7
Quantifying Uncertainty: Part II

Key terms and concepts: hypothesis testing; alpha; p-value; false positives and false negatives; rejection region; F-test

Red state, green state

Are conservative states more, or less, reliant on federal spending than liberal states?

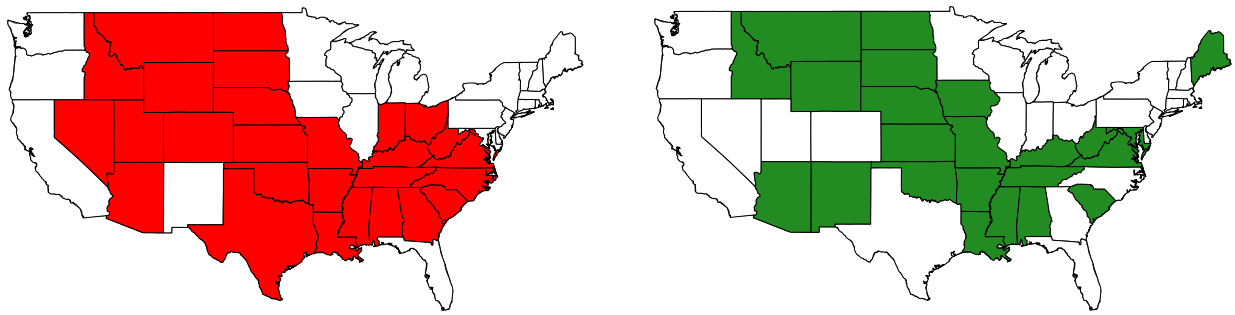

Figure 7.1: The red states where George W. Bush won more than 50% of the popular vote in 2004 (left); the green states that paid less in federal tax than they received in federal spending (right).

The two maps above would suggest that the answer is "more." In the red states, at left, more than 50% of the popular vote went to George W. Bush, the Republican candidate, in the 2004 presidential election. In the green states, at right, citizens received more in federal spending, at least in the aggregate, than they sent to the U.S. government in federal tax dollars.

The degree of overlap between the two groups is visually striking. Yet many explanations, other than the snarky "red state socialism" theories that circulate on political blogs[1], suggest themselves. These include: cherry-picked data, the reductionist binary distinction between red and blue states, the different political calculus involved with military spending versus social programs, the

[1] The Pulitzer-Prize-winning fact-checking organization PolitiFact weighs in here.

size-to-population ratio of the typical red state, and so forth.

Of course, one other explanation is that the observed overlap is a coincidence. In fact, this should be considered before the search for a substantive explanation begins, for there might be nothing in need of explanation at all. Let's consider two possible theories:

1. The 25 green states that receive a disproportionate amount of federal spending are a random sample of all states—implying, essentially, that federal dollars are allocated at random without regard to the political leanings of a state.[2] If true, this would mean that the observed overlap with the group of Bush states is merely a coincidence.
2. The 25 green states are not a random sample of all states, but are disproportionately biased towards (or away from) the red states.

[2] All models are wrong, but some models are less wrong than others!

We call hypothesis 1 the *null hypothesis,* often denoted H_0. It states that nothing special is going on in our data, and that any trend we thought might have existed isn't really there at all. It is a term coined back 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! The statistical term stands frozen in time as ordinary English idiom has moved on. 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 trying to choose between these competing hypotheses, it helps to get a sense of what the map might look like if the null hypothesis were true. In other words, what can we expect to see if the 25 money states are actually a random sample of all states?

Simulation can help this task of visualization a great deal. In Figure 7.2, we see 16 different parallel universes, all simulated in R. Each miniature map of the United States shows a random sample of 25 states (minus Alaska and Hawaii), with no factor other than pure chance distinguishing the green states from the white ones. These "Top 25" maps are completely meaningless. But by inspecting them, we begin to get a sense of what the green map might look like if the null hypothesis were true, and a state's ratio of dollars in to dollars out were completely uncorrelated with its political leanings.

The question, then, can be phrased as follows. How often do we see a random green map that overlaps with at least as many

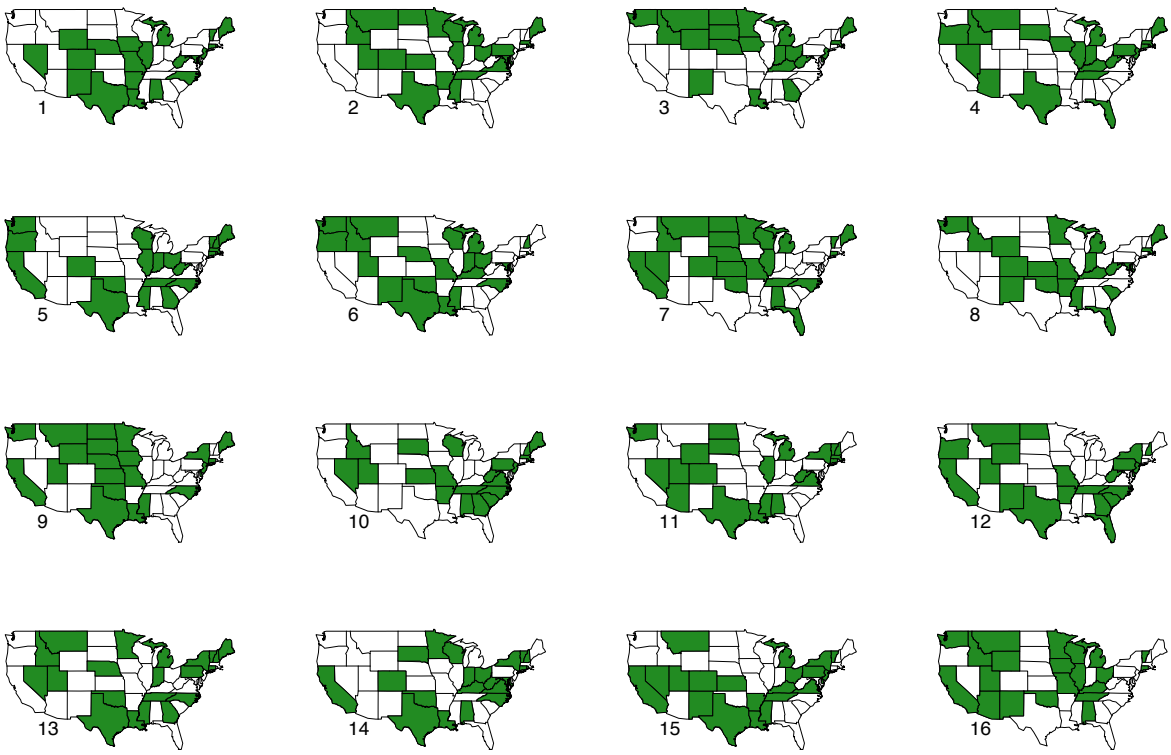

Figure 7.2: 16 different randomly generated maps corresponding to the null hypothesis that the green states are a random sample of all states.

