Logistics

1. PSET 4 graded - good job everyone!
   - Solutions posted
2. Zoom class Monday Nov 13 (will announce this on canvas)
3. Final exam date has not been set yet, trying to find out how this will be determined
4. Lab notebook: due date, clarifications posted on canvas (course main page)

Sampling from a Known DGP

In Topic 4A, we developed tools to think about theoretical probability distributions that tell us how random processes unfold.

We now pivot to thinking about how a dataset could be generated from a DGP that we modeled using a probability distribution.

The thought exercise we will do in this lecture is:

↪ Suppose we start with a probability distribution with specific parameters.

↪ Then, we take "draws" from this distribution.

↪ What does the "sample" we end up with look like? Does it resemble the original DGP? How would we know?

Sampling from a Known DGP

This may be confusing at first, but frequentist statistics is based on the idea of **sampling**.

↪ An opinion poll is a *sample* of the electorate.

↪ All US electoral districts in 2020 is a *sample* of electoral districts across all elections.

↪ Grad students in this classroom are a *sample* of all grad students at UC Davis.

↪ Grad students at UC Davis are a *sample* of all grad students in the US.

Sampling from a Known DGP

Theoretical approach : we start with a DGP and then have to figure out what the dataset looks like.

In empirical research, this is different !

We start with a dataset and then have to figure out the parameters of the DGP.

Typical approach in empirics: Make some assumptions about how the data set came to be, so you can make inferences about the DGP.

IID Samples

Recall that probability distributions represent random processes.

A **draw** from a probability distribution is one iteration of the random process, which yields one of the values of the random variable.

For example: a fair coin flip represented by a random variable X that can take a value of 1 (heads) or 0 (tails).

↪ One draw would be the number (1 or 0) corresponding to the outcome of one flip of a the coin.

↪ Recall in basic probability we referred to one iteration of a random process as an "experiment."

A collection of values from multiple draws is called a **sample**.

IID Samples

It's important to consider *how* the draws of a sample happen:

↪ Each draw in a sample could be independent or correlated with other draws.

↪ Draws in a sample could all be from the same probability distribution or different probability distributions.

The easiest kind of sample to work with is one in which each draw is **independent and identically distributed** (or **iid**).

↪ **Independent:** knowing the result of one draw tells you nothing about the result of another draw.

If we have a sample with iid draws, then we call it a **simple random sample**.

IID Samples

For a specific set of N observations that comprise a sample, we can write it as we did before: $x_1, x_2, ..., x_N$.

However, because we're taking draws from a distribution, each of these specific observations *could have* been different.

That leads to the next fact, which is sorta hard to get used to:

Each draw from a probability distribution is a random variable with its own distribution.

So we can think of a sample of size N as a collection of N random variables: $X_1, X_2, ..., X_N$.

IID Samples

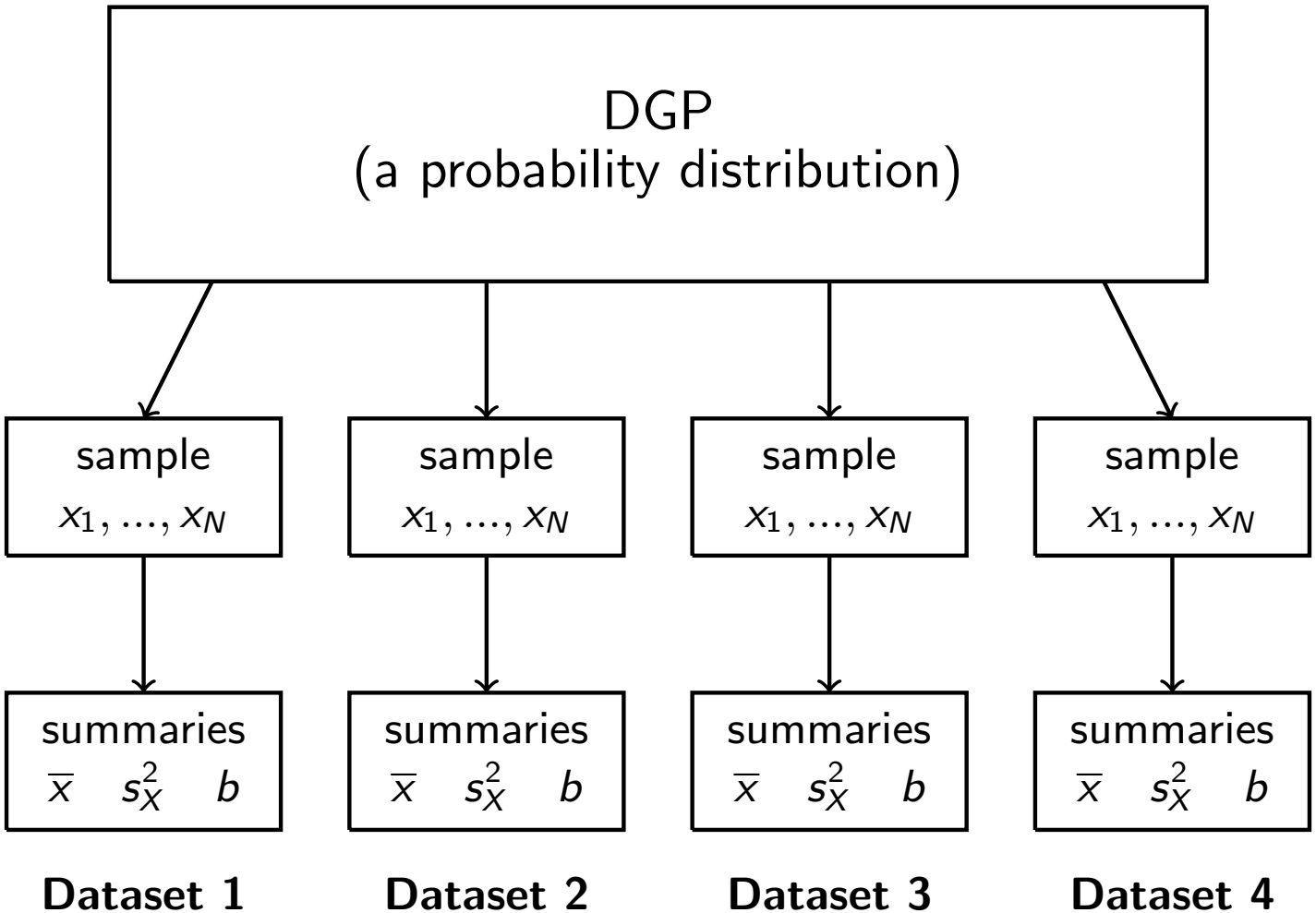
Let's get even more meta with another strange, but important, fact:

Since each draw in a sample is a random variable, then when you perform calculations on the sample, resulting statistics (means, standard deviations, regressions, etc.) are also random variables.

I wish to reiterate: for a *specific* sample (i.e., a specific dataset), you calculate only one mean, standard deviation, etc.

But, these can be thought of as random variables because they *could have been different* if the sample draws had been different.

IID Samples


DGP
(a probability distribution)
sample
$x_1, \ldots, x_N$
sample
$x_1, \ldots, x_N$
sample
$x_1, \ldots, x_N$
sample
$x_1, \ldots, x_N$
summaries
$\overline{x}$ s_X^2 b
summaries
$\overline{x}$ s_X^2 b
summaries
$\overline{x}$ s_X^2 b
summaries
$\overline{x}$ s_X^2 b
Dataset 1
Dataset 2
Dataset 3
Dataset 4

IID Samples

Suppose we have a DGP with $X \sim B(N = 1, p = 0.7)$, and we create seven samples, each with five iid draws.

	X_1 ↓	X_2 ↓	X_3 ↓	X_4 ↓	X_5 ↓	$\overline{X}$ ↓
	Draw 1	Draw 2	Draw 3	Draw 4	Draw 5	Sample Mean
Sample 1	1	1	1	1	0	0.8
Sample 2	0	1	1	1	1	0.8
Sample 3	0	0	1	1	1	0.6
Sample 4	1	0	1	0	0	0.4
Sample 5	0	0	0	1	1	0.4
Sample 6	1	1	0	1	1	0.8
Sample 7	0	0	1	1	1	0.6

IID Samples

Recall our goal is to learn about DGPs from samples.

↪ To keep things clear, we'll call a DGP probability distribution a **population distribution**.

↪ We'll also label the mean and standard deviation of the population distribution as μ and σ, respectively.

Recall also that in real life, we learn about the DGP by calculating things like sample means, standard deviations, regressions, etc.

Because we use these sample statistics to learn about parameters of a DGP, we call them **estimators** for the DGP parameters.

↪ For example, the sample mean is an *estimator* for the mean of the population (DGP) distribution.

IID Samples

If we can think of all of our sample statistics (mean, standard deviation, regression coefficient) as random variables.

↪ They each have their own probability distributions, which we call their **sampling distributions**.

For the remainder of this topic, we're going to explore what sampling distributions look like for a few statistics you might calculate on real life datasets.

The reason we do this is because it will clarify when and how we can learn about DGPs from samples.

We will assume we know a DGP and then examine how its sampling distribution behaves. But remember: we never know a DGP! We're doing this for pedagogical reasons.

Sampling Distribution: Sample Means

Suppose we have a sample with iid draws. We know that $\overline{X}$ is a random variable with a sampling distribution.

What is the expected value of this sampling distribution?

$$
\begin{aligned}
E(\overline{X}) &= E\left(\frac{X_1 + X_2 + \cdots + X_N}{N}\right) \\
&= \frac{1}{N}E(X_1 + X_2 + \cdots + X_N) \\
&= \frac{1}{N}E(X_1) + \frac{1}{N}E(X_2) + \cdots + \frac{1}{N}E(X_N) \\
&= \frac{1}{N}\mu + \frac{1}{N}\mu + \cdots + \frac{1}{N}\mu \\
&= \mu
\end{aligned}
$$

Sampling Distribution: Sample Means

Intuition: we usually calculate quantities like sample means.

↪ The math below shows: the expected value of the "sample mean" is the population mean

↪ Under some conditions, we can learn about the population mean from the sample mean.

$$
\begin{aligned}
E(\overline{X}) &= E\left(\frac{X_1 + X_2 + \cdots + X_N}{N}\right) \\
&= \frac{1}{N} E(X_1 + X_2 + \cdots + X_N) \\
&= \frac{1}{N} E(X_1) + \frac{1}{N} E(X_2) + \cdots + \frac{1}{N} E(X_N) \\
&= \frac{1}{N}\mu + \frac{1}{N}\mu + \cdots + \frac{1}{N}\mu \\
&= \mu
\end{aligned}
$$

Sampling Distribution: Sample Means

The math is pretty tedious, but the idea behind it is profound.

The expected value of the sampling distribution for $\overline{X}$ is simply the mean of the population (DGP) distribution!

This means that the sample mean $\overline{X}$ is an **unbiased estimator** for the population (DGP) mean, μ.

Unbiased: $E(\bar{X}) = \mu$

Why is this important? It gives us some faith that a sample mean will be "pretty close" to the population mean so that we can learn about the DGP mean from the sample mean.

However: – we still need to know how "close" the sample mean is to the population mean.

Sampling Distribution: Sample Means

Okay, what about the standard deviation of the sampling distribution for $\overline{X}$? Let's start with its variance:

$$
\begin{aligned}
Var(\overline{X}) &= Var\left(\frac{X_1 + X_2 + \cdots + X_N}{N}\right) \\
&= \frac{1}{N^2} Var(X_1 + X_2 + \cdots + X_N) \\
&= \frac{1}{N^2} Var(X_1) + \frac{1}{N^2} Var(X_2) + \cdots + \frac{1}{N^2} Var(X_N) \\
&= \frac{1}{N^2}\sigma^2 + \frac{1}{N^2}\sigma^2 + \cdots + \frac{1}{N^2}\sigma^2 \\
&= \frac{N\sigma^2}{N^2} \\
&= \frac{\sigma^2}{N}
\end{aligned}
$$

Recall: $Var(a(X_1 + X_2)) = a^2 Var(X_1) + a^2 Var(X_2)$ for iid draws.

Sampling Distribution: Sample Means

If we take the square root of variance of the sampling distribution we get the standard deviation of the sampling distribution.

We have a special name for this: the **standard error**.

$$SE(\overline{X}) = \sqrt{Var(\overline{X})} = \sqrt{\frac{\sigma^2}{N}} = \frac{\sigma}{\sqrt{N}}$$

The standard error of a sampling distribution gives us a sense for how confident we should be in our estimate of the mean.

Looking at the formula, our cofidence depends on two things:

↪ How much variation is there in the population distribution (σ)?

↪ How big of a sample (N) did we take?

Sampling Distribution: Sample Means

What if we want to know more about the sampling distribution of $\overline{X}$ than just its expected value and its standard error?

It turns out that if N is large, the sampling distribution of $\overline{X}$ looks approximately like a normal distribution, as long as the population (DGP) distribution has finite mean and variance.

$\hookrightarrow$ What is "large"? Depends on how unusual the DGP distribution is... but a rule of thumb: $N > 30$.

$\hookrightarrow$ It holds regardless of the shape of the population (DGP) distribution.

This is known as the **central limit theorem**.

Sampling Distribution: Sample Means

This might seem like statistical wizardry, but it really isn't.

