Contents

1 How Many Kinds of People Are There?

> There are 10 kinds of people in this world. Those who understand binary code and those who don't.
>
> — seen on a T-shirt

Things are about to get meta right from the start

I'm going to start off this first chapter in a book about data science with an unsubstantiated claim. My claim is this: People love to categorize themselves and others. They love to take quizzes online that tell you "what kind of person you are" in some way or another. They love to make statements that begin with, "there are two kinds of people in this world. . . " and so on. Ok? That's my claim. It's a bit of a mouthful.

Now, I just made a claim in support of which data *can absolutely* be brought to bear. But I won't use data to support it. What? Why not, for crying out loud?! This is a book about data science!!! The reason is this: this book encourages you to think critically and skeptically about all kinds of ideas, claims, and questions. It tries to show you how to talk about these ideas precisely and not succumb to fallacies and bad intuition. But while trying to develop these skills, it is important to know when we are in turbo critical thinking mode (that's a technical term[1]) and when we're not. Sometimes, we need to be able to say common-sense things and not have to support them.

What *exactly* am I even saying in my claim, you might be thinking? What do you mean by, "people love to" do X, where X, like ________ ["blank"], is a stand-in for some of the specific things I mentioned. That everybody does X? Most people? That people who do X derive pleasure above some pleasure threshold, thus designating "love" as opposed to "like?" You see, I could have tried to make my claim more precise. And I could have found polls and published reports that estimate just how many people have, by choice, taken some kind of person-category-test-thing, or posted funny jokes about "two kinds of people." But I'm just letting my claim stand as a common-sense claim. Just like if I said, people love going to the movies. I wouldn't feel the need to cite a scientific study to support that claim.

Now, if someone is making what to *them* appears to be a common-sense claim but to you appears false or at least non-obvious, you have a few options. You can challenge the assumption and ask for evidence. Or you can accept the assumption, *for argument's sake*, to see where this is going. Hopefully, my claim feels common-sense enough to you too (i.e., we have that in common). If not, I'll just ask you to follow along to see where this is all going. . .

[1] Just kidding; it's not really a technical term.

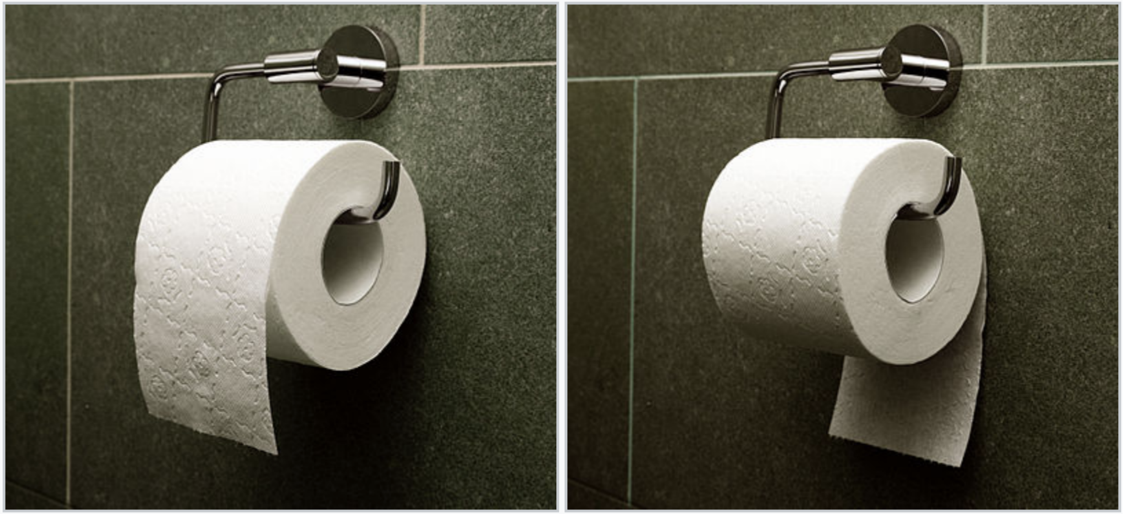


Figure 1: The great debate

Table 1: How people roll

	count
over	23
under	17

1.1 Categories, counts, and kinds

1.1.1 Two Kinds of People

“There are two kinds of people... which one are you?” questions have become something of an internet meme, particulary with the categorizations represented graphically or pictorially. There is a whole blog devoted to them by João Rocha. The images in Figure 1 probably need no explanation, as they concern the great toilet paper orientation debate.

source Wikimedia Commons User:Elya

Toilet paper orientation is a distinguishing **test question** that separates people into one of two “kinds” (or “types” or “categories”; sometimes English has several words that are used interchangeably). A fancy word for this “splitting into two” is dichotomy (die-COT-uh-mee), from the Greek. A **dichotomous question** has two possible answers. Here, you choose one way to orient the roll or the other. Let’s call this roll choice “over” (shown on left) or “under” (shown on right). Perhaps you have debated which is better with a friend or family member. Or perhaps you are lucky enough to have never thought about it at all. In any case, armed with this particular test question, we can go out and collect some data.

I went ahead and asked 40 people in Washington Square Park in New York City which kind of person they were, and the results are shown in Table 1. This being a book about data science, you might think I’m going to start calculating proportions right away, for example by saying that 57.5% of New Yorkers are over-hangers. Nope. Although you should be able to figure out that proportion conversion, it is not the point I want to focus on right now.

That point I want to focus on is that, based on our data, there *are* indeed two kinds of people here. If, for example, everyone in the world were an under-hanger (heaven forbid), then I couldn’t very well say that there were two kinds of people in this world. At least not with regard to toilet paper orientation. It would be like if I presented you with the data in Table 2. Looking at that, I can’t very well convince you that there are

Table 2: Kinds of people in Washington Square

	count
human	40
not human	0

Table 3: How people roll (alternate universe)

	count
under	40
over	0

two kinds of people.

That all seems pretty obvious, in part because I made up a *tautology* in the second example there. Being a human being is automatically associated with everyone who can be a *kind of person*.

But what if I had gotten exactly the same results for the toilet paper question? What if the data looked like Table 3. In this **alternate universe**, everyone I ask in Washington Square is an under-hanger. Yes, it's one of those scary alternate universes, like the Twilight Zone. Anyway, does that mean that there is only one kind of person when it comes to toilet paper orientation? Well… not necessarily. After all, this was just a **sample** of people in Washington Square. It was not the whole **population** of Washington Square, even, let alone New York City, let alone the world.

1.1.2 Samples and Populations

Samples and populations are sort of a big deal in statistics and data science, where these words have somewhat specialized meanings. Consider the following utterances, both of which make sense:

```
The population of New York City is 8.6 million
The population of New York City is ethnically diverse
```

Which is the population of New York City? In common usage, population often refers to the number or count of people, in a town, area, or country. Among statisticians and data scientists, population refers to a set or collection under consideration. It doesn't have to be a set of people. It could be a set of rats, non-governmental organizations, or domestic flights originating Chicago. But let's suppose the population does refer to a set of people. The number of those people is just one summary about the population, also known as the total *count*. The proportion of over-hangers is another summary of the population, as is the most-common birth-month.

If we always had access to all of the members in a population (the set or collection under consideration), the field of statistics wouldn't exist. We would just know a bunch of facts about, say, everyone in the whole world. And that would be that. While it is true that data are becoming more and ubiquitous, don't start betting on the demise of statistics. Even if we did have complete data for everyone in the world today, our population of interest might extend to the world as it will be next month, next year, or ten years from now. That is, we might want to make predictions about the future. In which case, we would want to draw *inferences* and to generalize from the data we have on hand—our sample—to data we don't have—the rest of the population. Making inferences from samples to populations will always be a compelling and challenging problem.

Since Washington Square is the center of my universe, that's where I sample. Even if we agreed that our population of interest were confined to Washington Square, we would still find it difficult to collect data on everyone there. There are a lot of them, many of them are on skateboards, and new people keep leaving and entering the park. It turns out, that's okay. We don't actually have to reach everybody to be able to do data science. However, we need to understand that when we sample 40 particular people in Washington Square, we might not get the same exact answers as if we had sampled 40 *different* people. The sampling process introduces an element of **uncertainty** into our process.

Coming back to our toilet paper debate, if we did find zero over-hangers in one sample, it doesn't guarantee that the number of over-hangers will also be zero in the next. The number may vary from sample to sample. Uncertainty does not, however, mean that the information derived from one sample is useless. In fact, soon we'll see that we can actually learn a lot from a sample simply by recognizing that sample values will vary. We can simulate samples on a computer to see how much they will vary. And then, using our simulations, we will be able to give probabilistic answers to questions like, "what are the chances that there really are no over-hangers in Washington Square?"

1.1.3 Summarizing data

When I presented my survey results to you in Table 1, notice that I did not present you with the raw data, but rather with a summary of the data. The particular summary I used was called "counts", that is, a total count of how many people responded "over" or "under." The raw data, in contrast, would have contained each individual response I collected, labled either with a name of the individual, or perhaps with some other unique identifier (such as a random number), or—if I don't need to keep track of particular individuals—with just a row number. Something like this, if we examine at the first six responses rather than all 40 of them. Raw data:

randomID	response
9246	under
1478	over
8831	under
8194	over
4178	under
4243	under

Counts is an example of a **summary statistic**, which is a fancy term for a number that is derived from the raw data. The count summary is as simple as it gets. It is literally the number of times that each response appears. We might note as well that,

```
count(under) + count(over) = total number of responses.
```

This mathematical statement is true because there are only two possible responses. If there were more than two responses, then I would need to add the counts for each possible response.

