(1) We have a *null hypothesis*, 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.

(3) There is a way of calculating the probability distribution of the test statistic, assuming that the null hypothesis is true. Here, we just ran a Monte Carlo simulation of coin flips, assuming an unbiased coin.

(4) Finally, we used this probability distribution 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 cases, we can't just run a simple simulation of coin flips. Luckily, there is a very general way of proceeding here, called the permutation test, which we will now learn about.

Permutation tests

Is gun violence correlated with gun policy?

GUN policy is an important and emotionally charged topic in 21st-century America, where gun violence occurs with far higher frequency than it does in other rich countries. Many people feel strongly that certain types of guns, like military-style assault weapons, should be banned, and that all gun purchases should be subject to stronger background checks. Others view gun ownership as both an important part of their cultural heritage and a basic right protected by the U.S. Constitution. Like with many issues, there seems to be little prospect of a national consensus.

Both gun laws, and the likelihood of dying violently as a result of gun crime, vary significantly from state to state. Figure 7.2 shows some of this variation in a *chloropleth map*, where discrete areas on the map are shaded according to the value of some numerical variable. Notice that the states are shown as a gridded tile of equal-sized hexagons, rather than as an actual map of the United States. This is common technique used to avoid the visual imbalances due to large differences in the states' total area.

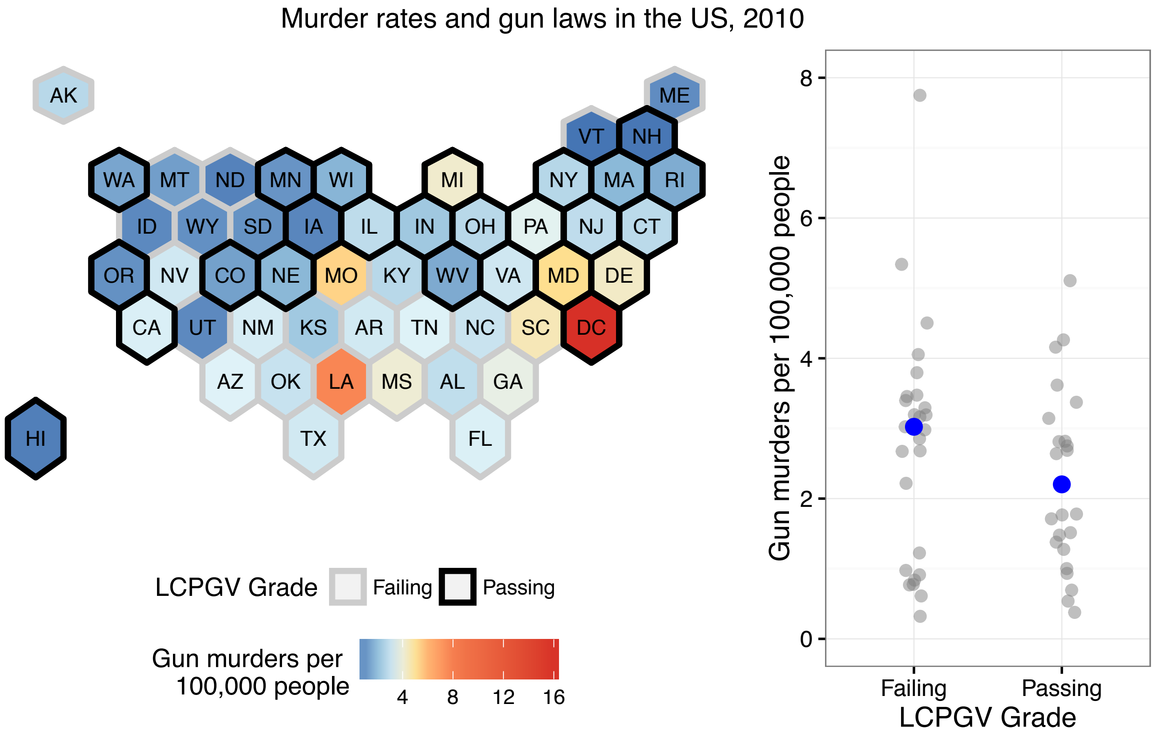

Figure 7.2: Left panel: a chloropleth map of murder rates versus gun laws across the U.S. states. The shaded color shows the state's gun-murder rate; blue is lower, and red is higher. The outline indicates whether a state's gun-control laws received a passing or a failing grade from the Law Center to Prevent Gun Violence (black for passing, grey for failing). The right panel shows a dot plot of the gun-murder rates across the two groups, together with the median for each group in blue. Washington (D.C.), at 16.2 gun murders per 100,000 people, is far off the top of the plot, but is still included in all calculations. According to its website, `http://smartgunlaws.org`, the LCPGV is "a national law center focused on providing comprehensive legal expertise in support of gun violence prevention and the promotion of smart gun laws that save lives." You can read a full description of the methodology used to grade states at this link.

In the chloropleth map in Figure 7.2, the fill color indicates each state's gun-murder rate in 2010: blue is lower, red is higher. The outline color indicates whether a state's gun-control laws received a passing or failing grade from the Law Center to Prevent Gun Violence (LCPGV). The center graded each state's gun laws on an A–F letter-grade scale; here "failing" means a grade of F. In the figure, a black outline means a passing grade, while a grey outline means a failing grade.

