observation and the smallest. But the range can be very sensitive to really big observations—the range of wealth is very different depending on whether Jeff Bezos is in the room. So it's not a great measure.

Instead, the most common percentile-based measure of variation you'll tend to see is the *interquartile range*, or IQR.[14] This gives the difference between the 75th percentile and the 25th percentile. The IQR is handy for a few reasons. First, you know that the value given by the IQR covers exactly half of your sample. So for the half of your sample closest to the median, the IQR gives you a good sense of *how* close to the median they are. Second, unlike the variance, the IQR isn't very strongly affected by big tail observations. So, as always, it's a good way of representing observations rather than values.

[14] Another one you might see is the "90/10" ratio, or the 90th percentile divided by the 10th percentile. This is a measure commonly used in studies of inequality, to give a sense of *just how different* the top and bottom of the distribution are.

Figure 3.8 shows where the IQR comes from on a distribution. In this case, the 25th and 75th percentiles are at \$21,761 and \$44,879, respectively, giving us an IQR of $44,879 - 21,761 = 23,118$. So the 50% of the cohorts nearest the median have a range of average incomes of \$23,118.

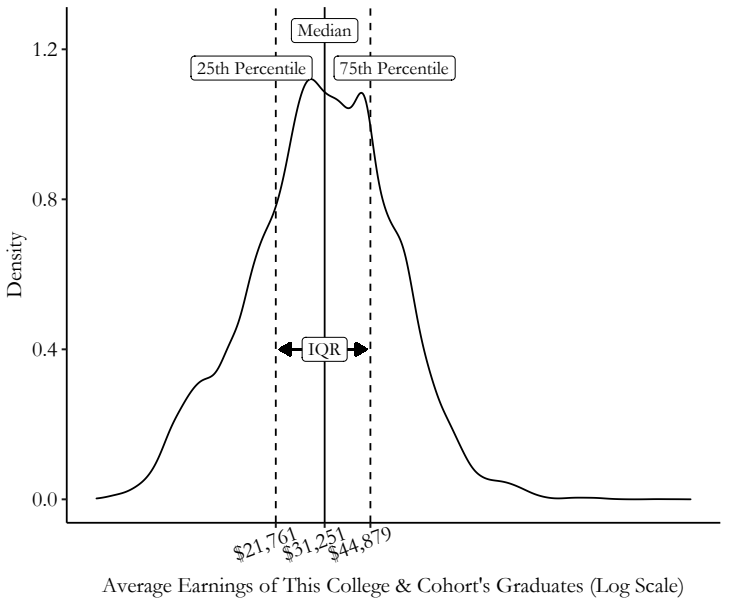

Figure 3.8: Distribution of Average Earnings across US College Cohorts

Beyond the variation there are of course a million other things we could describe about a distribution. I will cover only one of them here, and that's the *skew*.

Skew describes how the distribution *leans* to one side or the other. For example, let's talk about annual income. Most people have an income in a relatively narrow range—somewhere between \$0 and, say, \$150,000. But there are some people—and a fair number of them, actually, who have *enormous* incomes, *way* bigger than \$150,000, perhaps in the millions or tens or hundreds of millions.

So for annual income, the *right tail*—the part of the distribution on the right edge—is very big. Figure 3.9 shows what I'm talking about. Most of the weight is down near 0, but there are people with millions of dollars in income making the right tail of the distribution stretch way far out. The same isn't true on the left side—at least in this data, we're not seeing people with negative incomes.

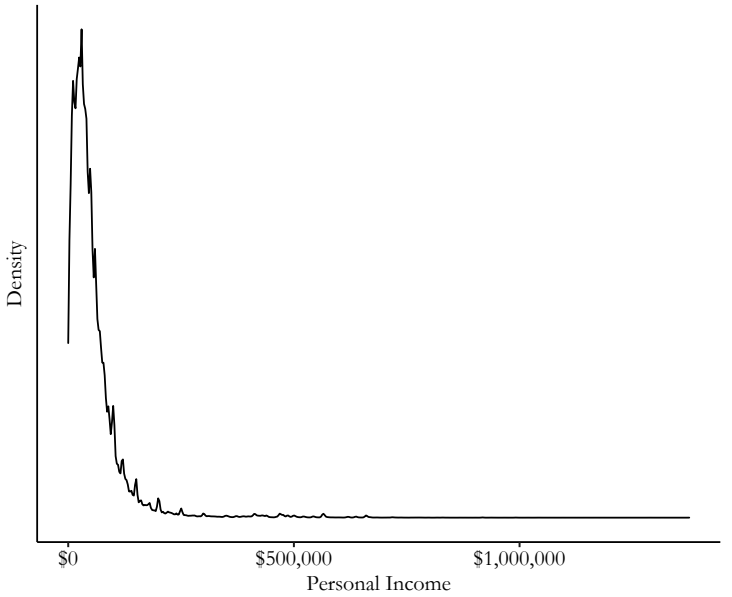

Figure 3.9: Distribution of Personal Income in 2018 American Community Survey

We say that distributions like this one, with a heavy right tail but no big left tail to speak of, has a "right skew" since it has a lot of right-tailed observations. Similarly, a distribution with lots of observations in the left tail would have a left skew. A distribution with similar tails on both sides is *symmetric*.

Right-skewed variables pop up all the time in social science. Basically anything that's unequally distributed, like income, will have a lot of people with relatively little, and a few people with a lot, and the people with a lot have *a lot*.

Skew can be an important feature of a distribution to describe. It can also give us problems if we're working with means and variances, since those really-huge values will affect any measure that tries to represent values.

One way of handling skew in data is by *transforming* the data. If we apply some function to the data that shrinks the impact of those really-big observations, the mean and variance work better. A common transformation in this case is the *log* transformation, where we take the natural logarithm of the data and use that instead. This can make the data much better-behaved.[15]

Figure 3.10 shows that once we take the log of income, there's still a bit of a tail remaining (on the left this time!), but in general we have a roughly symmetric distribution that our mean will work a lot better with.[16]

[15] The "natural" logarithm uses a logarithmic base of e. This is so common in statistics that if you just see "log" without any detail on what the base is, you can assume it's base-e.

[16] How about downsides of a log transformation? It can't handle negative values, or even 0 values, since log(0) is undefined. There are in fact a lot of 0 incomes in this data that aren't graphed. If we wanted to include them, we couldn't use a log transformation. If you have negative values things get a lot trickier, but if it's just 0s worrying you (and often it is), you might want to try the inverse hyperbolic sine transform (asinh), as that's similar to log for large values but sets 0s to 0.

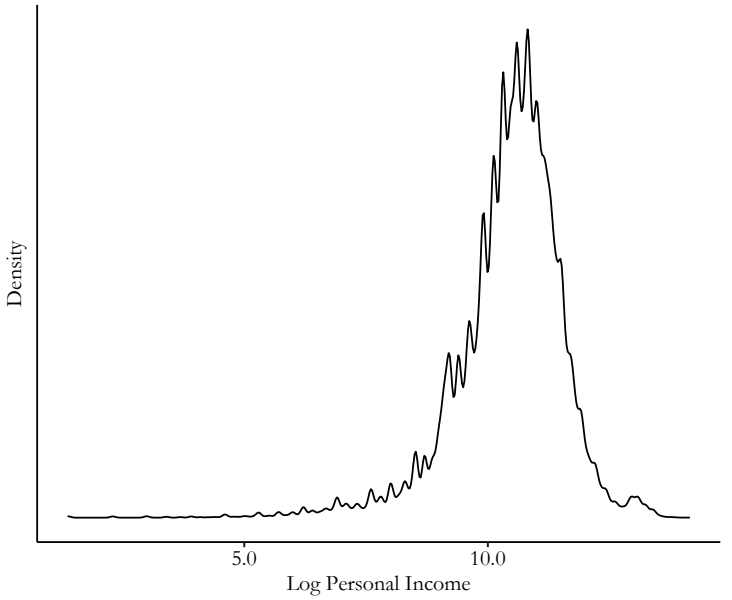

Figure 3.10: Distribution of Logged Personal Income in 2018 American Community Survey

One reason the natural log transformation is so popular is that the transformed variable has an easy interpretation. A log increase of, let's say, .01, translates to *approximately* a .01 × 100 = 1% increase in the original variable. So an increase in log income from 10.00 to 10.02 in log terms means a (10.02 − 10.00) = .02 ≈ 2% increase in income itself. This approximation works pretty well for small increases like .01 or .02, but it starts to break down for bigger increases like, say, .2. Anything above .1/10% or so and you should avoid the approximation.[17]

There might be trouble brewing if you take the log and it *still* looks skewed. This can be the case when you have *fat tails*, i.e., observations far away from the mean are very common. When you have fat tails on one side but not the other, this can make your data very difficult to work with indeed. I'll cover this a little bit all the way at the end of the book in Chapter 22.

[17] For bigger increases, an increase of p in the log is actually equivalent to a $(e^p - 1) \times 100\%$ increase in the variable. Or, a $X\%$ increase in the variable is actually equivalent to a $\log(1 + X)$ log increase. The approximation works because $e^p - 1$ and p are very close together for small values of p.

3.5 Theoretical Distributions

The difference between reality and the truth is that reality is always with you, but no matter how far you walk, truth is still on the horizon. So it's much more convenient to squint, shrug, go "eh, close enough," and head home.

Statistics makes a very clear distinction between the *truth* and the data we've collected. But isn't the data truly what

we've collected? Well, sure, but what it is supposed to *represent* is some broader truth beyond that.

Let's say you want to understand the average age at which children learn to share toys. So you interview 1,000 parents about when their kids started doing that. You calculate a mean and get that kids in your sample start to share easily around 4.2 years old.

Of course, what you *actually* have is that *the 1,000 kids in your sample* started to share easily around 4.2 years old. And you didn't set out to learn something about those 1,000 kids, right? You set out to learn something about kids in general! So the *true* average age at which kids in general start to share is one thing, and the average age you calculated in your data is another.

That's the whole point of doing statistics. We can never check every kid who ever existed on the age they started sharing. So given the real data we actually have, *what can we say about that true number?*

Figuring this out will require us to think about how data behaves under different versions of the truth. If the truth is that kids learn to share on average at 3.8 years of age, what kind of data does that generate? If the truth is that kids learn to share at 5 years of age, what kind of data does that generate? We'll need to pair our observed distributions, the ones we've been talking about so far in this chapter, with *theoretical distributions* of how data behaves under different versions of the truth.

