Power Rules

Power Rules*

Practical Statistical Power Calculations

Carlisle Rainey†

June 7, 2024

Draft *very much* in progress.
Comments welcomed! (crainey@fsu.edu)
Get latest version from GitHub.

Abstract

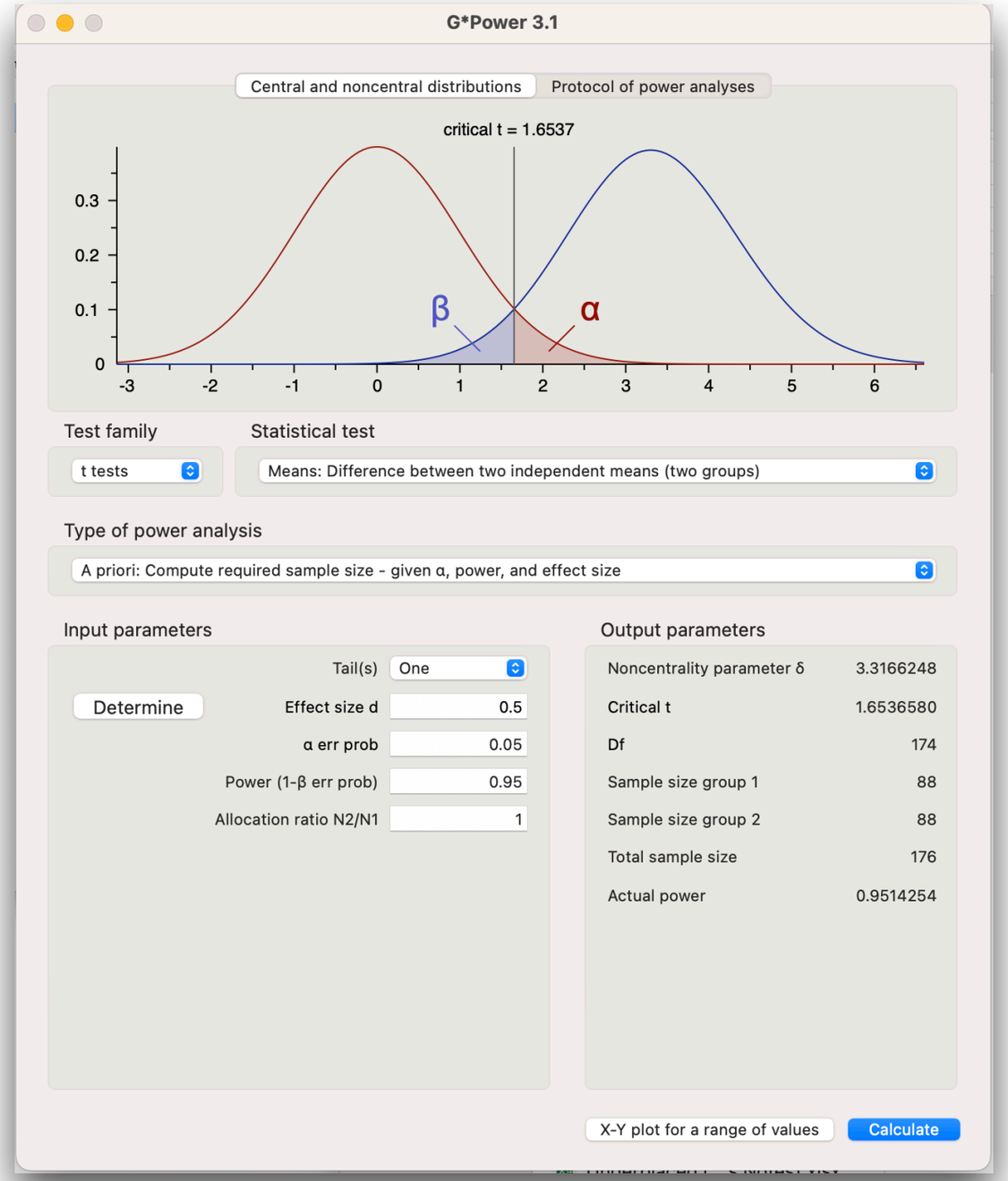
Recent work emphasizes the importance of statistical power and shows that power in the social sciences tends to be extremely low. In this paper, I offer simple rules that make statistical power more approachable for substantive researchers. The rules describe how researchers can compute power using (1) features of a reference population, (2) an existing study with a similar design and outcome, and/or (3) a pilot study. In the case of balanced, between-subjects designs (perhaps controlling for pre-treatment variables), these rules are sufficient for a complete and compelling power analysis for treatment effects and interactions using only paper-and-pencil. For more complex designs, these rules can provide a useful ballpark prediction before turning to specialized software or complex simulations. Most importantly, these rules help researchers develop a sharp intuition about statistical power. For example, it can be helpful for readers and researchers to know that experiments have 80% power to detect effects that are 2.5 times larger than the standard error and how to easily form a conservative prediction of the standard error using pilot data. These rules lower the barrier to entry for researchers new to thinking carefully about statistical power and help researchers design powerful, informative experiments.