The right panel of Figure 7.2 summarizes the relationship between gun laws and gun violence via a dot plot, together with the median for each group in blue. We use the median rather than the mean to estimate the center of each group, because the median is more robust to outliers; a clear example of an outlier here is Washington (D.C.), which at 16.2 gun murders per 100,000 people has a drastically higher rate than everywhere else in the country.

This dotplot shows that the median murder rate of states with a failing gun-laws grade is 3 murders per 100,000 people, while the median murder rate of states with a passing grade is 2.2 per

100,000. On the face of it, it would seem as the states with stricter gun laws have lower murder rates.

Let's aside for a moment the fact that correlation does not establish causality. We will instead address the question: could this association have arisen due to chance? To make this idea more specific, imagine we took all 50 states and randomly divided them into two groups, arbitrarily labeled the "passing" states and the "failing" states. We would expect that the median murder rate would differ a little bit between the two groups, simply due to random variation (for the same reason that hands in a card game vary from deal to deal). But how big of a difference between these two groups could be explained by chance?

Null and alternative hypotheses

Thus there are two hypotheses that can explain Figure 7.2:

(1) There is no systematic relationship between murder rates and gun laws; the observed observed relationship between murder rates and gun laws is consistent with other unrelated sources of random variation.

(2) The observed relationship between murder rates and gun laws is too large to be consistent with random variation.

We call hypothesis 1 the *null hypothesis*, often denoted H_0. Loosely, it states that nothing special is going on in our data, and that any relationship we thought might have existed isn't really there at all.[2] Meanwhile, hypothesis 2 is *alternative hypothesis.* In some cases the alternative hypothesis may just be the logical negation of the null hypothesis, but it can also be more specific.

In the approach to hypothesis testing that we'll learn here, we don't focus a whole lot on the alternative hypothesis.[3] Instead, we set out to check whether the null hypothesis looks plausible in light of the data—just as we did when we tried to check whether randomness could explain the Patriots' impressive run of 19 out of 25 coin flips won.

A permutation test: shuffling the cards

In the Patriots' coin-flipping example, we could easily simulate data under the null hypothesis, by programming a computer to repeatedly flip a virtual coin and keep track of the winner. But of course, most real-life hypothesis-testing situations don't involve

[2] "Null hypothesis" is a term coined in the early twentieth century, back when "null" was a common synonym for "zero" or "lacking in distinctive qualities." So if the term sounds dated, that's because it is.

[3] Specifically, this approach is called the *Fisherian* approach, named after the English statistician Ronald Fisher. There are more nuanced approaches to hypothesis testing in which the alternative hypothesis plays a major role. These include the Neyman–Pearson framework and the Bayesian framework, both of which are widely used in the real world, but which are a lot more complicated to understand.

