5
Quantifying uncertainty

Three questions involving uncertainty

In coming this far through the book, you've already learned many valuable skills: how to summarize evidence both graphically and numerically; how to fit basic group-wise and linear statistical models to data; how to combine grouping and numerical variables; how to use these models to explore trends and predict new outcomes; and how to summarize the information content of a model, by quantifying the model's predictive power in the context of the variation that remains unexplained.

But we're missing a crucial piece of the puzzle. Earlier we defined statistical modeling as the structured quantification of uncertainty. We've focused a lot so far on the "structure" part; now we'll begin to focus on the "uncertainty" part. Here are three archetypal questions that we will now take up in earnest.

(1) How confident are we in our estimate of an effect size? Take the following study of a new therapeutic regime for esophogeal cancer, from the New England Journal of Medicine in 2006:

> We randomly assigned patients with resectable adenocarcinoma of the stomach, esophagogastric junction, or lower esophagus to either perioperative chemotherapy and surgery (250 patients) or surgery alone (253 patients). . . . With a median follow-up of four years, 149 patients in the perioperative-chemotherapy group and 170 in the surgery group had died. As compared with the surgery group, the perioperative-chemotherapy group had a higher likelihood of overall survival (five-year survival rate, 36 percent vs. 23 percent).[1]

Thus the chemotherapy regime appears to save 1 additional person in 8, compared to surgery alone. But what if the physicians running the trial had enrolled a different sample of patients? Might the effect size have looked more like 1 patient in 6, or 1 in 16? Chemotherapy has nasty side effects

[1] Cunningham, et. al. "Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer." *New England Journal of Medicine*, 2006 July 6; 355(1):11-20.

and is very expensive. If you're a cancer patient or a Medicare administrator, uncertainty about the effect size matters.

(2) How confident are we in our forecast? We've touched on this one in the context of naïve prediction intervals. But in doing so we ignored a crucial fact: that uncertainty about a model's parameters translates into additional uncertainty about predictions, beyond the contribution of the residual.

(3) Could this association plausibly be due to chance? Several chapters ago, we noticed an apparent association for participants in the PREDIMED study between smoking and the risk of a cardiovascular event. Our 2x2 table looked like the one at right, and we calculated that the relative risk of a cardiovascular event was 1.94 for smokers, versus nonsmokers.

	Smoker?	
	No	Yes
No event	3778	2294
Event	114	138

But this is from a sample of people, not the entire global population. Suppose we had taken a different sample of people. Is it possible that we could have gotten a result so different that the association between smoking and cardiovascular events disappeared?

We use the phrase *statistical inference* to describe the framework and procedures we use to address questions like these. In this chapter, we'll address all three of these questions, using what is referred to as a resampling-based approach.

Sampling distributions, estimators, and alternate universes

DIFFERENT versions of these three questions come up again and again, in just about every data set you'll encounter. Luckily, once you know how to answer one of them, you will know how to answer all of them. That's because all three questions boil down to the same counter-factual: "if our data set had been different merely due to chance, would our answer have been different, too?" In fitting statistical models, we typically equate the trustworthiness of a procedure with its stability under the influence of luck, and we seek to measure the degree to which that procedure might have given a different answer if the forces of randomness had made the world look a bit different.

Confidence in your estimates $\iff$ Stability of those estimates under the influence of chance

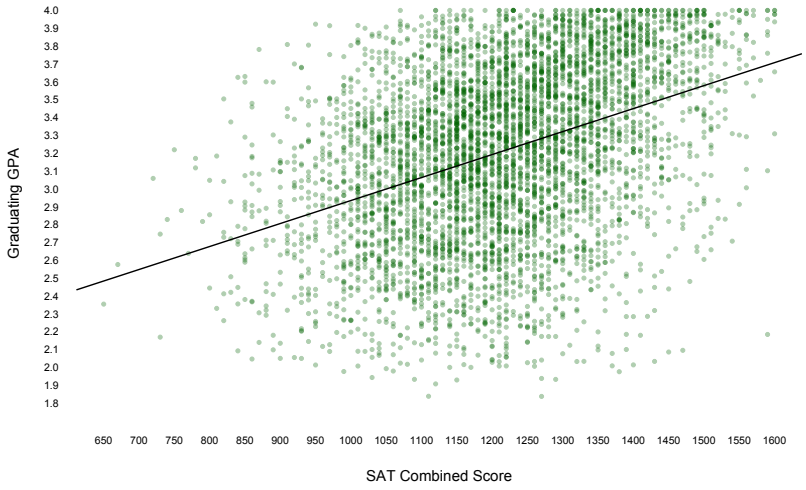


Figure 5.1: Graduating GPA versus high-school SAT score for all students who entered UT–Austin in the fall of 2000 and went on to earn a bachelor's degree within 6 years. The black line shows the least-squares fit.

You can see why it makes sense to equate stability with trustworthiness if you imagine a suspect who gives the police three different answers to the question, "Where were you last Tuesday night?" If the story keeps changing, there is little basis for trust.

Sources of instability. One obvious source of instability in our estimates is when our observations are subject to the forces of randomness. For example, suppose we wish to characterize the relationship between SAT score and graduating GPA for the entering class of 2000 at the University of Texas. Figure 5.1 shows the entire relevant population, yet there is still randomness to worry about—for, as the teacher in Ecclesiastes puts it, "time and chance happeneth to them all." If any of these 5,191 students had taken the SAT on a different day, or eaten a healthier breakfast on the day of their chemistry finals, we would be looking at a slightly different data set, and thus a slightly different least-squares line—even if the underlying SAT–GPA relationship had stayed the same.

Another source of instability is the effect of sampling variability, which arises when we're unable to study the entire population of interest. The key insight here is that a different sample would have led to different estimates of the model parameters. Consider the example above about the study of a new chemotherapy regime for esophogeal cancer. If doctors had taken a different sample of 503 cancer patients and gotten a drastically different estimate of the new treatment's effect, then the original estimate isn't very

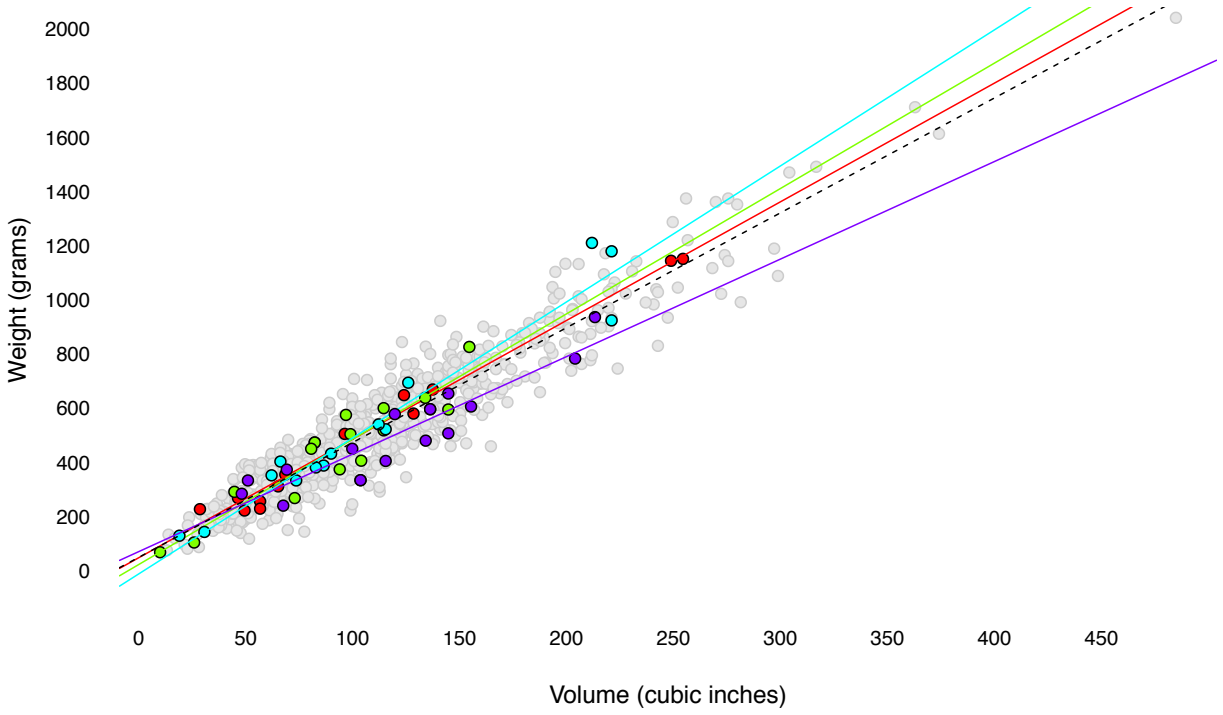


Figure 5.2: Four different days of fishing, coded by color, on an imaginary lake home to a population of 800 fish. On each day's fishing trip, you catch 15 fish, and end up estimating a slightly different weight–volume relationship. The dashed black line is the true relationship for the entire population.

trustworthy. If, on the other hand, pretty much any sample of 503 patients would have led to the same estimates, then their answer for *this particular* subset of 503 is likely to be accurate.