*This essay flows from many discussions with undergraduate students, PhD students, and colleagues. In these discussions, we have thought about power as a minor extension of estimation and inference rather than power as a distinct topic. In this essay, I try to articulate that perspective and help researchers extend their intuitions and expertise with estimation and inference to statistical power. Marco Aviña, Scott Clifford, Ryan Powers, and Chuck Smith provided valuable comments on earlier drafts.

†Carlisle Rainey is Associate Professor of Political Science, Florida State University, 540 Bellamy, Tallahassee, FL, 32306. (crainey@fsu.edu).

The Origin Story

FSU's RIBC Program



in the language of psychologists; but unfamiliar

Declaration 18.1 Canonical two-arm trial design.

```
declaration_18.1 <-
  declare_model(N = 100,
                U = rnorm(N),
                potential_outcomes(Y ~ 0.2 * Z + U)) +
  declare_inquiry(ATE = mean(Y_Z_1 - Y_Z_0)) +
  declare_assignment(Z = complete_ra(N, prob = 0.5)) +
  declare_measurement(Y = reveal_outcomes(Y ~ Z)) +
  declare_estimator(Y ~ Z, inquiry = "ATE")
```

very general and powerful; but complex

I want to start on the dry erase board.

Contributions

1. How-to guide for **political scientists**.

2. Develop **intuitions** rather than rote procedures.

3. Discuss the role **pilot data** can play in power calculations.

feel like "creative set of lecture notes" (but important?)

more novel contribution

Three Goals

1. **Undermine** your trust in experiments.
2. **Rebuild** your trust.
3. Discuss **a few handy rules**.

Goal #1

Undermine your trust in experiments

Name	Y(1)	Y(0)	Y(1) - Y(0)	W
	P.O. Under Treatment	P.O. Under Control	Unit Treatment Effect	In T (1) or C (0)?
Adam	7	5	2	1
Beth	2	3	-1	1
Carl	7	7	0	0
Donna	5	4	1	1
Eric	3	6	-3	0
Fiona	4	4	0	0
George	5	3	2	1
Haley	3	3	0	0
Isaac	4	2	2	0
Judy	1	3	-2	0
Kyle	7	1	6	1
Leigh	5	3	2	1

ATE = 0.75

Name	Y(1)	Y(0)	Y(1) - Y(0)	W
	P.O. Under Treatment	P.O. Under Control	Unit Treatment Effect	In T (1) or C (0)?
Adam	**7**	5	—	1
Beth	**2**	3	—	1
Carl	7	**7**	—	0
Donna	**5**	4	—	1
Eric	3	**6**	—	0
Fiona	4	**4**	—	0
George	**5**	3	—	1
Haley	3	**3**	—	0
Isaac	4	**2**	—	0
Judy	1	**3**	—	0
Kyle	**7**	1	—	1
Leigh	**5**	3	—	1

Treatment: 7 + 2 + 5 + 5 + 7 + 5 = 31 31/6 = 5.17

Control: 7 + 6 + 4 + 3 + 2 + 3 = 25 25/6 = 4.17

Estimate of ATE:
5.17 - 4.17 = 1.00

```
> # do simulations
> N <- length(Y1)
> n_iter <- 10
> set.seed(1234)
> tau_hat <- NULL
> for (i in 1:n_iter) {
+   W <- sample(rep(0:1, length.out = N))
+   tau_hat[i] <- mean(Y1[W == 1]) - mean(Y0[W == 0])
+   print_so_far(tau_hat)  # custom function
+   #pause(i) # custom function
+ }
Estimates so far
#1:      1.00
#2:      2.00
#3:      2.50
#4:      2.00
#5:      1.50
#6:      1.83
#7:      1.00
#8:      0.33
#9:     -1.00
#10:     1.83
```

Carlisle Rainey
@carlislerainey

There's value in talking about studies as "misleading and noisy" by default, where we can intentionally shrink that noise to tolerable levels through design.

Rather that vice versa, where a "scientific experiment" must surely produce the right answer (at least most of the time)

3:43 PM · Oct 30, 2023 · **67** Views

So why do we like randomization?

Theorem: $\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i \cdot Y_i(1)}{N_t/N} - \frac{(1-W_i)\cdot Y_i(0)}{N_c/N}\right)$ is an unbiased estimator of ATE.

Unbiased!