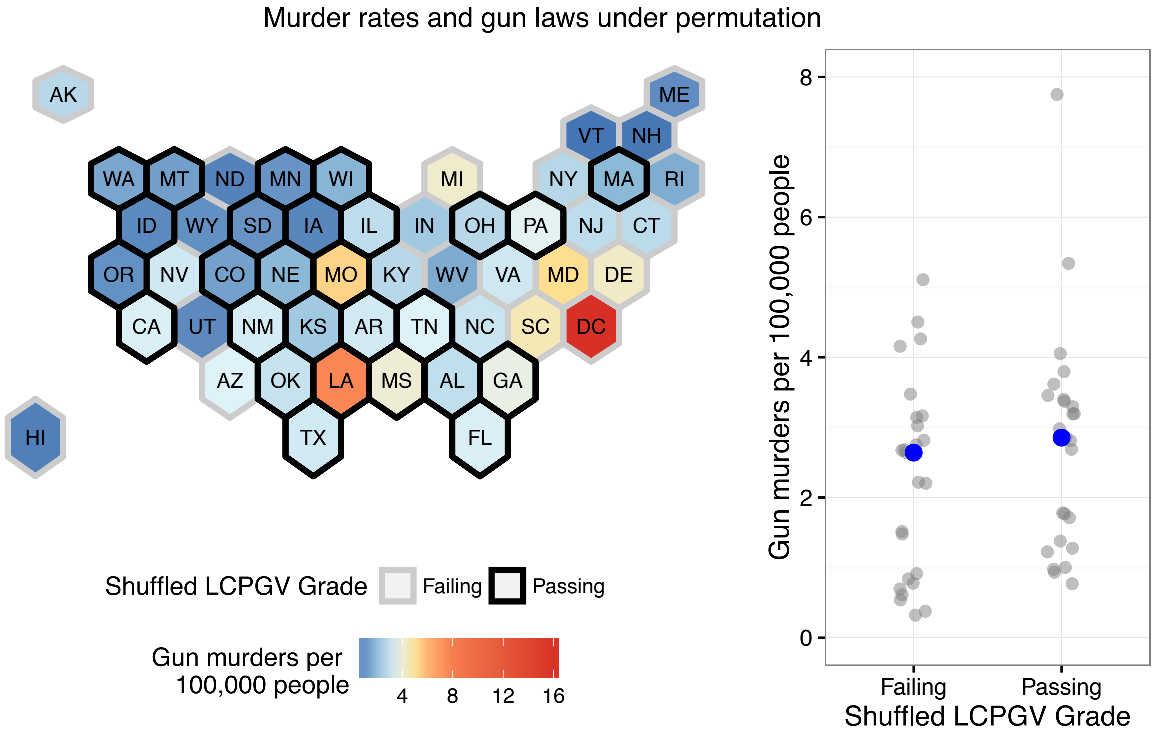

Figure 7.3: This map is almost identical to Figure 7.2, with one crucial difference: the identities of the states with passing and failing grades have been randomly permuted. There is still a small difference in the medians of the notionally passing and failing groups, due to random variation in the permutation process.

actual coin flips, which makes the virtual coin-flipping approach somewhat unhelpful as a general strategy.

It turns out, however, that in most situations, we can still harness the power of Monte Carlo simulation to understand what our data would look like if the null hypothesis were true. Rather than flipping virtual coins, we run something called a *permutation test*, which involves repeatedly permuting (or shuffling) the predictor variable and recalculating the statistic of interest.

To understand how this works, let's see an example. Figure 7.3 shows a map and dotplot very similar to those in Figure 7.2, with one crucial difference: in Figure 7.3, the identities of the states with notionally "passing" and "failing" gun laws have been randomly permuted. These grades bear no correspondence to reality. It's as though we took a deck of 51 cards, each card having some state's grade on it (treating D.C. as a state); shuffled the deck; and then dealt one card randomly to each state. The mathematical term for this is a *permutation* of the grades.