An example: simulating a sampling distribution by Monte Carlo. To get some intuition for this way of thinking, imagine that you go on a four-day fishing trip to a lovely small lake out the woods. The lake is home to a population of 800 fish of varying size and weight, depicted in Figure 5.2. On each day, you take a random sample from this population—that is, you catch (and subsequently release) 15 fish, recording the weight of each one, along with its length, height, and width (which multiply together to give a rough estimate of volume). You then use the day's catch to compute a different estimate of the volume–weight relationship for the entire population of fish in the lake. These four different days—and the four different least-squares fits—show up in different colors in Figure 5.2.

Four days of fishing give us some idea of how the estimates for β_0 and β_1 vary from sample to sample. But 2500 days of fishing,

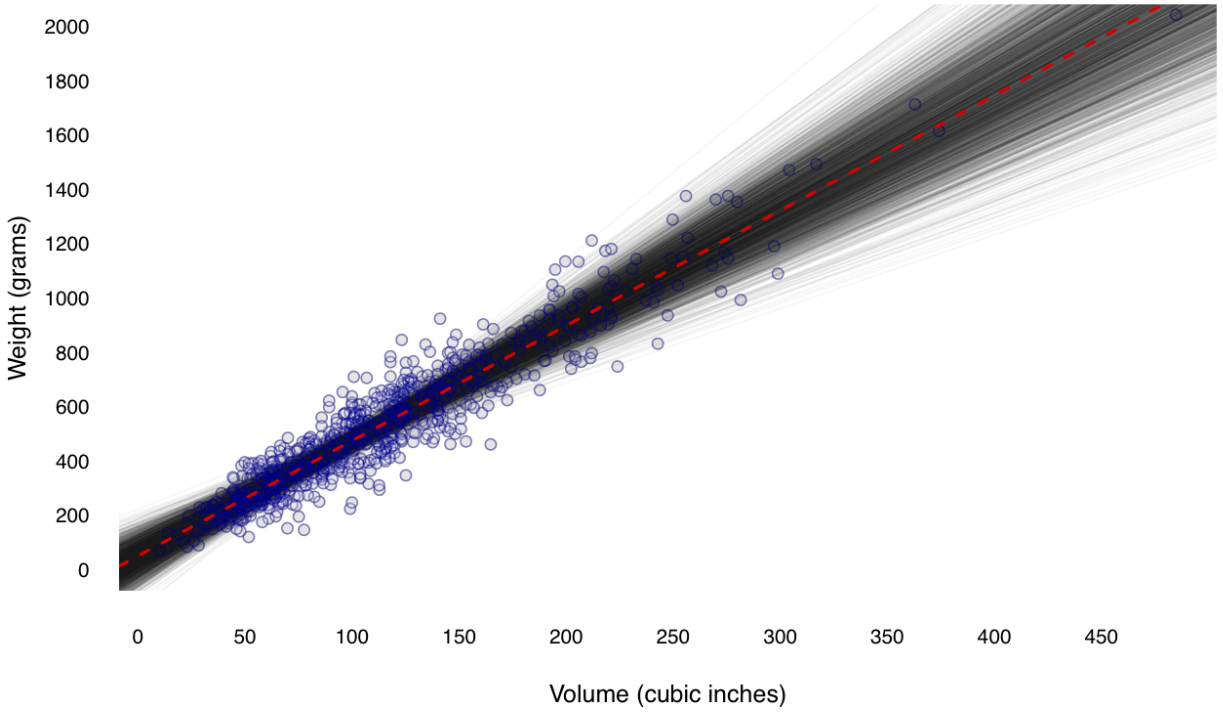
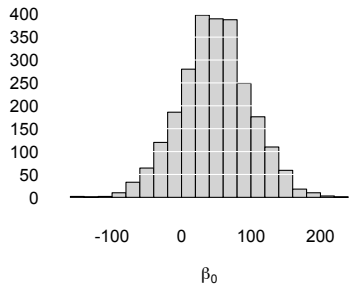
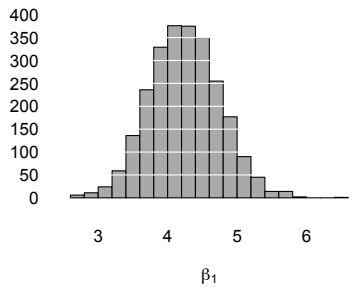


Figure 5.3: 2500 days of fishing, together with the 2500 different estimates of β_0 and β_1 (below), simulated by Monte Carlo.

simulated by computer, give us a better idea. Figure 5.3 shows just this: 2500 different samples of size 15 from the population, together with 2500 different least-squares estimates of the weight–volume relationship. This is an example of a *Monte Carlo simulation*, in which we run a computer program to repeatedly simulate a random process (in this case, sampling from a population).

These pictures show the *sampling distribution* of the least-squares line—that is, how the estimates for β_0 and β_1 change from sample to sample, shown in histograms in the right margin. In theory, to know the sampling distributions exactly, we'd need to take an infinite number of samples, but 2500 gives us a rough idea.





The sampling distribution. To understand the concept of a sampling distribution, it helps to distinguish between an *estimator* and an *estimate*. A good analogy here is that an estimator is to a court trial as an estimate is to a verdict. Just like a trial is a procedure for reaching a verdict about guilt or innocence, an estimator is a procedure for reaching an estimate of some population-level quantity on the basis of a sample. The least-squares procedure

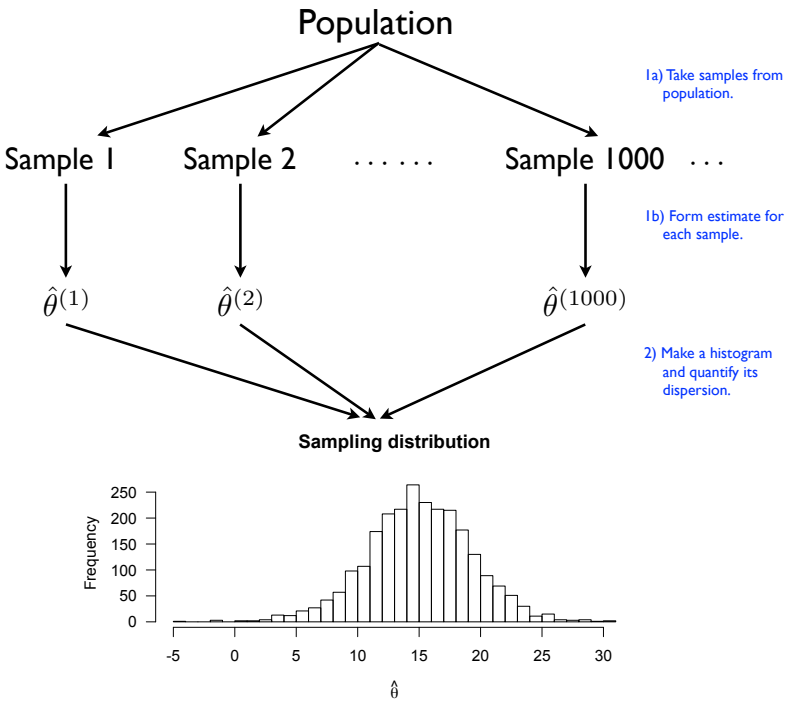


Figure 5.4: A stylized depiction of a sampling distribution of an estimator $\hat{\theta}$. To construct this distribution, we must imagine the following thought experiment. We repeatedly take many samples (say, 1000) from the population (step 1a). For each sample, we apply our estimator to compute the estimate $\hat{\theta}^{(r)}$ (step 1b). At the end, we combine all the estimates $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(1000)}$ into a histogram, and we summarize the dispersion of that histogram (step 2). Technically, the sampling distribution is the distribution of estimates we'd get with an infinite number of samples, and the histogram is an approximation of this distribution. The difference between the true distribution and the approximation generated by Monte Carlo is called *Monte Carlo error.*

is a specific set of steps (i.e. equations) that one applies to a data set. The procedure yields estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for the slope and intercept of a population-wide linear trend; while the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ you get for a specific data set are the estimates. An estimator's sampling distribution is the distribution of results (that is, the estimates) that one obtains from that estimator under repeated sampling from a population. Figure 5.4 shows graphically how, in principle, this distribution is constructed. Concrete examples of an estimator include the sample mean, the least squares procedure, and the residual standard deviation.