Randomized Experiment

Clinical Trial

Randomized Controlled Trial

"The Gold Standard"

How could a randomized experiment be misleading???

Let me be dramatic about this…

Randomization is a deal with the devil.

You guarantee unbiasedness.

But you also guarantee noise.

Theorem: $\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i \cdot Y_i(1)}{N_t/N} - \frac{(1-W_i)\cdot Y_i(0)}{N_c/N}\right)$ is an unbiased estimator of ATE.

Proof.

We're trying to show that $E_W\left[\widehat{\text{ATE}}\right] = \text{ATE}$.

$$E_W\left[\widehat{\text{ATE}}\right] = E_W\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i \cdot Y_i(1)}{N_t/N} - \frac{(1-W_i)\cdot Y_i(0)}{N_c/N}\right)\right] \tag{1}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(E_W[W_i]\cdot Y_i(1)\frac{1}{N_t/N} - E_W[1-W_i]\cdot Y_i(0)\frac{1}{N_c/N}\right) \tag{2}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) \tag{3}$$

$$= \text{ATE} \tag{4}$$

Done! $\widehat{\text{ATE}}$ is an unbiased estimator of ATE under the assumptions of the design. No parametric model of the data needed, just a (known) model of the the design.

Do not think of experiments as truth machines.

Instead, think of experiments as **noise machines**.

We inject noise (*bad!*) to get unbiasedness (*good!*).

This can be a **good trade** because we know* the amount of noise in the estimate.

We use that knowledge to create a **hypothesis test** or a **confidence interval**.

$$\widehat{\mathrm{Var}}(\widehat{ATE}) = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}$$

This is the SE from Welch's t-test or HC2 robust SEs.

$$\widehat{\mathrm{ATE}} \pm 1.64 \cdot \underbrace{\sqrt{\widehat{\mathrm{Var}}\left(\widehat{\mathrm{ATE}}\right)}}_{\text{estimate of SE}}$$

Hypothesis: The ATE is positive.

Test

1. Construct a 90% CI using the previous method.
2. Check that the lower bound of the 90% CI is greater than zero.
   (a) If the LB > 0, reject the null hypothesis.
   (b) If the LB < 0, do not reject the null hypothesis (i.e., make no claim).

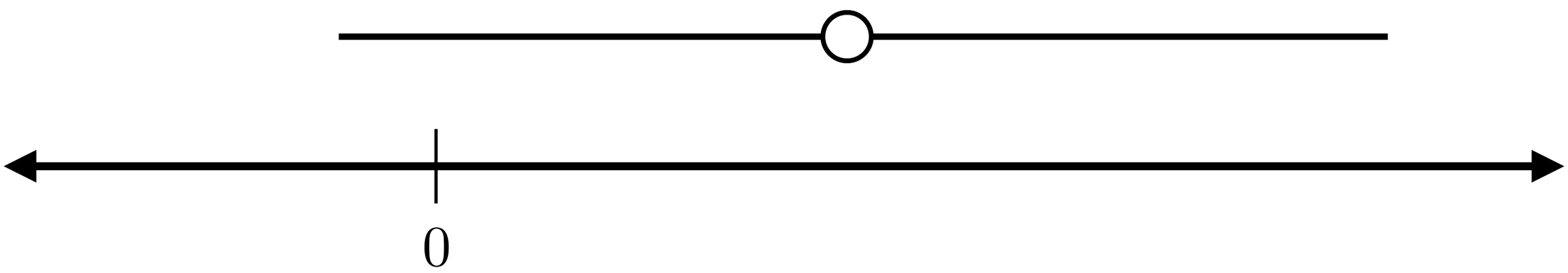
Hypothesis: The ATE is positive.

Test

1. Construct a 90% CI using the previous method.
2. Check that the lower bound of the 90% CI is greater than zero.
   (a) If the LB > 0, reject the null hypothesis.
   (b) If the LB < 0, do not reject the null hypothesis (i.e., make no claim).

Name	Label	Longer Label	Error Rate	Person in Charge
Type 1	“False Positive”	You claim that your research hypothesis is true, but it is not true.	5%	Statistician
Type 2	“False Negative” “Lost Opportunity”	You cannot make a claim about your research hypothesis.	???	You, the researcher

We should focus carefully on this “wasted effort” error.