Note that the *proportion* of "over" responses is also a summary statistic (which is just the counts of "over" divided by the total number of responses). Another summary statistic could be the ratio of "over" responses (counts) to "under" responses. For example, one way people use summary statistics in reporting data is through statements like, "twice as many people prefer chunky peanut butter to smooth."

No mean feat.

Whenever someone reports a mean (another word for average) value of some set of data, that is also a summary statistic. Does it make sense to construct an average from responses that are either "over" or "under"? No, it doesn't. That's because {over, under} is a categorical response, and you can't average over categories. Unless you're trying to make fun of statistics with a puerile joke. Different versions of this joke appear: "the average American has one tit and one testicle." At the risk of explaining the joke too much, here goes: Tits and testicles can certainly be treated as numerical data, and hence can be averaged. This joke hinges on the fact that the existence of testicles (or tits) is associated with a person's sex, which is categorical and not numerical. Assuming that half of all Americans are female (roughly true), we can't say that the average American is half male and half female. The real "punch" of this joke is to suggest that summary statistics about averages are just a bunch of nonsense. What do you think?

This is about as much as we need to say about summary statistics for the time being. But they'll be back.

Table 4: How people spread

	counts
chunky	17
smooth	23

1.1.3.1 Checkpoint

While focusing on the great toilet paper debate, we've managed to establish some important fundamental ideas.

- Dichotomous questions split people into two kinds, but only as long as it is actually possible for both answers to occur.
- In an alternate universe, people might give different answers than they do in this one. (Seriously, this is an important idea).
- Even when we casually refer to *people*, we may have a particular set of people, a population, in mind. Data about this population are likely to come from a sample, rather than from the whole population, and this fact introduces some uncertainty into claims about the whole population. Data science to the rescue!
- Clearly, we can ask people questions that prompt them to choose between more than two categories. But "two types of people" questions are more fun.[2] I mean there are so many of them! So... does that mean that there really are two types of people? To answer this, we will need to get into another great debate.

1.1.4 Dimensions

> "I always said if I had one breakfast to eat before I die, it would be Wonder Bread toasted, with Skippy Super Chunky melted on it, slices of overripe banana and fresh crisp bacon."
>
> — Michael Bloomberg

Former NYC mayor Michael Bloomberg is a chunky peanut butter kind of person. Are you? As peanut butter comes in "smooth" and "chunky" varieties (also known as creamy and crunchy, respectively), this question is also a dichotomous one. However, if we add this test question to our question pool, in addition to the one about toilet paper orientation, we will soon find that having two two-kinds-of-people questions begins to imply more than two kinds of people. Wait, what?

See, back when I went to talk to the people in Washington Square, I also asked them about the great peanut butter debate. As you can see from Table 4, smooth came out slightly ahead.

But this second question did not erase the first question about toilet paper. In fact the first few rows of our data from Washington Square are displayed below. Each row, representing one person, now has two columns, labeled "roll" (for toilet paper) and "spread" (for peanut butter):

```
##    roll spread
## 1 under chunky
## 2  over chunky
## 3 under smooth
## 4  over chunky
## 5 under smooth
## 6 under chunky
```

[2] That was another unsubstantiated claim.

Table 5: Two questions

	chunky	smooth
over	10	13
under	7	10

Table 6: PB preference

	counts
chunky	13
don't care	3
hate all	4
smooth	20

You may have noticed that among the first six people for whom I have shown data, none of them answered both over and smooth. But such response pairs exist. In fact, if we count each combination as it occurs–that is, under-chunky, over-chunky, under-smooth, and over-smooth–we get the results shown in Table 5. There are four combinations, because we have two questions with two possibilities (dichotomies) for each.

Before you read on, it's a good time to ask yourself if you can answer the following questions (answers in the footnote): (a) if there were two questions with three categories each, how many combinations could be observed? (b) if there were three dichotmous questions, how many combinations could be observed?[3]

Table 5 is an example of a kind of table that is so common in data science, it has its own name. Three of them, in fact. It is sometimes called a cross table (or crosstab), or a **two-way table** (makes sense), but most commonly it is known as a **contingency table** (wha? I'll explain later in Section 1.2.) I'm sorry that there are three names for the same thing. Really I am.

Ok, now things are about to get deep. The title of this chapter is "How Many Kinds of People are There?" And we've now explored how using two two-kinds questions leads to four types. You've probably figured out yourself that you take the product (i.e., multiply) of the number of categories in each of the questions, and that tells you how many "buckets" you can have overall. But still, there are different ways to arrive at different bucket numbers.

Consider Table 6 in contrast to 5. We've now given people four choices to express their peanut butter preference. In addition to chunky and smooth, they can also choose to say that they hate all peanut butter or don't care. We now have four kinds of people. But since we make the determination of what kind of person you are using just one question, we say that there is one **dimension** (in this case, peanut butter preference) along which people can be divided into four groups. In Table 5, there were two dimensions, a dimension of peanut butter and a dimension of toilet paper. Notice that this word, dimension, is used in much the same way as when we refer to geometric space as being two-dimensional (e.g., a drawing on flat sheet) or three-dimensional (e.g., a solid object, or sometimes a drawing that creates the illusion of looking at a solid object.) The three dimensions of space are often labeled something like (x, y, z). Here, our two dimensions could be labeled (pb, tp). The order doesn't matter. To summarize, in Table 5, we have two dimensions and four kinds. In Table 6, we have *one* dimension and four kinds.

So far so good: two questions, two dimensions, right? Well… maybe. We already saw that if a question does not actually divide people into kinds, because only one answer appears, then it doesn't really count. It is not a dimension. In our contingency table representation, this might look like the left side of Table 7. In an alternate universe, no one prefers smooth to chunky. Another way to say it is that the peanut butter question is not **informative** because it has no **variance**. Everyone in our sample is the same.

[3] (a) If the categories for each question are A, B, and C, we can get AA, AB, AC, BA, BB, … etc. We multiply the number of categories as many times as we have questions. So 3*3 = 9. (b) This time we have three questions, and for each one we have two options, so there are 2*2*2=8 possible combinations.

Table 7: Two questions (alternate universes)

	chunky	smooth
over	23	0
under	17	0

	chunky	smooth
over	0	23
under	17	0

Table 8: Alternate universe (two equivalent ways)

	chunky	smooth
over	0	23
under	17	0

	over	under
chunky	0	17
smooth	23	0

But now consider the alternate universe on the right of Table 7. In that case, everyone who is an over-hanger of toilet paper prefers smooth peanut butter, and everyone who is an under-hanger prefers chunky. If this is the case, there are only two kinds of people, at least in our sample. Those who over-hang *and* prefer smooth and those who under-hang *and* prefer chunky. But does it make sense to say there are two dimensions? We did ask two different questions!

You might reason about it the following way: in our sample, if I ask anyone just one of the two questions–about either toilet paper or peanut butter–then I immediately know the answer they would give to the other one. I don't actually have to ask two questions, other than to establish in the first place that I didn't have to. Since I only get information from one question, there is only one dimenion.

1.2 Independence, Association, and Contingency

> This section title sounds like a philosophy book by the late Richard Rorty. — inner voice

We just spent a little bit of time in an alternate universe, a bizarro world in which knowing how someone prefers to orient their toilet paper tells you what style of peanut butter they like, and *vice versa*. Notice that this knowing-about relationship is symmetric, and that in fact, the two representations as shown in Table 8 are informationally equivalent.

In our regular universe, however, this relationship was not observed. In Table 5, all four possible combinations occur. When knowledge about a person's answer to one question provides information about their answer to another question, we say that the two answers are **contingent** upon one another. This is the reason we called the two-way table a contingency table in the first place, although it is still called that even when two answers are not contingent. Go figure. Contingent is another word for **dependent**. To make matters worse, we *also* often say that the two responses are **associated**.

In our bizarro world scenario, one answer completely determines the other. This **deterministic** relationship is one extreme in the spectrum of association/dependence/contingency. At the other extreme, if the two responses are not at all associated/dependent/contingent, then we say that they are **independent**. To say that two responses are independent is to assert that knowing one of them does not give you any information about what the other one might be. This would have been my intuition, at least, about toilet paper and peanut butter. But whether they are independent or mildly associated with one another is an empirical question, which means we should try to answer it with data. In bizarro world, where they were deterministically related, we might reasonably want to know why. Could there be a gene that turns on toilet paper orientation and peanut butter preference at the same time?

It is worth noting that a single dataset often can't tell us for sure whether two variables are independent, associated/dependent/contingent, or deterministic. Suppose for a moment that we saw this contingency table:

You might think, wow! It looks like toilet paper preference is associated with peanut butter preference: People who prefer chunky peanut butter also seem more likely to be over-rollers, and people who prefer smooth

Table 9: Two questions

	chunky	smooth
over	17	4
under	3	16

peanut butter are more likely to be under rollers. How weird!