Good estimators are those that usually yield estimates close to the truth, with minimal variation. Therefore, we typically summarize a sampling distribution using its standard deviation, which we refer to as the *standard error*.[2] In quoting the standard error of an estimator's sampling distribution, you are saying: "If I were to take repeated samples from the population and use this estimator for every sample, my estimate is typically off from the truth by about this much." Notice again that this is a claim about a procedure, not a particular estimate. The bigger the standard error, the

[2] We are also sometimes interested in the mean of a sampling distribution. If the mean of an estimator's sampling distribution is equal to the true population value, we say that the estimator is *unbiased*. This term has a precise mathematical meaning, but also an unwarranted connotation of universal desireability that many statisticians find problematic. Alas, for historical reasons, we're basically stuck with the term. It turns out that unbiasedness is not always a good property of an estimator. There can be very good reasons to use estimators that we know to be biased. But that's for another book.

less stable the estimator across different samples, and the less you can trust the estimate for any particular sample. To give a specific example, for the 2500 samples in Figure 5.3, the standard error of $\hat{\beta}_0$ is about 50, while the standard error of $\hat{\beta}_1$ is about 0.5.

Of course, if you really could take repeated samples from the population, life would be easy. You could simply peer into all of those alternate universes, tap each version of yourself on the shoulder, and ask, "What slope and intercept did you get for *your* sample?" By tallying up these estimates and seeing how much they differed from one another, you could discover precisely how much confidence you should place in your own estimates of β_0 and β_1, and report appropriate error bars based on the standard error of your estimator.[3]

[3] Let's ignore the obvious fact that, if you had access to all those alternate universes, you'd also have more data. The presence of sample-to-sample variability is the important thing to focus on here.

Most of the time, however, we're stuck with one sample, and one version of reality. We cannot know the actual sampling distribution of our estimator, for the same reason that we cannot peer into all those other lives we might have lived, but didn't:

> Two roads diverged in a yellow wood,
> And sorry I could not travel both
> And be one traveler, long I stood
> And looked down one as far as I could
> To where it bent in the undergrowth. . . .[4]

[4] Robert Frost, *The Road Not Taken*, 1916.

Quantifying our uncertainty would seem to require knowing all the roads not taken—an impossible task.

Surprisingly, we can come close to performing the impossible. There are two ways of feasibly constructing something like the histogram in Figure 5.4, thereby approximating an estimator's sampling distribution without ever taking repeated samples from the population.

1) Resampling: that is, by pretending that the sample itself is the population, which allows one to approximate the effect of sampling variability by resampling from the sample.

2) Parametric probability modeling: that is, by assuming that the forces of randomness obey certain mathematical regularities, and by drawing conclusions about these regularities using probability theory.

In this chapter, we'll discuss the resampling approach, deferring the probability-modeling approach to a later chapter.

Bootstrapping: standard errors through resampling

At the core of the resampling approach to statistical inference lies a simple idea. Most of the time, we can't feasibly take repeated samples of size n from the population, to see how our estimate changes from one sample to the next. But we can repeatedly take samples of size n *from the sample itself*, and apply our estimator afresh to each notional sample. The idea is that the variability of the estimates across all these samples can be used to approximate our estimator's true sampling distribution.

This process—pretending that our sample is the whole population, and taking repeated samples of size n with replacement from our original sample of size n—is called *bootstrap resampling*, or just *bootstrapping*.[5] Each block of n resampled data points is called a bootstrapped sample. To bootstrap, we write a computer program that repeatedly resamples our original sample and recomputes our estimate for each bootstrapped sample. Modern software makes a non-issue of the calculational tedium involved.

You may be puzzled by something here. There are n data points in the original sample. If we repeatedly resample n data points from our "pseudo-population" of size n, won't each bootstrapped sample be identical to the original sample? If so, and every bootstrapped sample looks the same, then how can this process be used to simulate sampling variability?

This fact highlights a key requirement of bootstrapping: the resampling must be done *with replacement* from the original sample, so that each bootstrapped sample contains duplicates and omissions from the original sample.[6] These duplicates and omissions induce variation from one bootstrapped sample to the next, mimicking the variation you'd expect to see across the real repeated samples that you can't take.

To summarize, let's say we have a data set D, consisting of n cases. We want to understand how our estimator $\hat{\theta}$ might have behaved differently with a different sample of size n. To answer this question using bootstrapping, we follow two main steps.

(1) Repeat the following substeps many times (e.g. 1000 or more):

   a. Generate a new bootstrapped sample $D^{(r)}$ by taking n samples with replacement from D.

   b. Apply the estimator $\hat{\theta}$ to the bootstrapped sample $D^{(r)}$ and save the resulting estimate, $\hat{\theta}^{(r)}$.

[5] The term "bootstrapping" is a metaphor. It is an old-fashioned phrase that means performing a complex task starting from very limited resources. Imagine trying to climb over a tall fence. If you don't have a rope, just "pull yourself up by your own bootstraps."

[6] Imagine a lottery drawing, where there's a big urn with 60 numbered balls in it. We want to choose a random sample of 6 numbers from the urn. After we choose a ball, we could do one of two things: 1) put the ball to the side, or 2) record the number on the ball and then throw it back into the urn. If you set the ball aside, it can be selected only once; this is sampling without replacement, and it's what happens in a real lottery. But if instead you put the ball back into the urn, it has a chance of being selected more than once in the final sample; this is sampling with replacement, and it's what we do when we bootstrap.

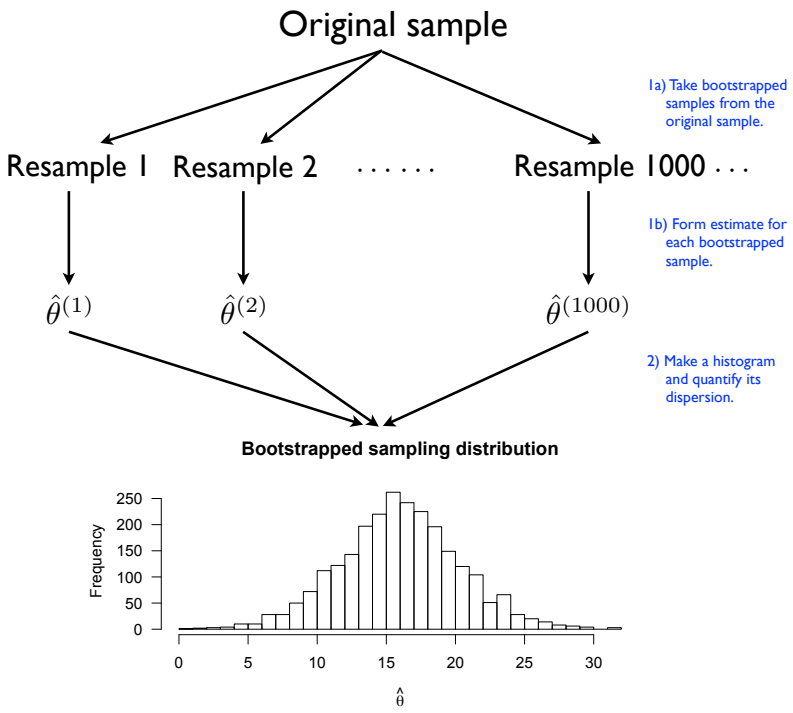


Figure 5.5: A stylized depiction of a bootstrapped sampling distribution of an estimator $\hat{\theta}$. We have a single original sample. We repeatedly take many bootstrapped samples (say, 1000) from the original sample (step 1a). For each resample, we compute the estimator $\hat{\theta}$ (step 1b). At the end, we combine all the estimates $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(1000)}$ into a histogram of the bootstrapped sampling distribution, and we summarize the dispersion of that histogram (step 2). Compare with Figure 5.4.

(2) Take all of the $\hat{\theta}^{(r)}$'s you've generated and make a histogram. This is your estimate of the sampling distribution.

See Figure 5.5, and compare with Figure 5.4.

Resampling won't yield the true sampling distribution of an estimator, but it is often good enough for approximating the standard error (which you'll remember is just the standard deviation of the sampling distribution). We use the term *bootstrapped standard error* for the standard deviation of the bootstrapped sampling distribution. The bootstrapped standard error is an estimate of the true standard error.

The quality of this estimate depends almost entirely on one thing: how closely the original sample resembles the wider population. This is a question of judgment best answered by someone with subject-area expertise relevant to the data set at hand. As a data analyst this often isn't under your control, and therefore it's almost worth remember that the bootstrap is not entirely free of assumptions. You can't magic your way to sensible estimates of the true sampling distribution by bootstrapping a biased, woefully small, or otherwise poor sample.