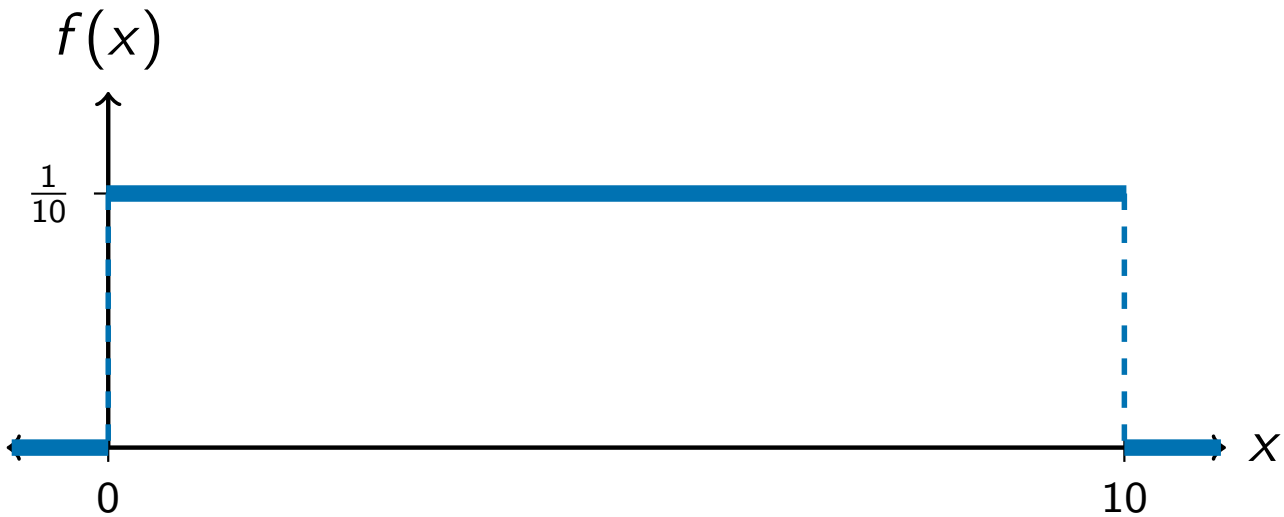
You can read Gailmard's explanation for why this happens, but the basic idea: averaging things tends to push toward the middle.

Instead let's jump right into an example.

Sampling Distribution: Sample Means

Let's assume this population (DGP) distribution: $X \sim \mathcal{U}[0, 10]$.

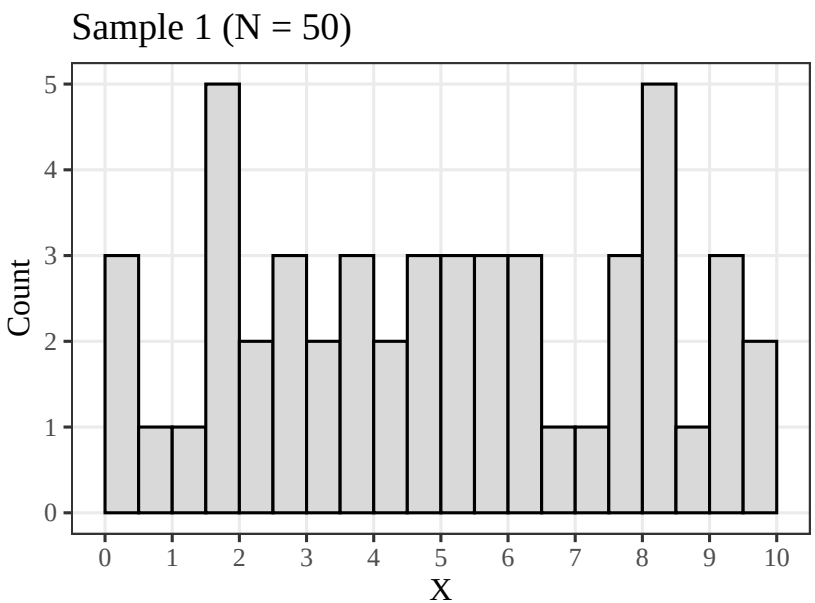
We can plot the PDF for this distribution:

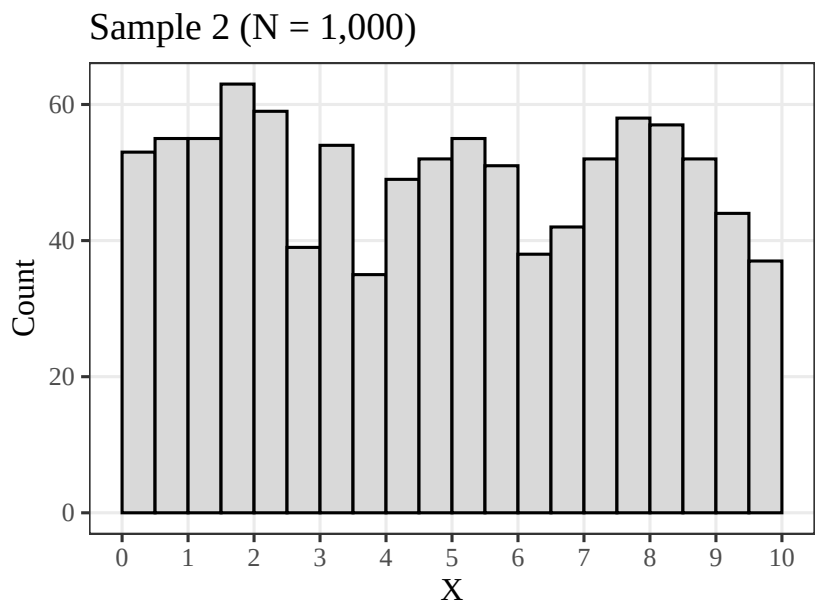


We know that $E(X) = 5$ and $Var(X) = \frac{25}{3}$. (Why?)

Sampling Distribution: Sample Means

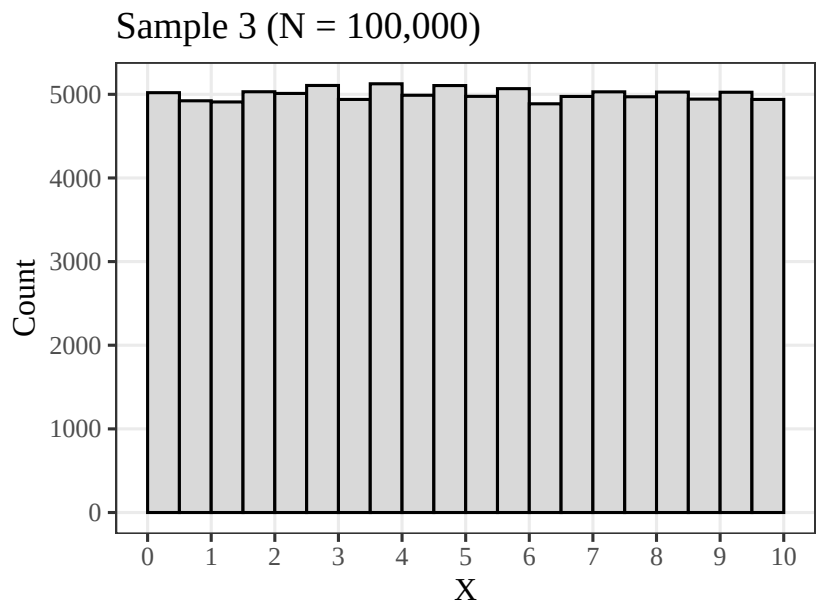
Now let's take two samples: one with 50 iid draws, and another with 1,000 iid draws. I will plot histograms for these two samples:





Note: the larger the sample size, the more the sample will "look like" the population distribution.

Sampling Distribution: Sample Means



Sampling Distribution: Sample Means

Now I want to see what the sampling distribution looks like for $\overline{X}$ when $N = 50$ and $N = 1000$.

The dilemma is that we can't actually see a sampling distribution because it is a probability distribution.

Instead we approximate it by running **Monte Carlo simulations**.

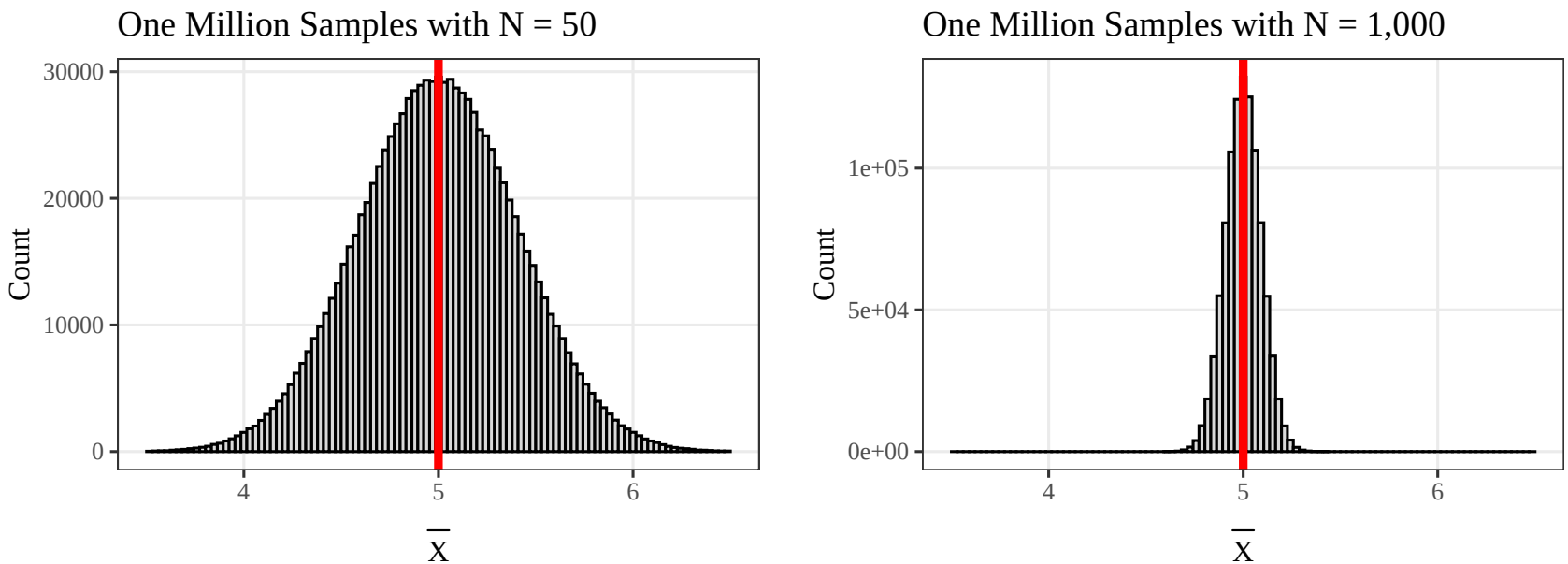
↪ Take a large number of samples and calculate the mean for each one. (I'll do 1,000,000 because I have a fast computer.)

↪ The resulting collection of 1,000,000 means will be an approximation of the sampling distribution

Technically: the sampling distribution is the one with infinite simulations, but my computer is not *that* fast.

Sampling Distribution: Sample Means

Here's a histogram where I plot the results of our two Monte Carlo simulations:



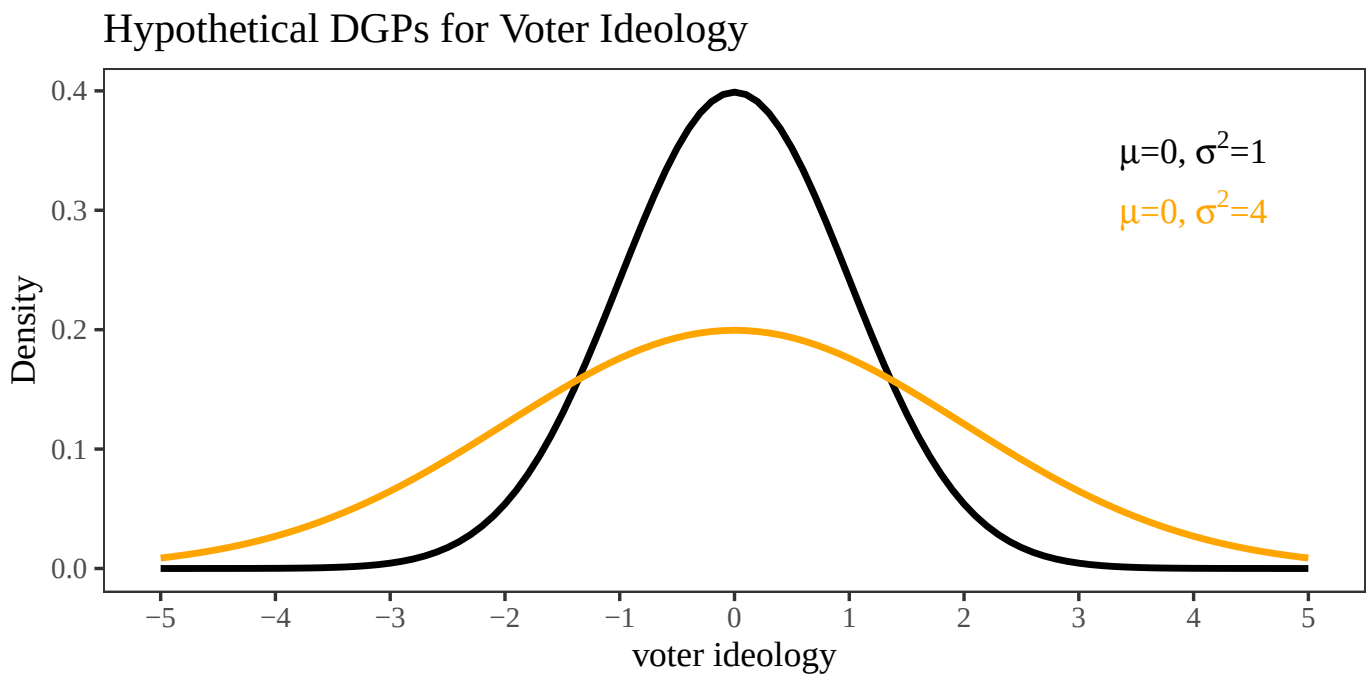
Both are approximately normal with means at approximately 5.

