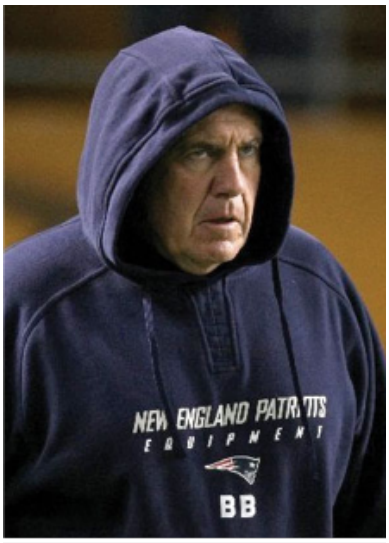
7
Finding anomalies (hypothesis testing)

Assessing the evidence for a hypothesis



Among professional football fans, the New England Patriots are a polarizing team. Their fan base is hugely devoted, probably due to their long run of success over more than a decade. Many others, however, dislike the Patriots for their highly publicized cheating episodes, whether for deflating footballs or clandestinely filming the practice sessions of their opponents. This feeling is so common among football fans that sports websites often run images like the one at right (of the Patriots' be-hoodied head coach, Bill Belichick), or articles with titles like "11 reasons why people hate the Patriots." Despite—or perhaps because of—their success, the Patriots always seems to be dogged by scandal and ill will.

But could even the Patriots cheat at the pre-game *coin toss*?

Believe it or not, many people think so! That's because, for a stretch of 25 games spanning the 2014-15 NFL seasons, the Patriots won 19 out of 25 coin tosses—that's a 76% winning percentage. Needless to say, the Patriots' detractors found this infuriating.

But before turning to religion, let's take a closer look at the evidence. Just how improbable is this anomaly? More specifically, how likely is it that one team could win the pre-game coin toss at least 19 out of 25 times, assuming that there's no cheating going on?

This question can be answered directly using probability theory. But it's even easier to answer using the Monte Carlo method. The Monte Carlo method, also known as Monte Carlo simulation, involves a computer program that simulates a random process (like, in this case, a sequence of 25 coin tosses). To use the Monte Carlo method to calculate the probability of some event A, we repeatedly simulate the random process many times, and each time we ask, "Did A happen?" The Monte Carlo estimate for $P(A)$ is the frequency with which A happens across all the simulations. So here,

