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.

Table 8: Alternate universe (two equivalent ways)

	chunky	smooth
over	0	23
under	17	0

	over	under
chunky	0	17
smooth	23	0

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:

Table 9: Two questions

	chunky	smooth
over	17	4
under	3	16

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 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:

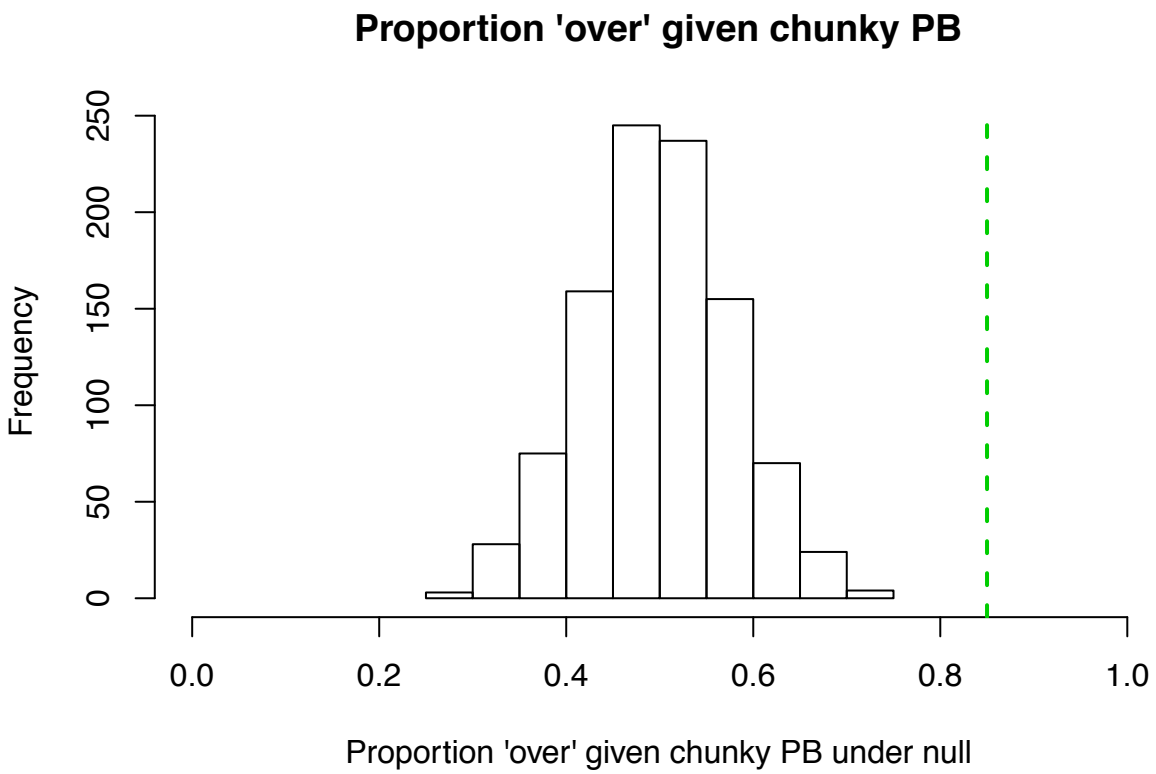
Peanutbutter	Toiletpaper
Smooth	Under
Chunky	Over
Chunky	Over
Chunky	Over
Smooth	Under

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!).

```
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.