Theoretical distribution. A distribution based on theoretical assumptions about the truth, rather than derived purely from data.

SOME QUICK NOTATION before we get much further.

If you've read any sort of statistics before, you may be familiar with symbols like β, μ, $\hat{\mu}$, $\bar{x}$. You may have memorized what means what. But it turns out you probably don't have to, as there's an order to all this madness. What do these all mean?

English/Latin letters represent *data.* So x might be a variable of actual observed data. That's our 1,000 surveys with parents about their kids' sharing ages.

Modifications of English/Latin letters represent *calculations* done with real data. A common way to indicate "mean" is to use a bar on top of the letter. So $\bar{x}$ is the mean of x we calculated in our data. That's the 4.2 we calculated from our survey.

Greek letters represent *the truth.*[18,19] We don't know what actual values these take, but we can make assumptions. Certain Greek letters are commonly used for certain kinds of truth—μ commonly indicates some sort of mean,[20] σ the standard devi-

[18] So we have a system where English is the down-and-dirty real world and Greek is the lofty and perfect truth. Who designed this?

[19] It's worth pointing out that a Greek letter in a *bad model* might not necessarily be "true." But it is meant to *imply* the truth, even if it doesn't actually get there.

[20] Technically, it represents "expected values" more often than means, but we won't be going much into that in this book.

ation, ρ for correlation, β for regression coefficients, ε for "error terms" (we'll get there), and so on. But the important thing is that Greek letters represent the truth.

Modifications of Greek letters represent *our estimate of the truth.* We don't know what the truth is, but we can make our best guess of it. That guess may be good, or bad, or completely misguided, but it's our guess nonetheless. The most common way to represent "my guess" is to put a "hat" on top of the Greek letter. So $\hat{\mu}$ is "my estimate of what I think μ is." If the way that I plan to estimate μ is by taking the mean of x, then I would say $\hat{\mu} = \bar{x}$.

THE THEORETICAL DISTRIBUTION IS WHAT GENERATED YOUR DATA. That's actually a good way to think about theoretical distributions. They're the distribution of *all* the data, even the data you didn't actually collect, and maybe could never actually collect! If you could collect literally an infinite number of observations, their distribution would be the theoretical distribution.[21]

[21] In statistics this is known as the "limiting distribution" because in the *limit* (i.e., as the number of observations approaches infinity) it's what you get. (Psst, that's another calculus thing. The calculus is coming for you!)

This fact tells us a few things. First, it tells us why we're interested in the theoretical distribution in the first place. Because that's where we get our data from! If we want to learn about the average age children share at, the place *that data comes from* is the theoretical distribution. So if we want to know the value of that number beyond the data we actually have, we have to use that data to claw our way back to the theoretical distribution. Only then will we know something really interesting!

Remember, we don't really care about the mean in our observed data, $\bar{x}$. We care about the *true average for everyone, μ!* The reason we bother gathering data in the first place is because it will let us make an *estimate* $\hat{\mu}$ about what the theoretical distribution it came from is like.

The second thing this "infinite observations" fact tells us is that the more observations we have, the better a job our observed data will do at matching that theoretical distribution. One observation isn't likely to do us much. But an infinite number would get the theoretical distribution exactly! Somewhere in the middle is going to have to be good enough. And the bigger our number of observations gets, the gooder-enougher we become.

This can be seen in Figure 3.11. No matter how many observations we have, the solid-line theoretical distribution always stays the same, of course. But while we do a pretty bad job at describing that distribution with only ten observations, by the

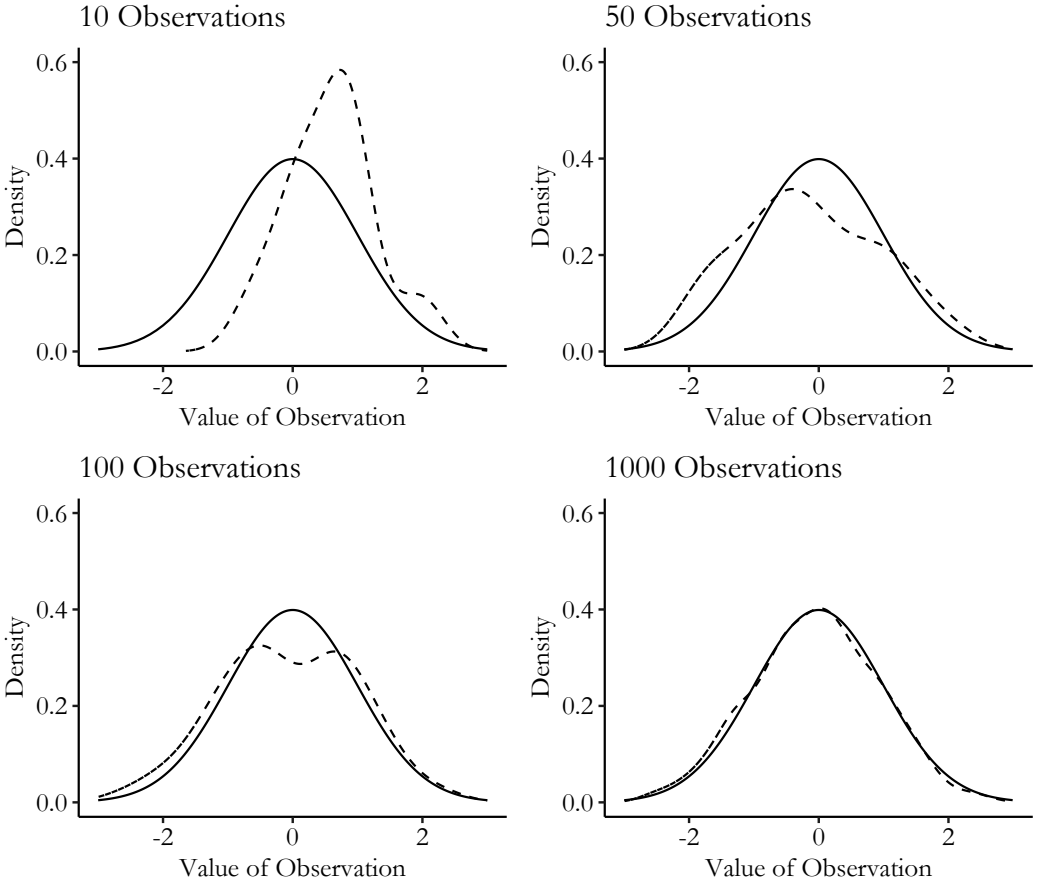

Figure 3.11: Trying to Match the Theoretical Distribution

time we're up to 100 we're doing a lot better. And by 1,000 we've got it pretty good! That's not to say that 1,000 is always "big enough to be just like the theoretical distribution." But here it worked pretty well.

This means as we get more and more observations, we're going to do a better and better job of getting an observed distribution that matches the theoretical one that we sampled the data from. Since that's the distribution we're interested in, that's a good thing! We just need to make sure to have plenty of observations.[22]

THERE ARE INFINITE DIFFERENT THEORETICAL DISTRIBUTIONS, but some pop up in applied work often. There are some well-known distributions that are applied over and over again. If we think that our data follows one of these distributions we're in luck, because it means we can use that theoretical distribution to do a lot of work for us!

[22] Although to be clear, the theoretical distribution we'll get is *the one our data came from.* This may not be the one we're actually interested in! Say our children-sharing data came only from children who go to daycare. Then, the distribution we'll get as we get more and more observations is the distribution *of daycare-going children.* If we're interested in *all* children, we won't get the result for all children no matter how many daycare-going children we survey.

I will cover only two that are especially important to know about in applied social science work, which are both depicted in Figure 3.12. There are many, many more I am leaving out: uniform distribution, Poisson, binomial, gamma, beta, and so on and so on. If you are interested, you may want to check out a more purely statistics-oriented book, like any of the eight zillion books I find when I search "Introduction to Statistics" on Amazon. I'm sure they're all nearly as good as my book. Nearly.

The first to cover is the *normal* distribution. The normal distribution is *symmetric* (i.e., the left and right tails are the same size, there's no skew/lean). The normal distribution often shows up when describing things that face real, physical restrictions, like height or intelligence.

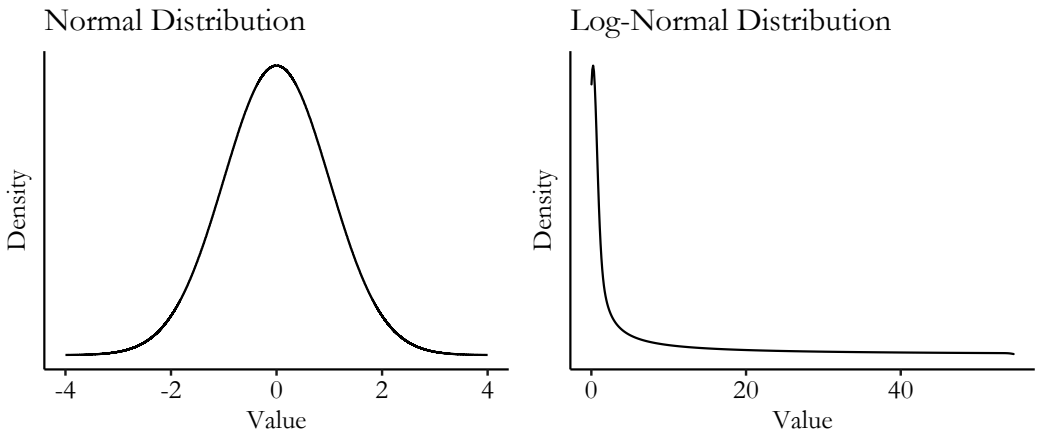

Figure 3.12: Normal and Log-Normal Distributions

The normal distribution also pops up a lot when looking at aggregated values.[23] Income might have a strong right skew, but if we take the *mean* of income in one sample of data, aggregating across observations, and then again in another sample of data, aggregating across different observations, and then again and again in different samples, *the distribution of the mean across different samples* would have a normal distribution.