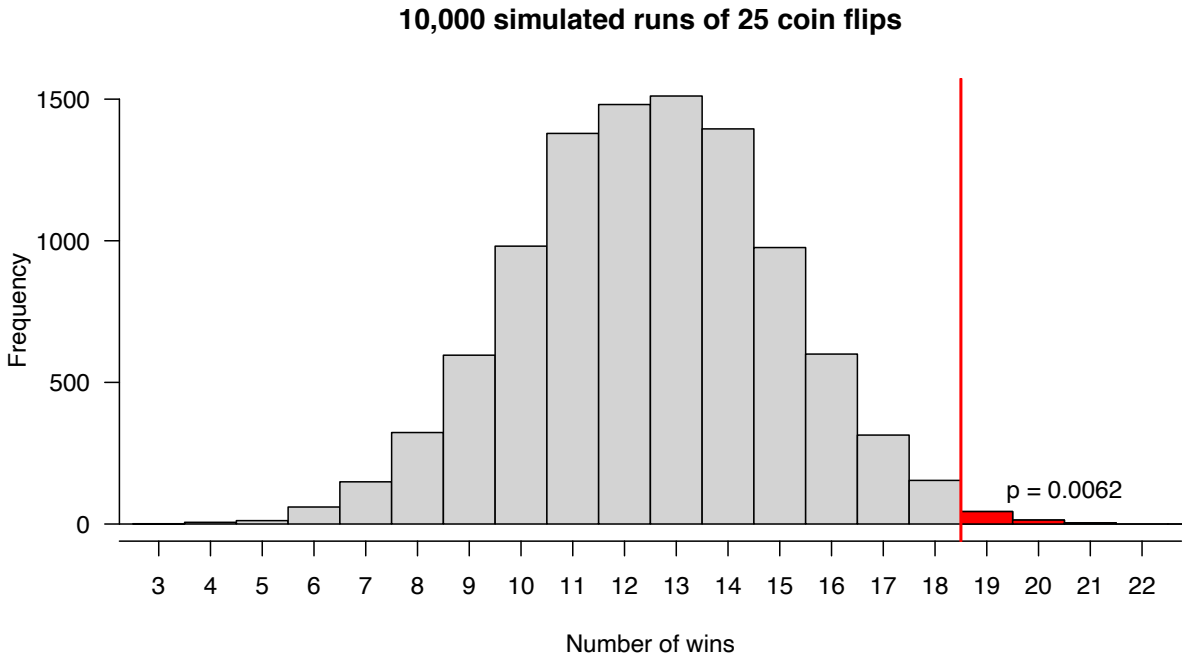


Figure 7.1: This histogram shows the results of a Monte Carlo simulation, in which we count the number of wins in 25 simulated coin flips over 10,000 different simulations. The red area (which has cumulative probability of 0.0062) approximates the probability of winning 19 or more flips, out of 25.

for example, we might run 10,000 Monte Carlo simulations, where each simulation is a sequence of 25 coin tosses. For each simulated sequence, we'd ask "Did the Patriots win 19 or more coin flips?"

In Figure 7.1, we see the results of precisely this Monte Carlo simulation for pre-game NFL coin tosses, assuming that the Patriots actually have a 50% chance of winning each toss. Specifically, we have repeated the following simple process 10,000 times:

1. Simulate 25 coin tosses in which the Patriots have a 50% chance of winning each toss.
2. Count how many times out of 25 that the Patriots won the toss.

If you're counting, that's 250,000 coin tosses: 10,000 simulations of 25 tosses each. Figure 7.1 shows the results: a histogram of the number of coin tosses won by the Patriots across our 10,000 simulations. Clearly 19 wins is an unusual, although not impossible, number under this distribution: in our simulation, the Patriots won at least 19 tosses only 62 of 10,000 times (probability $p = 0.0062$), shown as the red area in Figure 7.1.

So did the Patriots win 19 out of 25 coin tosses by chance? Well,

nobody knows for sure—I report, you decide.[1] But unless you're a hard-core NFL conspiracy theorist, let me encourage you to put aside the Patriots for a moment and focus instead on the process we've just gone through. This simple example has all the major elements of *hypothesis testing*:

(1) We have a *null hypothesis* H_0, that the pre-game coin toss in the Patriots' games was truly random.

(2) We use a *test statistic*, number of Patriots' coin-toss wins, to measure the evidence against the null hypothesis. It helps to have a letter to denote this test statistic as a general rule, so we'll use the Greek letter Δ for this purpose: Δ is our test statistic, in this case the number of coin toss wins.

(3) There is a way of calculating the probability distribution of the test statistic Δ, assuming that the null hypothesis is true. Here, we just ran a Monte Carlo simulation of coin flips, assuming an unbiased coin. We denote this distribution as $P(\Delta \mid H_0)$: that is, the probability distribution of our test statistic Δ, assuming that H_0 is true.

(4) Finally, we used this probability distribution $P(\Delta \mid H_0)$ to assess whether the null hypothesis looked believable in light of the data.

All hypothesis testing problems have these same four elements. Usually the difficult part is Step 3: calculating the probability distribution of the test statistic, assuming that the null hypothesis is true. The essence of the problem is that, in most real problems, we can't just run a simple simulation of coin flips. We'll have to work a bit harder when we revisit the idea of hypothesis testing in more complex settings.

There's also one subtle and potentially confusing point about the notation for hypothesis tests, and we want to make sure this point is clear before proceeding. It concerns the use of the letter Δ to refer to the test statistic. To understand hypothesis testing, you really have to under this test statistic in two parallel ways:

1. There's the actual value of the test statistic in the real world. In our example, that's 19 wins in the coin toss.
2. But then there's also all the different *possible* or *hypothetical* values of the test statistic that we *might* have gotten in a world where the null hypothesis is correct. In our example, these are shown in the histogram in Figure 7.1.

