8
Sampling variability

Quantifying uncertainty

A question that arises frequently in data science is: how confident are we in our estimate of some fact about a population, on the basis of data from a sample? Take the following examples of statistics derived from polling data, ripped from the 2020 headlines:

- "Almost one in four (23%) Americans know someone who has died from the coronavirus."[1]
- "Trump Stuck at 42% Job Approval."[2]
- "Poll finds Biden with a 9-point lead over Trump among registered voters and a 7-point lead among likely voters."[3]

All of these numbers are derived from samples, and they are therefore an imperfect guide to average opinion across the entire U.S. population. After all, the only way to be absolutely certain what *everyone* thinks is to... ask everyone! The point is that any conclusions derived from samples are inherently uncertain, due to the fact that a sample is an incomplete picture of a population. Our task is to provide an answer to the question of just *how* uncertain these numbers really are.

Sampling distributions and alternate universes

In statistics, we describe the uncertainty of our estimates by appealing to an idea called a *sampling distribution*. To define this concept, we need a bit of notation. Suppose we are trying to estimate a *parameter*, which just means some numerical feature of a population. Let's call this number θ. It might be a mean, or a proportion, or the slope of a straight-line trend—anything that we could, in principle, measure of the wider population, given enough time and resources.

[1] Axios-Ipsos poll of 1,019 adults, aged 18 or older.

[2] Gallup poll of 1,019 adults, aged 18 and older.

[3] Monmouth University poll of 758 likely voters.

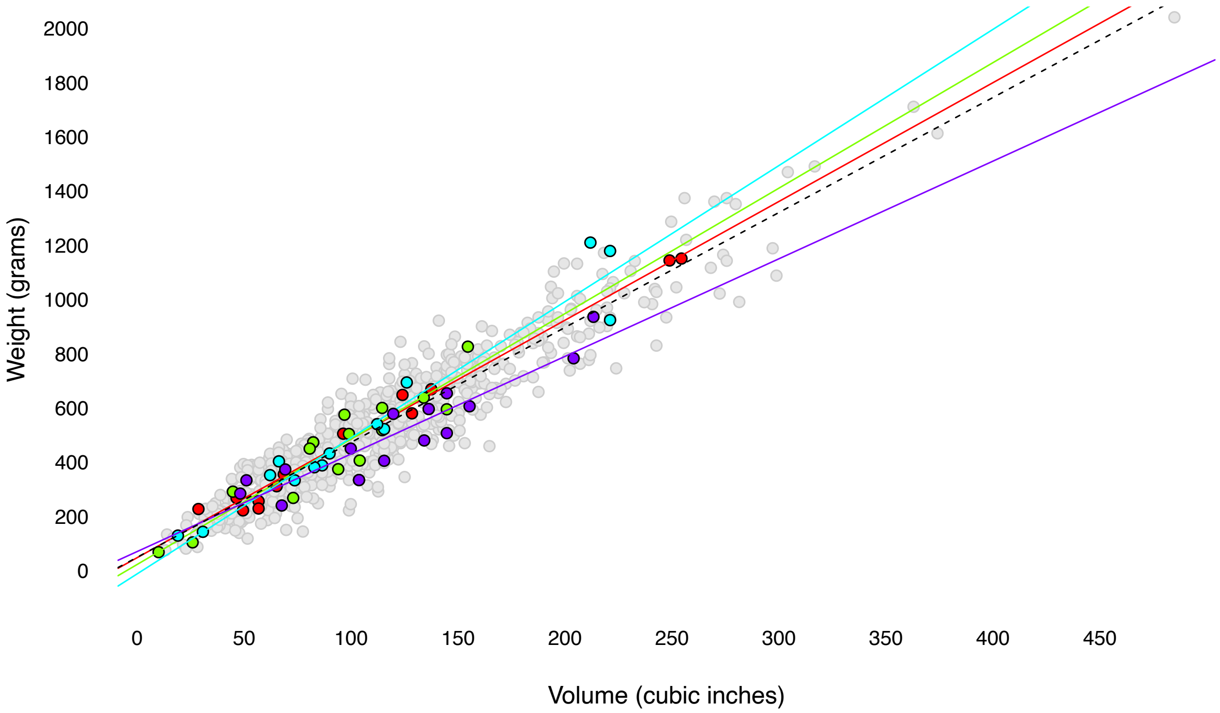


Figure 8.1: Four different days of fishing, coded by color, on an imaginary lake home to a population of 800 fish. On each day's fishing trip, you catch 15 fish, and end up estimating a slightly different weight–volume relationship. The dashed black line is the true relationship for the entire population.

