Statistics Minus The Math: An Introduction for the Social Sciences

Nathan Favero

2024-02-11

Table of contents

Introduction

This version (1.2) was updated 1/28/2023.

Version 1.1 available at https://nathanfavero.com/teaching-tools/. The only change from 1.1 is that the discussion of transforming variables now appears in Ch. 2 (rather than Ch. 3).

This book was largely adapted from the public domain resource *Online Statistics Education: A Multimedia Course of Study* (http://onlinestatbook.com/ Project Leader: David M. Lane, Rice University).

1 Models and Uncertainty

Before I leave my house each morning, I need to decide whether to take an umbrella. So I check my phone to see whether it's supposed to rain. Instead of giving me a direct yes or no answer, the weather tells me the percentage chance of rain for the day.

Why does the weather app give me a percentage? Because there's uncertainty. Science has done a lot to help us understand the weather. And as our understanding of the weather improves, our predictions get better. But we still can't predict rain perfectly.

Facing uncertainty is a common problem when we're looking at data. Whether we're trying to explain the weather, human behavior, or even plant growth, we can't make perfect predictions because there are things we can't fully explain with our current scientific knowledge.

In statistics, we have several tools that allow us to acknowledge uncertainty. This enables us to build models like the ones powering my weather app—models that give us a prediction that includes a description of how uncertain we are. Some days we are 100% sure it will rain, other days only 60%.

In order to build these models that acknowledge uncertainty, we need a way to talk about what we do know and what we don't know. Let me give a very simple example of a model that accounts for uncertainty:

$$happiness = 3.0 + 2.3 \times income + \varepsilon \tag{1.1}$$

This model attempts to explain one's level of happiness based on their income. You might notice that it looks very similar to the regression equations we saw in Chapter 3. That's because regression is one of the main tools used to estimate a model that includes uncertainty.

What does this model mean in practical terms? Well, there are no obvious units we can use to quantify the amount of happiness someone experiences, so the exact values of the numbers we see are not particularly meaningful. But the fact that there's a positive number (2.3) that is being multiplied by income implies that as income gets bigger, happiness gets larger.

The key part of this equation that I want to focus on is the little Greek letter at the end of the equation: ε. This letter is called "epsilon," and it is often used to represent what we call an **error term** (also sometimes called a **disturbance term**). The error term (ε) represents everything else besides income that affects happiness. By including an error term, we are acknowledging that we can't perfectly predict one's level of happiness based on their income.

We think that knowing one's income will help us predict their happiness, but we know there are other factors we won't be able to measure or identify that will also affect happiness. Thus, if all we know about someone is their income, we will have uncertainty about their exact level of happiness. By including an error term (ε) in the model, we make clear that we only claim to have a partial understanding of happiness, not a complete one.

Think for a moment about how few topics we could study if we didn't have the freedom to build models that include uncertainty. We'd only be able to build a model of a dependent variable after we had identified (and measured) *all* of the factors that affect that variable! We wouldn't be able to build a model of rain since we don't know all of the factors that affect the rain. We couldn't build a model of voting behavior since we don't know everything that affects how someone will vote. By including an error term in our model, we can build models even when our understanding of something is incomplete.

The first part of our model that appears on the right side of the equation ($3.0 + 2.3 \times income$) is sometimes described as the *systematic* part of our model. It's what we would use to build a prediction of happiness if all we know about some is their income level. Suppose, for example, that someone has an income of 4 units (perhaps income is measured in tens of thousands of dollars of annual income, so a salary of $40,000 is coded as a 4). According to our model, that person's happiness would be:

$$happiness = 3.0 + 2.3 \times (4) + \varepsilon$$

$$happiness = 12.2 + \varepsilon$$

We, therefore, predict that someone with an income of 4 will have a happiness of 12.2, but we also acknowledge that their actual happiness will likely be a bit different from our prediction since our model indicates that their actual happiness will equal 12.2 plus the value of the error term (ε).

The error term describes something unknown, so we can't measure it or directly observe it. But what we can do is talk about its characteristics using concepts from probability theory. Specifically, we're going to describe the value of the error term as being randomly selected. You may have dealt with randomness in math classes before using examples such as coin flips, die rolls, or drawing cards from a 52-card deck. Just as the likelihood of different outcomes from parlor games can be described using probability, we're going to use probability to describe different possible values for the error term of a statistical model.