red states as the *real* green map? We will measure overlap using an obvious *statistic*: the number of states that appear on both maps. (A *statistic* is a number, or set of numbers, that summarizes your sample. Here the number of states that overlap with the Bush map is a reduced summary of the data; we don't really care about the individual identities of the states that actually do overlap, just how many of them there are.)

Figure 7.2 gives us an intuitive idea of how much variation we can expect in the *sampling distribution* of this statistic under the hypothesis that the green states are a random sample of all states. (As before, the sampling distribution is simply the probability distribution of a statistic under repeated sampling from the population.) There are two possibilities here:

1. Suppose that we almost always get less overlap in a random map than we do in the real map. Then we will probably find it difficult to believe that the overlap in the real map

arose due to chance. We will, to use the well-worn statistical phrase, be forced to *reject the null hypothesis.*

2. Suppose, on the other hand, that we frequently get at least as much overlap in a random map as we do in the real map. Then there's no reason to be impressed by the overlap in the real map—it could have easily happened by chance! Hence we will *fail to reject the null hypothesis.*

Always remember that *failing to reject* the null hypothesis is not the same thing as *accepting* the null hypothesis as literal 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.

The logic of frequentist hypothesis testing

So far, we have contemplated the overlap between the red states and the green states using a combination of pictures and intuitive reasoning. A quick look at Figure 7.2 suggests that it's rare for a randomly generated map to beat or exceed the amount of overlap of the real maps in Figure 7.1. Can we quantify this judgment?

It turns out that we can, by performing something called a *hypothesis test*. A hypothesis test is simply a decision between two courses of action: to reject the null hypothesis on the basis of observed data, or not. To perform such a test, we must answer two questions. These same two questions will arise in all hypothesis-testing problems, so learn them well.

The first question is as follows. Suppose we believe that the null hypothesis is true, and that we plan to take some data to see if we're right. We will summarize the data using a statistic—in this case, the number of green states that overlap with the red-state map. The value of this statistic might end up surprising us, or it might not. How much surprise is too much surprise for us to merely say "no big deal" and go on believing in the null hypothesis?

Said a different way: what is our threshold of believable surprise, beyond which the data (as summarized by the statistic) will change our minds? We would clearly be surprised if a random sample of 25 states contained only Bush-voting states. But what if our sample contained 12 Bush states, or 15, or 18? What is

the threshold value of the statistic where we will cease to believe in the hypothesis that the real sample is a random sample of all states? This threshold is often called *critical value*, and the values of the statistic equal to or beyond the critical value are often referred to collectively as the *rejection region*. This is because we will reject the null if we observe a value of the statistic at least as extreme as the critical value.[3]

As an analogy, think of a criminal trial. In the United States, our null hypothesis is that someone who stands accused of a crime is innocent, until the prosecution provides evidence that proves guilt "beyond reasonable doubt." That threshold of reasonable doubt in a criminal trial is intuitively similar to the threshold of "believable surprise" in a statistical hypothesis test.

The second question is: does the real data fall at or beyond that threshold of believable surprise? If it does, we reject the null. If it doesn't, we fail to reject. Once we've chosen the threshold—which we must do before ever taking data—this yes-or-no question is easy to answer.

[3] You may notice some ambiguity in other definitions of a critical value. The issue is whether you reject the null if you observe a value of the statistic precisely equal to the critical value. Here, we assume that the critical value is inside the rejection region, rather than just outside it.

Choosing a critical value: the Neyman–Pearson, or frequentist, approach

How to choose a critical value is, in the end, a subjective matter. Probability theory can illuminate the consequences of a particular choice of threshold, but it does not provide a single, unambiguous right answer. I've used words like "believable" and "reasonable" not as a dodge, but because the choice truly is open for debate.

Having said that, it is important for you to learn the way in which generations of scientists, economists, and statisticians have chosen this threshold in their hypothesis tests. This is important not just so that you can understand the results of hypothesis tests performed by other people, but also so that you evaluate their reasoning and know when they have made a mistake.

This constellation of ideas is often referred to as Neyman–Pearson testing, and it forms the dominant approach to formal statistical hypothesis testing in most fields of human inquiry. The basic idea of Neyman–Pearson testing is to choose a threshold of believable surprise so as to control the frequency with which you will make errors in the repeated application of that threshold. This is why it is sometimes called *frequentist* testing, to distinguish it from Fisherian or Bayesian testing.[4]

Think of it this way. In choosing a threshold, we must strike a balance between two different kinds of error:

[4] The Neyman–Pearson approach to testing dates to the late 1920's and early 1930's, and was advocated primarily by two statisticians: Jerzy Neyman (a Polish–American) and Egon Pearson (an Englishman). The approach is not without controversy, however, and there are two alternative schools of thought about how to do hypothesis testing. The first is the so-called Fisherian approach, popularized by Ronald Fisher in the 1920's. Fisher was also English, and one of the great geniuses of the twentieth century—he almost single-handedly revolutionized both statistics and genetics. The second approach is the so-called Bayesian approach, advocated most strongly by Harold Jeffreys (yes, another Englishman) in the 1930's. We won't focus on either of these approaches in this course, but feel free to ask me for references if you are interested.

1. *False positives*, in which we wrongly reject a true null hypothesis. This is sometimes called a Type-I error.

2. *False negatives*, in which we wrongly fail to reject a false null hypothesis. This is sometimes called a Type-II error.