(that’s where power comes in)

Point #1

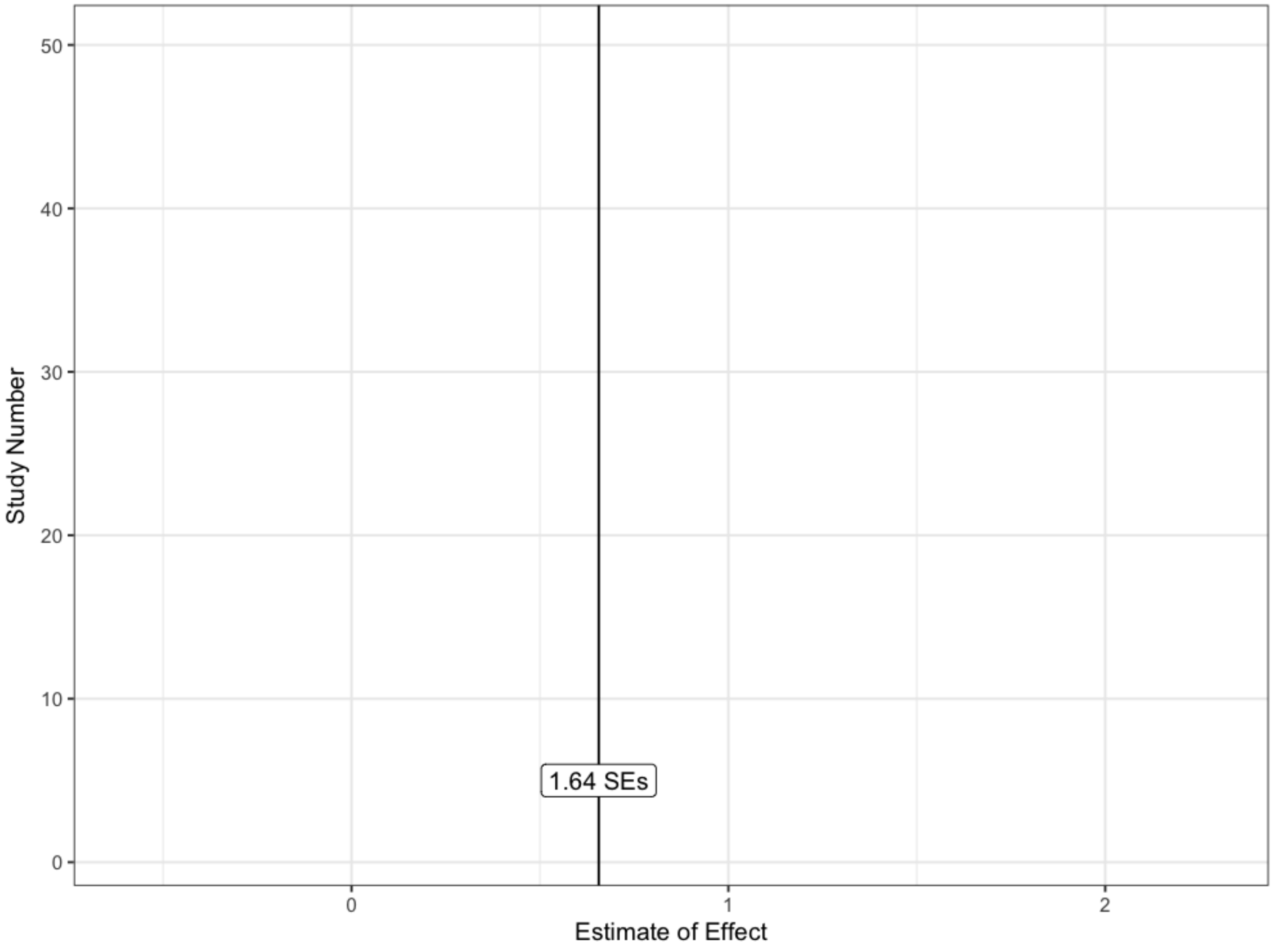
Experiments are NOISE MACHINES.