Now, you intuitively know that if you go back out onto Washington Square and you find 40 different people, these numbers probably won't be exactly the same. There is some element of randomness, and it's basically impossible to know what the data would look like if you could ask every single person in the world.

Instead, statisticians like to use something called **hypothesis testing**. This often involves coming up with a **null hypothesis** (in this case, the null hypothesis might be that peanut butter choice and toilet paper rolling are independent) and an **alternative hypothesis** (in this case, the alternative hypothesis might be that peanut butter choice and toilet paper rolling preference are dependent). Then, it is often possible to simulate what might happen by random chance under the null hypothesis and check if our observations are consistent with the null or not. This can lead us to either accept the null hypothesis, or "reject the null" in favor of the alternative.

Let's first take a look at what this dataset looked like before we tabulated it. There were 40 people sampled, and they each asked two questions. Let's look at the first five rows of data:

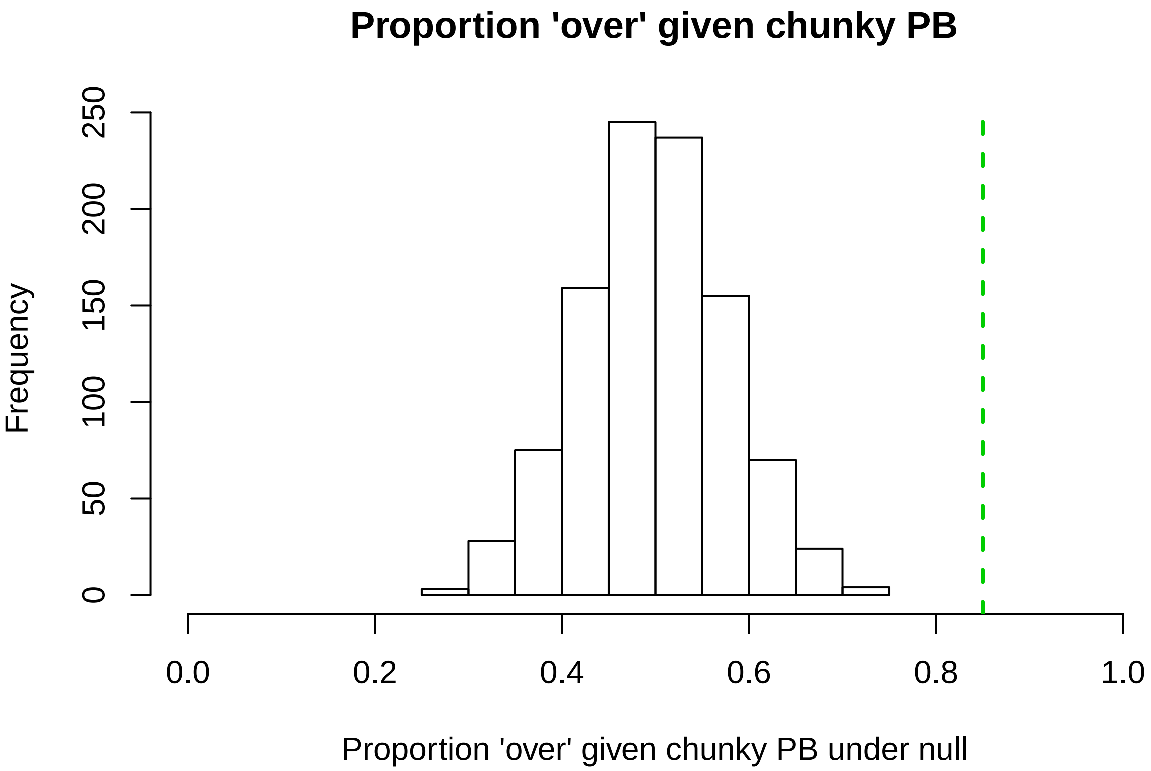
Peanutbutter	Toiletpaper
Smooth	Under
Chunky	Over
Chunky	Over
Chunky	Over
Smooth	Under

One way to think about the situation where toilet paper and peanut butter choices are independent is this: We keep the same proportions of answers to each question, but we randomly shuffle them separately. This is akin to writing every person's response to the peanutbutter question on an index card, shuffling them randomly, and then re-distributing them; then, doing the same for the toiletpaper question. If we shuffle responses to each question separately, there is no way for someone's randomly assigned answer to one question to influence their randomly assigned answer to another.

In the initial dataset, we found that 17/20 or 85% of people who preferred chunky peanutbutter also rolled their toilet paper "over". If we randomly re-shuffle people's answers 1000 times, then we can calculate the percentage of chunky peanutbutter people who roll their toilet paper "over" for each shuffled dataset. This will give us a sense for what kinds of values we might expect to see by random chance if the null were true (i.e., if these questions were actually independent). The code to do this is below (you do not need to perfectly understand it yet - we will come back to this later!).

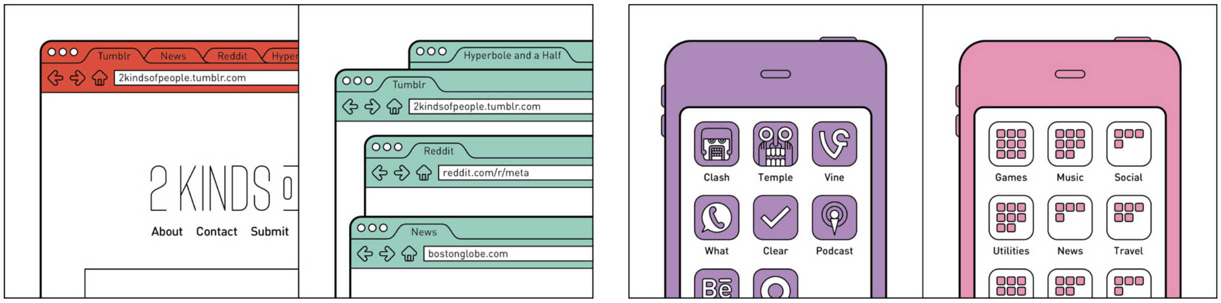
```r
simulated_proportions = vector()
for(i in 1:1000){
  data_example$Peanutbutter = sample(data_example$Peanutbutter)
  data_example$Toiletpaper = sample(data_example$Toiletpaper)
  simulated_proportions[i] = table(data_example)[1,1]/20
}

hist(simulated_proportions,
     main = "Proportion 'over' given chunky PB",
     xlab = "Proportion 'over' given chunky PB under null",
     xlim = c(0,1))
abline(v=0.85, lwd=2, col=3, lty=2)
```



We will cover histograms later in this chapter, so don't worry too much if this picture doesn't make sense to you yet. For now, the main takeaway is this: Even if we randomly reshuffle everyone's answers 1000 times (akin to going out on the street and asking 40 new people these questions 1000 different times, but assuming that the overall proportions of over vs. under rollers and chunky vs. smooth PB people will be the same as in our original dataset), we would not expect to see values as high as 85% (green dotted line) if the questions were truly independent.

1.2.1 Latent Factors and Measurement



source

Figure **??** shows two more two-kinds of people graphics from João Rocha's blog. I bet that you can identify yourself with one of the two images in each pair. I certainly can. But ask yourself, given our discussion above, do you think the choices a person would identify in each case above are independent or not independent (e.g., contingent, associated, dependent)?

In contrast to the toilet paper and peanut butter questions, which at least appear to be about totally different things, these two dichotomies have something similar going on in each of them. The choice on the left is about organizing your desktop browser, either in tabs or as separate windows. The choice on the right is

Table 10: Possible data for digital tidiness

	folders	apps		folders	apps
tabs	21	0	tabs	16	6
browser	0	19	browser	5	13

about organizing apps in your smartphone, either loose or in folders. We might say that both of them get at a tendency to organize your digital environment. Call it digitidness (short for digital tidiness). This tendency, we may imagine, might even carry over into non-digital environments, like your actual desk, bookshelf, or filing cabinet.

What we've done here is to try to explain the association between responses to the two questions (assuming that there is, i.e. that they are not independent) by appeal to some underlying **latent factor**. We say a factor is latent (meaning hidden) because we don't observe digitidiness itself directly, but we only observe tidy browsers or smartphone app folders. Perhaps you can think of another candidate factor besides digitidiness. In any case, we might propose that each of the two two-kinds questions in Figure **??** are in fact indirect **measurements** of the same factor. If so, this could explain why the two answers woudl be associated.

> Notice that a **factor** is also a dimension, in the sense we used before. We could have said "latent dimension", but we tend to use the word factor when we are drawing attention to the specific nature of the dimension rather than just counting. We also sometimes use the word **trait**. At least in psychology, trait tends to be reserved for stable psychological factors. Thus "stress" can be a factor but not a trait, whereas "social anxiety" may be a trait, if it is persistent. In this case, digitidiness might be considered a trait (and thus also a factor and a dimension).

Contrast this with toilet roll orientation, which we can observe directly just by looking in someone's bathroom. (We assume that they are telling the truth when they answered our questions, but we could in principle verify it.) It was only in the bizarro world when toilet roll orientation and peanut butter preference were perfectly related that we started to wonder if there maybe *was* an underlying genetic factor. Genetic factors were once not directly observable either, but we assumed them for explanatory value. Today we can of course observe specific genetic variation, although there are still many gaps in our understanding of the relationship between genes and observed behaviors.