The normal distribution technically has *infinite range*, meaning that every value is possible, even if unlikely. This means that any variable for which some values are impossible (like how height can't be negative) technically can't be normal. But if the approximation is very good, we tend to let that slide. One reason we let that slide is that the normal distribution has fairly *thin tails*—observations far from the mean are extremely

[23] This is due to something called the "central limit theorem." Really, read one of those eight zillion Introduction to Statistics textbooks I talked about! Interesting stuff.

unlikely. Notice how quickly the distribution goes to basically 0 in Figure 3.12. So sure, maybe saying that height follows a normal distribution means you're technically saying that negative heights are possible. But you're saying it's possible with a .00000001% chance, so that's close enough to count.

The second is a bit of a cheat, and it's the log-normal distribution. The log-normal distribution has a heavy right skew, but once we take the logarithm of it, it turns out to be a normal distribution! How handy.

The log-normal is a very convenient version of a skewed distribution, since we can take all the skew out of it by just applying a logarithm. Heavily skewed data comes up all the time in the real world. Anything that's unequally distributed and that doesn't have a maximum possible value is generally skewed (income, wealth…) as well as many things that tend to be "winner take all" or have some super-big hits (number of song downloads, number of hours logged in a certain video game, company sizes…). Notice how much of the weight is scrunched over to the left to make room for a very tiny number of really huge observations on the right. That's skew for you!

When we see a skewed distribution, we tend to hope it's log-normal for convenience reasons. However, there are of course many other skewed distributions out there. Skewed distributions with fat tails can be difficult to work with and can take specialized tools. So it's a good idea, after taking the log of a skewed variable, to look at its distribution to confirm that it does indeed look normal.[24] If it doesn't you may be wading into deep waters!

How can we use our empirical data to learn about the theoretical distribution?

Remember, our real reason for looking at and describing our variables is because we want to get a better idea of the theoretical distribution. We're not really interested in the values of the variable in our sample, we're interested in *using* our sample to find out about how the variable behaves in general.

We can, if we like, take our sample and look at its distribution (as well as its mean, standard deviation, and so on), figure that's the best guess we have as to what the theoretical distribution looks like, and go from there.[25] Of course, we know that's imperfect. Would we really believe that the distribution we happened to get in some data really represents the true theoretical distribution?

[24] One common family of skewed distributions that are difficult to work with are "power distributions," which pop up when you have data where big values are fairly common, like in the stock market, where days with huge upswings and downswings happen with some regularity. These can be tricky to work with—many theoretical power distributions don't even have well-defined means or variances.

[25] Or, if we're using Bayesian statistics, we can take our observed distribution and combine it with our best guess for the theoretical distribution before we collected our data.

One thing that is a bit easier to do is to learn *whether certain theoretical distributions are unlikely.* Maybe we can't figure out exactly what the theoretical distribution *is*, but maybe we can rule some stuff out.

How can we figure out how likely a certain theoretical distribution is? We follow these steps:

1. Choose some description of the theoretical distribution—its mean, its median, its standard deviation, etc. Let's use the mean as an example.

2. Use the properties of the theoretical distribution and your sample size to find the theoretical distribution *of that description in random samples*—means generally follow a normal distribution, and the standard deviation of that normal distribution is smaller for bigger sample sizes.

3. Make that same description of your observed data—so now we have the distribution of our theoretical mean, and we have the actual observed mean.

4. Use the theoretical distribution of that description to find out how unlikely it would be to get the data you got—if the theoretical distribution of the mean we're looking at has mean 1 and standard deviation 2, and our observed mean is 1.5, we're asking "how likely is it that we'd get a 1.5 or more from a normal distribution with mean 1 and standard deviation 2?"

5. If it's really unlikely, then you probably started with the wrong theoretical distribution, and can rule it out. If we're doing statistical significance testing, we might say that our observed mean is "statistically significantly different from" the mean of the theoretical distribution we started with.

Let's walk through an example.

Say we're interested in how many points basketball players make in each game. You collect data on 100 basketball players. Your observed data doesn't look particularly well-behaved and doesn't look like any sort of theoretical distribution you've heard of before. But you calculate its mean and standard deviation and find a mean of 102 with a standard deviation of 30.

Following step 1, you ask "could this have come from a distribution with a mean μ of 90?"—notice we haven't said anything here about it being from a normal distribution, or log-normal,

or anything else. We are going to try to rule out distributions with means of 90, that's all. Also notice we called the mean μ, a Greek letter (the truth!). We want to know if we can rule out that $\mu = 90$ is the truth.

Then, for step 2, we want to get the distribution of that description. As mentioned earlier in this chapter, means are generally distributed normally, centered around the theoretical mean itself. The standard deviation of the mean's distribution is just $\sigma/\sqrt{N}$, or the standard deviation of the overall distribution divided by $\sqrt{N}$, where N is the number of observations in the sample.

What's going on here? Well, what this is saying is: if you survey N basketball players and take the mean, and then survey another N basketball players, and then another N, and then another N, and so on, you'll get a different mean each time. If we take the mean from each sample as its own variable, the distribution of that variable will be normal, with a mean of the true theoretical mean, and a standard deviation of $\sigma/\sqrt{N}$. The $/\sqrt{N}$ is because the more players we survey each time, the more likely it is that we'll get very very close to the true mean. A mean of 10 basketball players could give you a result very far from the mean. But a mean of 1,000 basketball players is pretty likely to give you a mean close to the theoretical mean, and thus the smaller standard deviation.

σ is a Greek letter, and so of course that's the true standard deviation, which we don't know. But our best estimate of it is the standard deviation of 30 we got in our data. So we say that the mean is distributed normally with a mean of $\mu = 90$, and a standard deviation of $\hat{\sigma}/\sqrt{N} = 30/\sqrt{100} = 3$ (remember, $\hat{\sigma}$ means "our estimate of the true σ," which is just the standard deviation we got in our observed data).

Now for step 3, we get the same calculation in our observed data. As above, the mean in the observed data is 102.

Moving on to step 4, we can ask "how likely is it to get a 102, or even more, from a normal distribution with mean 90 and standard deviation 3?"[26] More precisely, we are generally interested in how likely it is to get a 102 *or something even farther away*, which could be more or less than μ.[27] So 102 is 12 away from 90, and thus we're interested in how likely it is to get a mean of 102 or more, or 78 or less (since 78 is also 12 away from 90).

By looking at the percentiles of the normal-with-mean-90-and-standard-deviation-3 distribution, we can determine that 78

[26] 3 and not 30? Remember, this is the standard deviation of the sampling distribution of the mean, not the standard deviation of the variable itself.

[27] This is a "two-sided test." A "one-sided test" would just ask how likely it would be to get 102 *or more*.

is not even the 1st percentile, it's more like the .004th percentile. So there's only a .004% chance of getting a mean of 78 or less in a 100-player sample if the true mean is 90 and the standard deviation is 30. Similarly, 102 is the 99.996th percentile, so again there's only a .004% chance of getting a mean of 102 or more in a 100-player sample if the true mean is 90 and the standard deviation is 30.

For step 5, we add those up and say that there's only a .008% chance of getting something as far off as 102 or more if the true mean is 90 and the true standard deviation is 30. That's pretty darn unlikely. So this data very likely did not come from a distribution with a mean of 90.

We could also frame all of this in terms of *hypothesis testing*. Following the same steps, we can say:

Step 1: Our *null hypothesis* is that the mean is 90. Our *alternative hypothesis* is that the mean is not 90.

Step 2: Pick a test statistic with a known distribution. Means are distributed normally, so we might use a Z-statistic, which is for describing points on normal distributions.

Step 3: Get that same test statistic in the data.

Step 4: Using the known distribution of the test statistic, calculate how likely it is to get your data's test statistic, or something even more extreme (p-value).

Step 5: Determine whether we can reject the null hypothesis. This comes down to your threshold (α)—how unlikely does your data's test statistic need to be for you to reject the null? A common number is 5%.[28] If that's your threshold, then if Step 4 gave you a lower p-value than your α threshold, then that's too unlikely for you, and you can reject the null hypothesis that the mean is 90.

[28] The choice of 5% is completely arbitrary and yet widely applied. Like your keyboard's QWERTY layout.

This section describes what hypothesis testing actually is and what it's for. Statistical significance is not a marker of being *correct* or *important*. It's just a marker of being able to *reject some theoretical distribution you've chosen.* That can certainly be interesting. At the very least, we've narrowed down the likely list of ways that our data could have been generated. And that's the whole point!

4
Describing Relationships

4.1 What Is a Relationship?

For most research questions, we are not just interested in the distribution of a single variable.[1] Instead, we are interested in the *relationship* we see in the data between two or more variables.

What does it mean for two variables to have a relationship? The relationship between two variables shows you *what learning about one variable tells you about the other.*

For example, take height and age among children. Generally, the older a child is, the taller they are. So, learning that one child is thirteen and another is six will give you a pretty good guess as to which of the two children is taller.

We can call the relationship between height and age *positive*, meaning that for higher values of one of the variables, we expect to see higher values of the other, too (more age is associated with more height). There are also negative relationships, where higher values of one tend to go along with lower values of the other (more age is associated with less crying). There are also null relationships where the variables have nothing to

[1] Get lost, Chapter 3, nobody likes you.

DOI: 10.1201/9781003226055-4

do with each other (older children aren't any more or less likely to live in France than younger children). All kinds of other relationships are positive sometimes and negative other times, or *really* positive at first and then only slightly positive later. Or perhaps one of the variables is categorical and there's not really a "higher" or "lower," just "different" (older children are more likely to use a bike for transportation than younger children). Lots of options here.

THE GOAL IN THIS CHAPTER is to figure out how to describe the relationship between two variables, so that we can accurately relay what we see in the data about our research question, which, once again, very likely has to do with the relationship between two variables. Once we know how to describe the relationship we see *in the data*, we can work in the rest of the book to make sure that the relationship we've described does indeed answer our research question.

Throughout this chapter, we're going to use some example data from a study by Emily Oster,[2] who used the National Health and Nutrition Examination Survey. Her research question was: do the health benefits of recommended medications look better than they actually are because already-otherwise-healthy people are more likely to follow the recommendations?