If we set a low threshold, then we are quite likely to observe a sample whose statistic is beyond that threshold, even if the null hypothesis is true. This means we are in very real danger of wrongly rejecting a true null hypothesis—that is, committing a Type-I error. In our example, suppose we decide that we are willing to reject the null hypothesis if, in the real data, we see 15 or more overlaps with the Bush map. But even in a random sample of states, we would expect to see something around 13-14 overlaps, and so it's not too much of a stretch to imagine that 15 could happen just by chance. If the null were true and we saw 15 overlaps anyway, we'd end up with a false positive.

Suppose, on the other hand, that we try to fix the problem by setting a high threshold—one that would be very unlikely to be met or exceeded under the null hypothesis. Let's say that we demand an overlap of at least 21 states in our map before we're willing to declare ourselves surprised by the data and to reject the null hypothesis.

To be sure, this conservative route will cut down on the chances of a false positive. But the tradeoff is that we might miss out on real differences. If, for example, the 25 green states were truly biased toward the Bush states but yielded an overlap of only 16 or 17 states, then we'd end up committing a Type-II error instead.

How is one to negotiate this tradeoff? You are free to make your own judgments, but the Neyman–Pearson approach prescribes the following basic steps.

1. Before you ever look at any data, choose a choose a rejection region R, corresponding to set of outcomes for the test statistic t that will lead you to reject the null hypothesis. The idea is that observating any value for t that lies within the rejection region R would be too surprising for you to go on believing in the truth of the null hypothesis.

2. Compute the probability that the test statistic will fall within the rejection region R, even if the null hypothesis is true. Denote this probability by α. Lower values of α (which is sometimes called the significance level) indicate less tolerance for rejecting true nulls, and therefore greater conservatism.

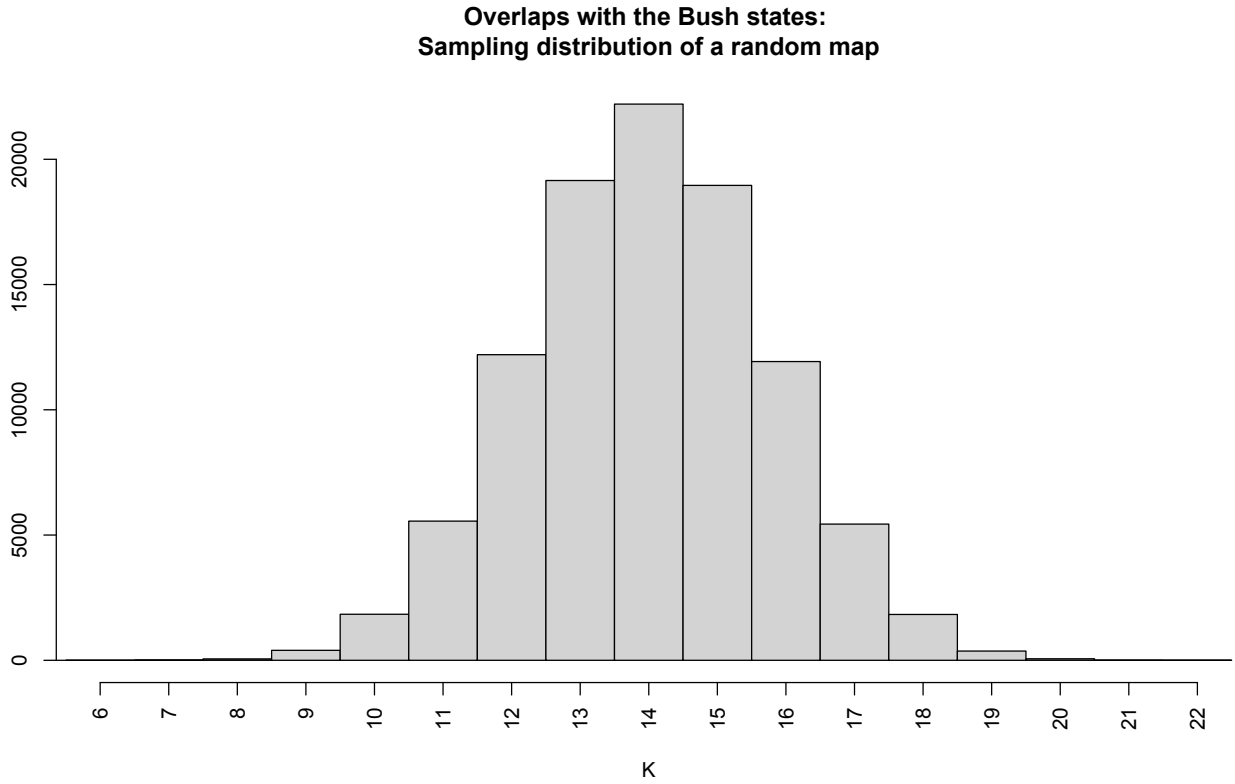

Figure 7.3: The number of states that overlapped with the Bush map in 100,000 simulated random maps just like those in Figure 7.2.

3. If the observed value of the test statistic t falls within the rejection region R, reject the null. Otherwise, do not.

To illustrate these ideas, let's now take a second look at the red-state/green-state maps, and conduct a formal Neyman–Pearson test of the null hypothesis that the 25 green states are a random sample of all states. We will follow the steps prescribed above.

1. Somewhat arbitrarily, let's choose a rejection region of $R = \{0, 1, \ldots, 9, 19, 20, \ldots, 25\}$. I repeated the Monte Carlo calculation that was used to produce the random maps in Figure 7.2. Instead of 16 random maps, however, I generated 100,000 of them. For each random map, I computed our summary statistic: the number of states that overlapped with the Bush map. Figure 7.3 shows the Monte Carlo estimate of this statistic's sampling distribution under the null hypothesis. Clearly, values of t in our chosen rejection region would be

fairly unlikely under the null hypothesis.

2. From Figure 7.3, it is easy to calculate α, the probability that t will fall in the rejection region R even if the null hypothesis is true. This turns out to be about 1%.

3. The value of the statistic for the real map is 21 overlaps. This falls inside our rejection region, and hence we reject the null hypothesis at the $\alpha = 0.01$ level.

Based on the results of the hypothesis test, it would appear that we have some causal explaining to do, after all, and that pure chance cannot explain the observed overlap in Figure 7.1.