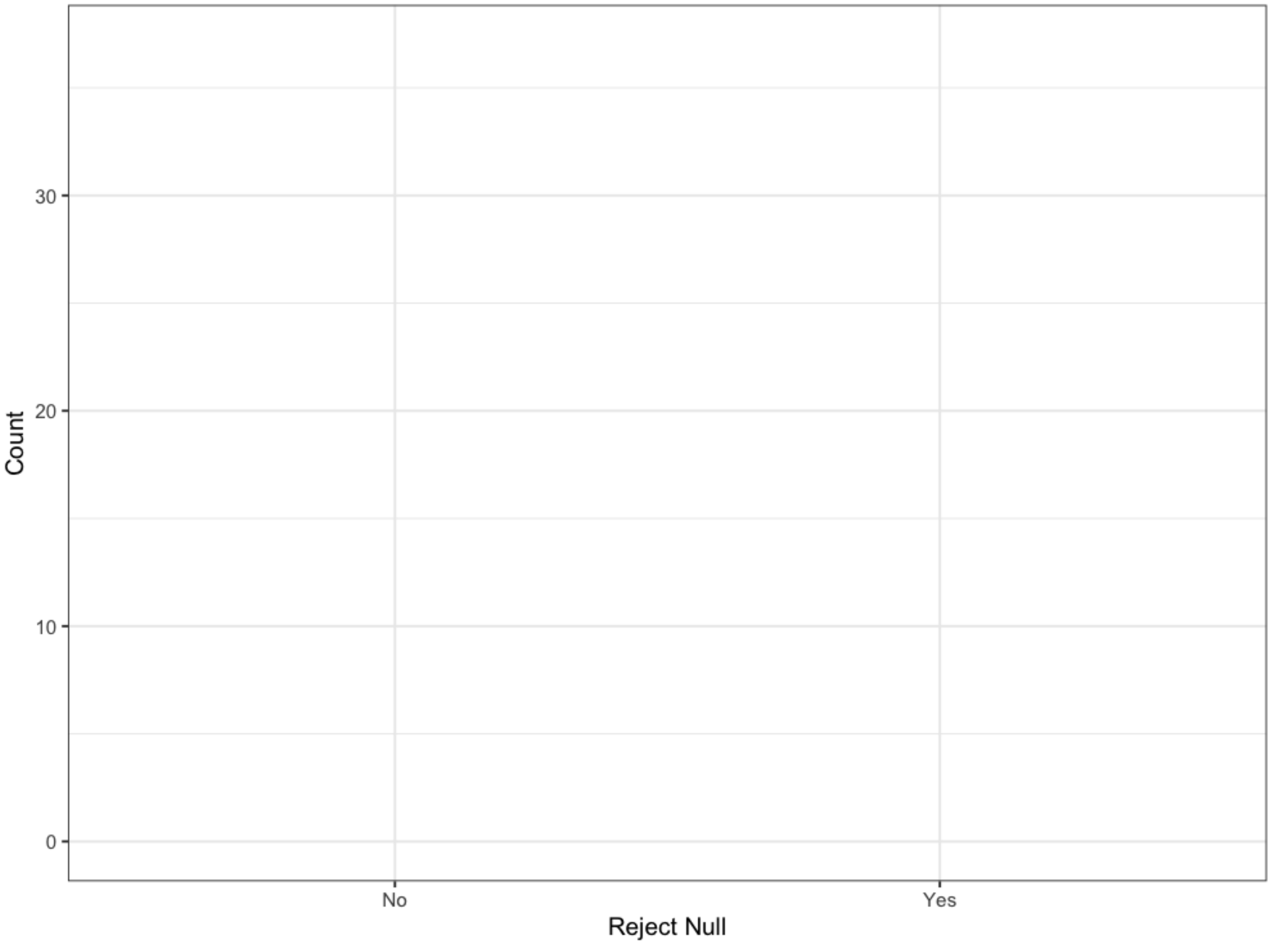
There's a real risk that you waste your time, money, and effort.