1.1 Assumptions about error terms

It's easy to write out an equation that includes an error term, but we are not going to be able to do much with our model unless we make some assumptions about the error term. One of the most important (and challenging) parts of doing statistical analysis is making assumptions

about the possible values of the error term. Different assumptions about the error term can result in very different conclusions.

Let's again consider the simple model of happiness that was introduced above:

$$happiness = 3.0 + 2.3 \times income + \varepsilon$$

We might assume the following things about the error term (ε):

1. The values of the error term (ε) can be described by a normal distribution with a mean of 0
2. Knowing someone's income doesn't help us predict the values of the error term (ε)

What do these two assumptions mean?

First, if the error term (ε) follows a normal distribution with a mean of zero, that means that (according to our model), people are just as likely to have a positive value of the error term as they are to have a negative value of the error term. In other words, all those factors we haven't accounted for in our model are equally likely to push people in the direction of being happier or in the direction of being less happy. Our model and assumptions tell us that if we predict happiness purely based on income, we'll *overestimate* some people's happiness, and we'll *underestimate* an equal number of people's happiness.[1]

Second, these assumptions allow us to describe how much individual observations will tend to deviate from our income-based predictions. We haven't specified in our assumptions what the standard deviation is for the normal distribution for the error term (ε), but statistical analysis will let us estimate the standard deviation of an error term. And we know that there is a 95% chance of drawing a value within two standard deviations of the mean for any normal distribution. So whatever the standard deviation of the error term (ε) is, we would expect that 95% of the time, the error term will take on a value that is within two standard deviations of zero. Conversely, 5% of the time, the error term will take on a value that is more than two standard deviations from zero. Suppose that the standard deviation of the error term (ε) happens to be three. If we have a dataset containing the income and happiness of 1,000 randomly selected people, we would expect that about 950 of these people will have a level of happiness that falls within six units of our income-based prediction. But for about 50 of these people, our prediction of their happiness will be off by more than six units.

Third, our assumptions imply that income is not tied in any consistent way to (the total sum of) factors other than income that also affect peoples' happiness. Remember, the error term (ε) represents all factors other than income that affect satisfaction. If income is related to these other factors, then the value of income should help us predict the value of the error term. For example, if having a stable environment in childhood tends to cause both higher incomes

[1] Note that these deviations from our prediction don't imply that our model is wrong; our model explicitly acknowledges that we'll get only imperfect estimates if we predict happiness based on income, since the unobserved error term (ε) also contributes to happiness.

and greater happiness in adulthood, the error term will partially reflect the effect of childhood stability on happiness, so high incomes (which are partially caused by childhood stability) will be probably be predictive of a more positive error term. This would constitute a violation of our assumptions since we specifically indicated that income wasn't predictive of the error term. As this example illustrates, our assumptions about error terms are often quite strict, making it rather difficult in practice to build good models that account for uncertainty.

1.2 Models and probabilistic thinking

Despite the difficulty inherent in building models that accommodate uncertainty, we have little alternative unless we wish to only build models of things we think we can predict with 100% accuracy. And fortunately, our models do not always have to be perfectly correct in order to generate useful predictions or explanations. As the statistician George Box famously said, "all models are wrong, but some are useful."

An important part of learning to do good statistical analysis is learning to think clearly about models so that you can pick out a model that is useful for whatever it is you want to accomplish. And the first step toward understanding many statistical models is learning to think about the world in probabilistic terms, as we've done here in this reading. Probabilistic thinking asks questions like:

- Based on what I do know and what I don't know, what can I predict?
- How does adding or removing different pieces of information change my prediction?
- How much uncertainty is there in my prediction?
- How often will my prediction differ greatly from what actually happens (even if my model is correct)?