Consider some data again, in two possible worlds, shown in Table 10. On the left, we have the deterministic scenario we saw before. As before, we identified this situation as having two kinds of people and really just one dimension. In contrast to before, where we had no real explanation for this coincidence, we attribute it now to some factor, like digital tidiness.

But now consider the possible results in the table on the right. Since all four possible quadrants have non-zero counts, we see that knowing whether someone organizes their browser using tabs does not completely (i.e., *deterministically*) specify whether or not they put their apps into folders. On the other hand, one answer *does seem to be associated* with the other. Notice that the values are still much higher in the "buckets" that we think of as indicating the presence or absence of digitial tidiness. These are the tabs-folders bucket (tidy) or the windows-loose bucket (not tidy). We say that the tidiness factor appears to explain much of the observed range, or **variance**, in responses to the two questions. But it doesn't explain all of it, since there are people (11 out of 40, in this case) who don't fall into one of these buckets.

This situation on the right is probably more realistic. After all, very few things in this world are absolute (unlike in bizarro world). So now the big question re-emerges: are there two kinds of people or four? One dimension, or two? It's sort of. . . like. . . in between. . . ?

Golda says: Although digitidiness explains a lot of what we see in our data, it doesn't explain it all. I believe that desktop tidiness and mobile tidiness are different, if related, tendencies. For example, when we use mobile phones, we're typically on-the-go and have less time. If we knew more about the people in our sample, we might see that these discrepancies in the organization of apps and tabs actually relate to other aspects of their lives. So, I say there are two dimensions.

Sidney says: Digitidiness is the only real factor here, but people may not always be consistent in these particular behaviors. Also some people are only sort-of-tidy, and apply this tidiness unevenly but randomly. These two-kinds of people choices don't leave room for shades of gray, so that's what we're seeing in the mixed categories where people are tidy in one environment and not in another. But ultimately there is really just one dimension here.

What do you think?

1.3 (An Infinite Number of) Shades of Gray (or Brown)

We've taken the two-kinds-of-people idea pretty far in this chapter already. But it's time to acknowledge the elephant in the room. Not every question about attributes, preferences, or behaviors can be answered in such an either/or manner. Digitidiness might be one of those things. Consider the following dialogue:

> Stacy: "There are liberal and conservative kinds of people, Trang. Which one are you?" Trang: "Well, you know I'm not sure I'm exactly one or the other. I think I'm somewhere in the middle."

Although we often use them as **discrete categories**, the words liberal and conservative might be better thought of as endpoints of **continuous scale**. In fact, they might even apply to different *dimensions* of political thought with respect to social issues or economic issues. If you think about it, it's not hard to come up with other examples of "categories" that really just describe one end or another of a continuous scale. Yes, there are short people and tall people, but everyone has a height, and a lot of people are "about average." Height is just a number on some scale. So it wouldn't necessarily make sense to put people into the categories of tall or short.

In the great toilet paper debate, we were able to identify two kinds of people based on two possible responses to the question of roll orientation. Two answers; two kinds. If instead of discrete categories, we have a number on a continuous scale, does that mean that there can't be "kinds" of people anymore? To answer this question, we'll need to understand what exactly we're talking about when we characterize people using a continuous scale.

Consider poopiness. On a scale where some people are really poopy (close to poopiness = 1), some aren't poopy at all (close to 0), and many are somewhere near the middle. That's not a very quantitative description. I used the words "some" and "most", but I didn't give you counts like I did in Table 1 about toilet roll orientation. I will try to do that in just a minute. Meanwhile, notice that scale here runs from 0 to 1, which I will also sometimes write as [0,1]. When it comes to height, we have established scales, like inches or centimeters. But when it comes to liberalism or poopiness, the scale does not necessarily refer to something we can see directly. Nevertheless we can use the scale to compare people and to see how a whole bunch of people "measure up." I've set the scale to [0,1], because it is a common scale, but it could have run from 1 to 10, for example, without significantly changing anything in what I'm about to say.

If I showed you the poopiness data for a sample of people, the list would look something like Table 11. As before, in this table each row stands for one person. To protect their identities, everyone is identified only by a number (e.g., 0083), which is shown in the first column. In the second column is each person's poopiness value.

Poopiness is shown as a decimal number. Part of the reason I've used this scale, instead of 1-100, is to emphasize that the data values can be arbitrarily close to one another. Two values may be different by 0.1 or 0.03, or even 0.000027, if we have enough precision in our data to say such a thing. These data are called **numerical** or **quantitative** as opposed to **categorical**. There are actually 148 values in the data set, but I've only shown the first six in Table 11.

It's not as easy to make sense of a bunch of decimal values like this as it is to look at simple counts of categories (like 17 for chunky, 23 for smooth). However, this sense-making problem has been solved by representing the same data using dot plots, stacked dot plots, frequency tables, and histograms, which you can read all about in any standard textbook (for example OpenIntro Statistics, Chapter 2). I'm going to go

Table 11: Don't ask me how I got these numbers.

	poopiness
0027	0.679
0083	0.473
0015	0.703
0024	0.691
0029	0.495
0014	0.833

Table 12: Frequency Table for Poopiness

Range	Frequency
0 - 0.05	0
0.05 - 0.1	0
0.1 - 0.15	0
0.15 - 0.2	0
0.2 - 0.25	2
0.25 - 0.3	9
0.3 - 0.35	12
0.35 - 0.4	9
0.4 - 0.45	9
0.45 - 0.5	14
0.5 - 0.55	14
0.55 - 0.6	9
0.6 - 0.65	19
0.65 - 0.7	9
0.7 - 0.75	13
0.75 - 0.8	16
0.8 - 0.85	7
0.85 - 0.9	5
0.9 - 0.95	1
0.95 - 1	0

straight into the **frequency table** and **histogram**, which you've probably seen before. These are the most commonly used representation for data of this kind.

Again, it is a bit awkward to count how many people have poopiness value of exactly 0.473. Maybe there is one, maybe none. How would we interpret that answer, anyway? Instead, we can group values into ranges, or "bins", e.g. 0-0.05, 0.05-0.1, 0.1-0.15, etc. and then count how many of our data fall into each bin.[4] This table of counts is typically called a frequency table. Frequency is just another word for counts.

A histogram is a bar plot of counts for poopiness values that fall into certain numerical ranges. So it's a bar plot of the data in Table 12. But oftentimes you'll just see the histogram without the frequency table.

Consider the range of poopiness values from 0.40-0.45. Our data set has 9 values in this range, as you can see in Table 12, so the height of the bar above this range of values on the x-axis (horizontal axis) is 9. I've colored it in pink only to help you see what I'm referring to. The y-axis in Figure 2 is labeled "Frequency", as in the table. Some more jargon: the numerical values that separate the bins are called "breaks." In Figure 2, the breaks are at increments of 0.05.

[4] Technically, each range is a semi-open interval, e.g. (0.1,0.15], so that any values exactly equal to 0.1 can only be included in one bin and not the ones on either side.

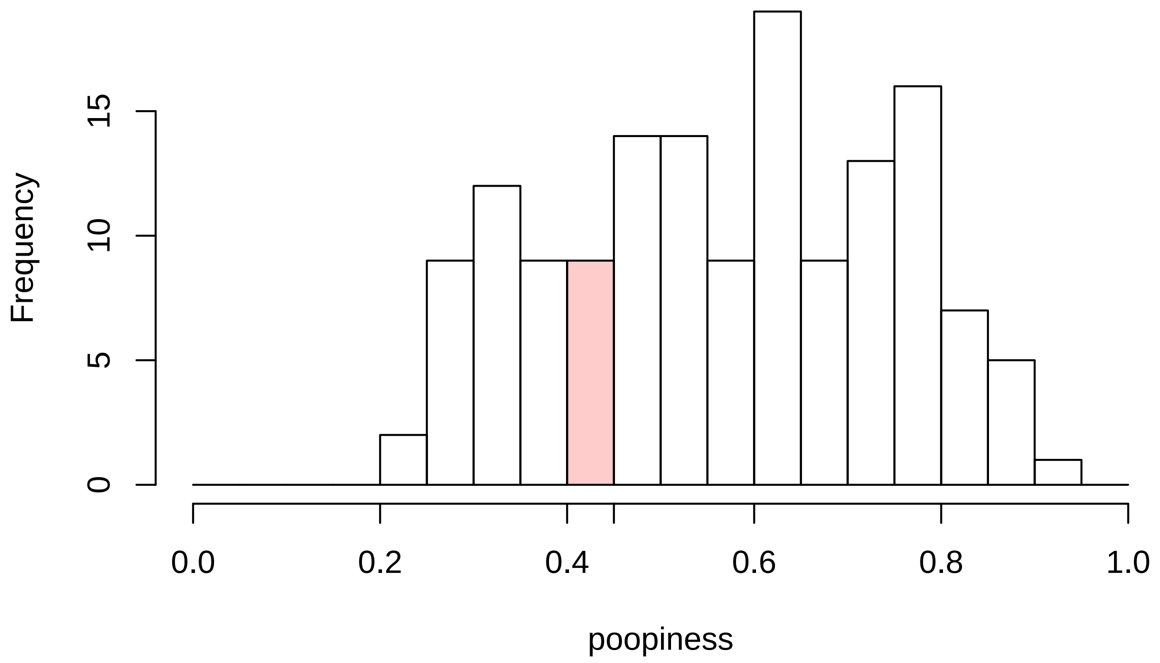


Figure 2: Histogram of Poopiness

> Question: Given that there are 20 possible bins in the histogram in Figure 2, but only some of them have non-zero counts, are there 20 kinds of people (in terms of poopiness) or 15 kinds of people?