Goal #2

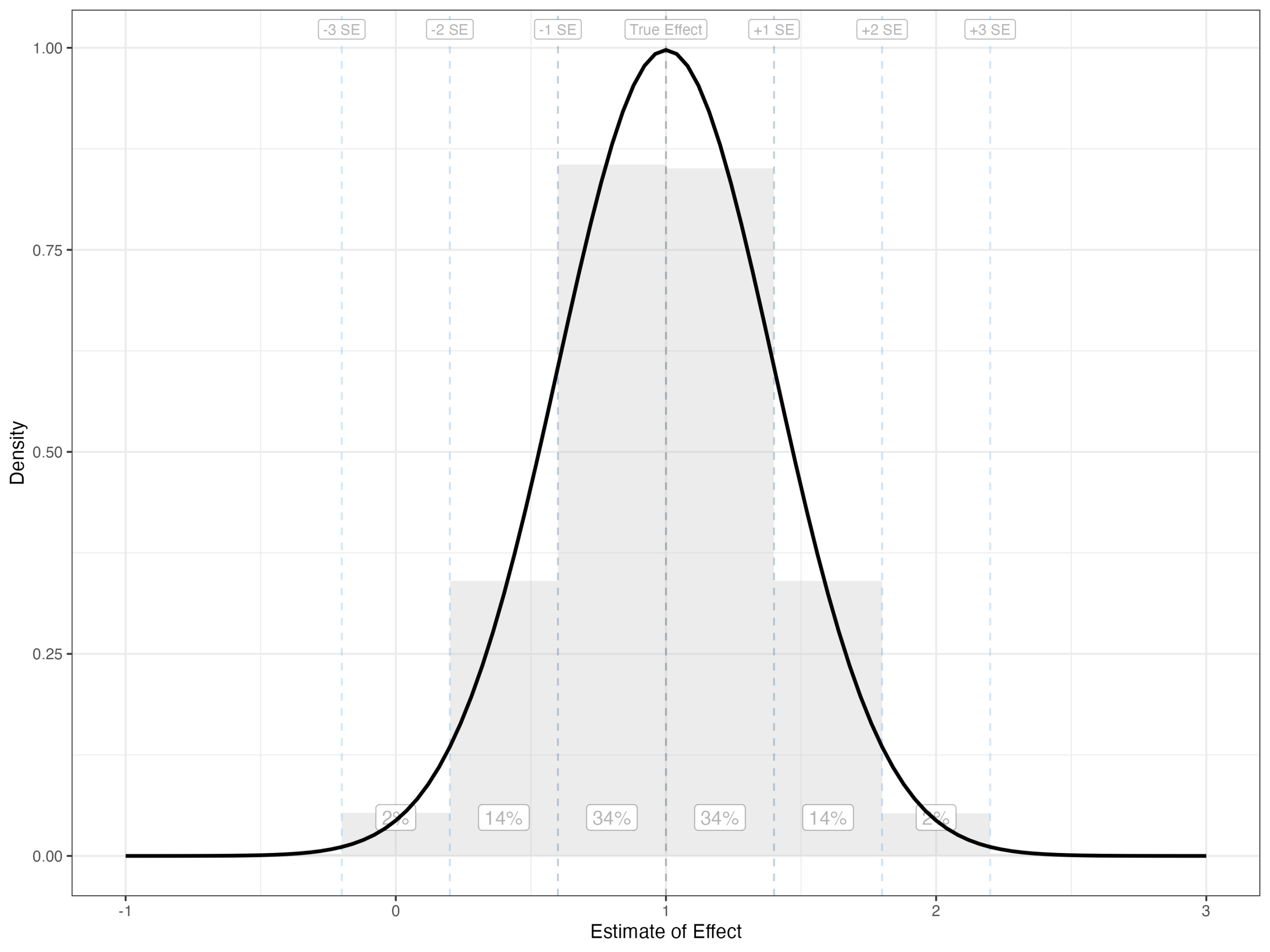
Rebuild your trust

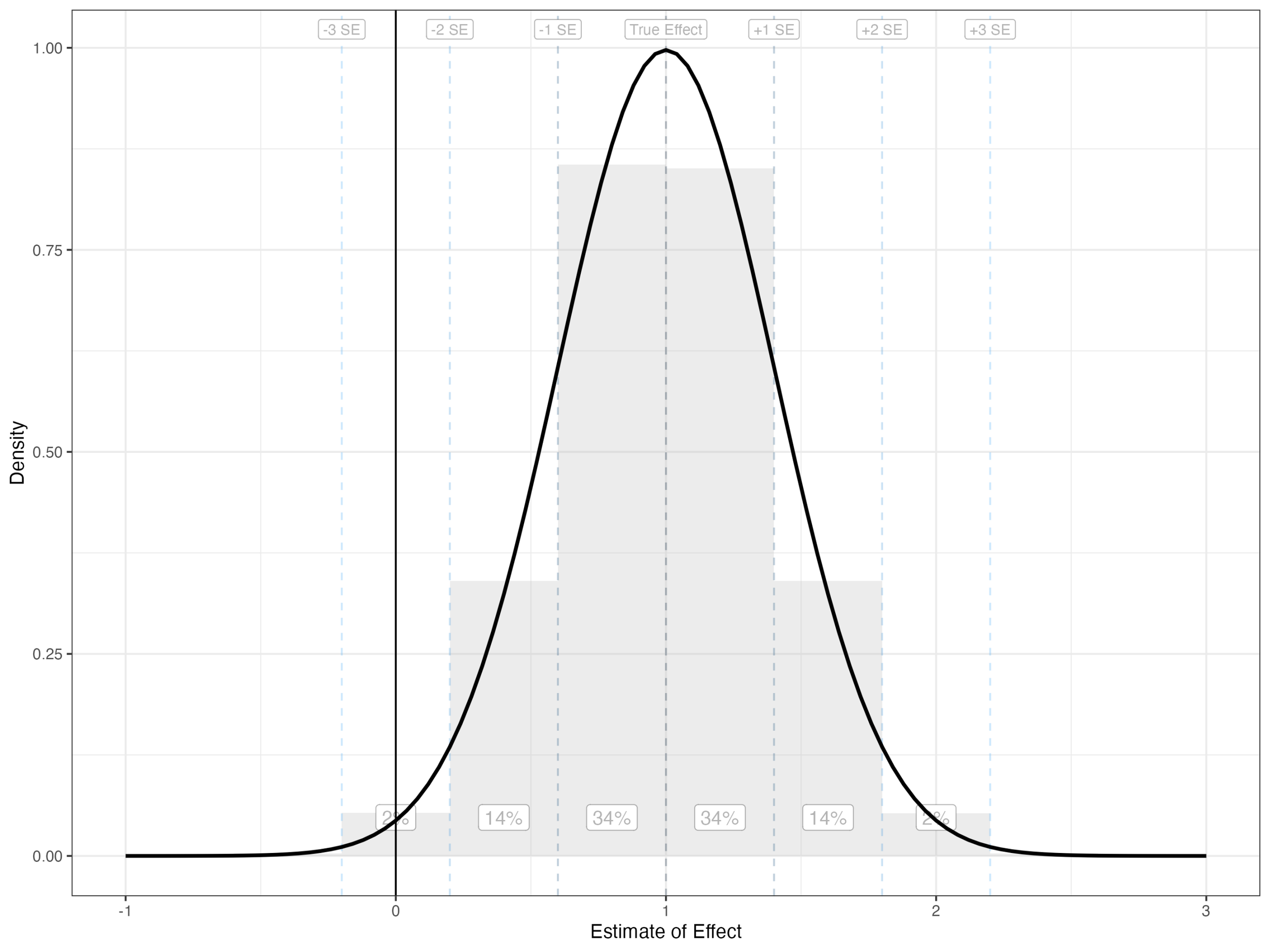
Make sure that our result is informative.