But: the standard error of the left is much larger.

Sampling Distribution: Sample Variance

In some situations, a researcher might want to learn something about the variance of the population (DGP) distribution.

↪ For example, have voters become more ideologically homogenous or heterogeneous over time?



Sampling Distribution: Sample Variance

If the population (DGP) distribution is normal and the sample is iid then: the sampling distribution for the sample variance is "approximately normal" with its expected value $E(S_X^2) = \sigma^2$.

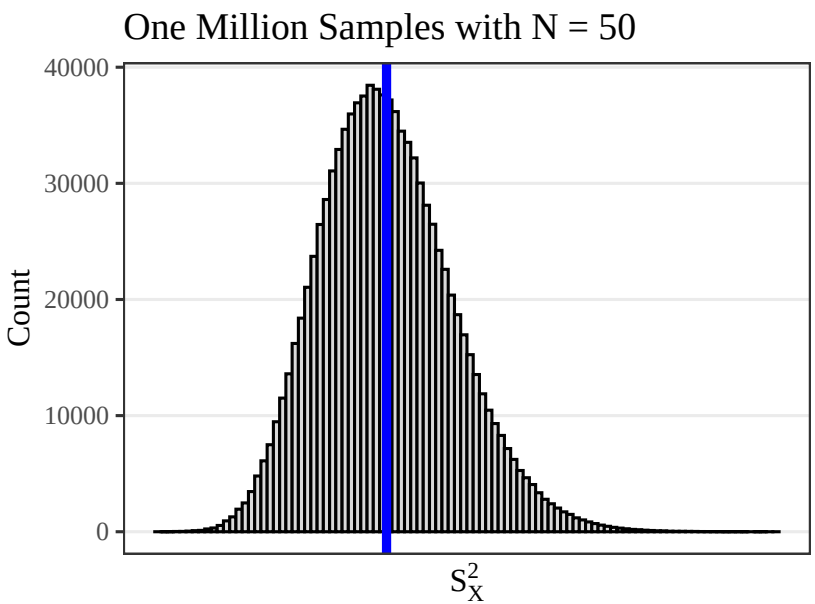
What do we mean by "approximately normal"?

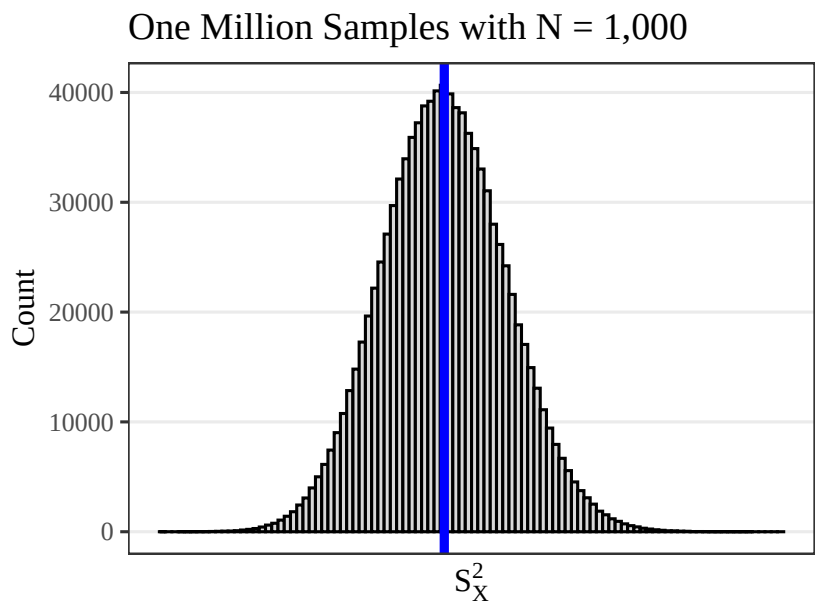
↪ Normal distributions always put some weight on negative numbers. But remember that variances can't be negative!

↪ Instead, the sample variance is distributed according to a χ^2 ("chi-squared") distribution with one parameter: $N - 1$.

↪ This parameter is called the **degrees of freedom**.

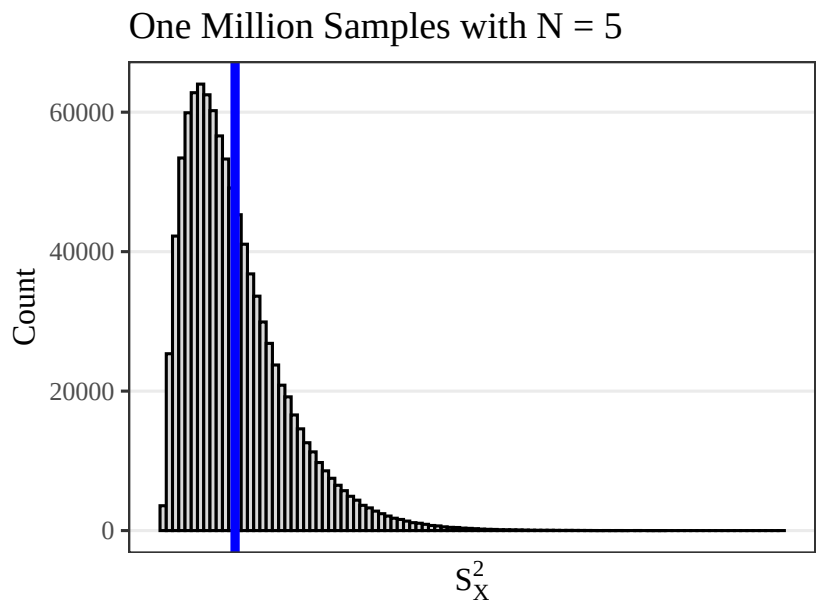
Sampling Distribution: Sample Variance





These look pretty close to normal.

Sampling Distribution: Sample Variance



This does not!

Summary so far

We usually assume a **population distribution** and then take a sample from it.

↪ Goal: learn about the population distribution from the sample.

↪ We can calculate quantities like sample means, and regression coefficients.

↪ If we use the sample mean, etc to learn about the population distribution, we call them **estimators**.

↪ Sample means, variances and regression coefficients are themselves random variables – they are functions of the data.

↪ Therefore, they have a **sampling distribution**.

Summary so far

We can think of the **sampling distribution** as the behavior of the **estimator** when we take many samples from the **population distribution**.

↪ Recall: the estimator can be the sample mean, regression coefficients, etc.

In particular, we can calculate the expected value and variance of the estimator

↪ We showed: for the sample mean $\overline{X}$, $E(\overline{X}) = \mu$ and $Var(\overline{X}) = \frac{\sigma^2}{N}$.

↪ Note: mu and σ are the mean and standard deviation of the **population distribution.**

Sampling Distribution: Difference in Means

We just examined the statistical link between the DGP mean/variance and the sample mean/variance for a single random variable.

Usually, we're interested in relationships between variables.

One way to look at the relationship between two variables is to calculate a difference in means.

↪ For example: what is the difference between the percent of court cases ending in a pro-plaintiff decision for cases heard by Democratic judges and cases heard by Republican judges?

Sampling Distribution: Difference in Means

Remember that the sample mean is a random variable.

An important fact: the sum or difference between two random variables is itself a random variable.

So, if we have two sample means $\overline{Y}_1$ and $\overline{Y}_2$, then the difference between them is also a random variable.

This random variable has its own sampling distribution that is normally distributed with expected value and standard errors:

$$E(\overline{Y}_1 - \overline{Y}_2) = \mu_1 - \mu_2 \qquad SE(\overline{Y}_1 - \overline{Y}_2) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

Sampling Distribution: Sample Regression Slope

Recall that another way to look at relationships between variables is the regression line.

We statistically model a linear relationship with the DGP regression with a linear link function:

$$E(Y|X) = \alpha + \beta X$$

The thing we're interested in is β (the slope).

In real datasets, we calculate regression slope b (see Topic 3).

This too is a random variable! Let's call the random variable $\hat{\beta}$. What's its sampling distribution?

Sampling Distribution: Sample Regression Slope

First, we need to re-write the DGP regression as:

$$Y = \alpha + \beta X + \varepsilon$$

The last term is known as the **disturbance** or **error**. Then:

$$E(\hat{\beta}) = \beta + E(\widehat{Cov}(X, \varepsilon))$$

If $X \perp\!\!\!\perp \varepsilon$, then $E(\widehat{Cov}(X, \varepsilon)) = 0$ and

$$E(\hat{\beta}) = \beta$$

If $X \perp\!\!\!\perp \varepsilon$, we say that X is **exogenous** or that it is an **exogenous regressor**. Otherwise, we say that X is **endogenous**.

Sampling Distribution: Sample Regression Slope

The standard error of the sampling distribution for $\hat{\beta}$ is:

$$SE(\hat{\beta}) = \sqrt{Var(\hat{\beta})}$$

If (1) each Y_i is independent and (2) $Var(Y|X)$ is constant (i.e., $Var(Y|X) = \sigma^2$), then

$$SE(\hat{\beta}) = \frac{\sigma}{s_X\sqrt{N-1}}$$

$\hookrightarrow$ Condition (2) is known as **homoscedasticity**.

Sampling Distribution: Sample Regression Slope

Suppose we have a dataset and we run a regression and calculate a regression slope b. We also calculate its standard error (note: we didn't do this in Topic 3).

We know that regression slope is an *unbiased* estimate of the DGP regression slope only if X is exogenous.

$\hookrightarrow$ Otherwise, estimate b will be biased and will cause us to draw incorrect conclusions about the "real" relationship in the DGP.

This standard error is only accurate if the Y values are not correlated and the conditional variance of Y (on X) is constant.

$\hookrightarrow$ Otherwise, we could have have a false sense of confidence in estimate b.

Estimands and Estimates are Different

An **estimand** is what we want to know about a population (DGP) distribution. For example: mean, regression slope, etc.

With real life data, we can only ever calculate an **estimate**.

Because we are working with samples, an estimate always involves *some* "error." There are two kinds:

↪ **Bias**: error that affects all observations in a systematic way.

↪ **Noise**: error that affects each observation separately due as a result of random chance.

BdM and Fowler's "favorite equation":

$$\text{estimate} = \text{estimand} + \text{bias} + \text{noise}$$

Estimands and Estimates are Different

A lot of active quantitative social science research involves thinking about and dealing with error.

In some sense, dealing with noise is "easier."

↪ Our confidence in our estimates increases when we increase our sample size: standard errors always decrease in N.

↪ Of course, this is only "easy" in theory... collecting more data is often costly and time-intensive.

↪ Plus, the DGP distribution could have high variance: this is a fundamental property of the world, and no amount of extra data will help you.

Estimands and Estimates are Different

Our mathematical discussion of sampling distributions so far hints at two major sources of bias:

1. Your sample is not iid. We call this **non-random sampling**.
2. Your X is not exogenous. We call this **omitted variable bias**.

There are many possible sources of bias beyond these two, which you will learn about over the next few years.

The basic lesson is: learning about a DGP from a sample requires some assumptions, *whether they're explicitly stated or not*.

The credibility of your claims about the DGP is a byproduct of how convincingly you are able to satisfy these assumptions.

Using Samples to Learn about an Unknown DGP

So far, Topic 4A has been devoted to thinking about known DGPs.

However, when we do real-world quantitative research, we only have access to one sample drawn from a DGP (our dataset).

Obviously, we hope to learn about the DGP from the sample.

How do we do this? There are two routes we can take:

↪ Try to evaluate how likely it would be for our sample estimate(s) to arise under a specific DGP we assume.

↪ Directly try to estimate features of the unknown DGP.

Quantifying How "Good" a Sample Estimate Is

So far, you have seen several examples of sampling distributions corresponding to statistics we care about:

↪ Sample means, differences in sample means, regression coefficients, etc.

For example, you now know that $\overline{X}$ is an estimator for μ that is a random variable distributed approximately normally with mean μ and standard error $\sigma/\sqrt{N}$.

It is useful to know a statistic's sampling distribution so that you can quantify how "good" a specific estimate is.

↪ For now: "good" = how far an estimate is from the DGP quantity we are interested in (i.e., the estimand).

Quantifying How "Good" a Sample Estimate Is

There are two ways to measure this:

1. How many standard errors is the estimate from the estimand? You can calculate this using **z-scores** or **t-scores**.
2. What is the probability of getting an estimate at least as extreme as the one you actually got? You can calculate this using **p-values**.

Quantifying How "Good" a Sample Estimate Is

Consider a hypothetical (and abstract) example.

You're interested in the expected value of a variable X, which you know is distributed normally with mean μ and standard deviation σ.