Trick question? You bet. The breaks (and thus bins) in a histogram are arbitrary. I can choose any breaks I want, as long as all of the data points fall into exactly one bin. (I can't just exclude some bins, though. That would be cheating.) The histograms in Figure 3 are both perfectly valid histograms. One of them has four bins, and one of them has only two bins.

It's tempting to take the counts on the right of Figure 3 and declare that there are two kinds of people. After all, this gets us back to familiar territory. Ta-dah!

As you can tell, because it says so right in the figure caption, this is terrible, horrible, no-good, very-bad thing to do. Why is it a bad thing to do?

a) The split was made at 0.5 on the poopiness scale, but that is not the average value of poopiness in the data set, which is closer to 0.57, as can be seen in Figure 2 (or from the "raw" data themselves).
b) You should always use at least 5 bins when you have numerical data
c) Representations of data should communicate honestly about the nature of the data themselves. In this case, poopiness is not a category.

What I did here was take a numerical/quantitative value (poopiness) and mis-represent it as a categorical value. I did it by *dichotomizing* it, i.e., by splitting off everyone above 0.5 and labeling them as "poopy". I could have alternately split at the mean or median value and labeled the resulting two groups as "low poopiness" and "high poopiness." But this would still have been a mis-representation. It would hide the fact that poopiness comes in a continuous range of values.

> ASIDE (*delivered in a hushed voice*): I won't be able to convince you of this now, but it turns out that if you do this—if you dichotomize numerical data—you will BREAK STATISTICS! Ok, that sounds a bit dramatic. But in all seriousness, one of the jobs of statistics is to understand associations between different variables, such as poopiness and, say, earning potential. If you treat poopiness (or other variables) as discrete when they are really continuous, you may very well get the wrong answers. As the man down the street from where I used to live often muttered to himself

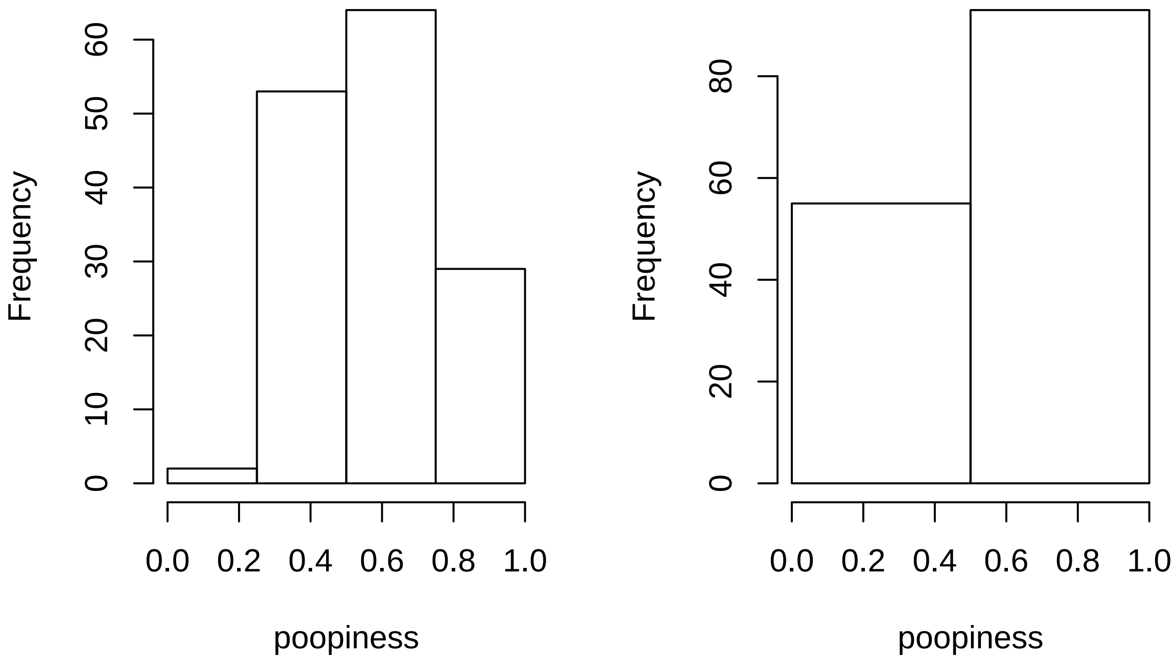


Figure 3: Other Histograms of Poopiness

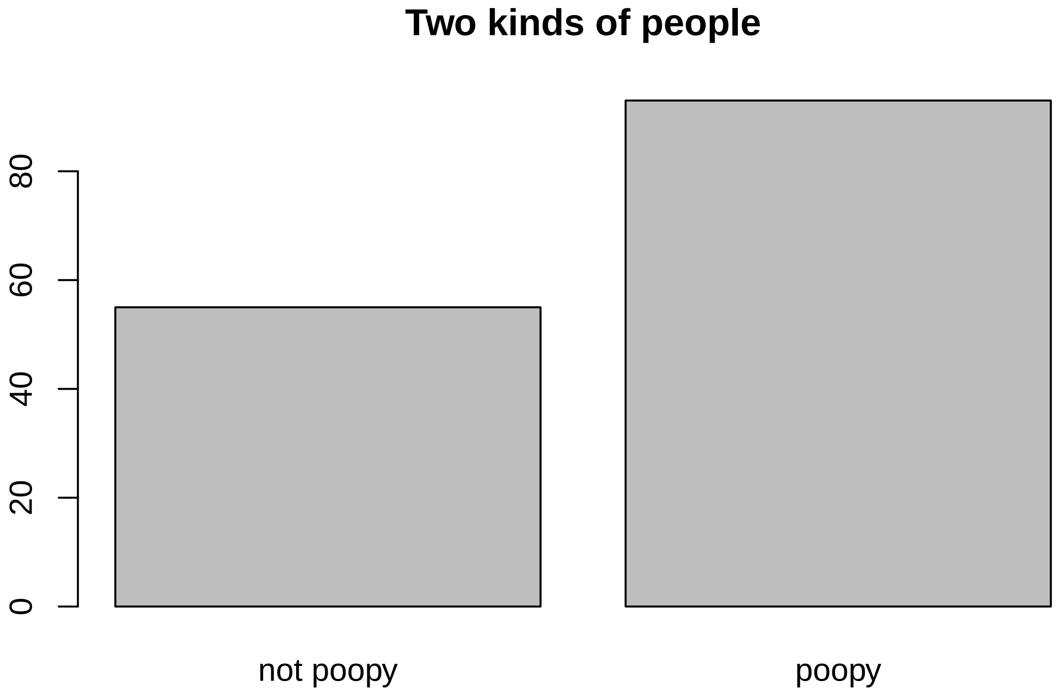


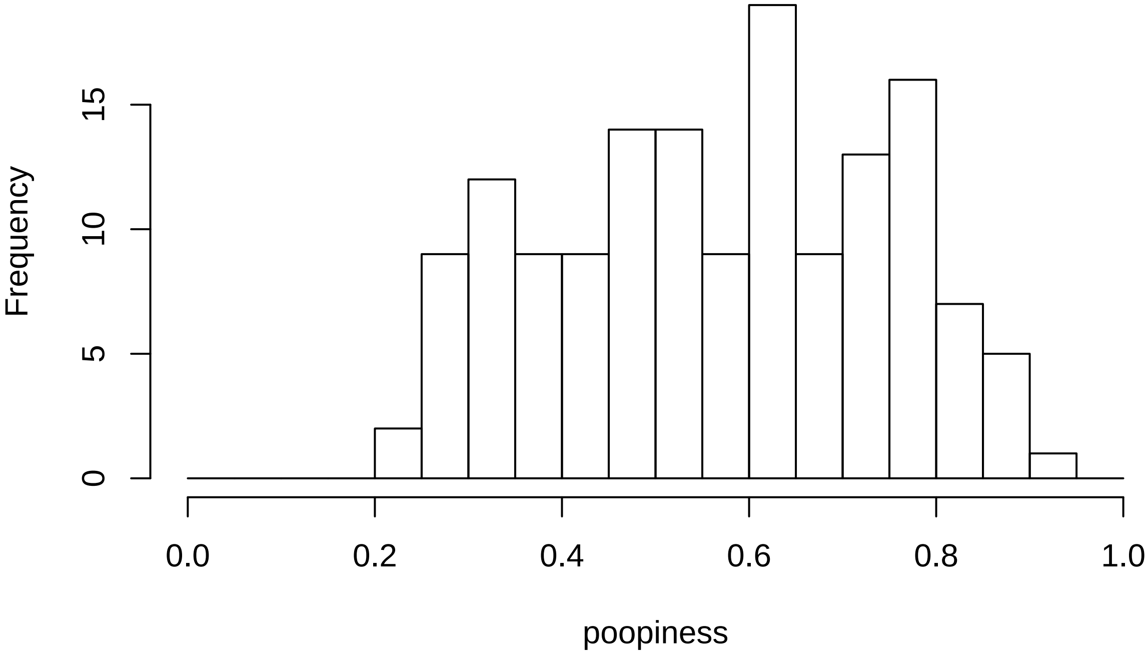
Figure 4: This is a terrible, horrible, no-good, very-bad thing to do.