Summarizing and interpreting a Neyman–Pearson test

What should you report as a final result when you conduct a Neyman–Pearson test? Following the example above, only three things are necessary:

1. the value of α that you have chosen (e.g. 0.10 or 0.05)
2. the observed value of the statistic that you are using to summarize your sample (e.g. 21 overlaps)
3. a yes-or-no answer to the question of whether the statistic from your sample falls into the rejection region corresponding to your chosen α

If you answer "yes" in Step 3—that is, if the observed statistic falls in the rejection region—then your result is said to be "statistically significant."

This procedure is sometimes called an α-level test, and it has the all-important *frequentist error bound*: if you pre-specify a value of α and then apply the associated α-level Neyman–Pearson test to lots of different data sets, then on average, you will reject no more than 100α% of the null hypotheses that you encounter. It is this guaranteed upper bound on the probability of falsely rejecting the null that makes the Neyman–Pearson approach so appealing to so many people.

Of course, the Neyman–Pearson approach is not without its limitations, since it doesn't allow you to say anything about some very important quantities:

- the unconditional probability of a false negative.[5]
- the probability that the null hypothesis is false, given the data.

[5] Under the Neyman–Pearson approach, it is possible to assess the probability of a false negative, conditional upon a specific formulation of the alternative hypothesis (e.g. specific alternative values for the parameters governing the probability distribution of the statistic). These are called power calculations.

- the probability that you have made the wrong decision for this particular data set.

The Neyman–Pearson approach also does not provide a numerical summary of the strength of evidence in the data against the null hypothesis. It doesn't allow you, for example, to compare two data sets and provide a quantitative answer to the question, "How much stronger or weaker is the evidence in data set 1 than in data set 2?" All you get is a binary result: reject, or don't. Of course, the binary nature of the result is what leads to the guarantee of the frequentist error bound, and so this can also be construed as a strength.

As for the choice of α: that is up to you. Remember, a hypothesis test is a formal decision about what to believe, and this decision should depend upon the consequences of the two types of error that you could make. In some situations these consequences may be highly asymmetric—just as they are in the decision of guilt or innocence in a criminal trial, where the difference between a false negative and a false positive is the difference between letting a guilty person go free and putting an innocent person behind bars. It is usually these kinds of utility considerations—How bad is a false positive? How bad is a false negative?—that go into the determination of α. Therefore, you should never let anyone assert that their α is objectively better than yours without making them explain why.

The importance of pre-specifying a significance level

The most important thing to remember about Neyman–Pearson testing is that you must choose α before you ever look at any data. This simply cannot be overemphasized.[6] It is invalid to take data first, then go searching for the smallest possible value of α for which your data would lead to a rejection, and in the end claim that you have rejected the null hypothesis after performing an α-level Neyman–Pearson test.

In fact, this fallacy of reporting a value of α that has been chosen after looking at the data is so common that it deserves a name. I call it the "fuzzy-α fallacy," owing to the fact that α should never be allowed to remain fuzzy beforehand, only coming into focus *after* the data have been collected and summarized. If you do this, you will lose the nice guarantee associated with the frequentist error bound, and your implied probability of falsely rejecting a true null may be very different from the claimed value of α. This

[6] Of course, we had already seen the data before we chose α in the previous example, even though we didn't use this to drive the choice itself. Nevertheless, take this as an example of how hypothesis testing works procedurally, not as a formally valid test.

is important enough to merit repetition: if you commit the fuzzy-α fallacy, your claimed frequentist error bound goes straight out the window. That's because the α-level is a property of a *procedure*, not a property of an individual data set.

In assessing statistical analyses performed by other people, you should be on the lookout for some obvious warning signs that are strongly associated with the fuzzy-α fallacy. One warning sign is when someone reports the results of two experiments and claims that they are significant at two different α-levels. For example: "We conducted two different trials for a new anti-inflammatory medicine. Both trials showed an improvement over existing drugs. The first trial was significant at the $\alpha = 0.05$ level, while the second was significant at the $\alpha = 0.01$ level." You should be immediately suspicious here—why else would there be two different α-levels in the same sentence if they weren't chosen in exactly the fallacious, data-dependent manner just described? Sometimes there's a legitimate reason, but often there's not.

A second warning sign is when someone describes a result using various adjectives designed to connote impressiveness: "very significant," "highly significant," "extremely significant," and so forth. The fallacy isn't in the terms themselves, which are perfectly fine as informal descriptions of evidence. (After all, some data sets do provide stronger evidence against the null than others.) Rather, the fallacy is in assuming that these terms have any strict mathematical meaning whatsoever, and that, upon hearing one, you should be impressed at how likely it is that the null hypothesis is wrong. Under a formal Neyman–Pearson test, the significance level is pre-specified, and the sample is either significant at this level, or it isn't—no further adjectives needed.

Let me give you an idea of how common this fallacy is, even in top-quality research journals. In the January, 2010 volume of the *Journal of the American Medical Association*, I found 12 scholarly articles that described original medical research and that quoted more than one "statistically significant" result. Of these 12 articles, 8 quoted all of their results at a single α-level (good), while 4 quoted their results at various different α-levels (bad).

This means that, in all likelihood, one-third of the articles in this (admittedly small and unscientific) sample do not have the frequentist error bound that they are claiming to have. Their results may be important, but from a frequentist perspective, they are nearly uninterpretable.

Interpreting p-values

YOU MAY have noticed that we've not yet mentioned one concept that is widely associated with statistical hypothesis testing: that of a *p-value*. The reason is that p-values play no formal role in Neyman-Pearson hypothesis testing. They are part of the Fisherian approach to hypothesis testing, rather than the frequentist view that we are learning in this course.

Having said that, it is important for any student of statistics to understand what p-values are—and more importantly, to understand what they aren't—even though they have no formal frequentist interpretation. This is because p-values are both widely used and widely misused.

Let's begin with a concise definition of a p-value, before we slowly unpack the definition to understand the concept a bit more deeply: *a p-value is the probability of observing a sample as extreme as, or more extreme than, the sample actually observed, given that the null hypothesis is true.* Now compare this definition side by side with the definition of α that we learned before.

α-level	p-value
The probability of observing a result as extreme as, or more extreme than, **the pre-specified critical value**, given that the null hypothesis is true.	The probability of observing a result as extreme as, or more extreme than, **the result actually observed**, given that the null hypothesis is true.