The quality of the Monte Carlo approximation also depends to a lesser extent on how many bootstrapped samples you take from the original sample. Simulating more bootstrapped samples help to reduce the variability inherent in any Monte Carlo simulation—up to a point. But taking more bootstrapped samples is never a substitute for having more actual samples in the real data set. Fundamentally, it is the size of your original sample that governs the precision of your estimates.

A natural question is: how well does bootstrapping work in practice? To see the procedure in action, let's reconsider the least-squares estimator of the slope (β_1) for the weight–volume line describing the fish in our hypothetical lake. The top row of Figure 5.6 shows three actual sampling distributions, corresponding to samples of size $n = 15$, $n = 50$, and $n = 100$ from the entire population. These were constructed using the Monte Carlo method described several pages ago, as depicted in Figures 5.3 and 5.4. For example, the top left panel (for $n = 15$) was constructed by taking 2,500 Monte Carlo samples from the true population in Figure 5.3, and computing the least-squares estimate of the slope for each sample as in Figure 5.4.

Below each true sampling distribution, we have focused on four of these 2500 samples. For each of these real samples, we ran the bootstrapping procedure by 2500 bootstrapped samples from the original sample of size n, treating it as a pseudo-population. For each bootstrapped sample, we compute the least-squares line for weight versus volume. These 2500 estimates of β_1 are what you see in each grey-colored panel of Figure 5.6. For example, the first grey panel in column 1 corresponds to the bootstrapped sampling distribution from the first sample of size 15; the second grey panel corresponds to the bootstrapped sampling distribution from the second sample of size 15; and so on for the rest of the grey panels.

If bootstrapping were perfect, each grey panel would look exactly like the corresponding orange panel above, regardless of the same size. But of course, bootstrapping isn't perfect. If you study these pictures closely, you'll notice a few things.

(1) The bootstrapped sampling distribution can differ substantially from one original sample to the next (top to bottom). The sample-to-sample differences are larger when the original sample size is small.

(2) The bootstrapped sampling distribution gets both closer to the truth, and less variable from one original sample

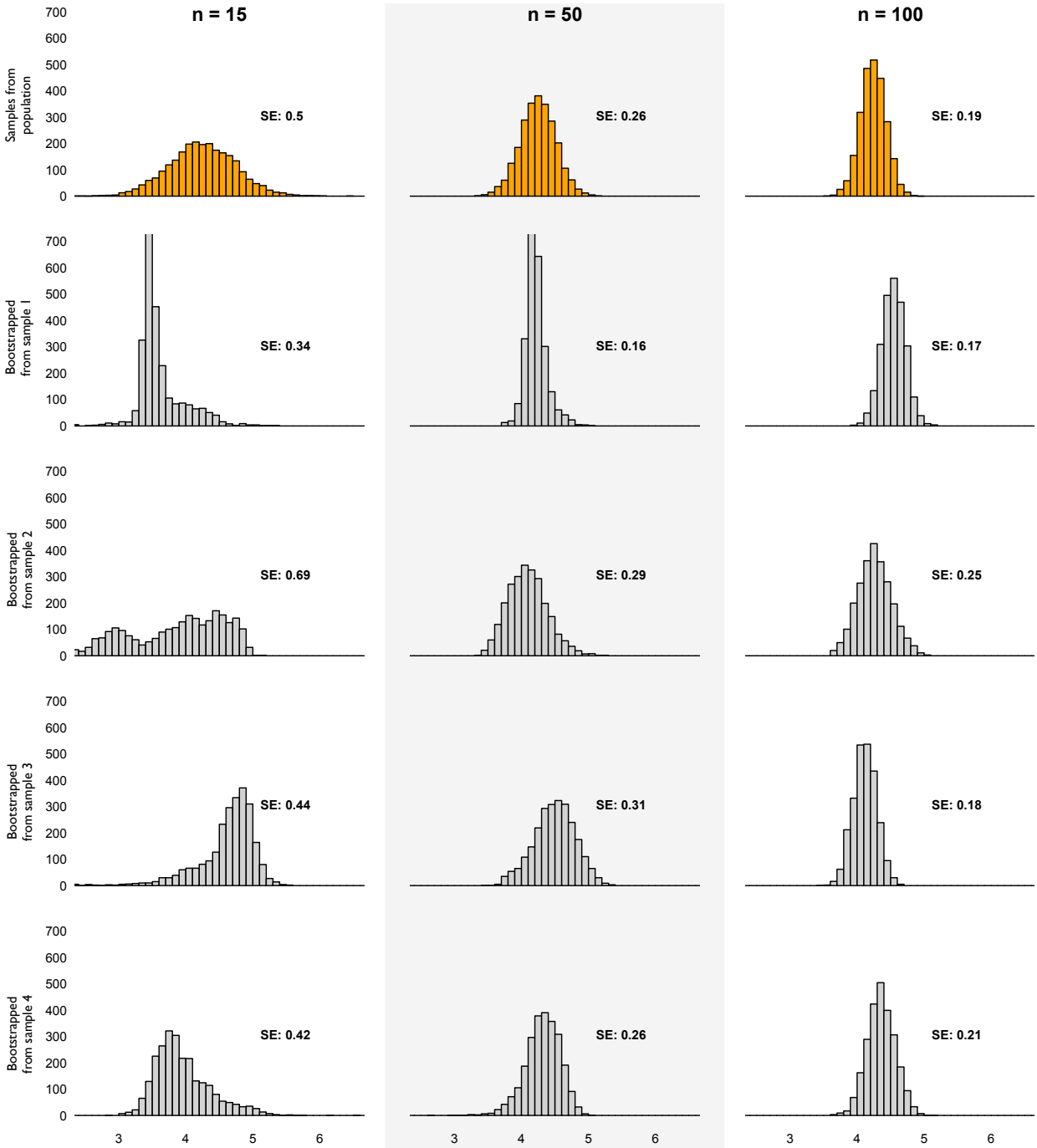


Figure 5.6: Actual (top, in orange) and bootstrapped sampling distributions (four replications) for the least-squares estimator of β_1 from Figure 5.2.

to the next, as the original sample size gets larger.

(3) The bootstrapped standard errors (printed next to each histogram) are often closer to the true standard error than you would expect, based on the visual correspondence of the bootstrapped sampling distribution to the true one.

Confidence intervals and coverage

Now that we've learned to approximate an estimator's sampling distribution via bootstrapping, what do we do with this information? The simplest answer is: we quantify the uncertainty of our estimate via a *confidence interval*, which is just is a range of plausible values for the truth, together with an associated *confidence level* between 0% and 100%. The width of a confidence interval conveys the precision with which the data have allowed you to estimate the underlying population parameter. If your interval actually contains the true population value, we say that the interval *covers* the truth. If it doesn't, the interval *fails to cover* the truth. In real life, you won't know whether your interval covers. The confidence level expresses how confident you are that it actually does.

There are many ways of generating confidence intervals from bootstrapped sampling distributions, ranging from the simple to the highly sophisticated (and mathematically daunting). We'll focus on two simple ways here, with the understanding that the more technical ways we don't discuss are a bit more accurate.[7]

First, there's the basic standard-error method. Here, you quote a symmetric error bar centered on the estimate from the original sample, plus-or-minus some multiple k of the bootstrapped standard error. To be precise, let's say that θ is some population parameter you're trying to estimate; that $\hat{\theta}$ is the estimate of θ generated by your actual sample; and that you've run the bootstrapping procedure on your sample and found that the bootstrapped standard error is $\hat{\sigma}$. Your confidence interval would then be

$$\theta \in \hat{\theta} \pm t^{\star}\hat{\sigma}\,,$$

where $t^{\star}$ is a chosen multiple. This number $t^{\star}$ is called the *critical value*. It is the number of standard errors you must go out from the center to capture a certain percentage of the sampling distribution. Typical values are $t^{\star} = 1$ (for an approximate 68% confidence interval) and $t^{\star} = 2$ (for an approximate 95% confidence interval).[8]

[7] If you want to get an introduction to the more technical ways of getting confidence intervals from the bootstrap, see the following article: "Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians." James Carpenter and John Bithell. *Statistics in Medicine* 2000; 19:1141–64.

[8] Precisely why $t^{\star} = 1$ corresponds to 68% and $t^{\star} = 2$ to 95% are beyond the scope of this chapter. It has to do with the normal distribution and something called the central limit theorem. For now, it is fine if you accept this is an empirical rule of thumb that statisticians have found gives a good approximation in situations where your bootstrapped sampling distribution looks approximately bell-shaped. Some of the more sophisticated ways of improving the bootstrap, mentioned in Footnote 7, are focused on improving the choice of $t^{\star}$ given by these simple guidelines.