To study this question, she looked at vitamin E supplements, which were only recommended for a brief period of time. She then answers her research question by examining the relationship between taking vitamin E, other indicators of caring about your health like not smoking, and outcomes like mortality, and how those relationships change before, during, and after the time vitamin E was recommended.[3]

We can start off with an example of a very straightforward way of showing the relationship between two continuous variables, which is a scatterplot, as shown in Figure 4.1. Scatterplots simply show you every data point there is to see. They can be handy for getting a good look at the data and trying to visualize from them what kind of relationship the two variables have. Does the data tend to slope up? Does it slope down a lot? Or slope down just a little like in Figure 4.1? Or go up and down?

A scatterplot is a basic way to show *all* the information about a relationship between two continuous variables, like the density plots were for a single continuous variable in Chapter 3.[4] And they're usually a great place to start describing a relationship.

[2] Emily Oster. Health recommendations and selection in health behaviors. *American Economic Review: Insights*, 2(2):143–60, 2020b.

Scatterplot. A graph that plots every data point for an x-axis variable and a y-axis variable.

[3] In this chapter, I'll add some analyses that weren't exactly in the original study but are in the same spirit, wherever it helps explain how to describe relationships. It's almost like she had other purposes for her study besides providing good examples for my textbook. Rude if you ask me.

[4] Unlike density plots, though, they tend to get very hard to read if you have a lot of data. That's why I only used 150 observations for that graph, not all of them.

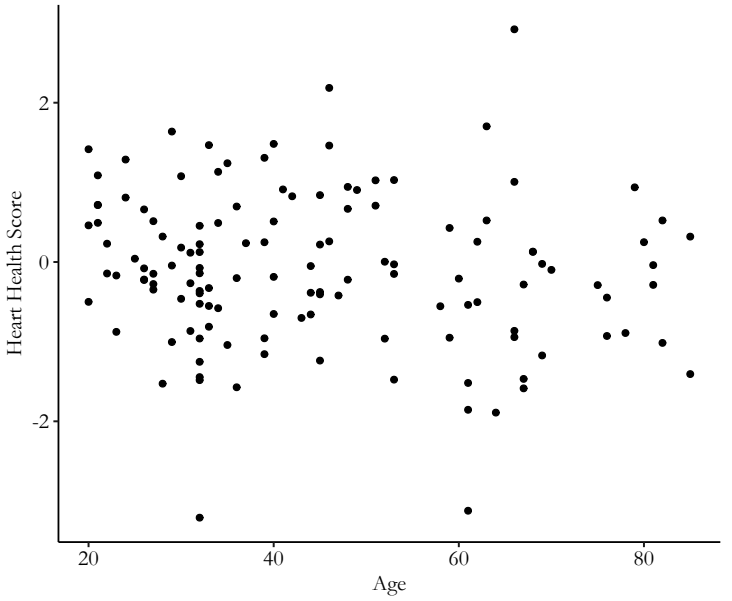

Figure 4.1: Age and Heart Health, 150 Observations

Scatterplots imply two things beyond what they actually show. One is bad, and one is good. The bad one is that it's very tempting to look at a relationship in a scatterplot and assume that it means that the x-axis causes the y-axis. Even if we know that's not true, it's very tempting. The good one is that it encourages us to use the scatterplot to imagine other ways of describing the relationship that might give us the information we want in a more digestible way. That's what the rest of this chapter is about.

4.2 Conditional Distributions

CHAPTER 3 WAS ALL ABOUT DESCRIBING the distributions of variables. However, the distributions in those chapters were what are called *unconditional* distributions.[5]

A *conditional* distribution is the distribution of one variable *given the value of another variable.*

Let's start with a more basic version—conditional probability. The probability that someone is a woman is roughly 50%. But the probability that someone *who is named Sarah* is a woman is much higher than 50%. You can also say "*among all Sarahs*, what proportion are women?" We would say that this is the "probability that someone is a woman conditional on being named Sarah."

[5] These are also called "marginal" distributions, but I really dislike this term, as I think it sounds like the opposite of what it means.

Conditional distribution. The distribution of a variable conditional on another variable taking a certain value.

Learning that someone is named Sarah changes the probability that we can place on them being a woman. Conditional distributions work the same way, except that this time, instead of just a single probability changing, an entire distribution changes.

Take Figure 4.2 for example. In this graph, we look at the distribution of how much vitamin E someone takes, among people who take any. We then split out the distribution by whether someone has engaged in vigorous exercise in the last month.

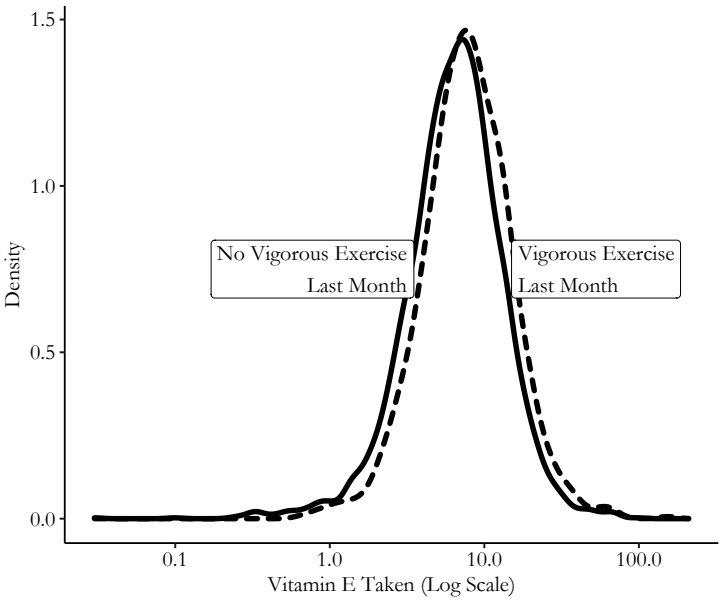

Figure 4.2: Distribution of Amount of vitamin E Taken by Exercise Level

We can see a small deviation in the distribution for those who exercise and those who don't.[6] In particular, those who exercise vigorously take larger doses of vitamin E when they take it. The distribution is different between exercisers and non-exercisers, telling us that vitamin E and exercise are *related* to each other in this data.

THE EXAMPLE I'VE GIVEN is for a continuous variable, but it works just as well for a categorical variable. Instead of looking at how large the doses are, let's look at whether someone takes vitamin E at all! Oster's hypothesis is that people who take vitamin E at all should be more likely to do other healthy things like exercise, because both are driven by how health-conscious you are.

Figure 4.3 shows an example of this. The distribution of whether you take vitamin E or not is shown twice here, once for those who currently smoke, and one for those who don't smoke.

[6] It doesn't look enormous, but this is actually how a lot of fairly prominent differences look in the social sciences. That rightward shift can be deceptively larger than it looks!

The distributions are clearly different, with a higher proportion taking vitamin E in the non-smoking crowd, exactly what Oster would expect.

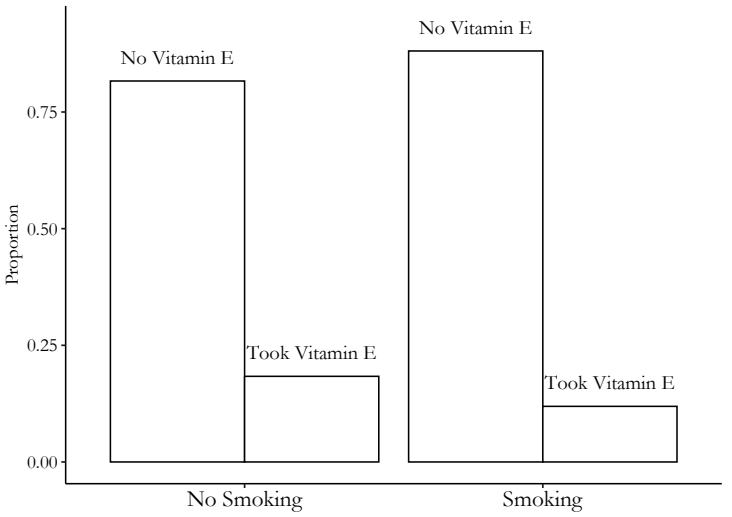

Figure 4.3: Distribution of Whether vitamin E is Taken by Whether you Smoke

4.3 Conditional Means

With the concept of a conditional distribution under our belt, it should be clear that we can then calculate *any* feature of that distribution conditional on the value of another variable. What's the 95th percentile of vitamin E taking overall and for smokers? What's the median? What's the standard deviation of mortality for people who take 90th-percentile levels of vitamin E, and for people who take 10th-percentile levels?

While all those possibilities remain floating in the air, we will focus on the conditional mean. Given a certain value of X, what do I expect the mean of Y to be?[7]

Conditional mean. The mean of one variable given that another variable takes a certain value.

Once we have the conditional mean, we can describe the relationship between the two variables fairly well. If the mean of Y is higher conditional on a higher value of X, then Y and X are positively related. Going further, we can map out all the conditional means of Y for each value of X, giving us the full picture on how the mean of one variable is related to the values of the other.

In some cases, this is easy to calculate. If the variable you are conditioning on is discrete (or categorical), you can just calculate the mean for all observations with that value.

[7] Why the mean? One reason is that the mean behaves a bit better in small samples, and once we start looking at things separately by specific values of X, samples get small. Another reason is that it helps us weight prediction errors and and so figure out how to minimize those errors. It's just handy.

See Figure 4.4, for example, which shows the proportion taking vitamin E conditional on whether the observations are from before vitamin E was recommended, during recommendation, or after.[8] I just took all the observations in the data from before the recommendation and calculated the proportion who took vitamin E. Then I did the same for the data during the recommendation, and after the recommendation.

[8] "Proportion" is the mean of a binary variable.