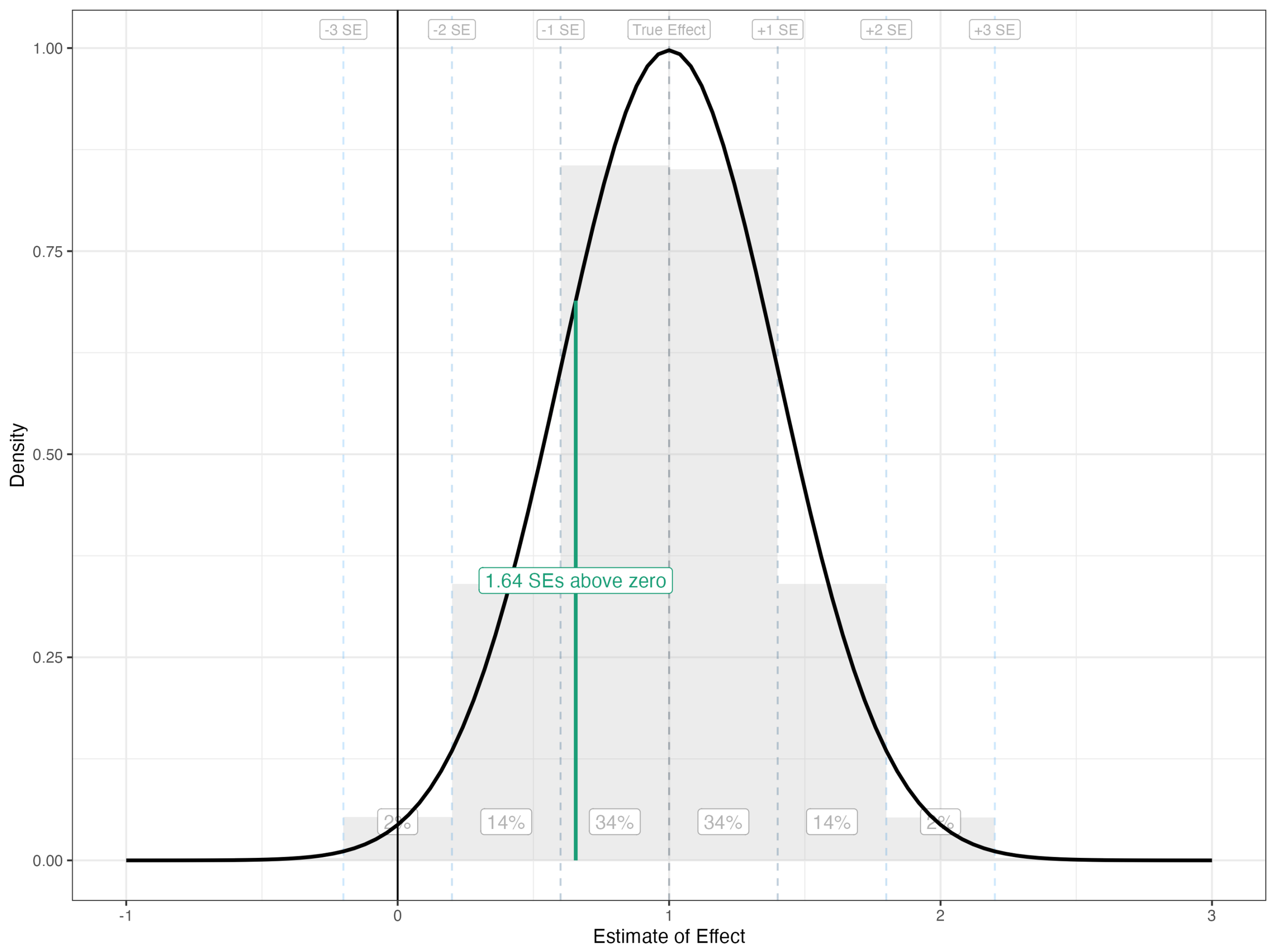

Study Number
50
40
30
20
10
0
1.64 SEs
0
1
2