2 Regression with Qualitative Independent Variables

Let's say I'm interested in studying how personality relates to gender. The most common personality measure in psychology is called the "Big Five" personality inventory. There is a standard set of 50 survey items that researchers can use to measure five aspects of personality. Figure 2.1 is an example of some of these questions and how they are formatted:

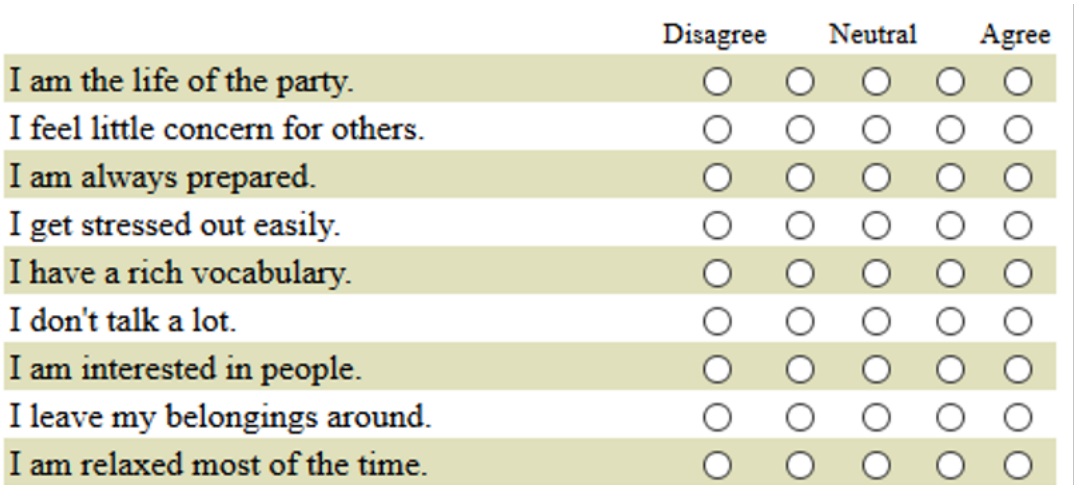
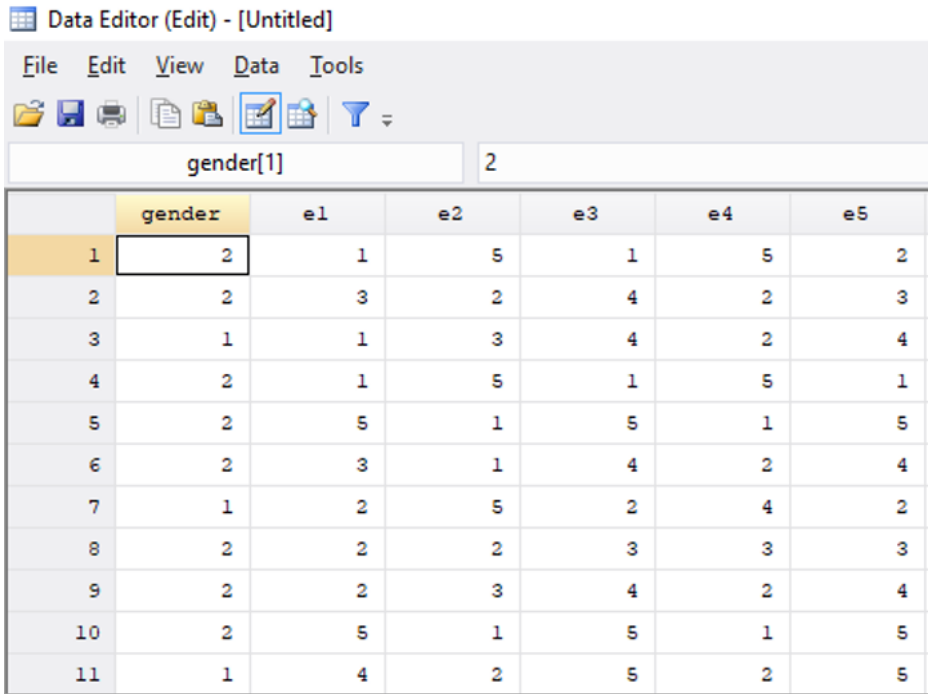


Figure 2.1

For now, I decide to focus on whether people are introverted or extroverted. Extroverts are outgoing and tend to enjoy interacting with others. Extroverts will tend to agree with the statement "I am the life of the party" while introverts will tend to agree with the item "I don't talk a lot."

I find a dataset that contains lots of responses to the Big Five personality questions as well as information on the gender of each respondent.[1] There are 10 different questions related to extroversion, and the dataset has one variable (column of data) for each of these 10 questions. The column labeled e1 shows responses to the item "I am the life of the party." A value of 1 means the respondent disagrees with this statement, while a 3 indicates neutral, and a 5 means they disagree.