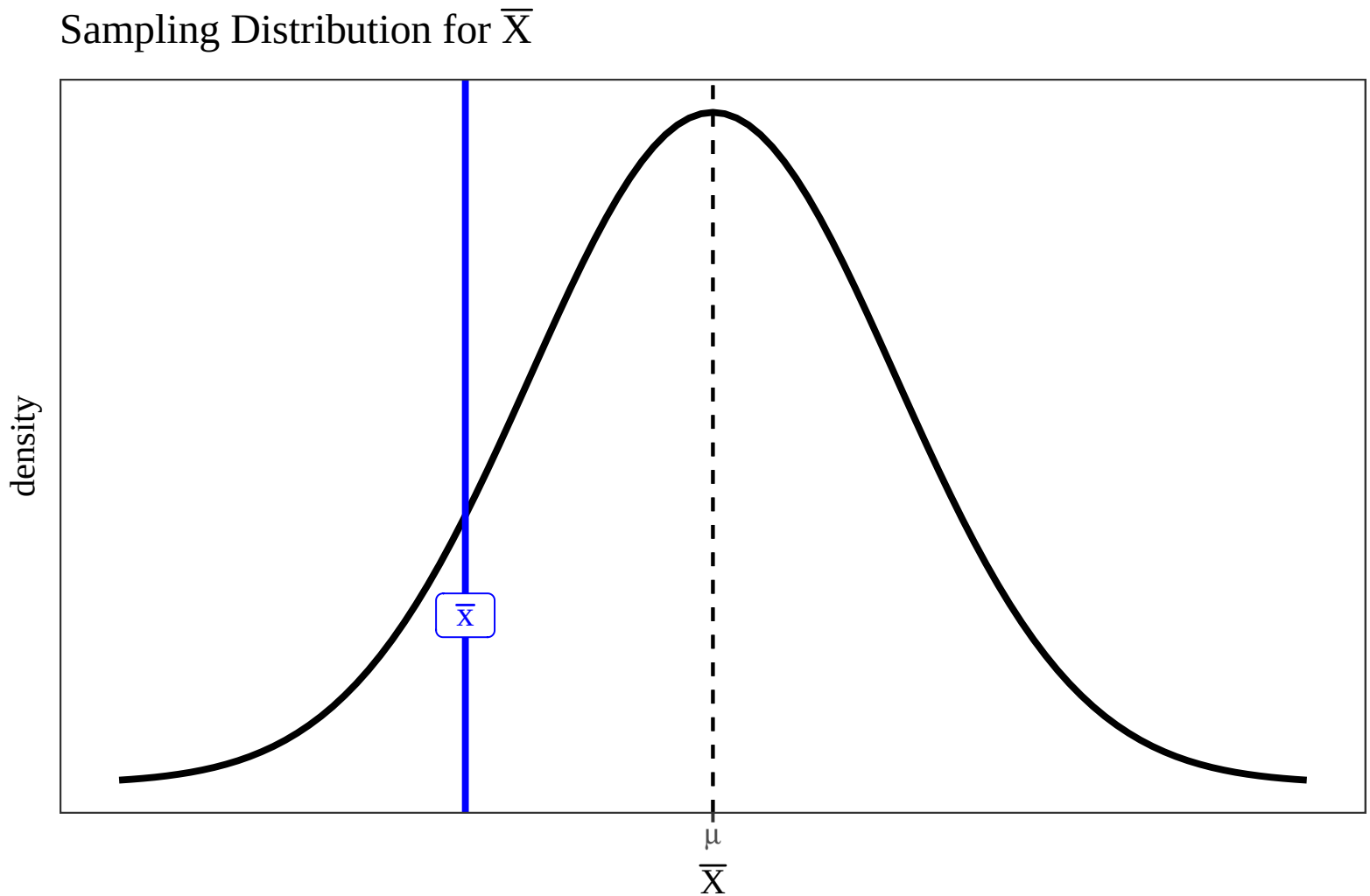
You draw a simple random sample of 500 from this distribution.

$\hookrightarrow$ (Elephant in the room: why are you learning from a sample when you know the DGP already??)

From previous lectures, you know that the sampling distribution of the sample mean $\overline{X}$ is distributed approximately normally with mean μ and standard deviation:

$$SE(\overline{X}) = \frac{\sigma}{\sqrt{N}}$$

Quantifying How "Good" a Sample Estimate Is



Quantifying How "Good" a Sample Estimate Is

Since we know the population (DGP) distribution, we know the sampling distribution for the sample mean.

↪ To repeat: we know it is (approximately) normally distributed with mean μ and standard error $\sigma/\sqrt{N}$.

We can draw it and see with our eyes that the estimate from our specific sample is some distance from the true DGP mean μ.

But, exactly how far?

We can measure the distance in standard errors: how many standard errors away from the true mean μ is our estimate $\overline{x}$?

Quantifying How "Good" a Sample Estimate Is

To calculate this, calculate a **z-score**:

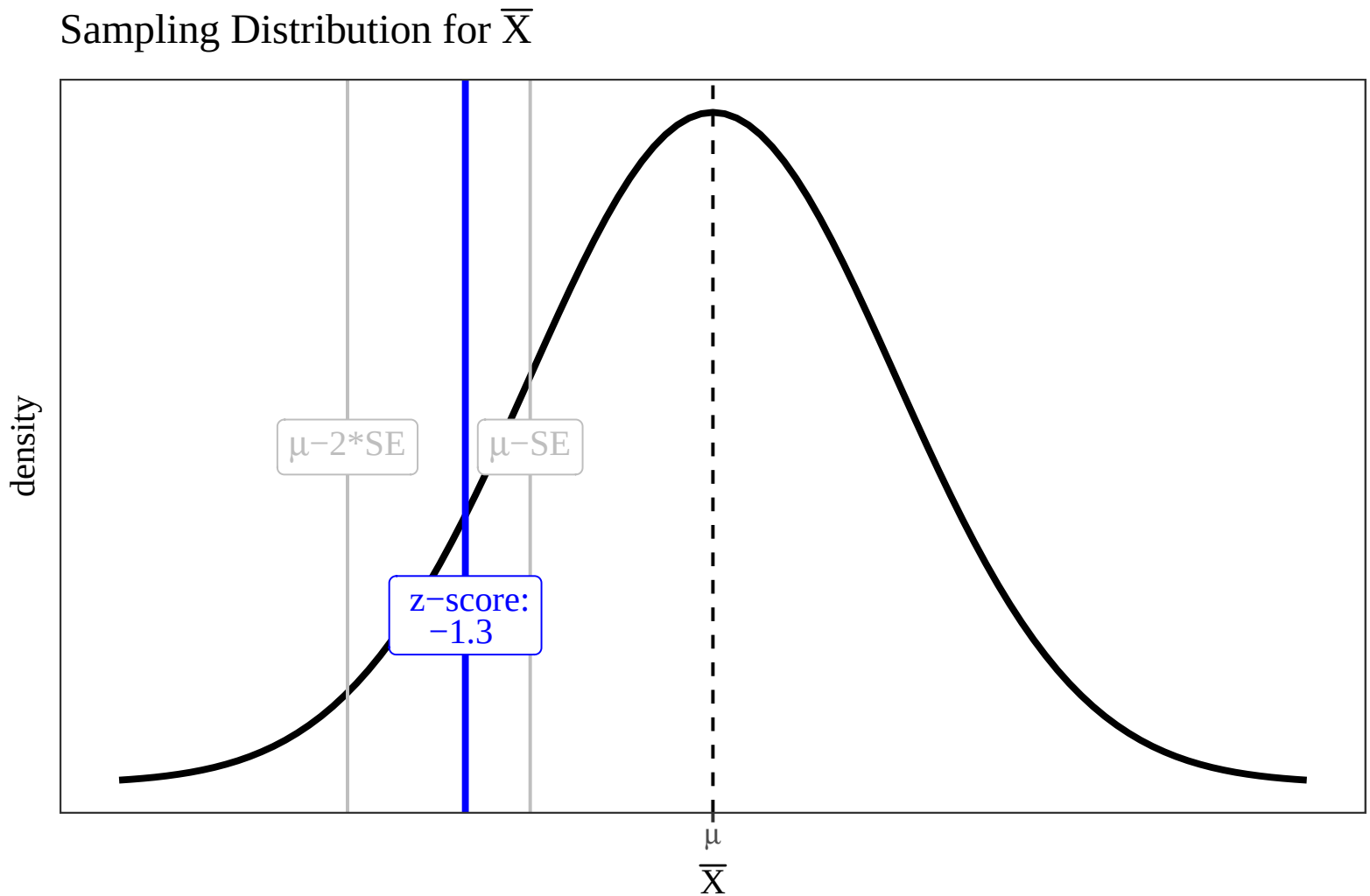
$$z = \frac{\overline{x} - \mu}{SE(\overline{X})} = \frac{\overline{x} - \mu}{\sigma/\sqrt{N}}$$

What's going on under the hood?

↪ A z-score "converts" a value drawn from any normal distribution into the equivalent value drawn from a standard normal distribution: $\mathcal{N}(0, 1)$.

↪ You can only use z-scores for normal distributions! Since sampling distributions are (approximately) normal, we can use it.

Quantifying How "Good" a Sample Estimate Is



Quantifying How "Good" a Sample Estimate Is

Z-scores are one way to measure how far an estimate is from μ.

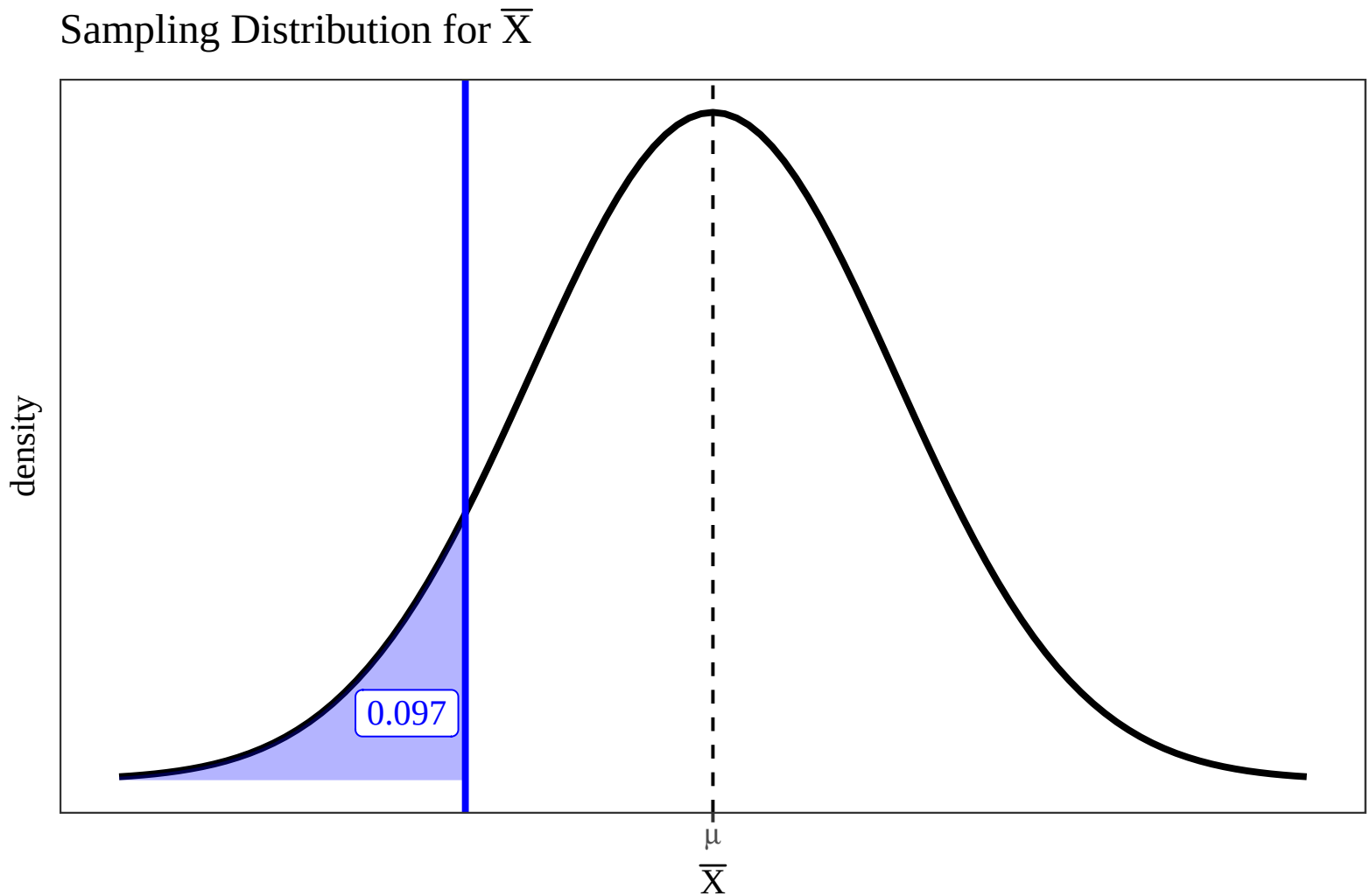
Another way to measure it is by calculating the probability of getting an estimate at least as far away as the one you actually did.

This is known as a **p-value**.

Before we talk about how to calculate this, let's visualize it on the sampling distribution we already drew.

Then, we'll talk about how to calculate it.

Quantifying How "Good" a Sample Estimate Is



Quantifying How "Good" a Sample Estimate Is

So: how do we calculate that p-value?

Since the sampling distribution is (approximately) normal, we just evaluate the CDF of the normal distribution at our estimate: $F(\overline{x})$.

This would be easier to do by hand if we were looking at a uniform distribution. (You should know how to do this!)

But here's the CDF of a normal distribution with mean μ and standard deviation $\sigma/\sqrt{N}$, evaluated at $\overline{x}$:

$$F(\overline{x}) = \frac{\sqrt{N}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\overline{x}} e^{-\frac{1}{2}\left(\frac{(\overline{x}-\mu)\sqrt{N}}{\sigma}\right)^2}$$

Yikes! Have fun calculating *that* by hand.

Quantifying How "Good" a Sample Estimate Is

We can ask R to help us out.

If we named the sample mean "`samp.mean`" and the true DGP mean and standard deviation "`dgp.mean`" and "`dgp.sd`" then:

```
> pnorm(samp.mean, dgp.mean, dgp.sd)
```

Alternatively, we can take advantage of the z-score!

Since z-score "converts" the sampling distribution into a standard normal, we can calculate the CDF of the standard normal evaluated at the z-score:

```
> pnorm(z.score, 0, 1)
```

The Core Problem(s)

Okay, there's a problem.