But of course, we rarely have the time and resources to conduct a *census*, by measuring θ for the whole population. So instead, we measure that same numerical property of a random sample from the population. Let's call the resulting number $\hat{\theta}$, to indicate that it's a sample-based estimate of θ. The sampling distribution asks the question: what kinds of estimates $\hat{\theta}$ might I have seen if I had taken different samples from the same underlying population? The logic here is roughly the following:

- If the sampling distribution is really concentrated around some value, then different samples all give pretty much the same answer. That means we can trust the estimate from any *specific* sample (i.e. ours).

- If the sampling distribution is really spread out, it means that different samples give wildly different estimates $\hat{\theta}$. That means we cannot trust the estimate from any specific sample, since it might be very far from the truth.

An example: simulating a sampling distribution by Monte Carlo. This is best illustrated by example. 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 8.1. 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 8.1.

Four days of fishing give us some idea of how the estimates for β_0 and β_1 (the two parameters of your straight line fit) vary from sample to sample. But 2500 days of fishing, simulated by computer, give us a better idea. Figure 8.2 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).

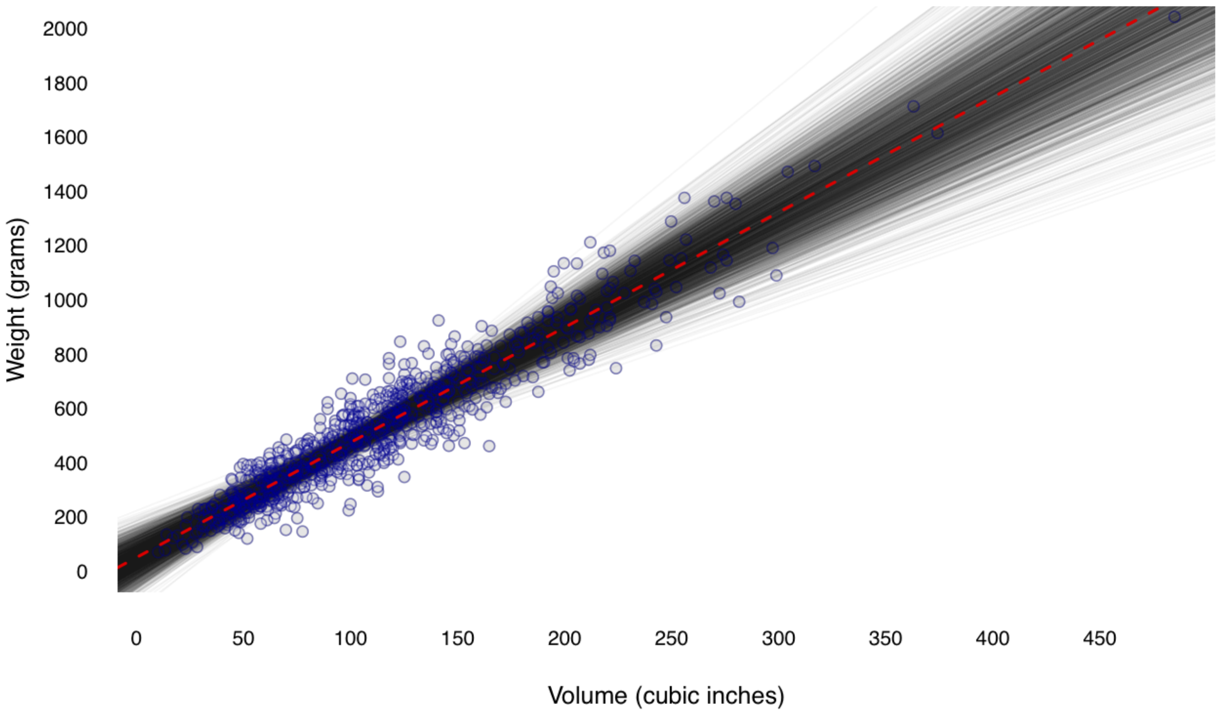
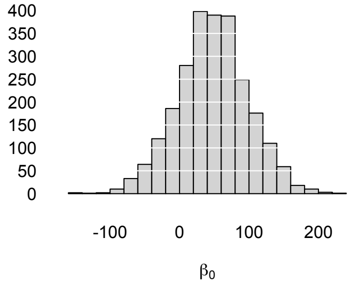
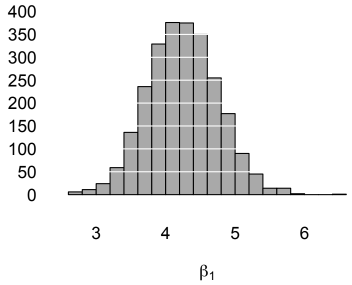