As expected, the median gun-murder rates of these two ran-

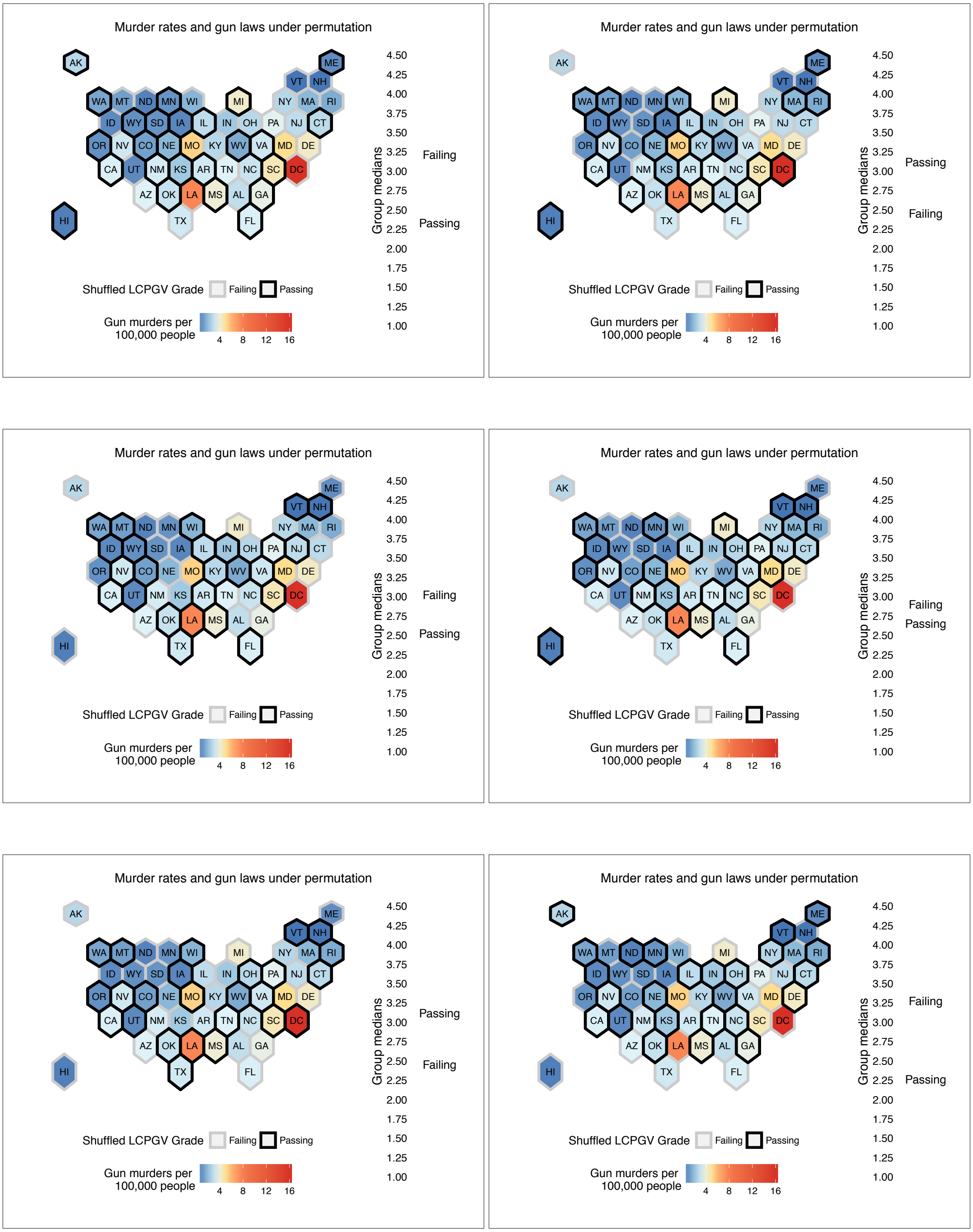
Murder rates and gun laws under permutation
AK ME
VT NH
WA MT ND MN WI MI NY MA RI
ID WY SD IA IL IN OH PA NJ CT
OR NV CO NE MO KY WV VA MD DE
CA UT NM KS AR TN NC SC DC
AZ OK LA MS AL GA
HI TX FL
Group medians
4.50
4.25
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
Failing
Passing
Shuffled LCPGV Grade
Failing
Passing
Gun murders per 100,000 people
4
8
12
16
Murder rates and gun laws under permutation
Passing
Failing
Murder rates and gun laws under permutation
Failing
Passing
Murder rates and gun laws under permutation
Failing
Passing
Murder rates and gun laws under permutation
Passing
Failing
Murder rates and gun laws under permutation
Failing
Passing