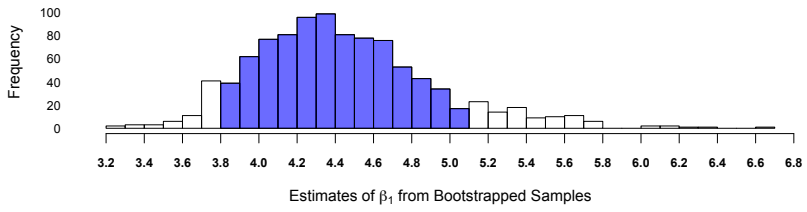


Figure 5.7: The estimated sampling distribution of $\hat{\beta}_1$ that arises from bootstrapping one sample of size 30 from the full fish population. The blue area reflects an 80% confidence interval generated by the coverage method, with symmetric tail areas of 10% above and 10% below the blue area.

Second, there's the coverage method, in which you simply calculate a coverage interval using the quantiles of your bootstrapped sampling distribution. For example, Figure 5.7 shows the bootstrapped sampling distribution for the slope of the weight–volume relationship arising from a single sample of 30 fish from the same lake as before. If you wanted to compute an 80% confidence interval based on this data, you would calculate the 10th and 90th percentiles of this histogram, giving you an interval that contains 80% of the bootstrapped estimates of the slope. In Figure 5.7, this interval is $(3.8, 5.1)$, shown in blue. This example highlights that, unlike the intervals generated by the standard-error method, the intervals generated by the coverage method need not be symmetric about the estimate $\hat{\theta}$ derived from your actual sample.

Is one of these two methods better? Not as a general rule. The coverage-interval approach is more common in practice, and it's a fine default option. The most conservative thing to do, assuming you don't want to go the very technical[9] route, is to compute both and report the wider interval.

[9] See Footnote 7.

What does "confidence" mean?

The word "confidence," as it is used in the phrase "confidence interval," has a notoriously tricky interpretation. To put it concisely but opaqely, confidence intervals are intervals generated by a method that satisfies the frequentist coverage principle. We state this principle roughly as follows.

> *The frequentist coverage principle:* If you were to analyze one data set after another for the rest of your life, and were to quote X% confidence intervals for every estimate you made, those intervals should cover their corresponding true values at least X% of the time. Here X can be any number between 0 and 100.

Let's unpack this a bit. Imagine that your interval was gener-

ated with a procedure that, under repeated use on one sample after the next, tends to yield intervals that cover the true value with a relative frequency of at least 80%. Then, and only then, may you claim a bona fide 80% confidence level for your specific interval. (You may, of course, aim for whatever coverage level you wish in lieu of 80%. Many people seem stuck on 95%, but it's entirely your choice.) Thus confidence intervals involve something of a bait-and-switch: they purport to answer a question about an individual interval, but instead give you information about some hypothetical assembly line that could be used to generate a whole batch of intervals. Nonetheless, there is an appealing "truth in advertising" property at play here: that if you're going to claim 80% confidence, you should be right 80% of the time over the long run.

An obvious question is: do bootstrapped confidence intervals satisfy the frequentist coverage property? If your sample is fairly representative of the population, then the answer is a qualified yes. That is, the bootstrapping procedure yields nominal X% intervals that cover the true value "approximately" X% of the time. Moreover, as the size of the original sample gets bigger, the quality of the approximation gets better. Alas, it is necessary to appeal to some very advanced probability theory to put both of these claims on firm footing. (This is best deferred to another, much more advanced book. If you're of a mathematical bent, the relevant branch of probability theorem is called empirical-process theory, which part of a wider area called stochastic processes.)

For our purposes, it is better to show the procedure in action. Figure 5.8, for example, depicts the results of running 100,000 regressions—1,000 bootstrapped samples for each of 100 different real samples from the population in Figure 5.2. The vertical black line shows the true population value of the weight–volume slope ($\beta_1 = 4.24$) for our population of fish. Each row corresponds to a different actual sample of size $n = 30$ from the population. Dots and crosses indicate the least-squares estimate of the slope arising from that sample, while the grey bars show the corresponding 80% bootstrapped confidence intervals generated by the coverage method (just like the blue region in Figure 5.7).

The nominal confidence level of 80% for each individual interval must be construed as a claim about the *whole ensemble* of 100 intervals: 80% should cover, 20% shouldn't. In fact, 83 of these intervals cover and 17 don't, so the claim is approximately correct.

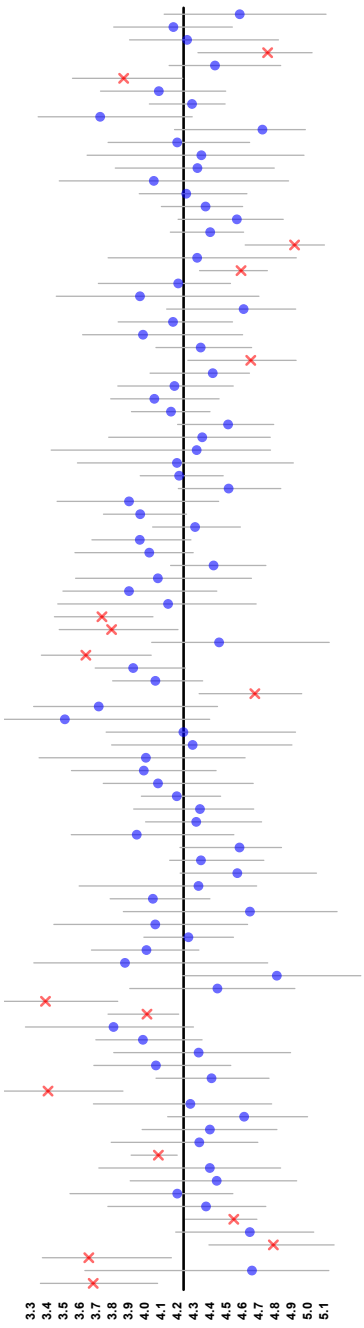


Figure 5.8: 100 different samples of size 30 from the population in Figure 5.2, along with each least-squares estimate of the weight–volume slope, and an 80% bootstrapped confidence interval, just like that at the top left. Blue dots show confidence intervals that cover; red crosses show those that don't.

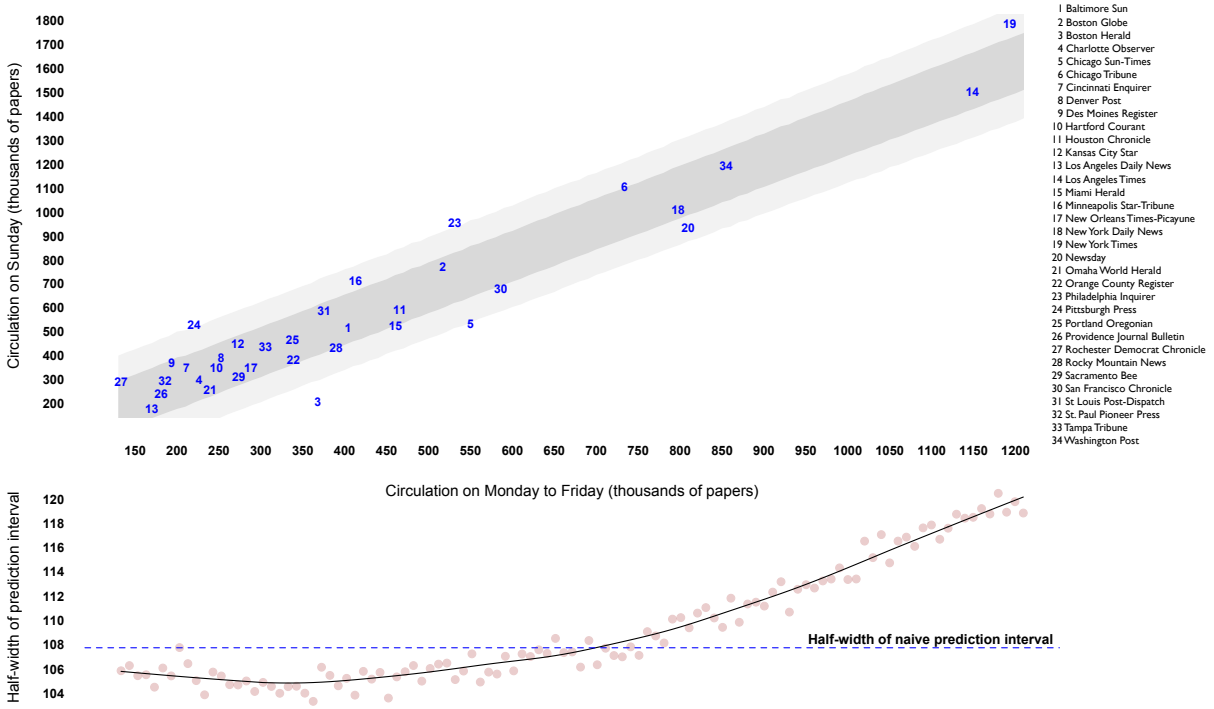


Figure 5.9: Sunday circulation versus daily circulation for 34 major metropolitan newspapers, together with one- and two-standard-deviation bootstrapped prediction intervals across the range of the X variable (top panel). Also shown is the half-width of the darker-grey prediction interval across the range of X (bottom panel), versus the half-width of the naïve prediction interval, shown by the dotted blue line.