> while waving his arms in the air, THAT IS AN ABSOLUTE IRONCLAD MATHEMATICAL FACT. No, but in all seriousness, there is a terrific paper on exactly this subject [@maccallum2002].

Dang it! you say. You've taken me down this rabbit hole of poopiness for too long. How many kinds of people are there? Are you saying that if one looks at properties that are described by numbers instead of categories, that there is only one kind of person? Is it all just shades of gray (or brown)?

1.3.1 Mixtures

Remember Figure 2? (Don't click it!) Here it is again so you don't have to scroll back. Data scientists like to say this picture shows you the **distribution** of poopiness in our sample. Statisticians use the word distribution in a more formal way that is best put off until we actually need it. We don't need it yet.



What if I told you that there ARE two kinds of people; you just can't see them unless I give you special glasses (or more information). If I gave you special glasses (or information), you would see this:

By what dark magic have you colorized the data! you say. Or, perhaps you just said, hm, interesting. In Figure 5, I've made a histogram with bars in two different colors, light green and pink. The colors are slightly transparent so that you can see both the green and pink distributions in their entirety even though they overlap. That's what the brownish bars mean. You're looking at the overlap of the green and pink bars, not another set of bars. Now, if you compare this histogram closely with the original, colorless histogram above, you'll see that the bin ranges are the same (width=0.05), and the the counts of green and pink bars add up to the total values that we had before. If there are green people and pink people, or in any case two different kinds of people, and if their poopiness is distributed as shown in Figure 5, then the poopiness of the mixture of these two groups of people will look just like Figure 2.

Ok, but that doesn't explain how you would know that there are two groups. If I didn't tell you. That's because *you wouldn't necessarily know. You would need to have more information.* Now you might suspect something if you saw a distribution that looked like this:

In Figure 6, the distribution has a double-hump like a Bactrian camel. In spite of that, it is not called a Bactrian distribution–which would make me happy–but a **bimodal** distribution. The point that I'm trying to make here is that a bimodal distribution makes you suspect that there could actually be two groups mixed together in our data.