Figure 7.4: Six maps with permutated gun-law grades, with the medians for the passing and failing groups.

dom chosen "passing" and "failing" groups aren't identical (right panel). The randomly chosen "failing" states have a median of 2.6, while the randomly chosen "passing" states have a slightly larger median of 2.8. Clearly we can get a difference in medians of at least 0.2 quite easily, just by random chance—that is, when the null hypothesis is true by design.

But Figure 7.3 shows the difference in medians for only a single permutation of the states' gun-law grades. This permutation is random, and a different permutation would have given as a slightly different answer. Therefore, to assess whether could we get a difference in group medians as large as 0.8 just by random chance, we need to try several more permutations.

Figure 7.4 shows 6 more maps generated using the same permutation procedure. For each map, we shuffle the grade variables for all the states and recompute the median murder rates for the notionally "passing" and "failing" groups. Each map leads to its own difference in medians. In some maps, the difference is positive ("passing" states are higher), while in others it is negative ("failing" are states higher). In at least one of the 6 maps—the bottom right one—the median for the "failing" states exceeds the median for the "passing" states by more than 1 murder per 100,000 people, just by chance. This is a larger difference than we see for the real map, in Figure 7.2.

Six permutations give us some idea of how much a difference in the medians we could expect to see if the null hypothesis were true. But ideally we'd have many more than 6. Figure 7.5 addresses this need, showing the result of a much larger Monte Carlo simulation in which we generated 5,000 random maps, each one with its own random permutation of the states' gun-law grades. For each of these 5,000 maps, we computed the difference in medians between the notionally passing and failing groups. These 5,000 differences in group medians across the 5,000 maps are shown as a histogram in Figure 7.5.

Hypothesis testing: a four-step process

Let's review the vocabulary that describes what we've done here. First, we specified a null hypothesis: that the correlation between rates of gun violence and state-level gun policies could be explained by other unrelated sources of random variation. We decided to measure this correlation using a specific statistic: the difference in medians between the states with passing grades and