You'll notice that the pink dots marking the half-width of each bootstrapped prediction interval wiggle up and down a bit from the black curve. This happens because we only took 2,500 bootstrap samples, which produces a bit of unwanted noise. Taking more bootstrapped samples would make the pink points fall closer to the black curve, but it wouldn't shift the black curve up or down.

Bootstrapped prediction intervals

Recall the problem of forecasting a future $y^\star$ corresponding to some predictor $x^\star$, using past data as a guide. (For example, how much should a used truck with 80,000 miles cost? How much can an Austin restaurant with a food rating of 7.5 charge for a meal?) Previously, we were content to quote a naïve prediction interval—for example,

$$\hat{y}^\star \in \hat{\beta}_0 + \hat{\beta}_1 x^\star \pm s_e \,,$$

or the best guess, plus-or-minus one residual standard deviation.

What made the prediction intervals naïve was the way we ignored uncertainty in our estimates for β_0 and β_1. For example, imagine that you work for a major metropolitan newspaper with a daily (Monday–Friday) circulation of 200,000 newspapers, and that your employer is contemplating a new weekend edition. You could certainly use the data in Figure 5.9, which correlates Sunday circulation with daily circulation for 34 major metropolitan newspapers, to inform your guess about the new Sunday edition's likely circulation. But the naïve prediction interval will mask real

sources of uncertainty. These may be large.

Luckily, now that we understand the logic of the bootstrap, we can try to account for this extra uncertainty. Just repeat the following steps a few thousand times:

(1) Take a single bootstrapped sample from the original sample, and compute the least-squares estimates $\hat{\beta}_0^{(r)}$ and $\hat{\beta}_1^{(r)}$. This gives you your best guess for the future y, given the information in the bootstrapped sample:

$$\hat{y}^{(r)} = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x^\star \,.$$

Here the superscript r denotes the r^{th} resample.

(2) Sample a residual $e^{(r)}$ at random from the bootstrapped least-squares fit, to mimic the unpredictable variation in the model.

(3) Set $y^{(r)} = \hat{y}^{(r)} + e^{(r)}$. This is your notional "future y" for the r^{th} bootstrapped sample.

If you take the standard deviation of all those $y^{(r)}$'s, you can directly quantify the uncertainty in your prediction at $x^\star$—for example, by quoting the dark- and light-grey prediction intervals in Figure 5.9, which stretch to one and two standard deviations (respectively) on either side of the least-squares line.

One noticeable feature of the bootstrapped prediction intervals is the way they bend outwards as they get further away from the center of the sample. This is a bit hard to see in the top panel of Figure 5.9. To show this effect more clearly, the bottom panel explicitly plots the half-width of the dark grey bootstrapped prediction intervals at 109 different hypothetical X points: every increment of 10,000 newspapers across the entire range of daily circulation, from 130,000 to 1.2 million.

The black curve shows an unmistakeable trend. Prediction uncertainty increases when you move away from the mean of X. Figure 5.3, several pages earlier, will give you some intuition for why this is so: small differences in the slope get magnified when you move further away from the middle of the sample. The naïve prediction interval fails to capture this effect entirely. On this problem, for example, the naïve interval understates prediction uncertainty by 10,000 newspapers or more for large values of X.

A final point worth noting: all of the previous warnings about bootstrapped standard errors also apply to bootstrapped prediction intervals. If the observed data is unrepresentative of the population, bootstrapping will mislead rather than inform.

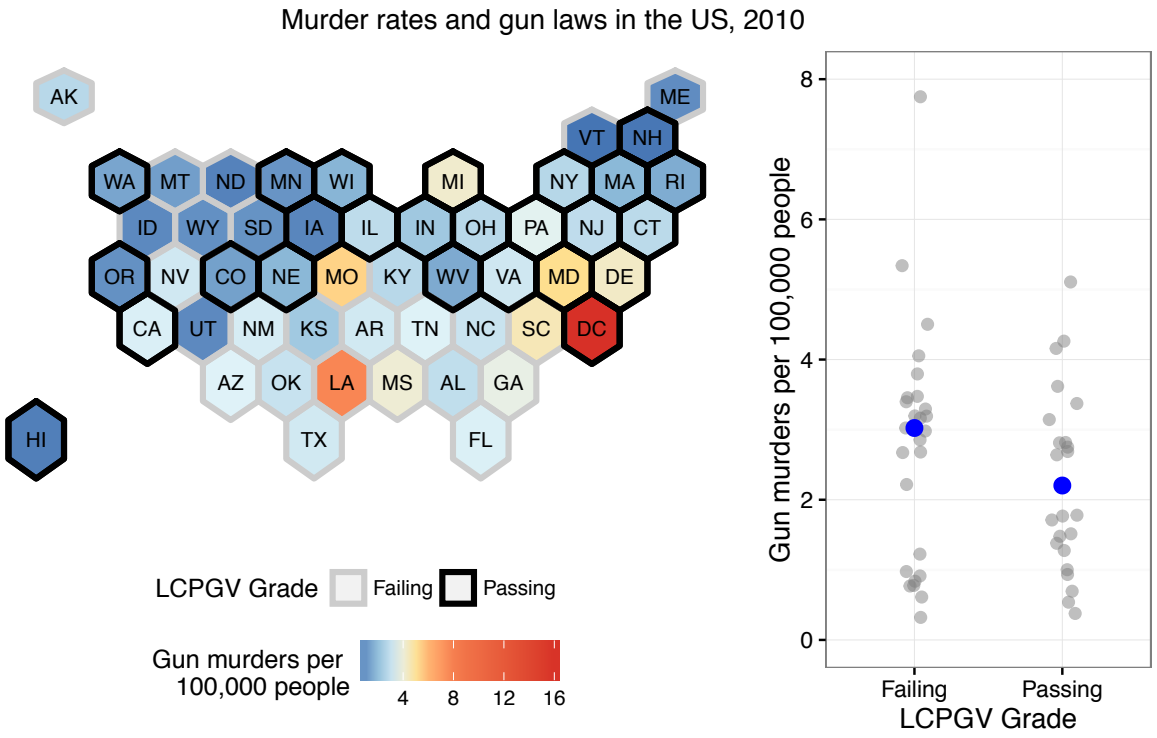


Figure 5.10: Left panel: a chloropleth map of murder rates versus gun laws across the U.S. states. The shaded color shows the state's gun-murder rate; blue is lower, and red is higher. The outline indicates whether a state's gun-control laws received a passing or a failing grade from the Law Center to 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.

Assessing the evidence for a hypothesis

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—especially 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 heritage and a basic right protected by the U.S. Constitution. As of 2016, 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 5.10 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.[10] In the chloropleth map in Figure 5.10, the

[10] 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.

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.[11]

The right panel of Figure 5.10 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 would this difference likely be?

Thus there are two hypotheses that can explain Figure 5.10:

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

(2) The observed correlation 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. 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.[12] 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.

[11] 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.

[12] "Null hypothesis" 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.