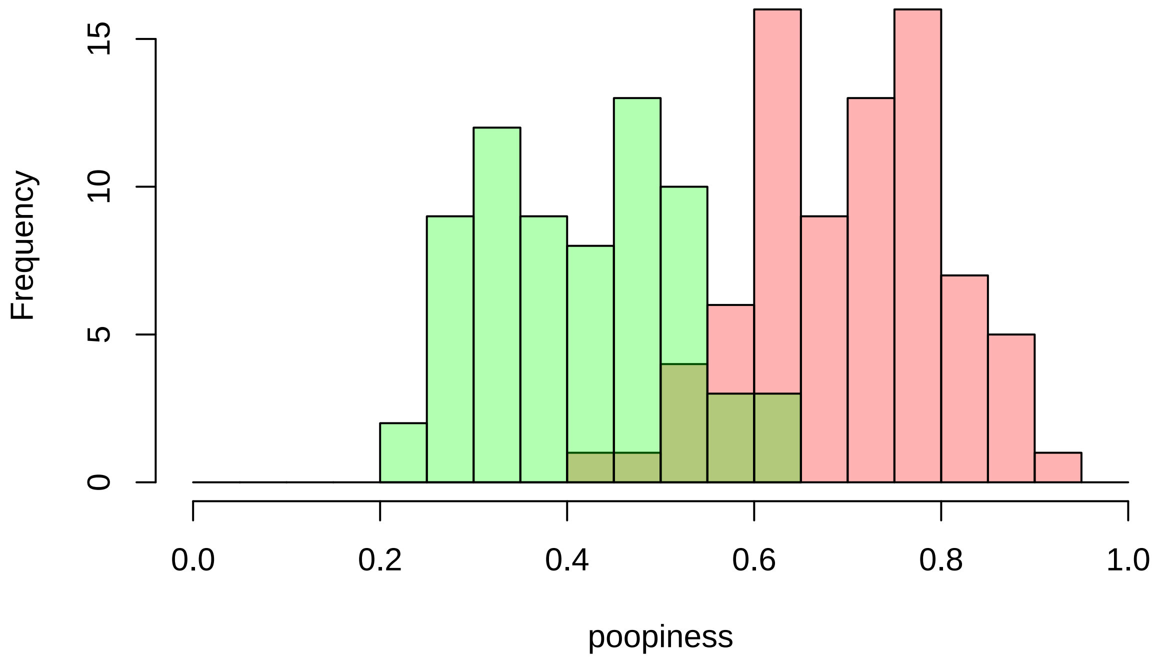


Figure 5: A mixture of poopiness

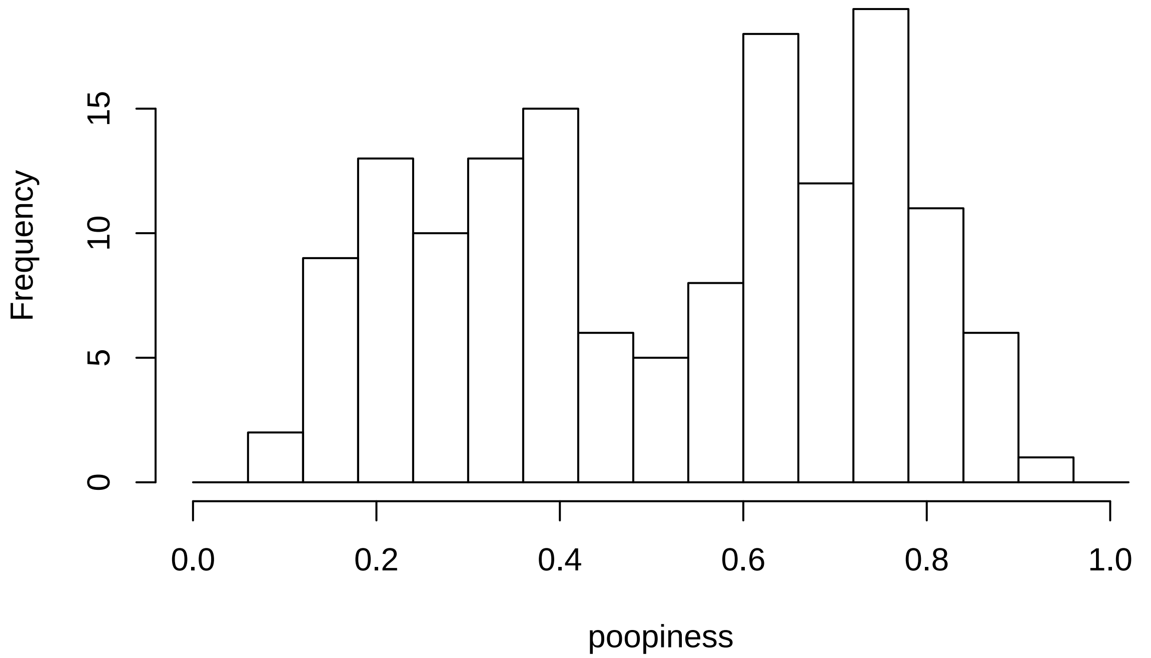


Figure 6: A suspicious mixture of poopiness

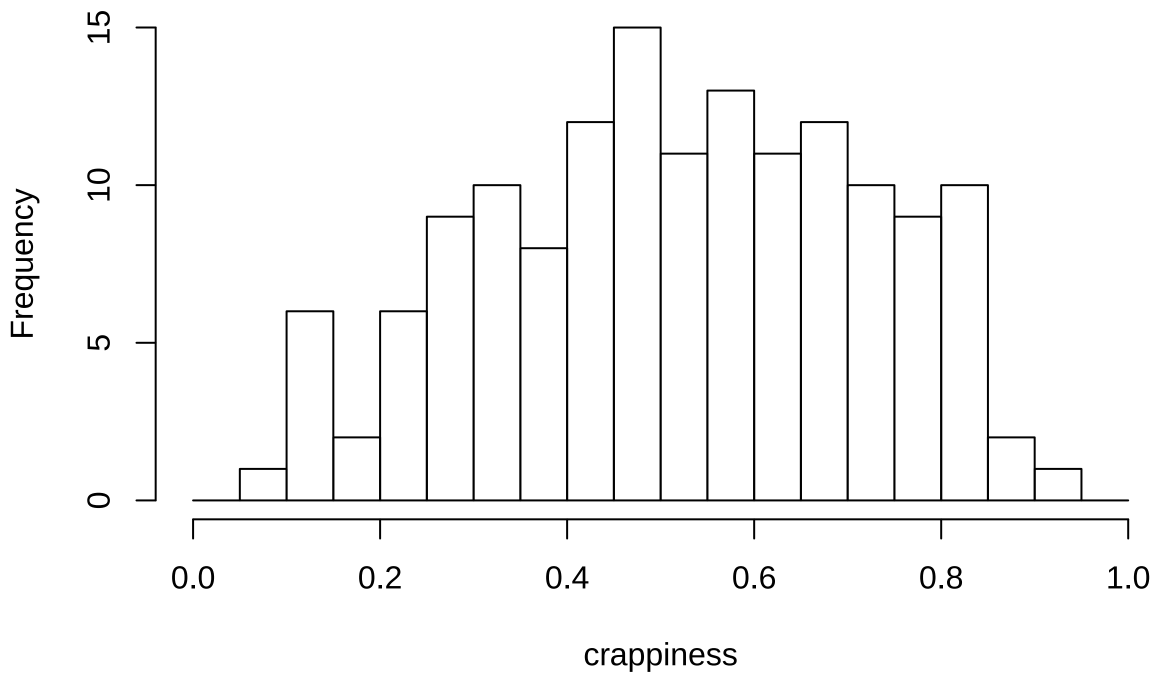


Figure 7: Histogram of Crappiness

But the original data for poopiness did not look bimodal. I suggested to you that you would need more information to determine if there are two groups. And so, I present you with… Crappiness! For each of the subjects in our poopiness data set, we have also collected data on their crappiness. Crappiness is also a numerical value ranging from [0,1]. It's sort of like poopiness, but different. Here are some values:

```
##      poopiness crappiness
## 0027     0.679      0.453
## 0083     0.473      0.627
## 0015     0.703      0.159
## 0024     0.691      0.519
## 0029     0.495      0.806
## 0014     0.833      0.147
```

And here…(drum roll please)… is a histogram of crappiness!

Hmm. I bet you were hoping that the crappiness data would look obviously bimodal, but it's not obvious. Nevertheless, hopefully you trust that I wouldn't lead you on a wild goose chase for no reason. Perhaps you can even see it coming. If we look at poopiness and crappiness separately, there is no clue that there might be distinct groups of people in our data set. But if we look at them together… there is.

When we looked at categorical data for two two-kinds-of-people questions, we made 2x2 contingency tables. We also used the word "dimension", for example to say that we were describing people along two dimensions (recall: toilet paper and peanut butter). Now that we are looking at numerical data (poopiness and crappiness), we can also use two dimensions, as in a two-dimensional scatterplot, to examine both variables at once. This scatterplot is shown in Figure 8. Each point represents data from one person, with their poopiness value on the x-axis and crappiness on the y-axis.

Alas, oh data! Your bimodal nature has revealed itself in the higher-dimensional plane!

How many kinds of people are there? When it comes to poopiness and crappiness, people exhibit a continuous range of values, so we can't neatly put them into buckets. Neither poopiness nor crappiness appear to be bimodally distributed on their own. However, when examined together, as in the scatterplot in Figure 8, a

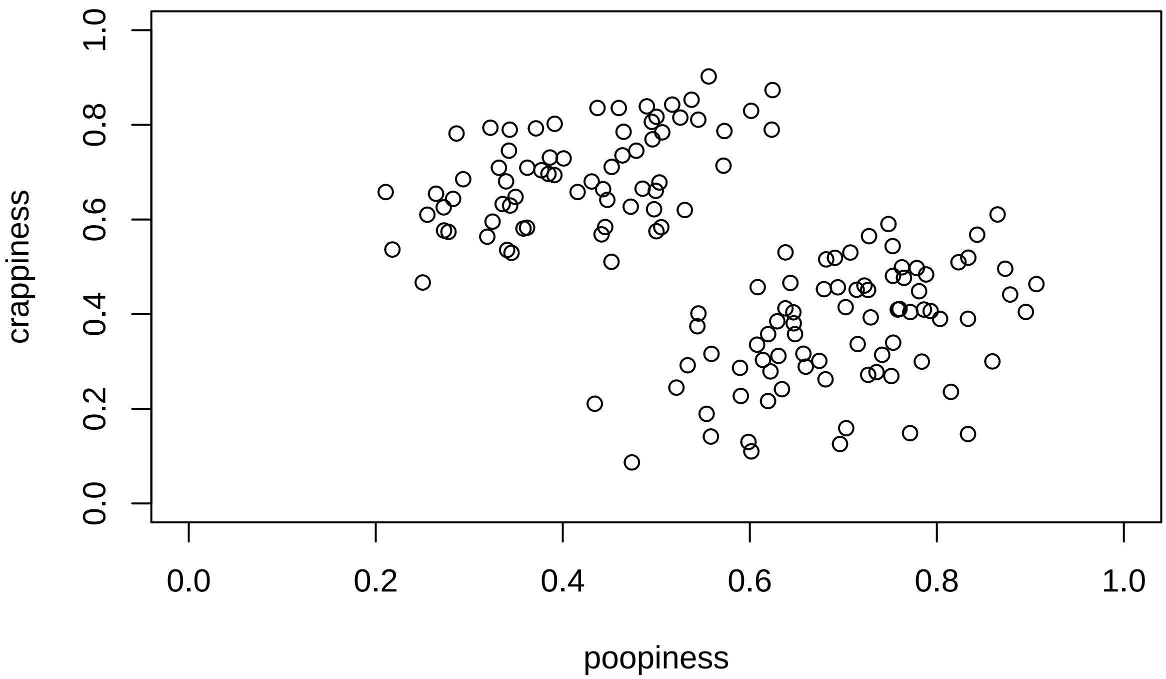


Figure 8: Scatterplot of Crappiness vs Poopiness

pretty suggestive pattern emerges in the data. There are two **clusters** of points, one group of which is lower in poopiness but higher in crappiness than the other. Interestingly, though, in both groups poopiness and crappiness tend to increase together. That is, they appear to be associated, not independent.

I do not mean to imply that clusters of points can always be found if we have data along many dimensions. That is certainly not always the case. The present example was concocted (I admit it!) to show that groups *can* emerge, even in numerical data. Cluster analysis [@kaufman2009] refers to set of data-science methods all about looking for the existence of groups in multidimensional data.

1.3.1.1 Check your understanding

1) Based on the scatterplot in Figure 8 and the grouped-by-color histogram for poopiness in Figure 5, describe what the equivalent grouped-by-color histogram for crappiness would look like. Would it look the same or different? Explain.

1.4 Cut Scores and Abnormality

> Because that's not what normal people do. — things my spouse says

If you read the aside in Section @ref{sec:shades}, you'll recall that I warned against possible negative consequences of setting arbitrary cut points to dichotomize a data set—that is, turning numerical data on a continuous scale into two categories by using a cutoff value. But now consider the following scenarios:

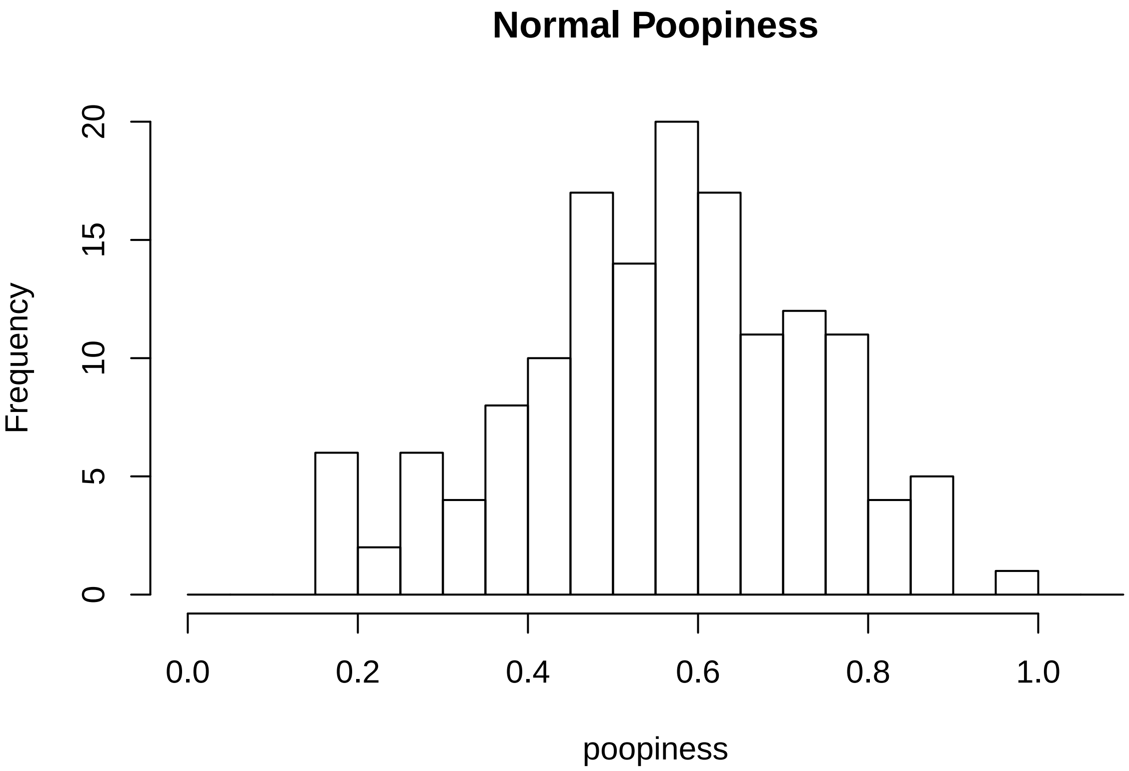
1) To pass the written test for your a driving learner's permit in California, you must answer at least 38 questions correctly out of 46. That's 82.6% correct. At 80.4% (37/46) or below, you fail and have to retake the test on another day.
2) A patient's blood test shows levels of ALT (alanine aminotransferase) at 77 units per liter. The lab report labels this as "abnormally" high, and the physician is concerned about possible liver damage or disease.