Take a good, long look at these two definitions. Notice what's the same and what's different. In both cases, we are explicitly assuming that the null hypothesis is true. In both cases, we are summing up the probability of the events that are at least as extreme as some threshold. The only thing that's different? For the definition of an α-level, that threshold is the pre-specified critical value, which is determined before the experiment ever takes place. But for the definition of a p-value, that threshold is the data you actually observed, which can only be determined after the experiment takes place.

For example, let's go back to the example of the overlap between the Bush states and the green states. There, our rejection region included all values of 9 or smaller, and all values of 19 or larger. This yielded an α-level of roughly $\alpha = 0.01$, ensuring that if

the null hypothesis were true, we would have no more than a 1% chance of falsely rejecting it.

But we actually observed 21 overlaps. The probability of getting 21 or more overlaps would have been extremely small under the assumption that the null hypothesis was true: 0.0003, or only 30 times in 100,000 random maps.

This is the *p*-value of our data, and it provides information about the strength of evidence in the observed sample. Smaller *p*-values tend to connote more extreme deviations from the kinds of samples that would be predicted if the null hypothesis were true, and therefore are "less consistent" with the null.

The *p*-value gives us information that is interesting, but that is very, very hard to interpret correctly. People make mistakes with *p*-values all the time—even the big boys and girls quoting *p*-values in original research papers—so it's worth warning you about a handful of common ones:

- The *p*-value is *not* the same thing as the α-level. Just compare their definitions above; the difference is in bold type. Someone who quotes the *p*-value but calls it the α-level is committing the fuzzy-α fallacy.
- The *p*-value is *not* the probability of having observed our sample, given that the null hypothesis is true. Rather, it is the probability of having observed our sample, *or any more extreme sample*, given that the null hypothesis is true.
- The *p*-value is *not* the probability that the null hypothesis is false, given the observed value of the sample. In fact, in one sense it is almost the opposite (see the previous item).
- The *p*-value is *not* the probability that you will falsely reject a true null hypothesis—that's the α-level!

To make matters worse, two *p*-values are not numerically comparable in the way that ordinary probabilities are. For example, a probability of 0.01 is ten times smaller than a probability of 0.10. But a *p*-value of 0.01 does not indicate evidence that is ten times "stronger" than a *p*-value of 0.10. This fact, more than anything else, is what makes *p*-values so hard to interpret. They sound so quantitative and exact, and yet you cannot even compare them in the same intuitive way you would compare ordinary numbers.

For example, let's say we assume that the null hypothesis is a normal distribution with mean $\mu = 0$ and standard deviation

$\sigma = 1$: $z \sim \mathrm{N}(0,1)$. You might reasonably assume that, if the null hypothesis were true, it would be ten times more likely that you would see a sample of z where $p = 0.10$ than one where $p = 0.01$. And you would probably also assume that it would be ten times more likely to see a sample where $p = 0.01$ than to see one where $p = 0.001$.

But these assumptions aren't right at all! In fact, if the null hypothesis is true, it turns out that the probability that you will get a sample where $p = 0.10$ is 6.58 times larger than the probability that you will get one with $p = 0.01$. And you will get a sample with $p = 0.01$ more often than one with $p = 0.001$ by a factor of 7.92. Go figure![7]

Always be careful when quoting or interpreting p-values. Remember: an α-level is a **formal** frequentist characterization of the error rate of a test. A p-value, on the other hand, is an **informal** description of the evidence in a specific sample—one whose interpretation is slippery and easily misconstrued.

[7] Caveats: (1) this refers to the probability density of the normal distribution, a notion that must be made precise using a branch of mathematics called measure theory; (2) the specific multiples quoted can be different if you apply a transformation to the test statistic; and (3) you're not responsible for understanding why, how, or even that this is the case. I include this fact purely for enrichment.

Hypothesis testing for regression coefficients

Recall that the t statistics are the standardized least-squares estimates of β_0 and β_1. To standardize, we subtract the mean and divide by the standard error:

$$t_0 = \left(\frac{\widehat{\beta}_0 - \beta_0}{\hat{\sigma}_0}\right) \quad \sim \quad t_{n-2}$$
$$t_1 = \left(\frac{\widehat{\beta}_1 - \beta_1}{\hat{\sigma}_1}\right) \quad \sim \quad t_{n-2}\,.$$

These depend upon the unknown values β_0 and β_1—which means that the t statistics themselves must be unknown. When reporting t-statistics, the typical software package assumes you are interested in the null hypothesis that the corresponding coefficient is zero. In other words, it is computing

$$t_0 = \frac{\widehat{\beta}_0}{\hat{\sigma}_0} \quad \text{and} \quad t_1 = \frac{\widehat{\beta}_1}{\hat{\sigma}_1}\,.$$

Why? The answer to this is the answer to the following hypothetical question: what kind of data would we expect to see if our x and y variable had no linear relationship whatsoever? Mathematically, this would mean that β_1, the true underlying slope parameter of the simple regression model, is exactly zero. Let's see what happens when we simulate from a model like this:

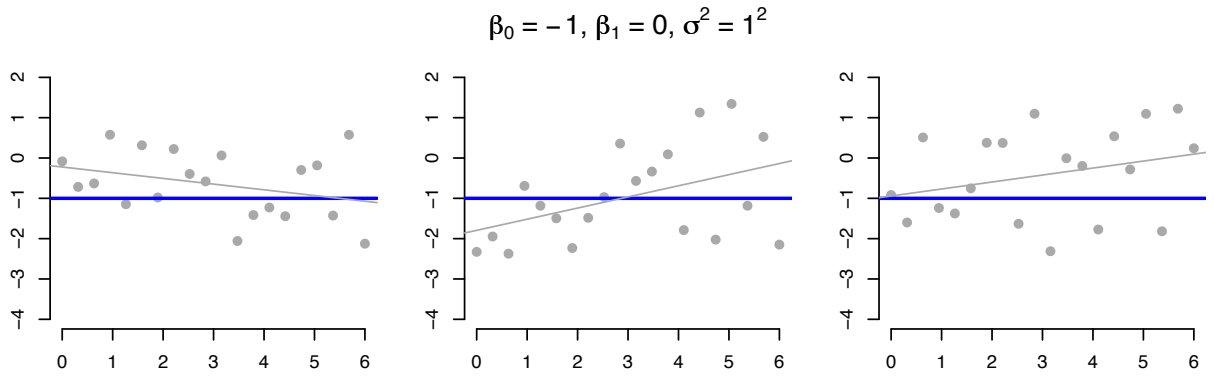

Figure 7.4: Three data sets of size $n = 20$ where the true slope is exactly zero.