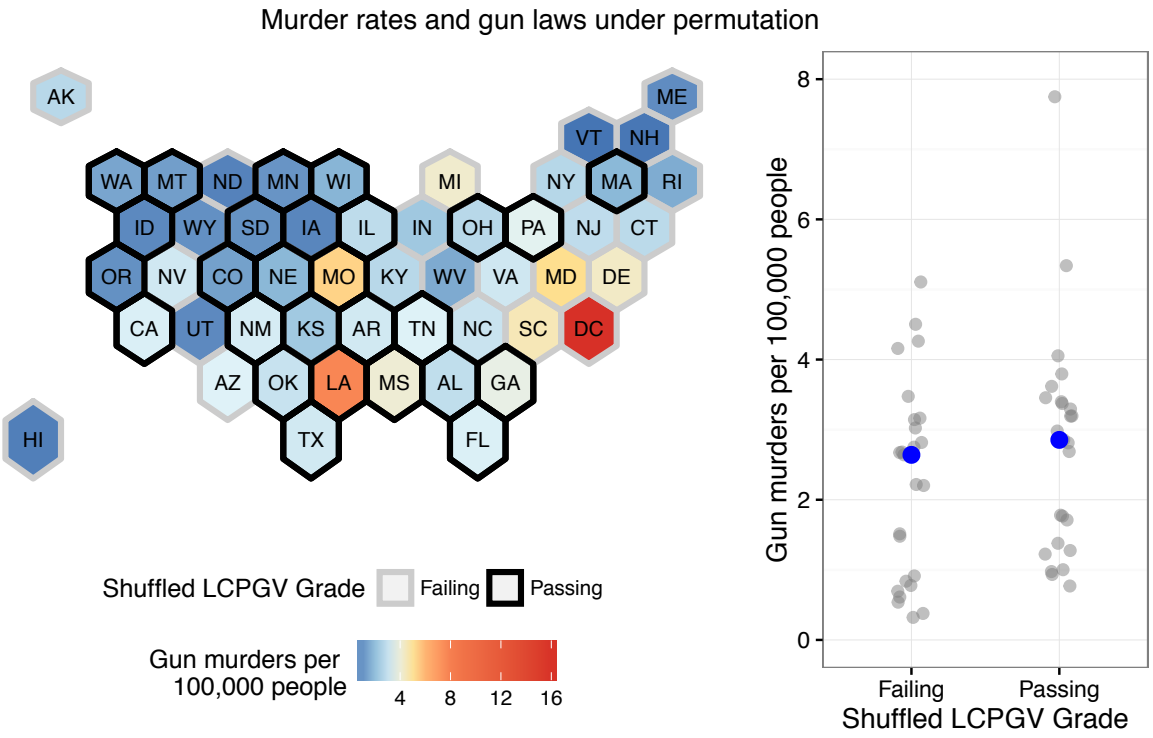


Figure 5.11: This map is almost identical to Figure 5.10, with one crucial difference: the identities of the states with passing and failing grades have been randomly permuted. There is still a small difference in the medians of the notionally passing and failing groups, due to random variation in the permutation process.

Permutation tests: shuffling the cards

Using the power of Monte Carlo simulation, we can assess which of these two hypotheses looks more plausible in light of the data. Figure 5.11 shows a map and dotplot very similar to those in Figure 5.10, with one crucial difference: in Figure 5.11, 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 random 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.

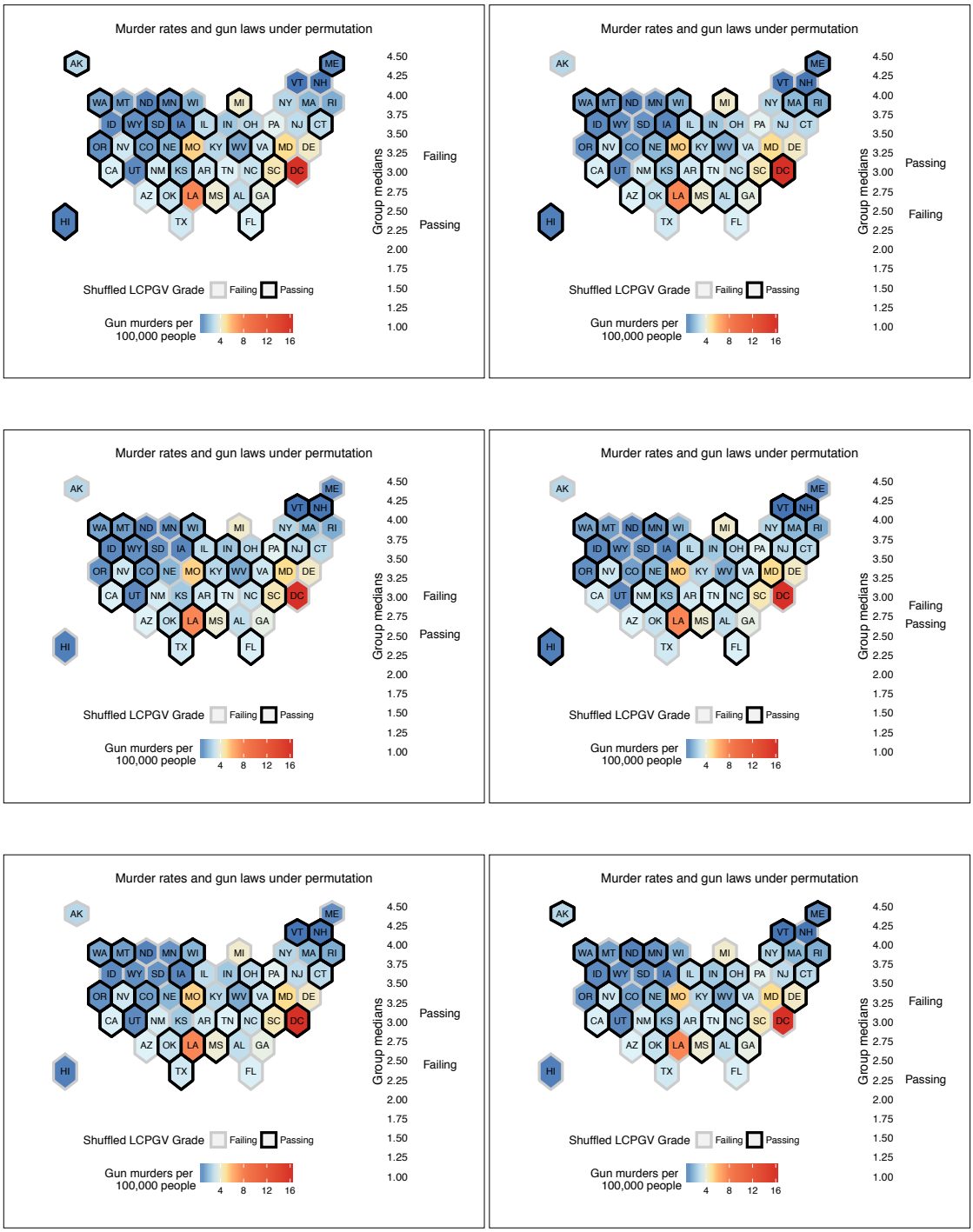


Figure 5.12: Six maps with permutated gun-law grades, with the medians for the passing and failing groups.

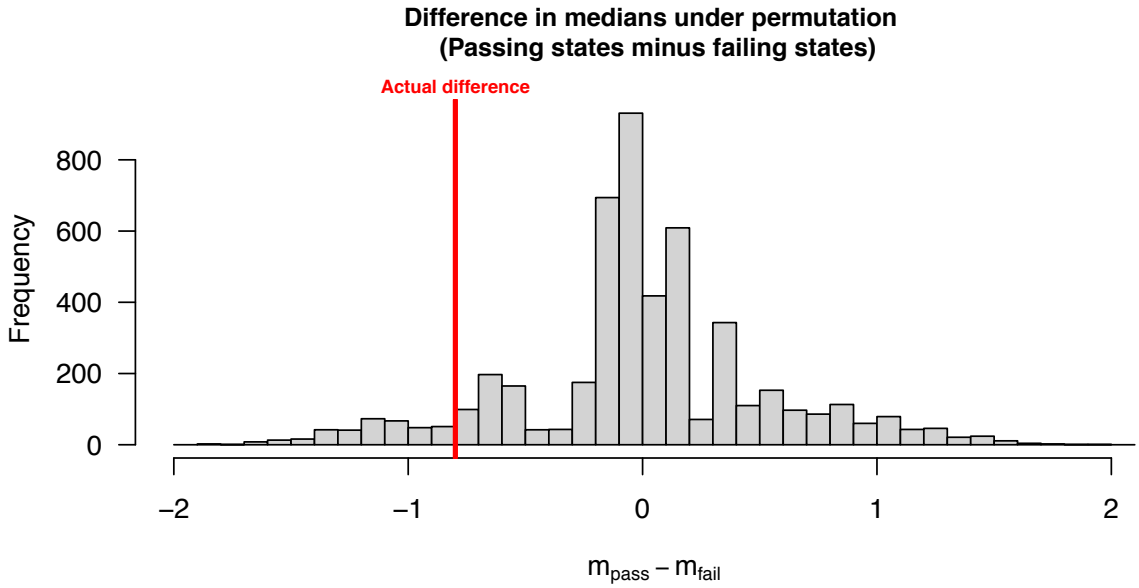


Figure 5.13: The histogram shows the difference in group medians for 5,000 simulated maps generated by the same permutation procedure as the 6 maps in Figure 5.12. Negative values indicate that the "failing" states had higher rates of gun violence than the "passing" states. The actual difference in medians for the real map in Figure 5.10 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.