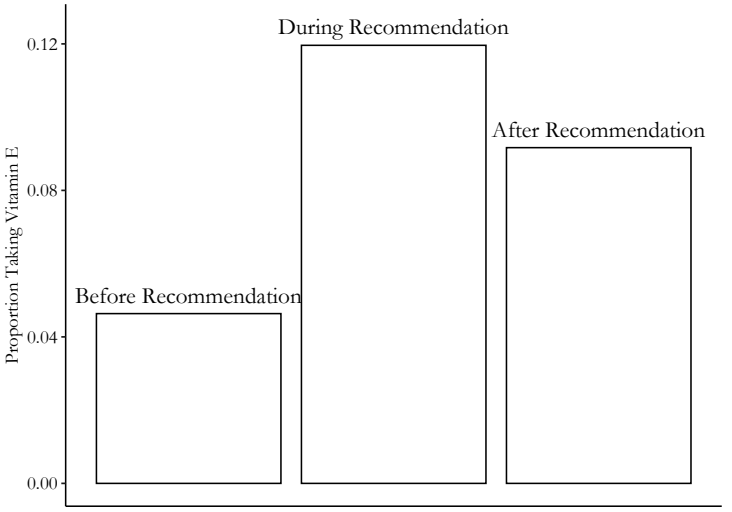

Figure 4.4: Proportion Taking Vitamin E Before It Was Recommended, During, and After

Figure 4.4 shows the relationship between the taking of vitamin E and the timing of the recommendation. We can see that the relationship between the taking of vitamin E and the recommendation being in place is positive (the proportion taking vitamin E is higher during the recommendation time). We also see that the relationship between vitamin E and *time* is at first positive (increasing as the recommendation goes into effect) and then negative (decreasing as the recommendation is removed).

THINGS GET A LITTLE MORE COMPLEX when you are conditioning on a continuous variable. After all, I can't give you the proportion taking vitamin E among those making $84,325 per year because there's unlikely to be more than one person with that exact number. For lots of numbers we'd have no data at all.

There are two approaches we can take here. One approach is to use a *range* of values for the variable we're conditioning on rather than a single value. Another is to use some sort of shape or line to fill in those gaps with no observations.

Let's focus first on using a range of values. Table 4.1 shows the proportion of people taking vitamin E conditional on body

mass index (BMI). Since BMI is continuous, I've cut it up into ten equally-sized ranges (bins) and calculated the proportion taking vitamin E within each of those ranges. Cutting the data into bins to take a conditional mean isn't actually done that often in real research, but it gives a good intuitive sense of what we're trying to do when we use other methods later.

BMI Bin	Proportion Taking Vitamin E
(11.6,20.6]	0.133
(20.6,29.5]	0.159
(29.5,38.4]	0.171
(38.4,47.3]	0.178
(47.3,56.2]	0.203
(56.2,65.1]	0.243
(65.1,74]	0.067
(74,83]	0.143

Table 4.1: Proportion Taking Vitamin E by Range of Body Mass Index Values

Those same ranges can be graphed, as in Figure 4.5. The flat lines reflect that we are assigning the same mean to every observation in that range of BMI values. They show the mean conditional on being in that BMI bin. We see from this that BMI has a positive relationship with taking vitamin E up until the 70+ ranges, at which point the conditional mean drops.

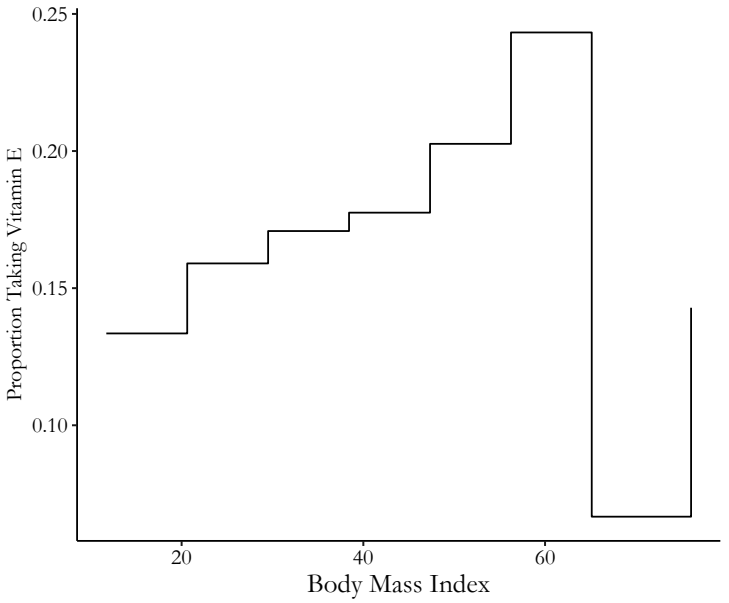

Figure 4.5: Proportion Taking Vitamin E by Range of Body Mass Index Values

Of course, while this approach is simple and illustrative, it's also fairly arbitrary. I picked the use of

ten bins (as opposed to nine, or eleven, or…) out of nowhere. It's also arbitrary to use evenly-sized bins; no real reason I had to do that. Plus, it's rather choppy. Do I really think that if someone is at the very top end of their bin, they're more like someone at the bottom of their bin than like the person at the very bottom end of the next bin?

Instead, we can use a range of X values to get conditional means of Y using *local means*. That is, to calculate the conditional mean of Y at a value of, say, $X = 2.5$, we take the mean of Y for all observations with X values *near* 2.5. There are different choices to make here—how close do you have to be? Do we count you equally if you're *very* close vs. *kind of* close?

Local mean. The mean of a variable Y calculated using only observations over a short range of another variable X.

A common way to do this kind of thing is with a LOESS curve,[9] also known as LOWESS.[10] LOESS provides a local prediction, which it gets by fitting a different shape for each value on the X axis, with the estimation of that shape weighting very-close observations more than kind-of close observations. The end result is nice and smooth.

[9] "Locally Estimated Scatterplot Smoothing"

[10] Depending on who you ask, LOESS and LOWESS might be the exact same thing, or might have slight differences in how they estimate their local prediction, with either name referring to either of the local-prediction variants.

Figure 4.6 shows the LOESS curve for the proportion taking vitamin E and BMI.

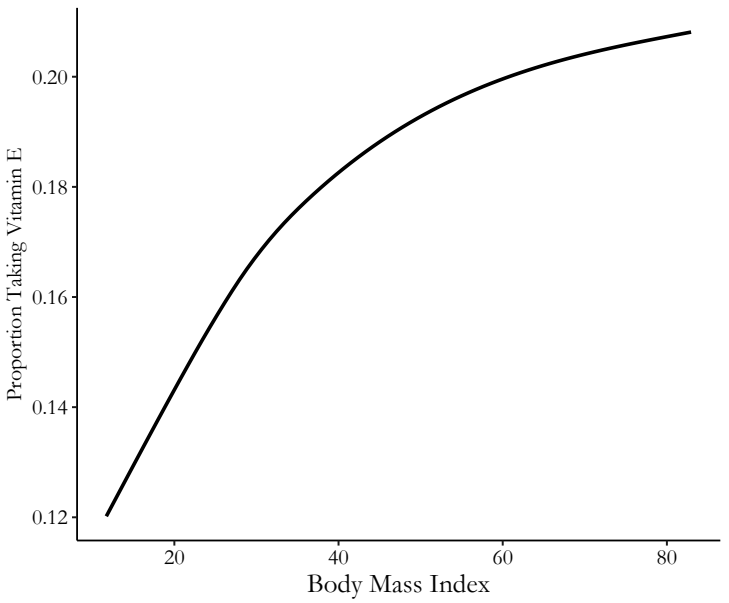

Figure 4.6: Proportion Taking Vitamin E by BMI with a LOESS Curve

LOESS. A curve that uses local averages to smooth out the relationship between two variables.

From Figure 4.6 we can see a clear relationship, with higher values of BMI being associated with more people taking vitamin E. The relationship is very strong at first, but then flattens out a bit, although it remains positive.[11] It got there by just

[11] Why doesn't this dip down at the end like Figure 4.5? There are very, very few observations in those really-high BMI bins. LOESS doesn't let that tiny number of observations pull it way down, and so sort of ignores them in a way that Figure 4.5 doesn't.

calculating the proportion of people taking vitamin E among those who have BMIs in a certain range, with "a certain range" moving along to the right only a bit at a time while it constructed its conditional means.

4.4 Line-fitting

SHOWING THE MEAN OF Y AMONG LOCAL VALUES OF X IS VALUABLE, and can produce a highly detailed picture of the relationship between X and Y. But it also has limitations. There still might be gaps in your data it has trouble filling in, for one. Also, it can be hard sometimes to concisely describe the relationship you see.[12]

Enter the concept of *line-fitting,* also known as *regression.*[13]

Instead of thinking locally and producing estimates of the mean of Y conditional on values of X, we can assume that the underlying relationship between Y and X can be represented by some sort of *shape*. In basic forms of regression, that shape is a straight line. For example, the line

$$Y = 3 + 4X \tag{4.1}$$

tells us that the mean of Y conditional on, say, $X = 5$ is $3 + 4(5) = 23$. It also tells us that the mean of Y conditional on a given value of X would be 4 higher if you instead made it conditional on a value of X one unit higher.

In Figure 4.7, we repeat the vitamin E/BMI relationship from before but now have a straight line fit to it. That particular straight line has a slope of .002, telling us that you are .2 percentage points more likely to take a vitamin E supplement than someone with a BMI one unit lower than you.

This approach has some real benefits. For one, it gives us the conditional mean of Y for *any* value of X we can think of, even if we don't have data for that specific value.[14] Also, it lets us very cleanly describe the relationship between Y and X. If the slope coefficient on X (.002 in the vitamin E/BMI regression) is positive, then X and Y are positively related. If it's negative, they're negatively related.

Those are pragmatic upsides for using a fitted line. There are more upsides in statistical terms in using a line-fitting procedure to estimate the relationship. Since the line is estimated using *all* the data, rather than just local data, the results are more precise. Also, the line can be easily extended to include more than one variable (more on that in the next section).

[12] Not to mention, it can be difficult, although certainly not impossible, to do what we do in the Conditional Conditional Means section with those methods.

[13] These two concepts are not the exact same thing, really. But they're close enough in most applications. Also, while I repeatedly mention conditional means in this section, there are versions of line-fitting that give conditional medians or percentiles or what-have-you as well.

Regression. The practice of fitting a shape, usually a line, to describe the relationship between two variables.

[14] Although if we don't have data anywhere near that value, we probably shouldn't be trying to get the conditional mean there.