In the real world, we do not know the population (DGP) distribution.

This means we don't know anything about the sampling distribution for $\overline{X}$ except that if our sample is big enough it should be approximately normal.

So how can we evaluate how "good" our estimate is without knowing the specific sampling distribution we have???

There are really two problems: we know neither the expected value nor the standard error of the sampling distribution.

The Core Problem(s)

The second is somewhat easier to deal with: we simply *estimate* the standard error using our sample.

So, instead of calculating this

$$SE(\overline{X}) = \frac{\sigma}{\sqrt{N}},$$

we calculate this

$$\widehat{SE}(\overline{X}) = \frac{S_X}{\sqrt{N}}.$$

Because this is calculated using our sample, it is known as a **bootstrapped standard error**.

The Student's t Distribution

For now, assume that you still know μ but have to estimate the standard error.

Calculating a bootstrapped standard error gives you an *estimate* of the sampling distribution, not the exact sampling distribution.

Unfortunately, because we've introduced some noise by estimating the sampling distribution, we can no longer consider it normally distributed. Sad!

So, since we're estimating σ, we can't calculate z-scores and instead have to calculate **t-scores** (or **t statistics**):

$$t = \frac{\overline{x} - \mu}{\widehat{SE}(\overline{X})} = \frac{\overline{x} - \mu}{S_X/\sqrt{N}}$$

The Student's t Distribution

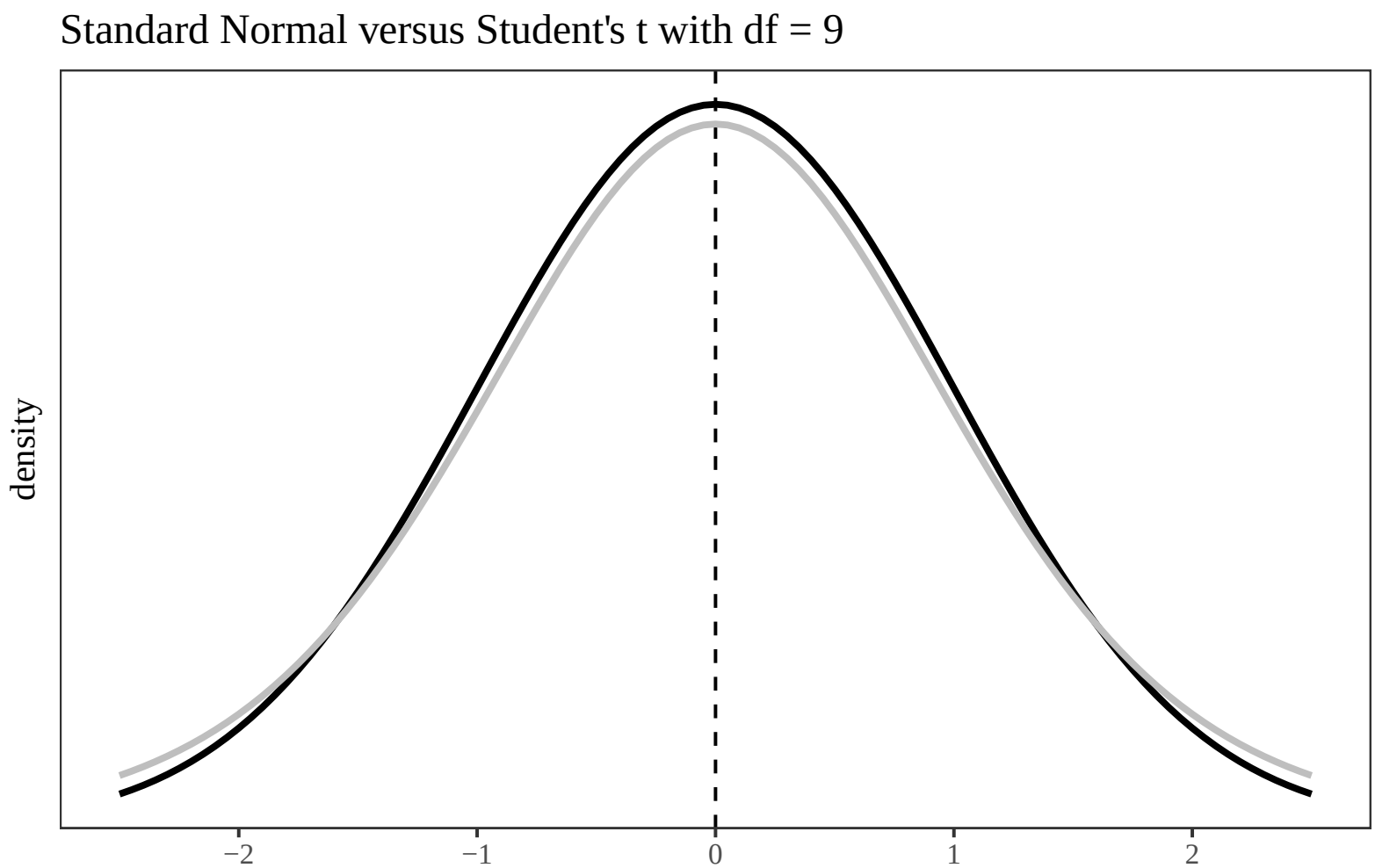
Similar to z-scores, t-scores "convert" values drawn from an estimated sampling distribution to a standardized distribution.

But since the estimated sampling distribution is not normal, neither is this standardized distribution.

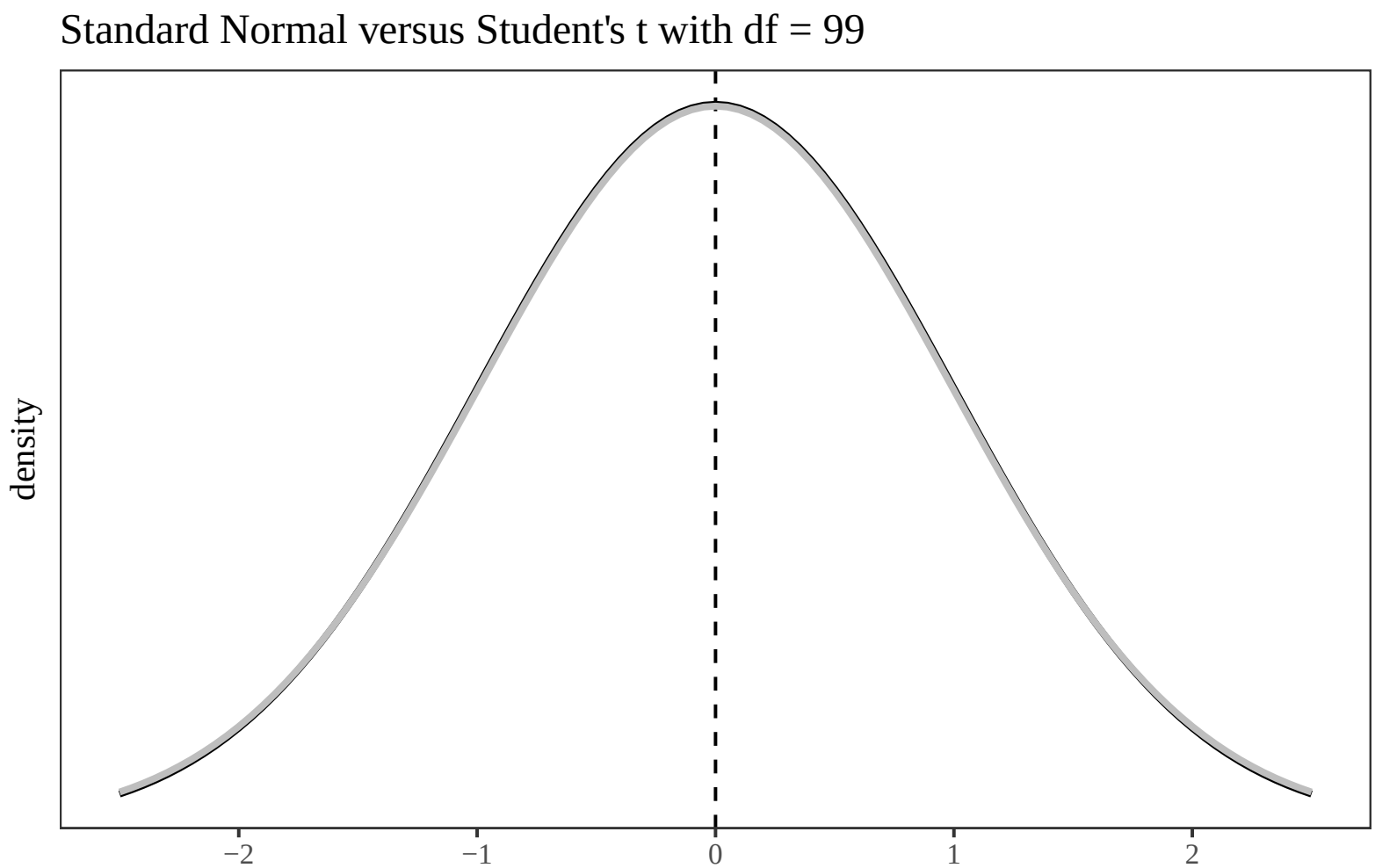
It is a **Student's t distribution**, which looks kind of like a standard normal distribution but with fatter tails.

- ↪ The fatness of the tails depends on the sample size N, and is caused by the noise introduced by estimating σ.
- ↪ As the sample size increases (in math: $N \to \infty$), the Student's t distribution becomes the standard normal distribution.
- ↪ This distribution has one parameter, the **degrees of freedom** (or **df**), which is equal to $N - 1$.

The Student's t Distribution



The Student's t Distribution



The Student's t Distribution

Just like with z-scores, we can use t-scores to calculate p-values when we have to bootstrap the standard error.

If you called the sample size `samp.size`, then the R code is:

```
> pt(samp.mean, samp.size - 1)
```

The Niceness of Normal Distributions

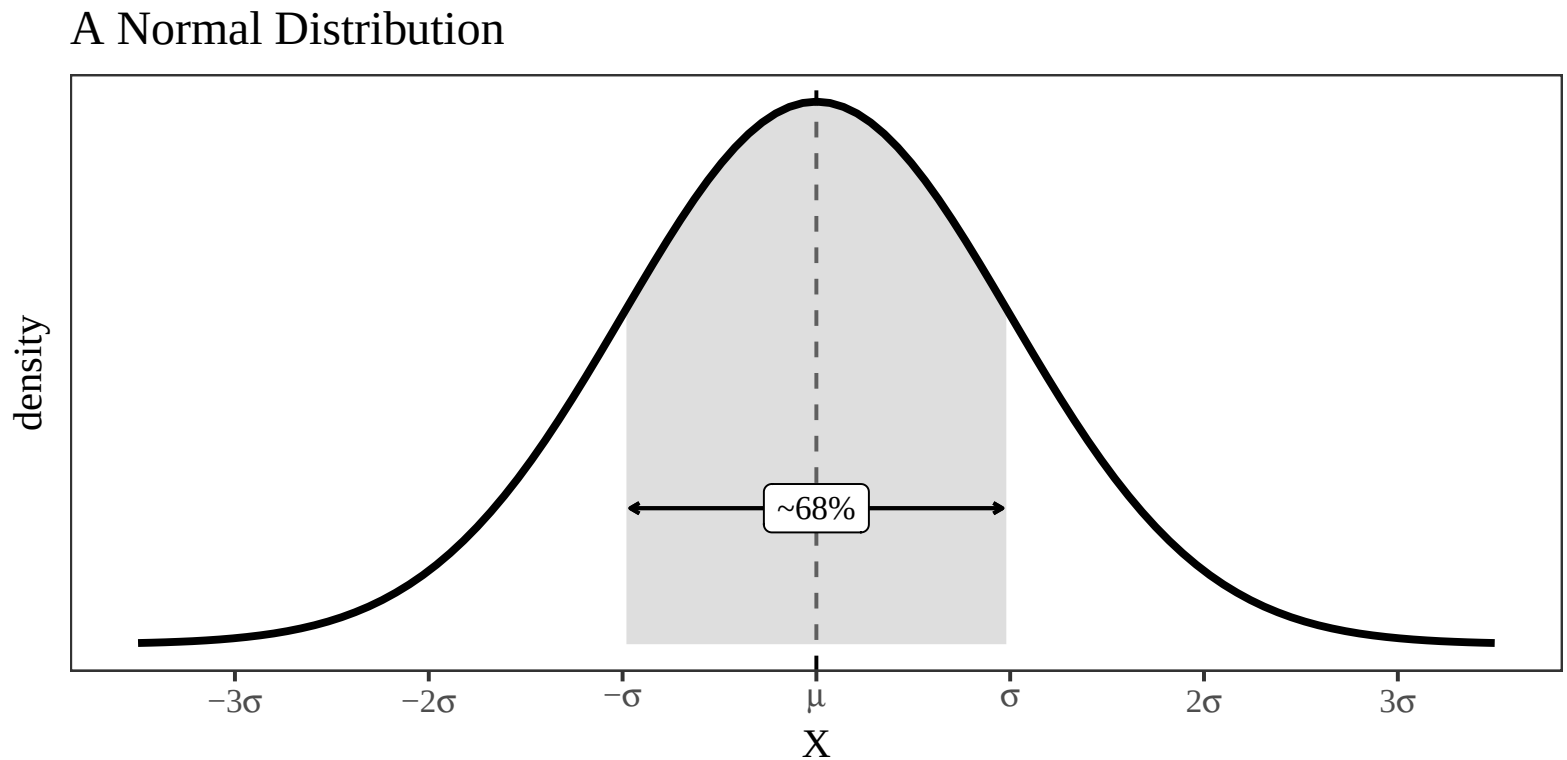
Sometimes you need to calculate a p-value but cannot use R! (Like on an exam...)