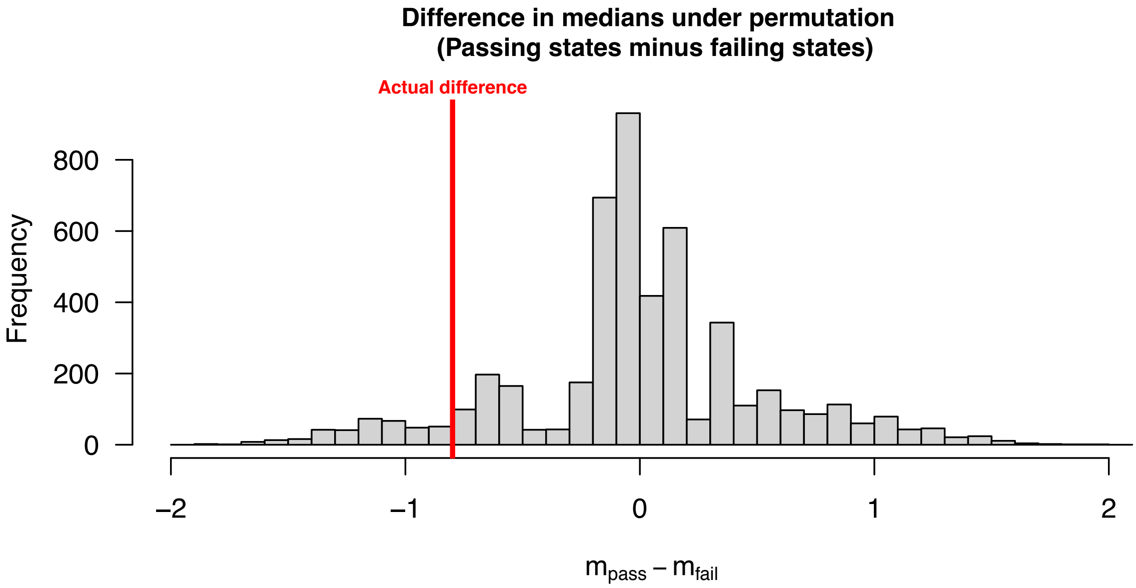

Figure 7.5: The histogram shows the difference in group medians for 5,000 simulated maps generated by the same permutation procedure as the 6 maps in Figure 7.4. Negative values indicate that the "failing" states had higher rates of gun violence than the "passing" states. The actual difference in medians for the real map in Figure 7.2 is shown as a vertical red line. This difference seems to be consistent with (although does not prove) the null hypothesis that other sources of random variation, and not necessarily state-level gun policy, explains the observed difference in murder rates.

those with failing grades. (Remember that a statistic is just some numerical summary of a data set.) To give this statistic a name, let's call it Δ (for difference in medians). It's intuitively clear that the larger Δ is, the less plausible the null hypothesis seems.

Figure 7.5 quantifies this intuition by giving us an idea of how much variation we can expect in the sampling distribution of our Δ statistic under the hypothesis that there is no systematic relationship between gun laws and rates of gun violence. As before, the sampling distribution is simply the probability distribution of the statistic under repeated sampling from the population—in this case, assuming that the null hypothesis is true.

There are two possibilities here, corresponding to the null and alternative hypotheses. First, suppose that we frequently get at least as extreme a value of Δ for a random map, like those in Figure 7.4, as we do in the real map from Figure 7.2. Then there's no reason to be especially impressed by the actual value of $\delta = -0.8$ we calculated from the real map.[4] It could have easily happened by chance. Hence we will be unable to reject the null hypothesis; it could have explained the data after all. (An important thing to remember is that *failing to reject* the null hypothesis is not the

[4] We use the lower-case δ to denote the value of the test statistic for your specific sample, to distinguish it from the Δ's simulated under permutation.

same thing as *accepting* the null hypothesis as truth. To use a relationship metaphor: failing to reject the null hypothesis is not like getting married. It's more like agreeing not to break up this time.)

On the other hand, suppose that we almost always get a smaller value of Δ in a random map than we do in the real map. Then we will probably find it difficult to believe that the correlation in the real map arose due to chance. We will instead be forced to reject the null hypothesis and conclude that it provides a poor description of the observable data.

Which of these two possibilities seems to apply in Figure 7.5? Here, the actual difference of -0.8 for the real map in Figure 7.2 is shown as a vertical red line. It's position on the histogram suggests possibility (1) here: $\delta = -0.8$ is consistent with (although does not prove) the null hypothesis that other sources of random variation unrelated to state-level gun policy can explain the observed difference in murder rates between the passing-grade and the failing-grade states.

To summarize, the four steps we followed above were:

(1) Choose a null hypothesis H_0, the hypothesis that there is no systematic relationship between the predictor and response variables.

(2) Choose a test statistic Δ that is sensitive to departures from the null hypothesis.

(3) Approximate $P(\Delta \mid H_0)$, the sampling distribution of the test statistic T under the assumption that H_0 is true.

(4) Assess whether the observed test statistic for your data, δ, is consistent with $P(\Delta \mid H_0)$.

For the gun-laws example, our test statistic in step (2) was the difference in medians between the "passing" states and the "failing" states. We then accomplished step (3) by randomly permuting the values of the predictor (gun laws) and recomputing the test statistic for the permuted data set. This shuffling procedure is called a permutation test when it's done in the context of this broader four-step process. There are other ways of accomplishing step (3)—for example, by appealing to probability theory and doing some math. But the permutation test is nice because it works for any test statistic (like the difference of medians in the previous example), and it doesn't require any strong assumptions.