If $\beta_0 = -1$ and $\sigma^2 = 1^2$, we get data that looks something like the above. The estimated slope will never be exactly zero, even if the true slope is. And the same will be true of the intercept: $\widehat{\beta}_0$ will never be zero, even if β_0 is.

By how much will the estimated coefficients vary when the true coefficients are zero? We now have a whole theoretical apparatus for answering this question! If β_1 is zero, then the value of t_1 from your data set is just a random draw from a t distribution with $n - 2$ degrees of freedom. To put this in terms that hark back to an earlier chapter: we know the *sampling distribution* of the t statistic under the null hypothesis that $\beta_1 = 0$.[8]

[8] Confusingly, t statistic is not short for test statistic, but refers to the t distribution. "Test statistic" is a more general term, applicable to any hypothesis-testing problem.

Using the t-statistic in a Neyman–Pearson test

This is all the information we need in order to conduct a Neyman–Pearson test of the null hypothesis that $\beta_1 = 0$. We could also test whether $\beta_0 = 0$, but let's focus on the slope; everything here also applies to testing the interception, for with $\widehat{\beta}_0$ and $\hat{\sigma}_0$ in place of $\widehat{\beta}_1$ and $\hat{\sigma}_1$.

To avoid confusion, let's use T_1 (with a capital letter) to denote a hypothetical value of the t statistic—that is, the value we might see in some parallel universe generated from the same underlying model. And let's use t_1 to denote the value of the t statistic for $\widehat{\beta}_1$ that you observed for your particular data set.

As in all Neyman–Pearson tests, there are three steps to follow:

1. Choose an α that encodes your tolerance for false positives. Recall that α is the probability that you will reject the null hy-

pothesis, given that the null hypothesis is true. The "industry standard" tends to be $\alpha = 0.05$, but as always, you should pick an α that you personally can live with.

2. Find the critical value $t^\star_\alpha$ corresponding to your chosen α. If your alternative hypothesis is that $\beta_1 \neq 0$, then you are performing a two-tailed test, and must find a $t^\star_\alpha$ such that

$$P(T_1 \geq t^\star_\alpha \mid \beta_1 = 0) = \alpha/2\,.$$

Since the t distribution is symmetric, this will ensure that

$$P(T_1 \leq -t^\star_\alpha \mid \beta_1 = 0) = \alpha/2$$

as well. In other words, you need to form a rejection region with $\alpha/2$ in the lower tail and $\alpha/2$ in the upper tail, thereby ensuring that the total probability of the rejection region is exactly α.

If $n \geq 30$ and $\alpha = 0.05$, you can always just use the rule of thumb from before: $t^\star_\alpha \approx 2$. Otherwise, you can find a critical value using the R command `qt(1-alpha/2, n-2)`.

If, on the other hand, your alternative hypothesis is that $\beta_1 > 0$, then you are performing a one-tailed test, and must find a $t^\star_\alpha$ such that

$$P(T_1 \geq t^\star_\alpha \mid \beta_1 = 0) = \alpha\,.$$

That's because your rejection region must only account for values of T_1 greater than zero, since you're only interested in values of the slope that are greater than zero. In this case, use the R command `qt(1-alpha, n-2)`.

3. Finally, answer the question: does t_1 from your data set fall in the rejection region? For a two-sided test, the rejection region is all values of t_1 such that $|t_1| \geq t^\star_\alpha$. For a one-sided test, the rejection region is either $t_1 \geq t^\star_\alpha$ or $t_1 \leq t^\star_\alpha$, depending on whether your alternative hypothesis is $\beta_1 > 0$ or $\beta_1 < 0$. Either way, if t_1 falls in the rejection region, reject at the specified α level. If not, don't.

Finally, remember our rule of thumb: if n is 30 or more, then the t distribution is pretty close to the normal. Therefore, if the null hypothesis is true, then t_1 will be within $2\hat{\sigma}_1$ of zero 95% of the time.

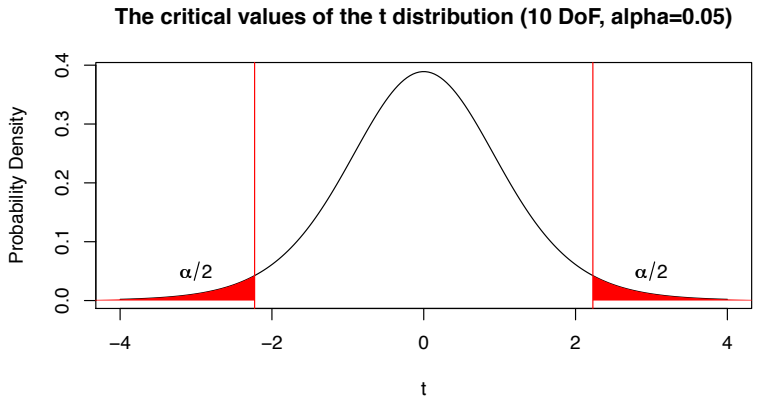

Figure 7.5: The critical values of the t_{10} distribution for a two-sided test at $\alpha = 0.05$.

This gives rise to a very simple guideline. If you're conducting a two-tailed test of $\beta_1 = 0$ at the $\alpha = 0.05$ level, then you should reject the null hypothesis if $\widehat{\beta}_1$ is more than twice its standard error. (Ditto $\widehat{\beta}_0$ and $\hat{\sigma}_0$.) Another way of phrasing this is: you should reject the null hypothesis at the $\alpha = 0.05$ level if the 95% confidence interval for β_1 fails to contain zero.

Testing for values other than zero

Regression software almost always assumes that you are testing the null hypothesis that $\beta_1 = 0$, and gives you the t statistics corresponding to this assumption. But with a simple calculation, you can easily test whether β_1 is equal to some value other than zero, instead.

