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.

Table 4: How people spread

	counts
chunky	17
smooth	23

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
```

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: Two questions

	chunky	smooth
over	10	13
under	7	10

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) I'm sorry that there are three names for the same thing. Really I am.

[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.

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.

Table 6: PB preference

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

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.

Table 7: Two questions (alternate universes)

	chunky	smooth		chunky	smooth
over	23	0	over	0	23
under	17	0	under	17	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.