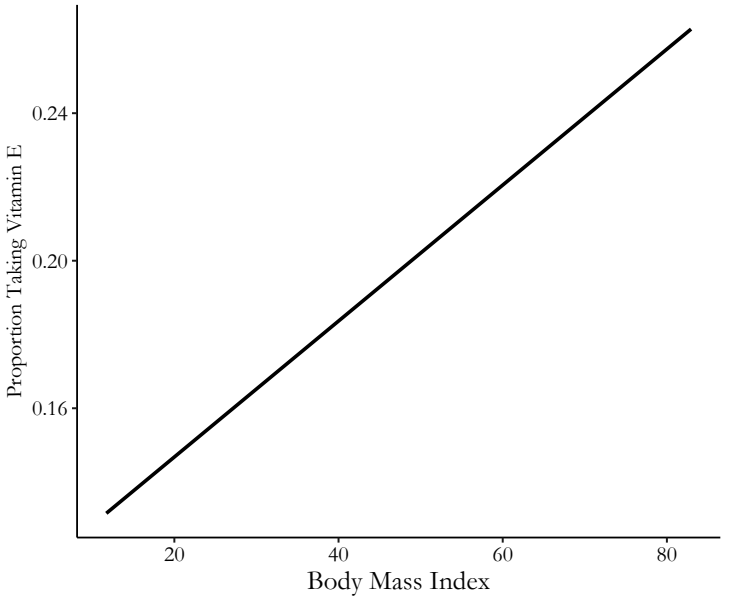

Figure 4.7: Proportion Taking Vitamin E by BMI with a Fitted Straight Line

There is a downside as well, of course. The biggest downside is that fitting a line requires us to *fit a line.* We need to pick what kind of shape the relationship is—a straight line? A curved line? A line that wobbles up and down and up and down? The line-fitting procedure will pick the best version of the shape we give it. But if the shape is all wrong to start with, our estimate of the conditional mean will be all wrong. Imagine trying to describe the relationship in Figure 4.6 using a straight line!

The weakness here isn't necessarily that straight lines aren't always correct—line-fitting procedures will let us use curvy lines. But we have to be aware ahead of time that a curvy line is the right thing to use, and then pick which kind of curvy line it is ahead of time.

That weakness is, naturally, set against the positives, which are strong enough that line-fitting is an extremely common practice across all applied statistical fields. So, then, how do we do it?

Ordinary Least Squares (OLS) is the most well-known application of line-fitting. OLS picks the line that gives the lowest *sum of squared residuals.* A residual is the difference between an observation's actual value and the conditional mean assigned by the line.[15]

Ordinary least squares. A regression method that uses a straight line and minimizes the sum of squared residuals.

[15] Or if you prefer, the difference between the actual value and the prediction.

Take that $Y = 3+4X$ line I described earlier. We determined that the conditional mean of Y when $X = 5$ was $3 + 4(5) = 23$. But what if we see someone in the data with $X = 5$ and $Y = 25$? Well then their *residual* is $25 - 23 = 2$. OLS takes that number,

squares it into a 4, then adds up all the predictions across all your data. Then it picks the values of β_0 and β_1 in the line $Y = \beta_0 + \beta_1 X$ that make that sum of squared residuals as small as possible, as in Figure 4.8.

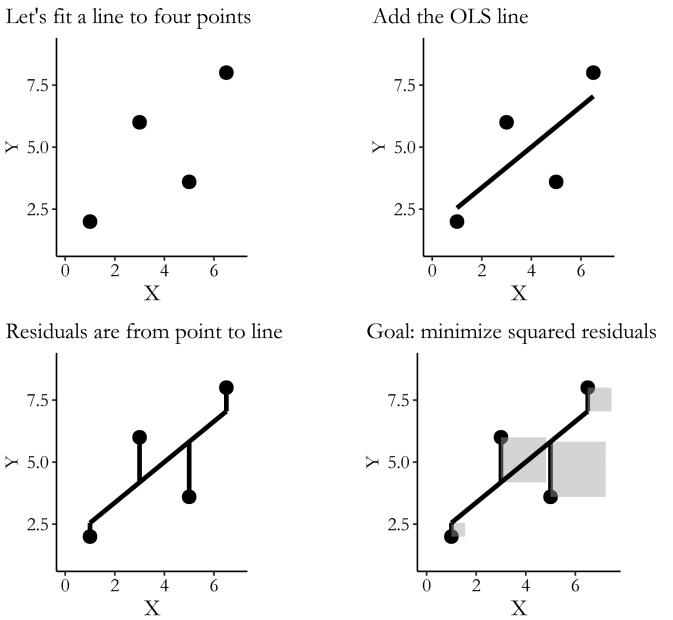

Figure 4.8: Fitting an OLS Line to Four Points

How does it do this?[16] It takes advantage of information about how the two variables move together or apart, encoded in the *covariance*.

If you recall the variance from Chapter 3, you'll remember that to calculate the variance of X, we: (a) subtracted the mean of X from X, (b) squared the result, (c) added up the result across all the observations, and (d) divided by the sample size minus one. The resulting variance shows how much a variable actually varies.

The covariance is the exact same thing, except that in step (a) you subtract the mean from *two* separate variables, and in step (b) you multiply the result from one variable by the result from the other. The resulting covariance shows how much two variables move together or apart. If they tend to be above average at the same time or below average at the same time, then multiplying one by the other will produce a positive result for most observations, increasing the covariance. If they have nothing to do with each other, then multiplying one by the other will give a positive result about half the time and a negative

[16] Calculus, for one. But besides that.

Covariance. A measurement of how much two variables vary with each other, as opposed to how much a single variable varies as in the variance. Technically, the average of the summed products of de-meaned variables.

result the other half, canceling out in step (c) and give you a covariance of 0.

How does OLS use covariance to get the relationship between Y and X? It just takes the covariance and divides it by the variance of X, i.e., $cov(X,Y)/var(X)$. That's it![17] This is roughly saying "of all the variation in X, how much of it varies along with Y?"[18] Then, once it has its slope, it picks an intercept for the line that makes the mean of the residuals (not the squared residuals) 0, i.e., the conditional mean is at least right on average.

The result from OLS is then a line with an intercept and a slope like $Y = 3 + 4X$. You can plug in a value of X to get the conditional mean of Y. And, crucially, you can describe the relationship between the variables using the slope. Since the line has $4X$ in it, we can say that a one-unit increase in X is associated with a four-unit increase in Y.

Sometimes we may find it useful to rescale the OLS result. This brings us to the concept of *correlation*. Correlation, specifically Pearson's correlation coefficient, takes this exact concept and just rescales it, multiplying the OLS slope by the standard deviation of X and dividing it by the standard deviation of Y. This is the same as taking the covariance between X and Y and dividing by both the standard deviation of X and the standard deviation of Y.

The correlation coefficient also relies on this concept of fitting a straight line. It just reports the result a little differently. We lose the ability to interpret the slope in terms of the units of X and Y.[19] However, we gain the ability to more easily tell how strong the relationship is. The correlation coefficient can only range from -1 to 1, and the interpretation is the same no matter what units the original variables were in. The closer to -1 it is, the more strongly the variables move in opposite directions (downward slope). The closer to 1 it is, the more strongly the variables move in the same direction (upward slope).

How about for vitamin E and BMI? OLS estimates the line

$$VitaminE = \beta_0 + \beta_1 BMI \tag{4.2}$$

and selects the best-fit values of β_1 and β_2 to give us

$$VitaminE = .110 + .002BMI \tag{4.3}$$

So for a one-unit increase in BMI we'd expect a .002 increase in the conditional mean of vitamin E. Since vitamin E is a binary

[17] For the two-variable version. We'll get to more complex ones in a bit.

[18] The sheer intuitive nature of this calculation might give a clue as to why we focus on minimizing the sum of squared residuals rather than, say, the residuals to the fourth power, or the product, or the sum of the absolute values. OLS gets some flak in some statistical circles for being restrictive, or for some of its assumptions. But the way that it seems to pop up everywhere and be linked to everything—it's the π of multivariate statistical methods, if you ask me. I could write a whole extra chapter just on cool stuff going on under the hood of OLS. Look at me, starstruck over a ratio.

[19] Why? Well, the slope of a straight line tells you the change in units-of-Y-*per*-units-of-X. You can read that "per" as "divided by." When we multiply by the standard deviation of X, that's in units of X, so the units cancel out with the per-units-of-X, leaving us with just units-of-Y. Then when we divide by the standard deviation of Y, that's in units of Y, canceling out with units-of-Y and leaving us without any units.

Correlation. A measurement of how two variables vary linearly together or apart, scaled to be between -1 and 1.

variable, we can think of a .002 increase in conditional mean as being a .2 percentage point increase in the proportion of people taking vitamin E.

Then, since the standard deviation of taking vitamin E is .369 and the standard deviation of BMI is 6.543, the Pearson correlation between the two is $.002 \times 6.543/.369 = .355$.

Regression slope coefficient. The linear relationship between two variables, estimated by regression. A one-unit change in one variable is associated with a (coefficient)-unit change in the other.

SOMETIMES BEING STRAIGHT IS INSUFFICIENT. OLS fits a straight line, but many sets of variables do not have a straight-line relationship! In fact, as shown in Figure 4.6, our vitamin E/BMI relationship is one of them. What to do?

Two heroes come to our rescue.[20]

[20] These are two heroes that will not really receive the attention necessary in this book, which in general covers regression just enough to get to the research design. See a little more in Chapter 13, or check out a more dedicated book on regression like Bailey's *Real Econometrics*.

The first of them is apparently also the villain, OLS. Turns out OLS doesn't actually have to fit a *straight* line. Haha, gotcha. It just needs to fit a line that is "linear in the coefficients," meaning that the slope coefficients don't have to do anything wilder than just being multiplied by a variable.

Asking it to estimate the β values in $Y = \beta_0 + \beta_1 X$ is fine, as before. But so is $Y = \beta_0 + \beta_1 X + \beta_2 X^2$—not a straight line! Or $Y = \beta_0 + \beta_1 \ln(X)$—also not a straight line! And so on. What would be something that's *not* linear in coefficients? That would be something like $Y = \beta_0 + X_1^{\beta}$ or $Y = \frac{\beta_0}{1+\beta_1 X}$.

So that scary-looking curved line in Figure 4.6? Not a problem, actually. As long as we take a look at our data beforehand to see what kind of shape makes sense (do we need a squared term for a parabola? Do we need a log term to rise quickly and then level out?), we can mimic that shape. For Figure 4.6 we could probably do with $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ to get the nice flexibility of the LOESS with the OLS bonuses of having fit a shape.