These two examples involve just the kind of dichotomization that I cautioned against, and yet they occur very commonly in practice. So what gives? Is it wrong to use cutoffs this way? Why do people do it?

The short answer is that we often find ourselves in need of a classification (pass or fail; diagnose liver disease or not) but without a perfect classification device. Rather we have only indirect measurements (of knowledge or liver function) in some quantitative measure. Perhaps you once found yourself on the “border” between letter grades for a course and were particularly perturbed (or relieved) by the imperfections of such a system. Or you may have found yourself with “slightly” abnormal levels in a blood test and wondered whether you should seek further tests.

Both the California department of motor vehicles and the physician in our scenarios need to make a decision based on imperfect evidence. They want to be able to say that the person’s test results show that they are ready to get behind the wheel of a car, in one scenario, or suffering from liver problems in the other. But all they can really do is express this belief using a **probability**. This probabilistic judgement is based on a mathematical **model** that relates traits like readiness-to-drive or liver-disease to certain test results. Understanding how these models come into existence is one of the learning objectives of this course.

The term **normal distribution** arose in statistics because the particular bell-shaped distribution occurs so frequently. If poopiness were normally distributed in our sample from before it might look like this.



Technically speaking, all of the values, including the maximal value of 0.962 that we observe in Figure **??** are normal. Poopiness varies in the population. It is impossible to be abnormally poopy, under the circumstances. By definition. some values at the extreme ends of a normal distibution are less likely to occur than values in the middle. But still they may occur rarely. It is only when extreme values (large or small) are associated with other conditions of interest, such as the relationship between elevated ALT and liver disease, that it makes sense to “flag” these extreme values.

Summary

1.4.0.1 Vocabulary

- kind, type, category

- dichotomous
- crosstab, two-way table, contingency table
- association, contingency, dependence
- latent factor, dimension, trait
- measurement, model, bimodal, cluster

We started out this chapter on a quest to answer our first big question: How many kinds of people are there? En route, we have examined both categorical data, such as from two-kinds-of-people questions, and numerical values like poopiness. The toilet paper and peanut butter orientation questions may seem silly and inconsequential to you. I can only imagine what you might think of the poopiness and crappiness dimensions that I completely made up (I admitted it!). However, in the next section, we will see that when it comes to personality psychology, there are real-world analogues of the discrete/categorical and continous/numerical multi-dimensional descriptions of people that we just saw.

1.5 Sixteen Personalities or Five Factors ?

Before you read this section, you might want to go ahead and take one of the personality tests based on the Meyers-Briggs Type Indicator (MBTI) categories and/or the five-factor model of personality (also called the Big Five). There is only one "official" MBTI, which is a commercial product. However, there are several free alternatives online which use the same typology classification. There are also several variations of

> Test yourself:
>
> - MBTI-style at 16personalities.com or here
> - Big Five here or here or here. General information about these test items.

I will only minimially describe the MBTI and the Five Factor Model (FFM, or Big Five) here, in terms of the topics we have been discussing. There are many resources for learning more about these personality tests. Some are referenced under further reading.

1.5.1 MBTI

The MBTI will categorize people, based on their responses, dichotomously along each of four dimensions, also called "scales." These are:

- Extraversion-Introversion (E-I)
- Sensation-Intuition (S-N)
- Thinking-Feeling (T-F)
- Judging-Perceiving (J-P)

Thus there are sixteen possible combinations, for example "INTP". Each person is assigned to one of these sixteen personalities. Many online tests will provide you with a report to help interpret your classification. That is, the four dimensions are understood to come together in some holistic picture of your "type."

1.5.2 Big Five

The term "Big Five" is a commonly used term for the five-factor model of personality. Based on responses to questionnaires, people are assigned a numerical score along five dimensions (also called scales or factors!)

- Neuroticism refers to the tendency to experience negative feelings.
- Extraversion is marked by pronounced engagement with the external world.
- Openness to Experience describes a dimension of cognitive style that distinguishes imaginative, creative people from down-to-earth, conventional people.
- Agreeableness reflects individual differences in concern with cooperation and social harmony. Agreeable individuals value getting along with others.

- Conscientiousness concerns the way in which we control, regulate, and direct our impulses.

  Fun fact: both OCEAN and CANOE are mnemonic devices that can help you recall the names of the Big Five dimensions.

Since the results of a Big Five test, such as the IPIP-NEO, are five numbers, you don't get assigned a personality "type" by these tests. Rather, you may be provided with an explanation of what it means to score high (or low) on, say, Extraversion. You may have noticed that extraversion (occasionally spelled "extroversion") appears on both the MBTI and the Big Five.

1.5.3 Twenty Questions (about Extraversion)

Suppose, for whatever reason, we want to identify a person's extraversion. We may want either (a) to classify them as extraverted or not (i.e., introverted), or (b) to quantify a degree of extraversion, say on a scale of 0-100. Why not then just pose the question in the following way. In the first case:

a) Choose the one that describes you: Extraverted | Introverted

or, in the second case,

b) Identify yourself on the following scale: Extraversion 0 – + – + – + – + – 50 – + – + – + – + – 100

Personality tests, such as those we've discussed above, do not ask questions like these. Rather, they include many different questions, sometimes twenty or even more, about things like going to parties, making friends, and drawing attention to oneself.

Why ask twenty questions instead of just one? Recall from the great toilet paper debate (Section 1.1) that no one ever felt it was necessary to ask twenty questions to know whether you were an over-hanger or an under-hanger. However, when we discussed digitidiness (Section 1.2.1), we suspected that two different questions may have both been getting at the same latent factor. The situation here, in the real-life domain of personality testing, is similar.

Psychologists ~~believe that extraversion is an underlying factor~~ invented the idea of extraversion to explain patterns of behavior, including patterns of responses to questions about how people feel in various situations. Such as enjoyment or lack thereof in being the center of attention. The use of indirect evidence such as questionnaire responses to make inferences about psychological traits is the main task of **psychological measurement** or psychometrics. The main challenge of psychometrics, perhaps even the reason for its existence, is that human beings are noisy.

Put another way, you cannot expect a deterministic relationships between how a person feels or acts in one situation and how they act in another. An extraverted person is not *always* extraverted. And an extraverted person might not always answer questions about their feelings in the same way. It is hard to observe or even self-report on extraversion directly, because extraversions manifests itself differently at different times and in different contexts. Whether this noise is due to some mysterious internal process, like a coin flip in your brain, or due to many unaccountable external factors, like whether you slept poorly that day, we can't say. What we can say is that human noisiness manifests itself as **measurement error**.[5] We can also say that, in spite of measurement error, some patterns do remain.

1.5.4 But what's the point?

Trying to describe people in terms of kinds or numerical scales is complicated. Why do we even bother? It's tempting to say that we just want to understand ourselves better, and that is certainly a reasonable answer. Sometimes, though, we want to predict how someone will act in the future, perhaps in a situation that differs from one that they have faced in the past. In that case, we can't exactly use the past to predict the future,

[5] The word **error** makes it sounds like there is a right answer, and that tests get it wrong. This is, indeed, one view. However, you don't have to believe that there is a right answer. For example, you can believe that human beings have some amount of inherent unpredictability.

unless we do so by making inferences about underlying traits from past behaviors and then predicting how someone with those particular traits would act in a new context. This purpose drives some uses of tests based on the MBTI and the Big Five, for example by employers or career counselors. However, although the MBTI is oftern used for these purposes, one should exercise caution in doing so [@pittenger1993]. You should certainly not assume that all personality tests do an equally good job of providing information for the desired inferences.

According to the standards of the American Psychological Association [@american1999], whenever psychological tests are used for some specific purpose (e.g., employment, admission to a school or hospital, or even in court) there must be a valid argument for the intended purpose of the test scores. This **validation argument** will usually involve many facets, including how consistent the results of the test are, whether it is a fair test for all groups of people, whether test scores really are associated with relevant outcomes in the domain of use, and so on. These arguments, and challenges to them, are all part of validity.

Still meta after all these pages

I'd like to point out that we still haven't even once asked a question of the form "what proportion of people. . . " or "what is the probability that a person. . . ?" There's nothing wrong with these questions, and we will have plenty of time to investigate questions of this kind in the remaining chapters about other Big Questions. But to be true to our question about how many kinds of people are there, we didn't need to know all of the specifics. Our discussion has been rather *metaphysical*, in the sense that we have tried to understand how differences that we observe among people can be expressed in terms of kinds (categories) or numbers, which are different *kinds* of data. Whoa.

Exercises

- Come up with a population (the members do not have to be human beings) and three *creative* summary statistics that can be derived about it (i.e., go beyond average weight).