[1] https://openpsychometrics.org/_rawdata/ (the file I used is called "BIG5.zip")



	gender	e1	e2	e3	e4	e5
1	2	1	5	1	5	2
2	2	3	2	4	2	3
3	1	1	3	4	2	4
4	2	1	5	1	5	1
5	2	5	1	5	1	5
6	2	3	1	4	2	4
7	1	2	5	2	4	2
8	2	2	2	3	3	3
9	2	2	3	4	2	4
10	2	5	1	5	1	5
11	1	4	2	5	2	5

Figure 2.2

For all of the odd-numbered extroversion questions (e1, e3, e5, etc.), agreement indicates extroversion. For the even-numbered items (e2, e4, e6, etc.), agreement indicates introversion. To create a single extroversion variable that combines responses from all 10 survey items, I create a tally, adding up all the values for odd-numbered questions and then subtracting the responses to the even-numbered questions. An extreme extrovert will have a 5 for all the odd-numbered questions and a 1 for all of the even-numbered ones, giving them a score of 20 (5x5-5x1=20). An extreme introvert will have a -20 since they will answer 1 to all the odd-numbered questions and 5 to all the even-numbered ones (5x1-5x5=-20).

Most people lie somewhere in the middle between introversion and extroversion:

Our gender variable was measured by asking respondents "What is your gender?" and they could choose from male, female, or other. In a moment, we'll consider those who responded "other," but for now, let's just look at those who chose either male or female.

2.1 Predicting extraversion using gender

If I want to describe differences in extraversion by gender in this dataset, I can compute the mean value of extraversion for males and for females. It turns out that males have an average extraversion of -0.46 while females' average level of extraversion is 0.53. Thus, the average

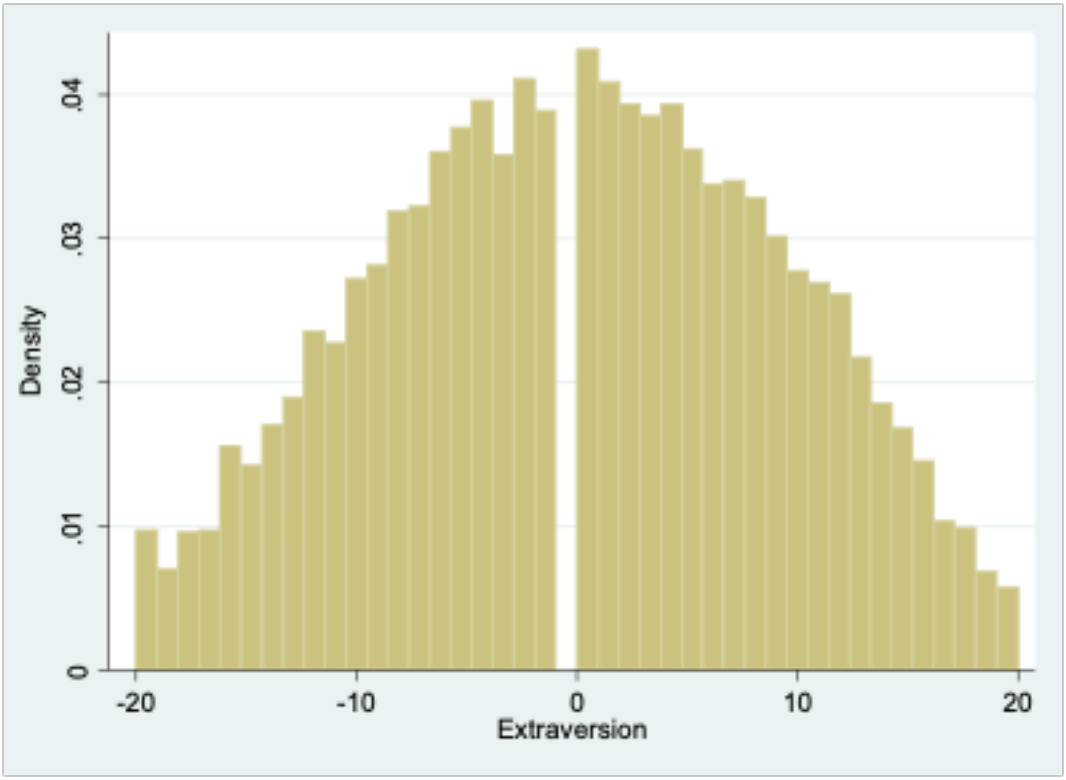


Figure 2.3

female is about 1-point more extraverted than the average male. But of course, there is lots of variation in extraversion among both groups:

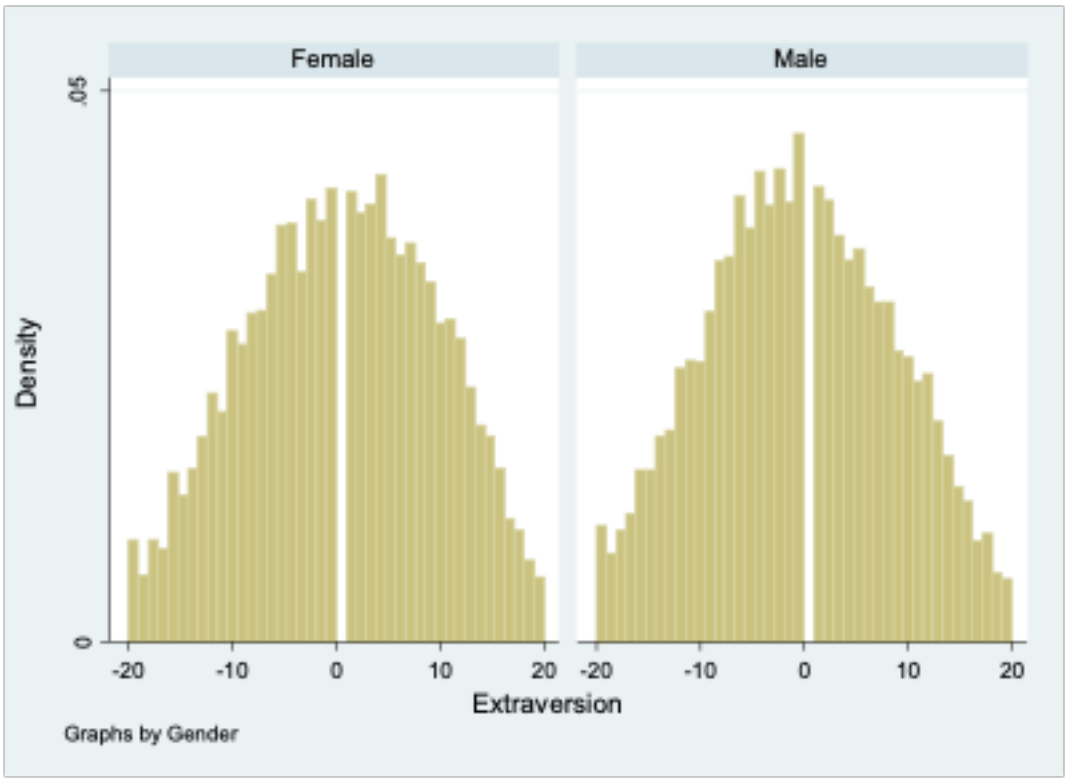


Figure 2.4

There are plenty of females who are introverts and plenty of males who are extroverts.

If you asked me to guess the extroversion level of someone and the only thing you told me about them was their gender, my best bet would probably be to guess the average extroversion level for someone of that gender. So for a female I knew nothing else about, I would guess their extroversion to be 0.53, while for a male I'd guess -0.46.

Social scientists use the **dependent variable** to describe the variable they're making a prediction about and **independent variable** to describe the variables that help them make that prediction. So in this example, extraversion is my dependent variable and gender is my independent variable.

When we're working with data, sometimes it's helpful to express how I would make a guess about a dependent variable (extraversion) based on other factors (gender) using a mathematical formula. In fact, this is exactly what we do when we run a regression. There are many ways I could write this formula, but I'll show just two for now. First, I could write:

$$\widehat{Extraversion} = 0.53 \times Female - 0.46 \times Male \tag{2.1}$$

Notice I've added a "hat" above the name of the variable *Extraversion*; this hat means that I'm making a guess about the value of that variable (I'm guessing the level of extraversion based on gender). The equation has two other variables *Female* and *Male*, and these two variables will take on a value of 1 if the person's gender is equal to the name of the variable and will otherwise take on a value of 0. For a female, *Female* will equal 1 and *Male* will equal 0, giving us:

$$\widehat{Extraversion} = 0.53 \times (1) - 0.46 \times (0) = 0.53$$

So our guess for the level of extroversion ($\widehat{Extraversion}$) of a female we know nothing about is 0.53.