Figure 8.2: 2500 days of fishing, together with the 2500 different estimates of β_0 and β_1 (below), simulated by Monte Carlo.





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 pretty good idea.)

The sampling distribution. Here's an analogy that might help you understand the concept of a sampling distribution. Let's think of our sampling and estimation procedure as like a trial in a courtroom: it's a formalized procedure that's designed to uncover the truth in the face of uncertainty and incomplete data. And we can think of the particular estimate we got for a particular data set as like the verdict in a specific trial. In both cases, the important thing to focus on is the *procedure.* In the justice system, we care deeply about whether someone received a fair trial. In data science, we care about whether our sampling and estimation procedure is capable of accurately estimating a population-level parameter from a sample of data.

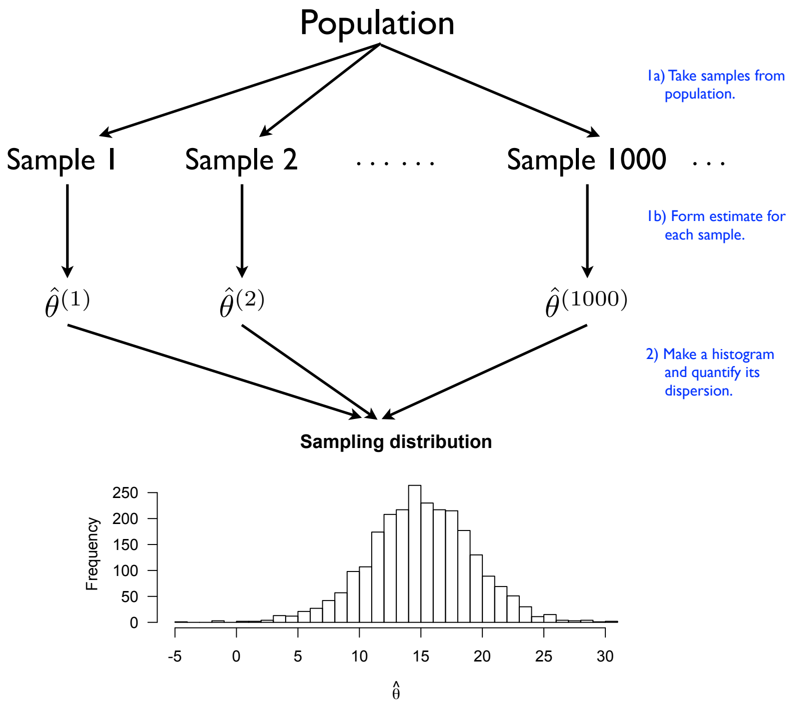


Figure 8.3: A stylized depiction of a sampling distribution of an estimator $\hat{\theta}$. To construct this distribution, we must imagine the following thought experiment. We repeatedly take many samples (say, 1000) from the population (step 1a). For each sample, we apply our estimator to compute the 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.*

Let's take the least-squares estimation procedure. This is a specific set of steps that you apply (or rather, get your computer to apply) to a data set. The procedure yields a "verdict": estimates $\hat{\beta}_0$ 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 8.3 shows graphically how, in principle, this distribution is constructed.

We typically summarize a sampling distribution using its standard deviation, which we refer to as the *standard error*.[4] 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 less stable the estimator across different samples, and the less you can trust the estimate for any particular

[4] 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.

sample. To give a specific example, for the 2500 samples in Figure 8.2, 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.[5]

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. . . .[6]

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 8.3, thereby approximating an estimator's sampling distribution without ever taking repeated samples from the population.

1) *Simulation:* that is, by simulating the sampling process on a computer, which allows one to approximate the effect of sampling variability.

2) *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 the next chapter, we'll discuss the first approach, deferring the second approach to another book.

[5] 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.

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