The second hero is "nonlinear regression" which can take many, many forms. Often it is of the form $Y = F(\beta_0 + \beta_1 X)$ where $F()$ is… some function, depending on what you're doing.

Nonlinear regression is commonly used when Y can only take a limited number of values. For example, we've been using all kinds of line-fitting approaches for the relationship between vitamin E and BMI, but vitamin E is *binary*—you take it or you don't. A straight line like OLS will give us something that doesn't really represent the true relationship—straight lines increase gradually, but something binary jumps from "no" to "yes" all at once! Even a line that obeys the curve like $VitaminE = \beta_0 + \beta_1 BMI + \beta_2 BMI^2$ will be a bit misleading. Even if we think about the dependent variable as the *probability* of taking vitamin E, which *can* change gradually like

a straight line, follow that line out far enough and eventually you'll predict that people with really high BMIs are more than 100% likely to use vitamin E, and people with really low BMIs are less than 0% likely. Uh-oh.

You can solve this by using an $F()$ that doesn't go above 100% or below 0%, like a "probit" or "logit" function. I'll cover more on these in Chapter 13.

There are many other functions you could use, of course, for all kinds of different Y variables and the values they can take. I won't be spending much time on them in this book, but do be aware that they're out there, and they represent another important way of fitting a (non-straight) line.

4.5 *Conditional Conditional Means, a.k.a. "Controlling for a Variable"*

LET US ENTER THE LAND OF THE UNEXPLAINED. By which I mean the residuals.

When you get the mean of Y conditional on X, no matter how you actually do it, you're splitting each observation into two parts—the part *explained by* X (the conditional mean), and the part *not explained by* X (the residual). If the mean of Y conditional on $X = 5$ is 10, and we get an observation with $X = 5$ and $Y = 13$, then the prediction is 10 and the residual is $13-10$. Figure 4.9 shows how we can distinguish the conditional mean from the residual.

Residual. The difference between the actual and predicted values of an observation.

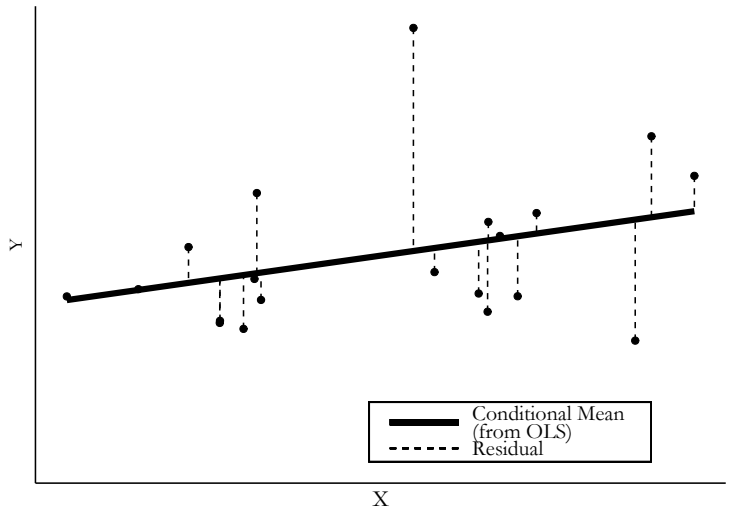

Figure 4.9: An OLS Line and its Residuals

It might seem like those residuals are just little nuisances or failures, the parts we couldn't predict. But it turns out there's a little magic in there. Because we can also think of the residual as *the part of Y that has nothing to do with X.* After all, if the conditional mean is 10 and the actual value is 13, then X can only be responsible for the 10. The extra 3 must be because of some other part of the data generating process.

Why would we want that? It turns out there are a number of uses for the residual. Just off the bat, perhaps we don't just want to know the variation in vitamin E alone. Maybe what we want is to know how much variation there is in vitamin E-taking that *isn't explained by BMI.* Looking at the residuals from Figure 4.7 would answer exactly that question.

Things get real interesting when we look at the residuals of two variables at once.

WHAT IF WE TAKE THE EXPLAINED PART OUT OF TWO DIFFERENT VARIABLES? Let's expand our analysis to include a third variable. Let's keep it simple with Y, X, and Z. So, what do we do?[21]

1. Get the mean of Y conditional on Z.
2. Subtract out that conditional mean to get the residual of Y. Call this Y^R.
3. Get the mean of X conditional on Z.
4. Subtract out that conditional mean to get the residual of X. Call this X^R.
5. Describe the relationship between Y^R and X^R.

[21] This particular set of calculations, when applied to linear regression, is known as the Frisch-Waugh-Lovell theorem and doesn't apply precisely to regression approaches that are nonlinear in parameters, like logit or probit as previously described. However, for those regressions the concept is still the same.

Now, since Y^R and X^R have had the parts of Y and X that can be explained with Z *removed*, the relationship we see between Y^R and X^R is *the part of the relationship between Y and X that is not explained by Z.*

In other words, were getting the *Mean of Y conditional on X* all conditional on Z. We're *washing out the part of the X/Y relationship that is explained by Z.*

In doing this, we are taking out all the variation related to Z, in effect not allowing Z to vary. This is why we call this process "controlling for" Z (although "adjusting for" Z might be a little more accurate).

Controlling for a variable. Removing all the variation associated with that variable from all the other variables.

Let's take our ice cream and shorts example. We see that days where more people eat ice cream also tend to be days where more

people wear shorts. But we also know that the temperature outside affects both of these things.

If we really want to know if ice cream-eating affects shorts-wearing, we would want to know *how much of a relationship is there between ice cream and shorts that isn't explained by temperature?* So we would get the mean of ice cream conditional on temperature, and then take the residual, getting only the variation in ice cream that has nothing to do with temperature. Then we would take the mean of shorts-wearing conditional on temperature, and take the residual, getting only the variation in shorts-wearing that has nothing to do with temperature. Finally, we get the mean of the shorts-wearing residual conditional on the ice cream residual. If the shorts mean doesn't change much conditional on different values of ice cream eating, then the entire relationship was just explained by heat! If there's still a strong relationship there, maybe we do have something.

THE EASIEST WAY TO TAKE CONDITIONAL CONDITIONAL MEANS IS WITH REGRESSION. Regression allows us to control for a variable by simply adding it to the equation. Now we have "multivariate" regression. So instead of

$$Y = \beta_0 + \beta_1 X \tag{4.4}$$

we just use

$$Y = \beta_0 + \beta_1 X + \beta_2 Z \tag{4.5}$$

and voila, the OLS estimate for β_1 will automatically go through the steps of removing the conditional means and analyzing the relationship between Y^R and X^R.

Even better, we can do things conditional on *more than one variable.* So we could add W and do…

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 W \tag{4.6}$$

and now the β_1 that OLS picks will give us the relationship between Y and X conditional on *both* Z and W.

Let's take a quick look at how this might affect our vitamin E/BMI relationship. Some variables that might be related to both taking vitamin E and to BMI are gender and age. So let's add those two variables to our regression and see what we get.

Before, with only BMI, we estimated

$$VitaminE = .110 + .002BMI \tag{4.7}$$

Now, with BMI, gender, and age, we get

$$VitaminE = -.006 + .001BMI + .002Age + .016Female \quad (4.8)$$

The effect of BMI has changed a bit, from .002 to .001, telling us that some of the relationship we saw between BMI and vitamin E was explained by age and/or gender. We also see that older people are more likely to take vitamin E—for each additional year of age we expect the proportion taking vitamin E to go up by .2 percentage points. Women are also more likely than men to take the supplement. A one-unit increase in "Female" (i.e., going from 0—a man—to 1—a woman) is associated with an increased proportion taking vitamin E of 1.6 percentage points.[22]

[22] Of course, OLS by itself doesn't know *which* variable is the treatment. So the Age effect "controls for" BMI and Female, and the Female effect controls for BMI and Age. However, we don't want to get too wrapped up in interpreting the coefficients on controls, as we generally haven't put much work into identifying their effects (see Chapter 5).

So how does regression do this? Put your mental-visualization glasses on.

One way is mathematically. If you happen to know a little linear algebra (and if you don't, you can skip straight to the next paragraph), the formula for multivariate OLS is $(A'A)^{-1}A'Y$, where A is a matrix of all the variables other than Y, including the X we're interested in. In other words, it washes out the influence of all the non-X variables on the X/Y relationship by dividing out a bunch of covariances.

Another way is graphically. If you can think of a two-variable OLS line $Y = \beta_0 + \beta_1 X$ as being a line, you can think of a three-variable OLS line as a *plane* in 3-D space (or with four variables, in 4-D space, and so on). We can visualize this by looking at each of the three sides of that 3-D image one at a time.

Figure 4.10 shows the X-Y axis on the top-left. Then to the right you can see the Z-Y axis, and below the Z-X axis. The coordinates are flipped on the Z-X axis—even though we're getting the mean of X conditional on Z here, I've put X on the x-axis to be consistent with the X-Y graph. The upward slope on the Z-Y and Z-X axes shows that Z is explaining part of both X and Y, and that we could take that explanation out to focus on the residuals.

Then, in Figure 4.11, we flatten out those explanations. The upward slopes get flattened out, moving the X and Y points with them. You can see how subtracting out the parts explained by Z literally leaves the X/Y relationship no part of Z to hold on to! Z has been flatlined in both directions, providing no additional "lift" to the points in the X/Y graph. What's still there on X/Y is there without Z.