[1] Despite the small probability of such an extreme result, it's hard to believe that the Patriots cheated on the coin toss, for a few reasons. First, how could they? The coin toss would be extremely hard to manipulate, even if you were inclined to do so. Moreover, the Patriots are just one team, and this is just one 25-game stretch. There are 32 NFL teams, so the probability that *one* of them would go on an unusual coin-toss winning streak over *some* 25-game stretch over a long time period is a lot larger than the number we've calculated. Finally, after this 25-game stretch, the Patriots reverted back to a more typical coin-toss winning percentage, closer to 50%. The 25-game stretch was probably just luck.

It's really, really important to distinguish these two: that's because the whole point of hypothesis testing is to determine whether your observed test statistic is consistent with the kinds of test statistics you might expect to see under the null hypothesis. So here's the convention we'll use: Δ refers to the test statistic as a hypothetical entity under the null hypothesis, whereas Δ_{obs} refers to the actual test statistic you observed in the real world. Thus the histogram in Figure 7.1 shows the distribution of possible values for Δ, whereas the red vertical line in that figure marks the observed value of $\Delta_{\text{obs}} = 19$.

Using and interpreting p-values

There's one obvious question we haven't really answered. How do we accomplish step (4) in the hypothesis test? That is, how can we measure whether the observed statistic for your data is consistent with the null hypothesis?

The typical approach here is to compute something called a *p-value*. Although we didn't call it by the name "p-value," this is exactly what we did for the Patriots' coin-flipping example.

Let's begin with a concise definition of a p-value, before we slowly unpack the definition (which is dense and non-intuitive). *A p-value is the probability of observing a test statistic as extreme as, or more extreme than, the test statistic actually observed, given that the null hypothesis is true.* The way to compute the p-value is to calculate a *tail area* indicating what proportion of the sampling distribution, $P(\Delta \mid H_0)$, lies at or beyond the observed test statistic.

Using p-values has both advantages and disadvantages. The main advantage is that the p-value gives us a continuous measure of evidence against the null hypothesis. The smaller the p-value, the more unlikely it is that we would have seen our data under the null hypothesis, and therefore the greater the evidence the data provide that H_0 is false.

The main disadvantage is that the p-value is hard to interpret correctly. Just look at the definition—it's pretty counterintuitive! To avoid having to think too hard about what a p-value actually means, people often take $p \leq 0.05$ as a very important threshold that demarcates "significant" ($p \leq 0.05$) from "insignificant" ($p > 0.05$) results. While there are some legitimate reasons[2] for thinking in these terms, in practice, the $p \leq 0.05$ criterion can feel pretty silly. After all, there isn't some magical threshold at which a result becomes important: in all practical terms, $p = .049$ and

[2] If you are interested in these reasons, you should read up on the Neyman–Pearson school of hypothesis testing.

$p = .051$ are nearly identical in terms of the amount of evidence they provide against a null hypothesis.

Because of how counterintuitive *p*-values are, people make mistakes with them all the time, even (perhaps especially) people with Ph.D's quoting *p*-values in original research papers. Here is some advice about a few common misinterpretations:

- The *p*-value is *not* the probability that the null hypothesis is true, given that we have observed our statistic. Remember, the *p*-value assumes that the null hypothesis is true.
- The *p*-value is *not* the probability of having observed our statistic, given that the null hypothesis is true. Rather, it is the probability of having observed our statistic, *or any more extreme statistic*, given that the null hypothesis is true.
- The *p*-value is *not* the probability that your procedure will falsely reject the null hypothesis, given that the null hypothesis is true.[3]

The moral of the story is: always be careful when quoting or interpreting *p*-values. In many circumstances, a better question to ask than "what is the *p*-value?" is "what is a plausible range for the size of the effect?" This question can be answered with something called a confidence interval, which we'll turn to in the very next chapter.

[3] To get a guarantee of this sort, you have to set up a pre-specified rejection region for your *p*-value (like 0.05), in which case the size of that rejection region—and not the observed *p*-value itself—can be interpreted as the probability that your procedure will falsely reject the null hypothesis, given that the null hypothesis is true. As above: if you're interested, read about the Neyman–Pearson approach to testing—totally optional here.