The nice thing about normal sampling distributions, we can take advantage of the "niceness" of normal distributions.

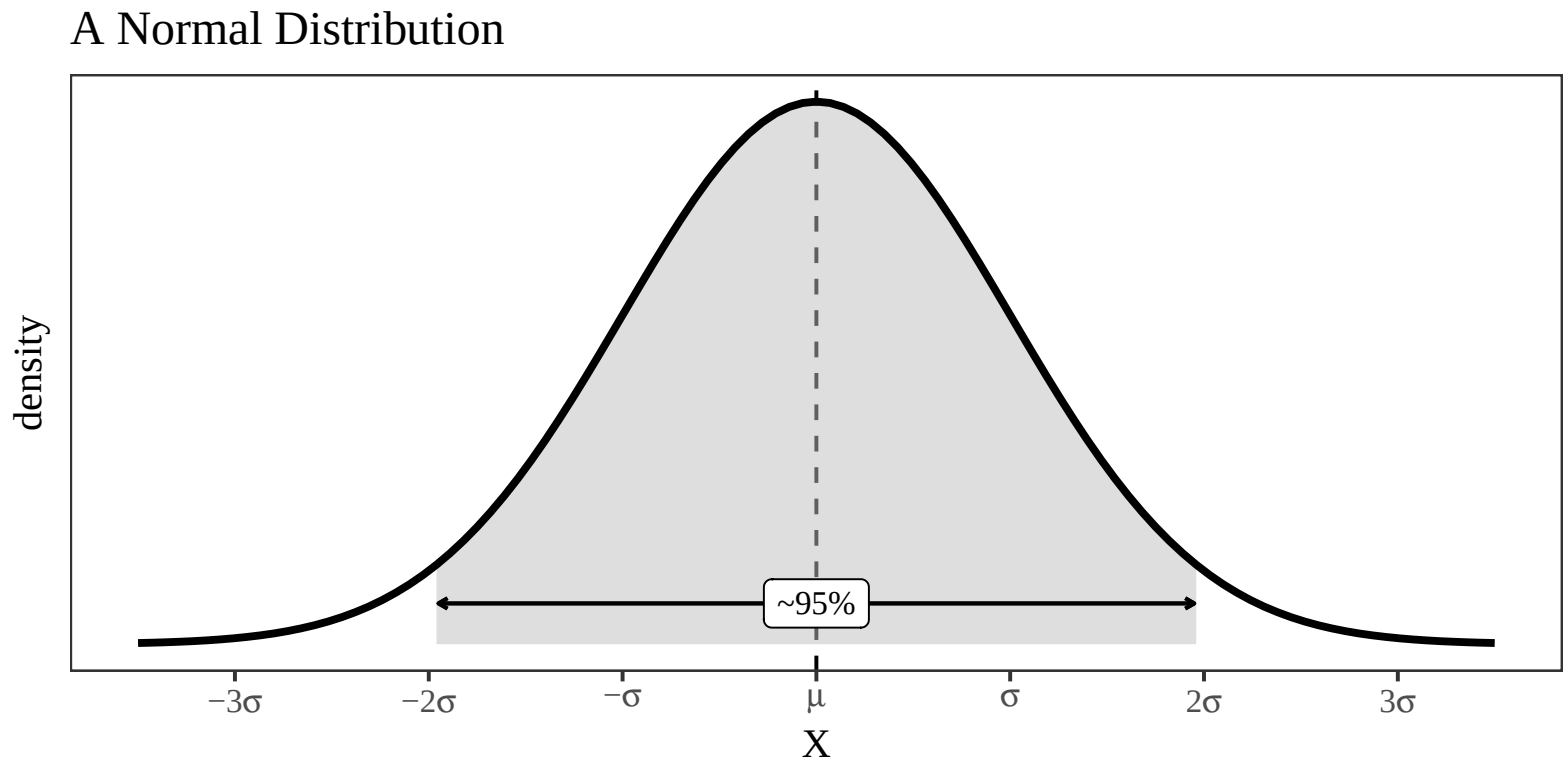
Especially the **1-2-3 Rule**:

- ↪ About two-thirds of the values are within one standard deviation of the mean: $\Pr(\mu - \sigma \leq \mu \leq \mu + \sigma) \approx 0.68$
- ↪ About 95% of the values are within two standard deviations of the mean: $\Pr(\mu - 2\sigma \leq \mu \leq \mu + 2\sigma) \approx 0.95$
- ↪ Almost all of the values are within three standard deviations of the mean: $\Pr(\mu - 3\sigma \leq \mu \leq \mu + 3\sigma) \approx 0.997$

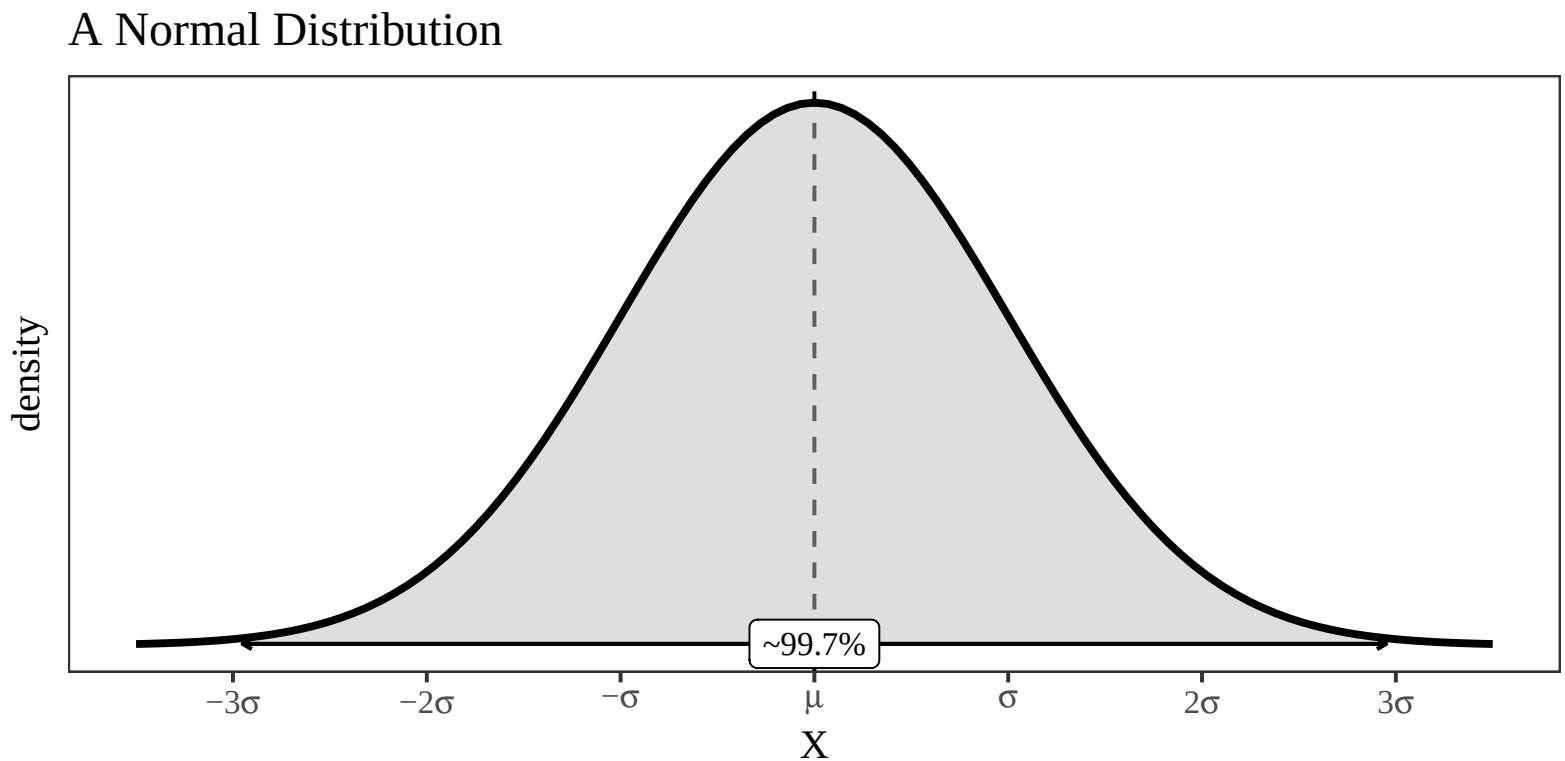
The Niceness of Normal Distributions



The Niceness of Normal Distributions

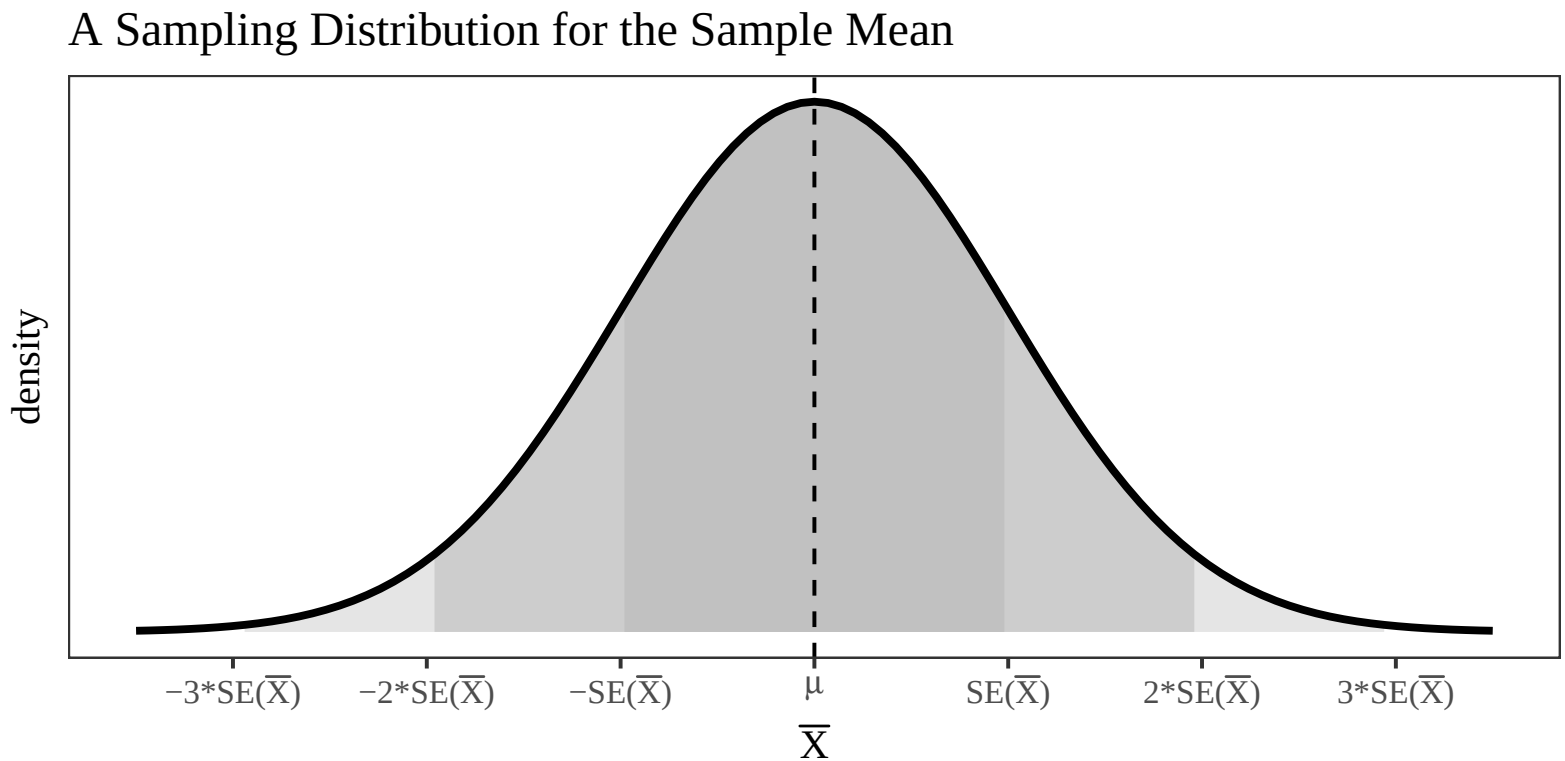


The Niceness of Normal Distributions



The Niceness of Normal Distributions

Recall: many statistics we care about (e.g., means, regression coefficients) have normally distributed sampling distributions.



The Niceness of Normal Distributions

Recall that we can use z-scores to calculate how many standard deviations away from the mean a particular value is.

So, if we get a z-score of around -2 in a sampling distribution, we know that the value is around 2 standard error below the mean.

The probability of getting a value at least 2 SEs away from the mean (in either direction) is approximately $1 - 0.95 = 0.05$.

This gives us some nice rules of thumb: based on the z-score, you can approximate whether your p-value is above or below some "important" thresholds.

$\hookrightarrow$ Like 0.05, which you will see is important.

The Niceness of Normal Distributions

This is nice and all, but sometimes you need to calculate a specific p-value without using R.

Whenever we can convert a value from a normal distribution into a z-score, we can just use a **z table**.

Every intro statistics text book (except ours LOL) has a z table.

But in the modern era, you can find them all over the internet!

https://en.wikipedia.org/wiki/Standard_normal_table

When you can only calculate a t-score, then you can use a **t table**, but these are usually set up a little differently.

What If We Don't Know the DGP Mean?

Recall: when we do not know the DGP, we don't know the mean or standard error of the sampling distribution.