But Figure 5.11 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 5.12 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 5.10.

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 5.13 ad-

dresses 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 5.13.

Conclusions

Let's use some more specific vocabulary to describe 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 those with failing grades.[13] Just 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.

[13] Remember that a statistic is any numerical summary of a data set.

Figure 5.13 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 5.12, as we do in the real map from Figure 5.10. Then there's no reason to be especially impressed by the actual value of $\delta = -0.8$ we calculated from the real map.[14] 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.[15]

[14] 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.

[15] Always remember that *failing to reject* the null hypothesis is not the 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 5.13? Here, the actual difference of -0.8 for the real map in Figure 5.10

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

In the previous example, we accomplished step (3) by randomly permuting the values of the predictor (gun laws) and recomputing the test statistic for the permuted data set. This is called a *permutation test* when done in the context of these four steps. 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.

The logic of frequentist hypothesis testing

EVERY part of the four-step recipe we followed in our permutation test seems straightforward, until we get to step (4): check whether the observed test statistic is consistent with the null hypothesis. In our analysis of the gun-laws data, we accomplished this step simply by making a qualitative judgment about where the actual value $\delta = -0.8$ sat within the histogram of values simulated under H_0 and shown in Figure 5.13. But ideally we could place this decision upon a firmer footing. Thus the question is: how we do quantify the degree to which our observed statistic is consistent

with the sampling distribution under the null hypothesis, calculated in step (3)?

We can think of this question as having two sub-questions. First, suppose that we go into the permutation test believing in the null hypothesis. What is our threshold of believable surprise, beyond which the data—as summarized by the statistic Δ—will change our minds? We would clearly start to feel uncomfortable with the null hypothesis if the actual difference in group medians had turned out to be $\delta = -3$. Such an extreme difference is far off the left-hand side of the histogram in Figure 5.13. Likewise, if the observed difference were merely $\delta = -0.1$, we would obviously have no reason to change our minds; the null hypothesis is capable of explaining differences much larger than this. Somewhere in between -0.1 and -3, there is room for doubt. What, precisely, is the threshold value of Δ that would lead us to abandon our belief in the hypothesis that observed difference in group medians is due to random variation?

This threshold is called the *critical value* of the test, and the values of the statistic equal to or beyond the critical value are referred to collectively as the *rejection region*. This is because we will reject the null hypothesis if we observe a value of Δ at least as extreme as the critical value.[16]

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.

Once we've chosen a critical value, the second sub-question is much simpler: yes or no, does the value of the statistic from the real data (δ) fall in the rejection region? 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.

[16] 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 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 under the repeated application of that threshold. This is why it is sometimes called *frequentist* testing, to distinguish it from Fisherian or Bayesian testing.[17]

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

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 to reject the null hypothesis if, in the real data, we see a difference in medians of -0.5 or larger.[18] But even under a random permutation of each state's notional gun laws, we would expect to see differences as large is -0.5 with reasonable frequency. If the null hypothesis were true and we saw $\Delta = -0.5$ 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 a difference in medians of at least $\Delta = -2$ 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 real difference were only -0.5, then we'd stand a very real chance of committing

[17] 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.

[18] Here "larger" means more negative, i.e. larger in the expected direction of lower gun-murder rates in states with passing gun-law grades.

a Type-II error instead.

How is one to negotiate this tradeoff? Let's recall the four basic steps from above:

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

The Neyman–Pearson approach prescribes the following sub-steps for step (4).

(4a) Upon inspecting the sampling distribution from step (3), choose a choose a rejection region R, corresponding to set of outcomes for the test statistic Δ that will lead you to reject the null hypothesis. The idea is that observating any value for Δ that lies within the rejection region R would be too surprising for you to go on believing in the truth of the null hypothesis.

(4b) 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.

(4c) If the observed value of the test statistic (δ) 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 gun-laws map, and conduct a formal Neyman–Pearson test of the null hypothesis that random variation can explain the observed difference in gun-violence rates between the states with passing gun-law grades and those with failing grades. We will follow the steps prescribed above, using the sampling distribution in Figure 5.14 (which is the same as the one in Figure 5.13) as a guide.

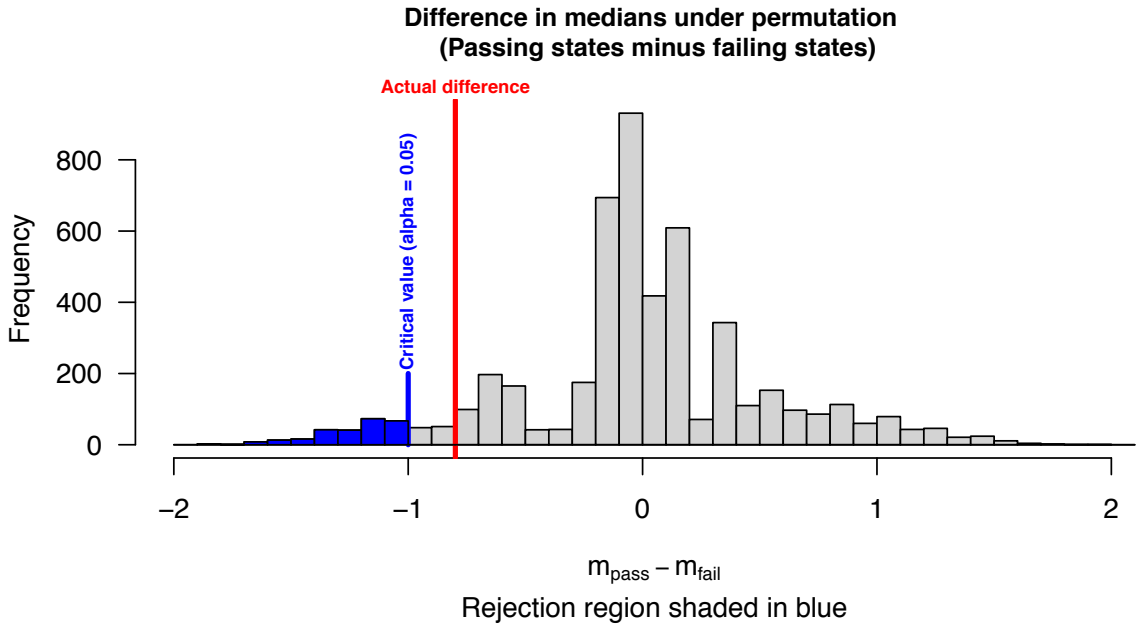


Figure 5.14: In our analysis of the relationship between gun laws and rates of gun violence at the state level, our test statistic of $\delta = -0.8$ falls outside the rejection region at the $\alpha = 0.05$ level.

(4a) Based in the samping distribution of Δ under the null hypothesis we simulated earlier, let's choose a rejection region of $R = \{\Delta \leq -1\}$.

(4b) From Figure 5.14, it is easy to calculate α, the probability that Δ will fall in the rejection region R even if the null hypothesis is true. This turns out to be about $\alpha = 0.05$. This area is shaded in blue in Figure 5.14.

(4c) The value of the statistic for our real data set is $\delta = -0.8$. This falls inside our rejection region, and hence we fail to the null hypothesis at the $\alpha = 0.05$ level.

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. $\delta = -0.8$)

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" at the end—that is, if the observed statistic falls in the rejection region—then your result is said to be *statistically significant* at the stated α level.

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 $100\alpha\%$ 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.[19]
- the probability that the null hypothesis is false.
- 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 feature rather than a bug.

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 type of considerations—How bad is a false positive? A false negative?—that go into the choice of α.

[19] 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 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 cannot be overemphasized.[20] 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.

[20] 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.

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 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 for this, 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," "cataclysmically 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. Adjectives only muddle the picture.

Let me give you an idea of how common this fallacy is, even in top-quality research journals. In a volume of the *Journal of the American Medical Association* from several years ago (January 2010), 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 scientifically 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, in the table on the next page.

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

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.

Table 5.1: Definitions of the p-value and α level.

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 gun-laws example. There, our rejection region included all values of Δ that were -1 or larger in the negative direction (blue region in Figure 5.14). This yielded an α-level of $\alpha = 0.05$, ensuring that if the null hypothesis were true, we would have no more than a 5% chance of falsely rejecting it. But we actually observed a difference in the medians of $\delta = -0.8$. Under the assumption that the null hypothesis were true, the probability of getting $\delta = -0.8$ (or something more extreme in the negative direction) is $p = 0.072$.

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. It's therefore worth warning you up front about a handful of common mistaktes:

- 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 that some test statistic z follows 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.[21]

The moral of the story is: 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 best thought of as an **informal** description of the evidence in a specific sample—one whose interpretation is slippery and easily misconstrued.

[21] 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.