Let's say your null hypothesis is that $\beta_1 = \theta$, where θ is a number other than zero. For example, in some situations, you might want to test whether a regression coefficient is exactly equal to one. How should you proceed?

Let's return to the definition of the t-statistic. If the true value of β_1 is really θ—in other words, if your null hypothesis is really true—then the following relationship holds:

$$t_1 = \left(\frac{\widehat{\beta}_1 - \theta}{\hat{\sigma}_1} \right) \sim t_{n-2} \, .$$

This particular t_1 is not part of the standard regression output; the software is assuming that your θ is zero, and in this case it's

not. But since you know θ, you can conduct a Neyman–Pearson test by just computing t_1 yourself according to the above formula: take the least-squares estimate, subtract θ, and divide by the standard error. With that one simple modification, you can then conduct a Neyman–Pearson test using the same steps above. Just remember to use your own "hand-calculated" t_1, and not the t_1 that the software calculates for you, in determining whether the data falls in the rejection region!

You can use a confidence interval to conduct a Neyman–Pearson test here, as well. If you want to test whether β_1 is significantly different from θ, then just check whether θ falls inside the $100(1-\alpha)\%$ confidence interval for β_1.

Statistical versus practical significance

Remember not to confuse statistical significance with practical significance. Let's imagine the following example to illustrate the difference. Suppose we take some data on an x variable and a y variable, and run a regression to understand the x–y relationship. We compute the least-squares estimate of the slope as $\widehat{\beta}_1 = 0.01$, and the standard error as $\hat{\sigma}_1 = 0.001$. Then if we test the hypothesis that $\beta_1 = 0$, our t statistic will be

$$t_1 = \frac{0.01}{0.001} = 10\,.$$

This is an enormous t statistic, and would lead us to reject the null hypothesis at pretty much any α level that wasn't absurdly small.

Should we reason, therefore, that x has a big effect on y—an effect that has been subjected to rigorous empirical scrutiny, and is substantiated by an enormous t statistic and a microscopic p value? Of course not! The 95% confidence interval for β_1 will be approximately $(0.008, 0.012)$. The data are telling us that, in all likelihood, x is useless for predicting y. After all, enormous changes in x are associated with only minuscule changes in y.

The moral of the story is: big t statistics and small p values can still happen even for tiny, insignificant effect sizes. This is especially likely to happen when you have a whole lot of data, since large samples allow you to estimate even tiny slopes with great precision.

Therefore, don't look merely at the t statistic or p value in summarizing an analysis. These quantities are not designed to distinguish a mountain from an anthill; they will only tell you that the

ground isn't flat. Confidence intervals matter, too, and can save you the embarrassment of championing a result that is statistically significant, but practically useless.

F tests

Return to the regression model for a student's college GPA in terms of SAT scores and undergraduate college:

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.691e+00  9.624e-02  17.566   <2e-16 ***
SAT.V                  1.486e-03  8.515e-05  17.455   <2e-16 ***
SAT.Q                  1.186e-03  9.098e-05  13.041   <2e-16 ***
SchoolBUSINESS         5.784e-03  7.827e-02   0.074   0.9411
SchoolCOMMUNICATIONS   8.565e-02  8.088e-02   1.059   0.2896
SchoolEDUCATION        4.492e-02  8.552e-02   0.525   0.5994
SchoolENGINEERING     -1.890e-01  7.851e-02  -2.408   0.0161 *
SchoolFINE ARTS        8.423e-03  8.443e-02   0.100   0.9205
SchoolLIBERAL ARTS    -1.374e-01  7.763e-02  -1.770   0.0767 .
SchoolNATURAL SCIENCE -1.495e-01  7.789e-02  -1.920   0.0549 .
SchoolNURSING          2.423e-02  1.022e-01   0.237   0.8126
SchoolSOCIAL WORK     -3.787e-02  1.391e-01  -0.272   0.7854
```

You might notice that only a few of the individual dummy variables looked close to significant at the $\alpha = 0.05$ level. We could throw out the others, and just keep in the ones for which we reject the null hypothesis that the regression coefficient is zero (at some pre-specified α level, of course).

But what if we'd like to test the significance of our dummy variables *as a block*, rather than one at a time? After all, we'd like to think of the category itself as a variable, and the bunches of dummy variables as just a trick for letting the regression model take category membership into account. We might decide that it doesn't make sense to allow some categories to enter the regression model, but not others.

One way of doing this is look at how much R^2 improved after adding the category. With just SAT Math and SAT Verbal, R^2 was 0.155. Now with the nine dummy variables added in, R^2 is 0.185. This looks like a modest change in absolute terms, but does represent a relative change of about 20%. Is this change significant? We know by now, of course, that adding *any predictor at all* to a regres-

sion model, even a random one, will make R^2 jump by at least a little bit.

Hence the test of our hypothesis that category membership provides additional predictive power can be phrased as follows: could the addition of random predictors have plausibly made R^2 jump by at least as much as it jumped when we added the real predictors?

This is where the F-test comes in handy. As it turns out, it is possible to quantify precisely how much R^2 is expected to increase when we add predictors to a model that are uncorrelated with the response! This requires defining something called the F statistic:

$$f = \frac{\Delta R^2}{1 - R_F^2} \cdot \frac{n - p_F - 1}{p_F - p_R}, \tag{7.1}$$

where:

- R_F^2 is the value of R^2 under the full model—that is, the one that includes the block of variables you're interested in testing.
- $\Delta R^2 = R_F^2 - R_R^2$ is the gain in R^2 in moving from the Reduced model (without the block of tested coefficients) to the Full model (with the block of tested coefficients).
- p_F and p_R are the number of parameters, not including the intercept, in the Full and Reduced models, respectively.
- n is the number of observations in the sample.

This formula has a lot of pieces, but you can recognize F as essentially a rescaled version of ΔR^2. The rescaling takes into account how many variables are at stake, how many observations are in the sample, and how big R^2 was to begin with (that is, before adding the block of questionable variables).

The F statistic is the analogue of the t statistic for testing a single regression coefficient. Under the null hypothesis that the entire block of coefficients is zero, this statistic has what is known as an F distribution with $(p_F - p_R, n - p_F - 1)$ degrees of freedom. That is, if H_0 is true, then

$$(f \mid H_0) \sim F(p_F - p_R, n - p_F - 1).$$

(Yes, that is indeed two distinct degrees-of-freedom parameters, compared to one for the t distribution.) Larger values of the F statistic correspond to greater evidence against the null hypothesis; if f is large enough, you reject the null. To find the critical value of the appropriate F distribution at a specified α level, use

the R function qf$(1-\alpha, d_1, d_2)$, where d_1 and d_2 are the two degrees of freedom parameters.

For example, in our data on UT students from the entering class of 2000:

- $R_F^2 = 0.185$ from the full model with all nine dummy variables in it
- $\Delta R^2 = R_F^2 - R_R^2 = 0.03$ is the gain in R^2 over the reduced model without the nine dummy variables
- $p_F = 11$ (two SAT scores, plus nine dummy variables), and $p_R = 2$ (just the two SAT scores)
- $n = 5191$

Putting these pieces together, we compute that

$$f = \frac{\Delta R^2}{1 - R_F^2} \cdot \frac{n - p_F - 1}{p_F - p_R} = \frac{0.03}{1 - 0.185} \cdot \frac{5191 - 11 - 1}{11 - 2} = 21.2$$

Suppose we want to conduct a Neyman–Pearson test with $\alpha = 0.01$. We must compare $f = 21.2$, the observed value of our test statistic, to $f^\star$, the 1% critical value of an $F(9, 5179)$ distribution. Here $f^\star$ turns out to be roughly 2.41. Therefore $f > f^\star$; we reject the null hypothesis at the $\alpha = 0.01$ level and conclude that the block of dummy variables should be in the model, after all.

You can also use the F test to determine whether the R^2 value for a multiple regression model is significant compared to a model with no predictors at all. Think of it as an omnibus test for all the coefficients at once. To conduct this test, just use the same formulas above, plugging in $p_R = 0$ and $R_F^2 = \Delta R^2 = R^2$ (since the "reduced model" has no parameters and an R^2 of zero by definition). Everything else, including the computation of the critical value, is the same.

Occam's Razor: comparing more than two models

F TESTS are useful for comparing a complex model to a simpler one. Often, however, we must compare more than two models, all of which look plausible. For this, we need some tools for multiple-model comparisons.

If your only goal is to choose from among different transformations of the same set of variables, then R^2 will do just fine; the best transformation will yield the most precise fit. But often you'll face the problem of comparing models with different numbers of predictors, or with different powers of a single predictor. For

example, you might need to choose between y versus x (with one slope coefficient), and y versus x and x^2 (with two different slope coefficients). The model with x^2 will always fit the data better, but it might not be a better model, because it might end up overfitting noise in the data.

For these kinds of cross-dimensional comparisons, R^2 is useless. In fact, in many ways it is worse than nothing: R^2 will always go up when we add new predictors, even if those predictors have nothing to do with the response. This is why it is crucial not to think of R^2 as measuring "goodness of fit." You'll just end up tying yourself into a mental knot pondering how a "good" model can still be rotten.

So let's put aside R^2. Instead, we need a criterion that balances the twin virtues of fit and simplicity. Put another way, we need a quantitative version of Occam's Razor, the philosophical principle that instructs us to make explanations only as complex as they need to be. When it comes to regression models, fit and simplicity are both easily operationalized. Models that fit more precisely have higher R^2, and lower residual variance. Simpler models have fewer free parameters to estimate.

In many ways its the same as the principle behind good writing: make your explanations precise, but also make them simple. Consider this passage from Ecclesiastes:

> I returned and saw under the sun, that the race is not to the swift, nor the battle to the strong, neither yet bread to the wise, nor yet riches to men of understanding, nor yet favour to men of skill; but time and chance happeneth to them all.

And now George Orwell's parody:

> Objective considerations of contemporary phenomena compel the conclusion that success or failure in competitive activities exhibits no tendency to be commensurate with innate capacity, but that a considerable element of the unpredictable must invariably be taken into account.[9]

[9] "Politics and the English Language," 1946

Orwell's parody is the literary version of an overfitted model. In regression, as in writing: try to say it as simply and elegantly as possible.

But now? Here are two commonly used criteria for balancing fit and simplicity in regression models.

1) Adjusted R^2, denoted R^2_A. We recall that the original R^2 was defined as

$$R^2 = 1 - \frac{UV}{TV}\,.$$

Adjusted R^2 is defined in almost the same way, but with a subtle modification to punish models that have larger numbers of parameters:

$$R^2_A = 1 - \frac{UV}{TV} \cdot \left(\frac{n-1}{n-p-1} \right) = 1 - \left\{ (1 - R^2) \cdot \left(\frac{n-1}{n-p-1} \right) \right\} .$$

Here n denotes the sample size, and p the number of regressors in the model (not counting the intercept), just as for the F test. Higher values of R^2_A are (ostensibly) better. But unlike R^2, R^2_A can sometimes go down when you add another predictor to the model. Adjusted R^2 is part of the standard output in most regression software, but even if you can't find it, it's easy to calculate from regular R^2.

2) AIC, which stands for "Akaike information criterion." Don't pay too much attention to the full name; what's important is not the derivation of AIC, but the manner in which it balances fit and simplicity:

$$\text{AIC} = n \log \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right\} + 2p \, .$$

The first term (the one involving the residual sum of squares) measures lack of fit; models with larger residuals fit the data less precisely. The second term measures complexity; p is the number of free parameters in the model, and is therefore larger for more complex models. From this, we reason that lower values of AIC are (ostensibly) better.[10]

Of course, the best Occam's Razor of all is to test models by having them predict y's for x's they've never "seen" before—that is, on fresh data that hasn't itself been used to fit the models. This is called out-of-sample predictive validation, and is very effective at winnowing down your list of good models. The only problem is that data is sometimes expensive, and you might not be able to collect enough extra data to run a good out-of-sample test. That's when these other Occam's Razors can be very useful.[11]

[10] The definition I've given for AIC may differ from the definitions given by other sources by a constant additive term. But since we use AIC to compare models, adding a constant to every model's AIC doesn't change which one is lowest.

[11] In particular, one can justify AIC as an approximation to the out-of-sample test error that you would get if you actually were able to collect a fresh data set and make predictions.