We can easily deal with the second problem by calculating a bootstrapped standard error and using t-scores to measure how many SEs our estimate is from the DGP mean.

But when we don't know the DGP mean, the problem is more fundamental:

If we don't know "the truth," then what's the appropriate benchmark to judge whether we have a "good" estimate?

There are two directions we can go...

What If We Don't Know the DGP Mean?

First: we can come up with a conjecture about the DGP and then *test* whether our estimate provides good evidence in favor of that conjecture.

↪ This is known as **hypothesis testing**.

Second: we can take advantage of some of the math of sampling distributions to simply *estimate* the unknown features of the DGP we want to know.

↪ This is known as **statistical estimation**.

In some ways, estimation is more intuitive than hypothesis testing.

So we'll start with hypothesis testing and then move to estimation.

Randomly Assigned Judges?

Here's a scenario:

↪ There are two judges sitting in a court.

↪ The clerk of the court claims that all cases are randomly assigned to a judge.

↪ However, she won't share details of the exact process.

↪ You managed to collect a dataset of all 100 cases heard by these two judges over the course of two years.

↪ For each case, you know which judge was assigned.

↪ You see that 43 cases were assigned to Judge 1 and 57 were assigned to Judge 2.

Let's assess the clerk's claim that cases are randomly assigned.

Randomly Assigned Judges?

First note that the clerk is making a claim about the DGP!

Mathematically, she's saying that whether case i will be assigned Judge 1 is a random variable (call it J_i) distributed according to:

$$J_i \sim \text{Bin}(1, 0.5)$$

However, since we have a dataset of 100 "draws" from the DGP, we can think of it as a random variable J where

$$J \sim \text{Bin}(100, 0.5)$$

What have we done? We've just written down the clerk's claim in more precise mathematical terms.

Randomly Assigned Judges?

We call a claim (or "conjecture") about a DGP a **hypothesis**.

We have real world data: 43/100 cases were assigned to Judge 1.

We can use the data to **test** the hypothesis.

Let's ask: what's the probability that we would observe Judge 1 being assigned 7 cases fewer or more than half (i.e., $\Pr[J \leq 43 \text{ or } J \geq 57]$) if the claim about the DGP is true?

I can calculate this by hand, but it sucks and it is boring!

```
> pbinom(43, 100, 0.5) + (1 - pbinom(56, 100, 0.5))
[1] 0.1933479
```

Randomly Assigned Judges?

We call a claim (or "conjecture") about a DGP a **hypothesis**.

We have real world data: 43/100 cases were assigned to Judge 1.

We can use the data to **test** the hypothesis.

Let's ask: what's the probability that we would observe Judge 1 being assigned 7 cases fewer or more than half (i.e., $\Pr[J \leq 43 \text{ or } J \geq 57]$) if the claim about the DGP is true?

I can calculate this by hand, but it sucks and it is boring!

```
> sum(dbinom(c(seq(0,43), seq(57,100)), 100, 0.5))
[1] 0.1933479
```

Randomly Assigned Judges?

We call a claim (or "conjecture") about a DGP a **hypothesis**.

We have real world data: 43/100 cases were assigned to Judge 1.

We can use the data to **test** the hypothesis.

Let's ask: what's the probability that we would observe Judge 1 being assigned 7 cases fewer or more than half (i.e., $\Pr[J \leq 43 \text{ or } J \geq 57]$) if the claim about the DGP is true?

Our test revealed that if cases are truly randomized, the probability of Judge 1 getting 7 fewer or more than half is 0.193.

How confident should we be that cases are randomly assigned?

Generalizing Past the Example

Tests like this are very, very common in quantitative research.

They're so common people often don't always realize that they're doing a hypothesis test!

For example when someone asks "is this estimate statistically significant?" they are doing a hypothesis test.

There are many different kinds of hypothesis tests, but they all have the following in common: a null hypothesis, a test statistic, and a p-value.

And they all work in the same basic way.

Generalizing Past the Example

A hypothesis test is about a "parameter" of an (unknown) DGP.

You've seen examples of various parameters that researchers usually care about: means, regression slopes, etc.

$\hookrightarrow$ You've seen these called "estimands" too.

There are three ingredients/steps in the process.

Ingredient 1: Hypotheses

A **hypothesis** is a claim that a parameter takes specific value(s).

All hypothesis tests have two hypotheses:

↪ A **null hypothesis**: a claim the researcher argues against.

↪ An **alternative hypothesis**: a claim the researcher argues for.

Setting up hypotheses in this way is *conservative*: forcing a high burden of proof on the researcher.

Researcher picks hypotheses, but they should never (I repeat: never) be chosen based on the observed data.

Ingredient 1: Hypotheses

A couple examples will help. When a researcher...

1. ... runs a regression, typically:

$\hookrightarrow$ The null hypothesis: $H_0 : \beta = 0$ and $H_a : \beta \neq 0$.

2. ... compares means of two distributions, typically:

$\hookrightarrow$ The null hypothesis: $H_0 : \mu_1 - \mu_2 = 0$ and $H_a : \mu_1 - \mu_2 \neq 0$.

Notice that these set up "hard tests" for finding relationships.

For a univariate distributions, you need a benchmark, for example:

$\hookrightarrow$ The null hypothesis: $H_0 : \mu = 0$ and $H_a : \mu \neq 0$.

Ingredient 1: Hypotheses

You do not need to have hypotheses that cover the whole "parameter space." For example: $H_0 : \mu = 0$ and $H_a : \mu > 0$.

↪ Practically, it's not common to specify hypotheses where some parameter values are not covered by either hypothesis.

For a **two-tailed test**, the alternative hypothesis conjectures that the parameter is on either side of H_0.

↪ For example: $H_0 : \beta = 0$ and $H_a : \beta \neq 0$.

For a **one-tailed test**, the alternative hypothesis conjectures that the parameter is on one side of H_0.

↪ For example: $H_0 : \mu \leq 3$ and $H_a : \mu > 3$.

Ingredient 1: Hypotheses

A little pedantic, but important:

↪ Hypotheses are (mathematical) statements about the DGP, and so they are either true or false.

↪ But, a researcher can never know for sure whether a hypothesis is true or false because we cannot see the DGP!

↪ Hypothesis tests involve making calculations from observed data and then evaluating the *strength* of the evidence.

Ingredient 2: Test Statistics

When you "do" the test, you compute a **test statistic**, which provides the evidence you use to adjudicate hypotheses.

These are typically the sample analogues of the parameter. (That is: the estimators for the estimand.)

For example:

$\hookrightarrow$ For a hypothesis about β, the test statistic will be $\hat{\beta}$.

$\hookrightarrow$ For a hypothesis about μ, the test statistic will be $\overline{X}$.

Ingredient 2: Test Statistics

In reality: most of the time we convert our sample estimate to a z-score or a t-score, which then becomes the test statistic.

For example, suppose that we want to test a null hypothesis that the sample regression coefficient is zero.

If we satisfy the assumptions, then we know the sampling distribution for β is approximately normal with $E(\hat{\beta}) = \beta$ and $\widehat{SE}(\hat{\beta})$ estimated using the formula 7.16 in Gailmard (2014).

If we are assuming $\beta = 0$, we'll convert an estimated regression slope it into a t-score (using $H_0 : \beta = 0$ instead of the real β):

$$t = \frac{b - 0}{\widehat{SE}(\hat{\beta})}$$

Ingredient 3: Decision Rules and Significance

We say that you **reject** a hypothesis if the test statistic is in a pre-specified "rejection range."

In practice: if you set up hypotheses to "partition the parameter space" then rejecting H_0 means accepting H_a and vice versa.

Because we never know for sure if H_0 or H_a is true (we don't see the DGP!), then we can make errors when we reject hypotheses:

A **Type I** error: rejecting H_0 when it is actually true.

A **Type II** error: accepting H_0 when it is actually false.

Ingredient 3: Decision Rules and Significance

How do you pick the rejection range?

↪ Because the test statistic is a random variable with a sampling distribution, there's some noise in the estimate.

↪ So, you can't be *too* demanding with your test.

↪ If $H_0 : \beta = 0$, it would be sort of useless to come up with a rejection range where you reject H_0 unless $\hat{\beta} = 0$.

Usually, people choose rejection ranges based on **significance level** (or **size**) they choose for their test.

The significance level of a test is the probability of making a Type I error (rejecting H_0 when it is actually true).

Ingredient 3: Decision Rules and Significance

We usually denote the significance level with a Greek alpha: α.

In practice, most researchers use $\alpha = 0.01$, $\alpha = 0.05$ or $\alpha = 0.10$.

$\hookrightarrow$ It is falling out of fashion to use $\alpha = 0.1$. Do not use this significance level!

$\hookrightarrow$ When someone says, "This statistic is statistically significant" they almost always mean that H_0 was rejected using a significance level of $\alpha = 0.05$.

$\hookrightarrow$ You should get in the habit of saying "This statistic is statistically significant at the XX significance level."

The values of the test statistic that lead to a rejection are derived from the significance level the researcher chooses in advance.

Ingredient 3: Decision Rules and Significance

How do you actually decide to accept or reject based on data?

First, we calculate the relevant p-value based on the actual estimate from our data.

Then, we assess the p-value relative to the level of significance we pre-specified.

Recall the judge randomization example: $H_0 : p = 0.5$ and $H_a : p \neq 0.5$.

In the judge randomization example: p-value of 0.193.

If we use $\alpha = 0.05$, then, we cannot reject the null hypothesis.

Another Example

To R!

Downsides of Learning from Hypothesis Tests

Recall: we never know the DGP, but want to know it.

Hypothesis testing involves making a conjecture about a DGP's parameter and then evaluating whether an estimate provides good evidence in favor of that conjecture.

But this is a round about way of learning about the DGP! How do you know what hypothesis to set up??

$\hookrightarrow$ In practice: we make a "conservative" conjecture about the DGP and see if our estimate can reject that conjecture.

Downsides of Learning from Hypothesis Tests

Many real-life empirical projects involve conducting several hypothesis tests.

For example, in a study of judges decision-making:

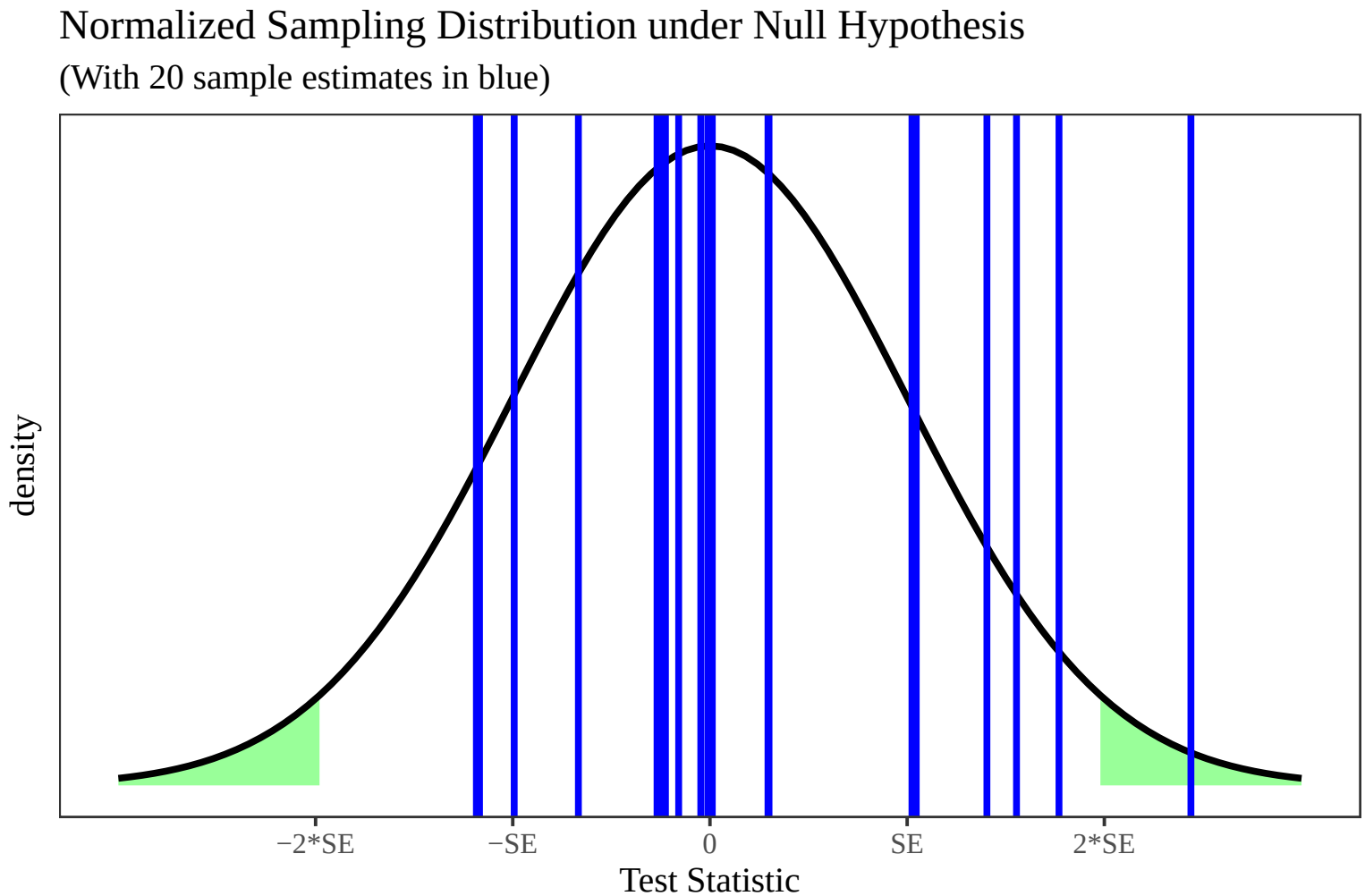
↪ Does judge partisanship have a statistically significant effect on whether cases settle? Whether they are dismissed? Etc.