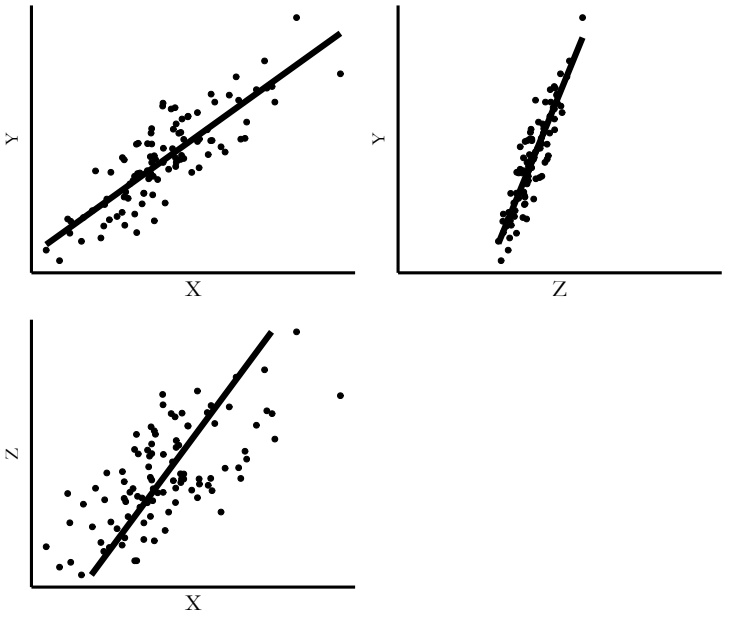

Figure 4.10: A Three-Variable Regression from All Three Dimensions

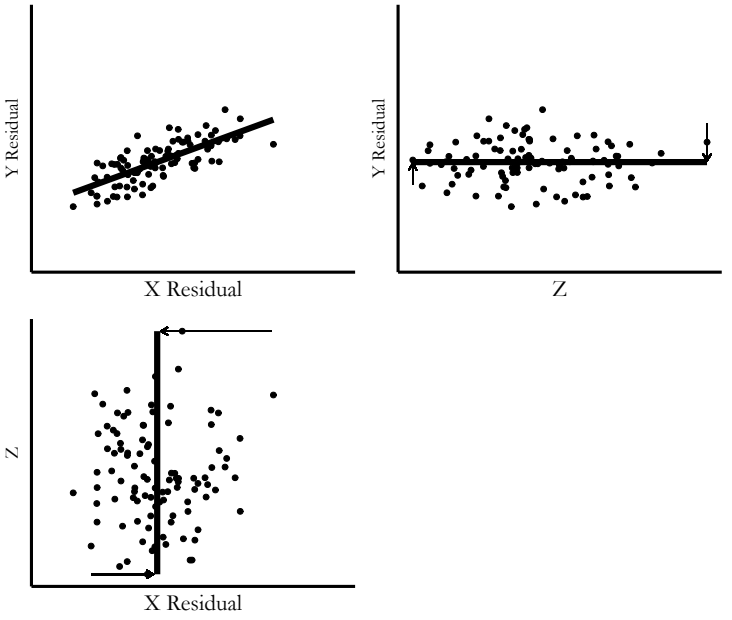

Figure 4.11: A Three-Variable Regression from All Three Dimensions After Removing the Variation Explained by Z

4.6 What We're Not Covering

In the previous chapter, on describing variables, we did a pretty good job covering a lot of what you'd want to know when de-

scribing a variable. This chapter, however, leaves out a whole lot more.

This is largely for reasons of focus. This book is about research design. Once you've got research design pinned down, there are certainly a lot of statistical issues you need to deal with at that point. But things like specific probability distributions (normal vs. log-normal vs. t vs. Poisson vs. a million others we didn't cover), functional form (OLS vs. probit/logit vs. many others), or standard errors and hypothesis testing can be a distraction when thinking about the broad strokes of how you're going to answer your research question.

In one case, omission is less for focus and more to cover it more appropriately later. Notice how I introduced the Oster paper as being all about how the relationship between vitamin E and health indicators changed over time… but then I never showed how the BMI relationship changed over time? There are a number of research designs that have to do with *how a relationship changes* in different settings.[23] However, a proper treatment of this will have to wait until Part II of the book.

[23] Controlling for time would not achieve this. Controlling for time would remove the part of the relationship explained by time, but would not show how the relationship changes over time.

To be clear, you want to know all this stuff. And I will cover it more in this book in Chapter 13, and many of the other Part II chapters. You can also check out a more traditional econometrics book like Bailey's *Real Econometrics* or Wooldridge's *Introductory Econometrics*. But for observational data, most of the time these are things to consider *after* you have your design and plan to take that design to actual data.

For now, I want you think about *what you want to do* with your data—what kinds of descriptions of variables your research design requires, what kinds of relationships, what kinds of conditional means and conditional conditional means. Figure out how you want your data to *move*. Figure out the journey you're going to take first; you can pack your bags when it's actually time to leave.

4.7 Relationships in Software

In this section, I'll show you how to calculate or graph the relationship between variables in three different languages: R, Stata, and Python.

These code chunks may rely on *packages* that you have to install. Anywhere you see `library(X)` or `X::` in R or `import X` or `from X import` in Python, that's a package **X** that will need to be installed if it isn't already installed. You can do this with

`install.packages('X')` in R, or using a package manager like **pip** or **conda** in Python. In Stata, packages don't need to be loaded each time they're used like in R or Python, so I'll always specify in the code example if there's a package that might need to be installed. In all three languages, you only have to install each package once, and then you can load it and use it as many times as you want.

The data sets for all the examples in this book can be found in the **causaldata** package, which I've made available for all three languages. Do `install.packages('causaldata')` in R, `ssc install causaldata` in Stata, or `pip install causaldata` (if using **pip**) in Python.

So let's do those code examples! The Oster data, while free to download, would require special permissions to redistribute. Instead, I will be using data from Mroz (1987),[24] which is a data set of women's labor force participation and earnings from 1975.

[24] Thomas A Mroz. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, 55(4): 765–799, 1987.

In each of these languages, I'm going to:

1. Load in the data
2. Draw a scatterplot between log women's earnings and log other earnings in the household,[25] among women who work
3. Get the conditional mean of women's earnings by whether they attended college
4. Get the conditional mean of women's earnings by different bins of other household earnings
5. Draw the LOESS and linear regression curves of the mean of log women's earnings conditional on the log amount of other earnings in the household
6. Run a linear regression of log women's earnings on log other earnings in the household, by itself and including controls for college attendance and the number of children under five in the household

[25] Why am I using the log of earnings for most of these steps? Think carefully about what we learned about logarithms in Chapter 3.

```
# R CODE
library(tidyverse); library(modelsummary)

df <- causaldata::Mroz %>%
    # Keep just working women
    filter(lfp == TRUE) %>%
    # Get unlogged earnings %>%
    mutate(earn = exp(lwg))

```

```r
# 1. Draw a scatterplot
ggplot(df, aes(x = inc, y = earn)) +
    geom_point() +
    # Use a log scale for both axes
    # We'll get warnings as it drops the 0s, that's ok
    scale_x_log10() + scale_y_log10()

# 2. Get the conditional mean by college attendance
df %>%
    # wc is the college variable
    group_by(wc) %>%
    # Functions besides mean could be used here to get other conditionals
    summarize(earn = mean(earn))

# 3. Get the conditional mean by bins
df %>%
    # use cut() to cut the variable into 10 bins
    mutate(inc_cut = cut(inc, 10)) %>%
    group_by(inc_cut) %>%
    summarize(earn = mean(earn))

# 4. Draw the LOESS and linear regression curves
ggplot(df, aes(x = inc, y = earn)) +
    geom_point() +
    # geom_smooth by default draws a LOESS; we don't want standard errors
    geom_smooth(se = FALSE) +
    scale_x_log10() + scale_y_log10()
    # Linear regression needs a 'lm' method
    ggplot(df, aes(x = inc, y = earn)) +
    geom_point() +
    geom_smooth(method = 'lm', se = FALSE) +
    scale_x_log10() + scale_y_log10()

# 5. Run a linear regression, by itself and including controls
model1 <- lm(lwg ~ log(inc), data = df)
# k5 is number of kids under 5 in the house
model2 <- lm(lwg ~ log(inc) + wc + k5, data = df)
# And make a nice table
msummary(list(model1, model2))
```

```
* STATA CODE
* Don't forget to install causaldata with ssc install causaldata
* if you haven't yet.
causaldata Mroz.dta, use clear download
* Keep just working women
keep if lfp == 1
* Get unlogged earnings
g earn = exp(lwg)
* Drop negative other earnings
drop if inc < 0

* 1. Draw a scatterplot
twoway scatter inc earn, yscale(log) xscale(log)

* 2. Get the conditional mean college attendance
table wc, c(mean earn)

* 3. Get the conditional mean by bins
* Create the cut variable with ten groupings
egen inc_cut = cut(inc), group(10) label
```

```
table inc_cut, c(mean earn)

* 4. Draw the LOESS and linear regression curves
* Create the logs manually for the fitted lines
g loginc = log(inc)
twoway scatter loginc lwg || lowess loginc lwg
twoway scatter loginc lwg || lfit loginc lwg

* 5. Run a linear regression, by itself and including controls
reg lwg loginc
reg lwg loginc wc k5
```

```
# PYTHON CODE
import pandas as pd
import numpy as np
import statsmodels.formula.api as sm
import matplotlib.pyplot as plt
import seaborn as sns
from causaldata import Mroz

# Read in data
dt = Mroz.load_pandas().data
# Keep just working women
dt = dt[dt['lfp'] == True]
# Create unlogged earnings
dt.loc[:,'earn'] = dt['lwg'].apply('exp')

# 1. Draw a scatterplot
sns.scatterplot(x = 'inc',
y = 'earn',
data = dt).set(xscale="log", yscale="log")
# The .set() gives us log scale axes

# 2. Get the conditional mean by college attendance
# wc is the college variable
dt.groupby('wc')[['earn']].mean()

# 3. Get the conditional mean by bins
# Use cut to get 10 bins
dt.loc[:, 'inc_bin'] = pd.cut(dt['inc'],10)
dt.groupby('inc_bin')[['earn']].mean()

# 4. Draw the LOESS and linear regression curves
# Do log beforehand for these axes
dt.loc[:,'linc'] = dt['inc'].apply('log')
sns.regplot(x = 'linc',
            y = 'lwg',
            data = dt,
            lowess = True)
sns.regplot(x = 'linc',
            y = 'lwg',
            data = dt,
            ci = None)

# 5. Run a linear regression, by itself and including controls
m1 = sm.ols(formula = 'lwg ~ linc', data = dt).fit()
print(m1.summary())
# k5 is number of kids under 5 in the house
m2 = sm.ols(formula = 'lwg ~ linc + wc + k5', data = dt).fit()
print(m2.summary())
```