For a male, our guess is:

$$\widehat{Extraversion} = 0.53 \times (0) - 0.46 \times (1) = -0.46$$

There's a second way I can write my formula, which will turn out to be more useful in the future when we come to consider multiple factors at the same time that might help us predict the value of a dependent variable. Rather than having two variables to represent gender in my equation, I can just use one:

$$\widehat{Extraversion} = 0.53 - 0.99 \times Male \tag{2.2}$$

In Equation 2.2, we start from female as our baseline. Notice that the first number we see (0.53) is our guess for the value of extraversion for a female. When we're considering a female, Male=0, so:

$$\widehat{Extraversion} = 0.53 - 0.99 \times 0 = 0.53$$

Thus, we get the right prediction for females from this equation, even though we didn't include a variable specifically for females. If we have a male, Male=1, so we get:

$$\widehat{Extraversion} = 0.53 - 0.99 \times 1 = -0.46$$

This is the same prediction we got before. Remember, I decided to initially just analyze respondents who selected either male or female. Since we are only considering two categories (male or female), and each respondent is either a male or a female, saying $Male = 1$ lets me know that $Female = 0$. It's actually repetitive in this context to both say that $Male = 1$

and $Female = 0$. Similarly, saying $Male = 0$ implies that $Female = 1$. So I can simplify my equation by just including one variable to indicate binary gender.

Notice that in Equation 2.2, the number next to *Male* is equal to the difference between the average level of extraversion for females and the average level for males (0.53-(-0.46)=0.99). This is because Equation 2.2 starts with females as the baseline, so to get our prediction for males, we have to adjust our baseline prediction by the average difference for males.

Equation 2.2 is also typically how we will arrange our equation when we're running a regression.

2.2 Prediction with more than two categories for gender

I now move beyond the gender binary and consider the "other" category in survey responses. I'll refer to this other category as "non-binary" gender. The average level of extraversion among those with non-binary gender is -5.66. So non-binary people tend to be quite a bit more introverted than those who identify as male or female. As with males and females, there is considerable variation among non-binary people:

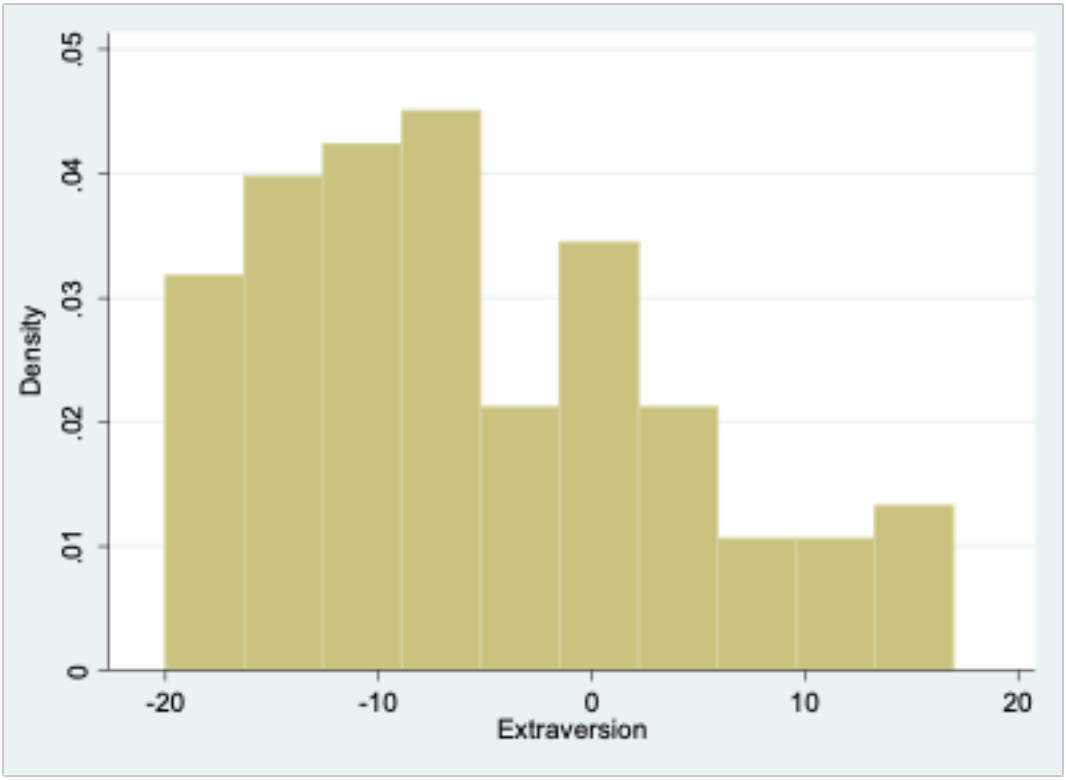


Figure 2.5

The number of non-binary respondents is relatively small (102), so it's not terrible surprising that this histogram looks a bit choppier than the ones we saw before.

Again, if we had to make a guess about the level of extraversion of someone, and all we knew about that person was that their gender was non-binary, we would probably want to guess the mean value among non-binary respondents (-5.66). Modifying Equation 2.1 to incorporate a third category is relatively straightforward:

$$\widehat{Extraversion} = 0.53 \times Female - 0.46 \times Male - 5.69 \times Other \tag{2.3}$$

For someone who identifies as female, we would plug in $Female = 1$, $Male = 0$, and $Other = 0$:

$$\widehat{Extraversion} = 0.53 \times (1) - 0.46 \times (0) - 5.66 \times (0) = 0.53$$

If someone identifies as non-binary, we would use $Female = 0$, $Male = 0$, and $Other = 1$:

$$\widehat{Extraversion} = 0.53 \times (0) - 0.46 \times (0) - 5.66 \times (1) = -5.66$$

We can also return to the format of Equation 2.2 but modify it to include the other category. This is how we will typically write our equation if we are doing a regression:

$$\widehat{Extraversion} = 0.53 - 0.99 \times Male - 6.19 \times Other \tag{2.4}$$

Now that there are three possible values for gender (female, male, and other), knowing the value of *Male* doesn't necessarily allow us to conclude what the vale of female is. If , the individual could identify as either female or non-binary. So we have to include a second variable. In this case, we chose to include the variable *Other*. If we know the values of *Male* and *Other*, we can always figure out the value of *Female* by process of elimination.

For a non-binary person, we plug in $Male = 0$, and $Other = 1$:

$$\widehat{Extraversion} = 0.53 - 0.99 \times (0) - 6.19 \times (1) = -5.66$$

When considering a female, we use $Male = 0$, and $Other = 0$:

$$\widehat{Extraversion} = 0.53 - 0.99 \times (0) - 6.19 \times (0) = 0.53$$

Equation 2.3 and Equation 2.4 communicate an equivalent method of making a prediction about extraversion based on gender; they just offer this information in two different formats.

Equation 2.4 might be a bit trickier to understand for now, but it will become very useful in the future.

Notice that we can talk about gender either as one qualitative variable with three possible values (female, male, or other), or we can talk about it as a series of three dummy variables (*Female*, *Male*, and *Other*) that can take each on a value of either 0 or 1. This can make things a bit confusing, but the important thing to remember is that when we have a qualitative variable with more than two categories, we'll need to break out the categories into a set of dummy variables for purposes of representing the qualitative variable in an equation.

However, as Equation 2.2 and Equation 2.4 illustrate, we don't necessarily need a dummy variable for every single category. Specifically, whenever we want to create an equation with a qualitative independent variable in a format like Equation 2.2 or Equation 2.4, the number of dummy variables should be equal to the number of categories minus one. Since our gender variable can take on three possible values in this example, we included two independent variables in Equation 2.4. No dummy variable is included for female, so we call female the **omitted category** or the **baseline category**. Remember, the first number in Equation 2.4 is 0.53, which represents our guess for females—the baseline category. If we instead had a qualitative variable with five categories, we would include four dummy variables in our equation.