↪ Does a judge's race have a statistically significant effect on how cases end? Their gender? Their religion? Etc.

Thinking a little about how sampling works, this creates a **multiple comparisons (or testing) problem**.

↪ Solution: either don't do it or use a correction. Or do nothing. See Gailmard (2014) or the internet for more.

Downsides of Learning from Hypothesis Tests



Interval Estimation

There's another way to think about how we might learn about a DGP from a sample estimate.

Instead of coming up with one specific null hypothesis about a parameter, why not just ask:

What null hypotheses about the value of a DGP parameter would we not reject given the estimate we actually have?

Alternatively: in light of sampling uncertainty, what range of values above and below our estimate would be "reasonable" guesses about the true value of the parameter.

This is the motivation for **interval estimation**.

Interval Estimation

Just like with hypothesis tests, interval estimation involves specifying how "confident" you want to be.

You need to pick your confidence level, which is just $1 - \alpha$ (one minus the significance level).

Suppose you are interested in the value of a DGP parameter θ and you use an estimator $\hat{\theta}$ to come up with an estimate T.
(For example, θ could be a sample mean or regression slope.)

A **95% confidence interval** is the range of possible values above and below an estimate T that have a 0.95 probability of arising under the distribution of the estimator $\hat{\theta}$.

Interval Estimation

The easiest situation arises when you can convince yourself that your estimator $\hat{\theta}$ is distributed normally.

From before: we know that there's a 0.95 probability that T is no more than 1.96 standard errors away from the true value of θ.

$\hookrightarrow$ Think back to the 1-2-3 rule.

But we see T and not θ, so we invert the problem: there is a 0.95 probability that θ is no more than 1.96 standard errors away from the estimate T. In math:

$$\Pr\left(\hat{\theta} - 1.96 \times SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + 1.96 \times SE(\hat{\theta})\right) = 0.95$$

A CI is just the interval of values $\leq$ 1.96 SEs away from T!

Interval Estimation

But of course, we are always ever working with estimates from a sample. So, we don't know $SE(\hat{\theta})$ and have to bootstrap: $\widehat{SE}(\hat{\theta})$.

After bootstrapping, we know our sampling distribution isn't exactly normal anymore; we have to use the Student's t distribution.

Practically speaking, this involves two changes:

$$\Pr\left(\hat{\theta} - \hat{t} \times \widehat{SE}(\hat{\theta}) \leq \theta \leq \hat{\theta} + \hat{t} \times SE(\hat{\theta})\right) = 0.95$$

where $\hat{t}$ is the critical value of t required for 95% confidence given that you have $N - 1$ degrees of freedom. You can find $\hat{t}$ in R:

```
> qt(1-(0.05/2), N-1)
```

Interval Estimation

Or using a t-table, like this one from Wikipedia:

One-sided	**75%**	**80%**	**85%**	**90%**	**95%**	**97.5%**	**99%**	**99.5%**	**99.75%**	**99.9%**	**99.95%**
Two-sided	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.816	1.080	1.386	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869

...

80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291
One-sided	**75%**	**80%**	**85%**	**90%**	**95%**	**97.5%**	**99%**	**99.5%**	**99.75%**	**99.9%**	**99.95%**
Two-sided	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**

Interval Estimation

Then, to calculate the 95% CI, we just slot in our estimates:

$$\underbrace{T - \hat{t} \times \widehat{SE}(\hat{\theta})}_{\text{lower bound of CI}} \qquad \underbrace{T + \hat{t} \times \widehat{SE}(\hat{\theta})}_{\text{upper bound of CI}}$$

where T is our specific estimate for θ from the sample we drew.

Interval Estimation

Example: suppose we're thinking about a regression slope.

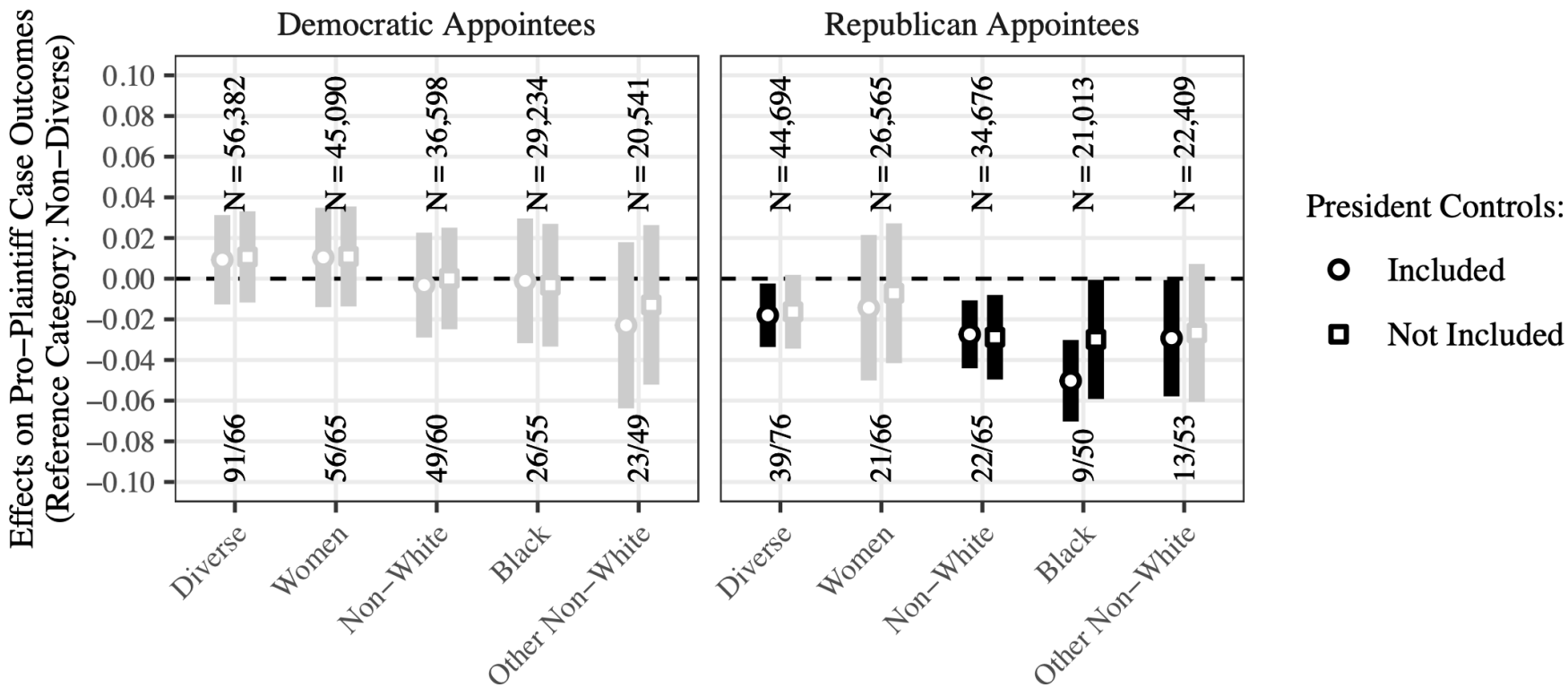
We have a dataset with $N = 120$ and R calculates $b = 0.72$ and $\widehat{SE}(\hat{\beta}) = 0.35$.

Since we have 119 degrees of freedom, $\hat{t} \approx 1.98$ (see table). Then:

$$0.72 - 1.98(0.35) = 0.027 \qquad 0.72 + 1.98(0.35) = 1.413$$

If we take into consideration our uncertainty, then $b = 0.72$ is our best guess about β. But we wouldn't be surprised if it were anywhere between 0.027 and 1.413, where we defined "not being surprised" as having 95% confidence.

Interval Estimation



Interval Estimation

	Republican Appointees									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Diverse	–0.018*	–0.016	—	—	—	—	—	—	—	—
	(0.008)	(0.009)								
Women	—	—	–0.014	–0.007	—	—	—	—	—	—
			(0.016)	(0.015)						
Non-White	—	—	—	—	–0.027*	–0.029*	—	—	—	—
					(0.007)	(0.010)				
Black	—	—	—	—	—	—	–0.05*	–0.03*	—	—
							(0.009)	(0.012)		
Other Non-White	—	—	—	—	—	—	—	—	–0.029*	–0.027
									(0.012)	(0.014)
Bush41	0.037	—	0.059	—	0.023	—	0.067	—	–0.033	—
	(0.028)		(0.032)		(0.030)		(0.039)		(0.029)	
Bush43	0.039	—	0.044	—	0.035	—	0.039	—	0.032	—
	(0.023)		(0.027)		(0.024)		(0.028)		(0.021)	
Cases (observations)	44,694	44,694	26,565	26,565	34,676	34,676	21,013	21,013	22,409	22,409
Treatment judges	39	39	21	21	22	22	9	9	13	13
Control judges	76	76	66	66	65	65	50	50	53	53
Randomization blocks	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Min. Units/Trt. Arm	1	1	1	1	1	1	1	1	1	1
Drop Preska	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Drop Chiefs										
Drop NYSD (09-16)										

Desirable Properties of Estimators

Confidence intervals are nice because they explicitly present estimates in a way that conveys uncertainty.

But we almost always want to give people a **point estimate** of the particular estimand we're interested in.

So far, we've sort of danced around this, presuming an estimate we get from a sample is our point estimate for the estimand.

But this relied heavily on your intuitions about specific kinds of sample calculations: means, regression slopes, etc.

Let's dig a little deeper and connect some loose threads.

Ideal Properties 1: Unbiasedness

First, recall: "estimator" is the phrase we use to describe the "process" of getting estimates.

For example, if we are (theoretically) interested in the expected value of a DGP represented by a univariate probability distribution, then the estimator we've focused on is the sample mean.

We like this estimator because some math tells us:

$$E(\overline{X}) = E(X)$$

In other words, the sample mean is an **unbiased estimator** for the DGP expected value.

Ideal Properties 1: Unbiasedness

We're not always interested in expected values of single variables.

We also saw that the sample (OLS) regression slope is an unbiased estimator for the DGP (OLS) regression slope if X and ε are not correlated:

$$E(\hat{\beta}) = E(\beta) \quad \text{if} \quad E(\widehat{Cov}(X, \varepsilon)) = 0$$

More generally, an estimator $\hat{\theta}$ is unbiased if:

$$E(\hat{\theta}) - E(\theta) = B(\hat{\theta}) = 0$$

If an estimator is biased, then it produces systematically larger or smaller estimates.

Ideal Properties 2: Precision

But that's not the only thing we care about.

Recall we can calculate the standard error of the sampling distributions for $\overline{X}$ and $\hat{\beta}$.

This tells us how **precise** (or noisy) the estimator is. (Mathematically: precision is $1/Var(\hat{\theta})$.)

What do we know about the estimators you know and love?

$\hookrightarrow$ The precision of $\overline{X}$ is $1/Var(\overline{X}) = N/Var(X)$.

$\hookrightarrow$ Assuming Y_is are independent and homoskedacity holds, then the precision of $\hat{\beta}$ is $1/Var(\hat{\beta}) = (N-1)s_X^2/Var(Y|X)$.

Ideal Properties 2: Precision

Ideally the estimator you use will be unbiased and precise.

But in practice, this is not always easy to achieve.

The **bias-variance tradeoff** captures the idea that, in many situations, the only way to reduce bias increases variance.

One criterion for evaluating how a specific estimator "balances" bias and variance is to look at it's **mean square error**:

$$MSE(\hat{\theta}) = E((\hat{\theta} - \theta)^2) = Var(\hat{\theta}) + \underbrace{(E(\hat{\theta}) - E(\theta))^2}_{\text{bias squared}}$$

Some people like estimators that minimize MSE, but others are queasy about increasing bias to reduce variance.

Ideal Properties 3: Consistency

A final way to evaluate an estimator is to think about its **consistency**.

This is a little mathematically complicated, but the basic idea can be stated pretty easily using prose.

An estimator is **consistent** if the probability of getting estimates "far" from the estimand goes to zero as the sample size gets infinitely big.

Roughly speaking: it is good to know an estimator is consistent because it means that in smaller samples that we typically have access to, we can expect it to behave in the ways it would in larger samples.

Ideal Properties 3: Consistency

We haven't talked about the consistency of the $\overline{X}$ estimator yet.

But, as you might imagine, it is consistent.

This is known as the **(weak) law of large numbers**.

You'll have to wait to learn more about the consistency of $\hat{\beta}$.

But, whenever we're talking about an estimator's "consistency" we're basically talking about how the estimator behaves in very very very large samples.

Issues around consistency are often referred to as **asymptotics**.

Ideal Properties 3: Consistency

A lot of this stuff is abstract, so let's consider a purposefully absurd example.

Suppose I'm interested in Biden approval in the population.

I draw a simple random sample from the population.

Instead of the sample mean, suppose I use this estimator, which I will just call $\widetilde{C}$: $\widetilde{C} = 0.6$.

$\hookrightarrow$ This just says: no matter what is in the sample, estimate Biden approval as 60%.

This is a very precise estimator! But it sucks: it is very biased.

Quick Summary

To emphasize how this works:

↪ Researcher specifies an estimand they care about: a parameter of the DGP that we cannot actually observe.

↪ Researcher has access to a sample from the DGP, and decides on an estimator to use to calculate an estimate.

Every particular estimator can give "better" or "worse" estimates for the estimand: bias, precision and consistency.

Why do you need this stuff? Because most research involves trying to figure out what estimator you need for your situation.

In practice: the thing people worry most about is bias...