Estimate of Effect

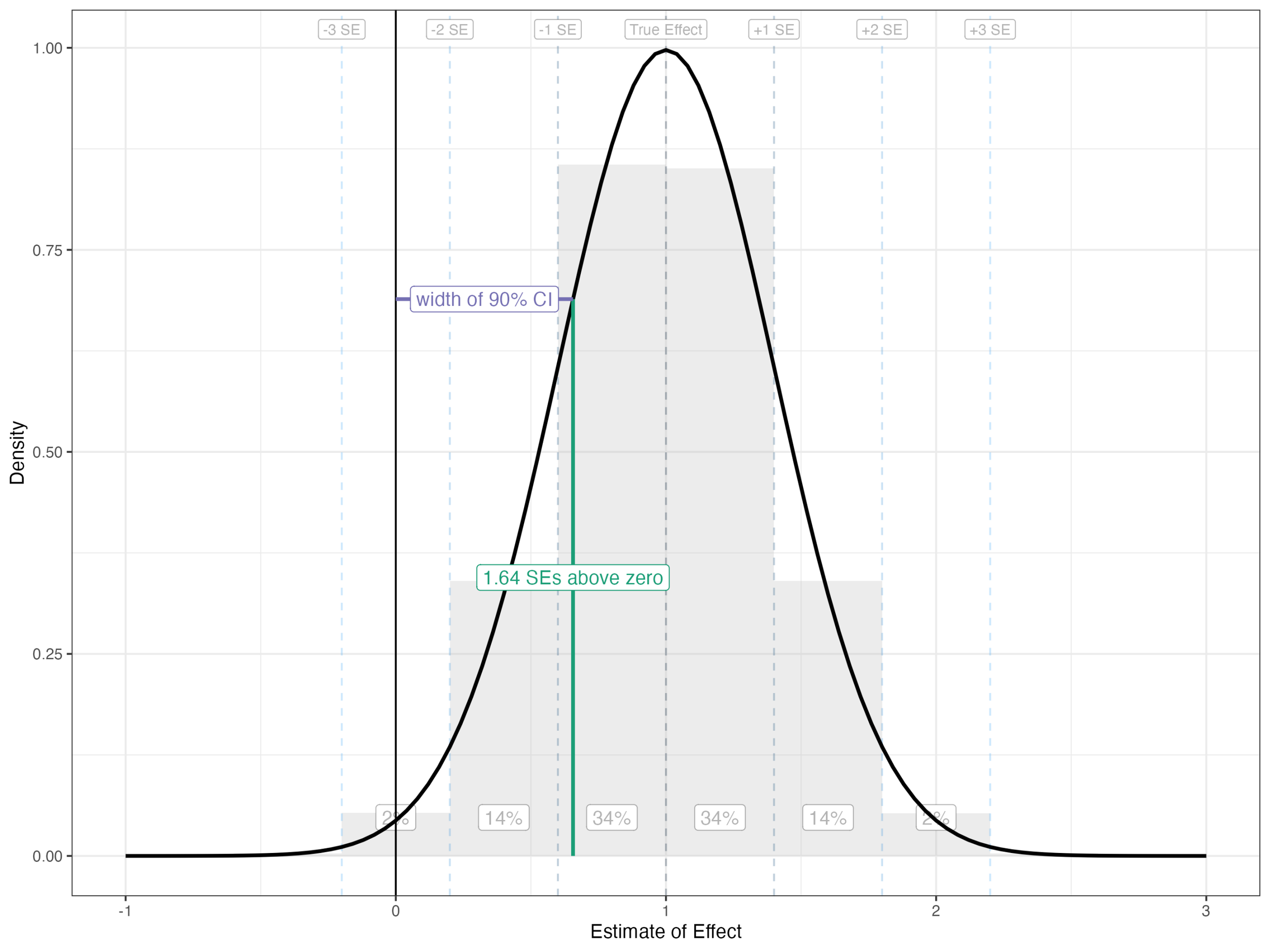

Count
30
20
10
0
No
Yes
Reject Null

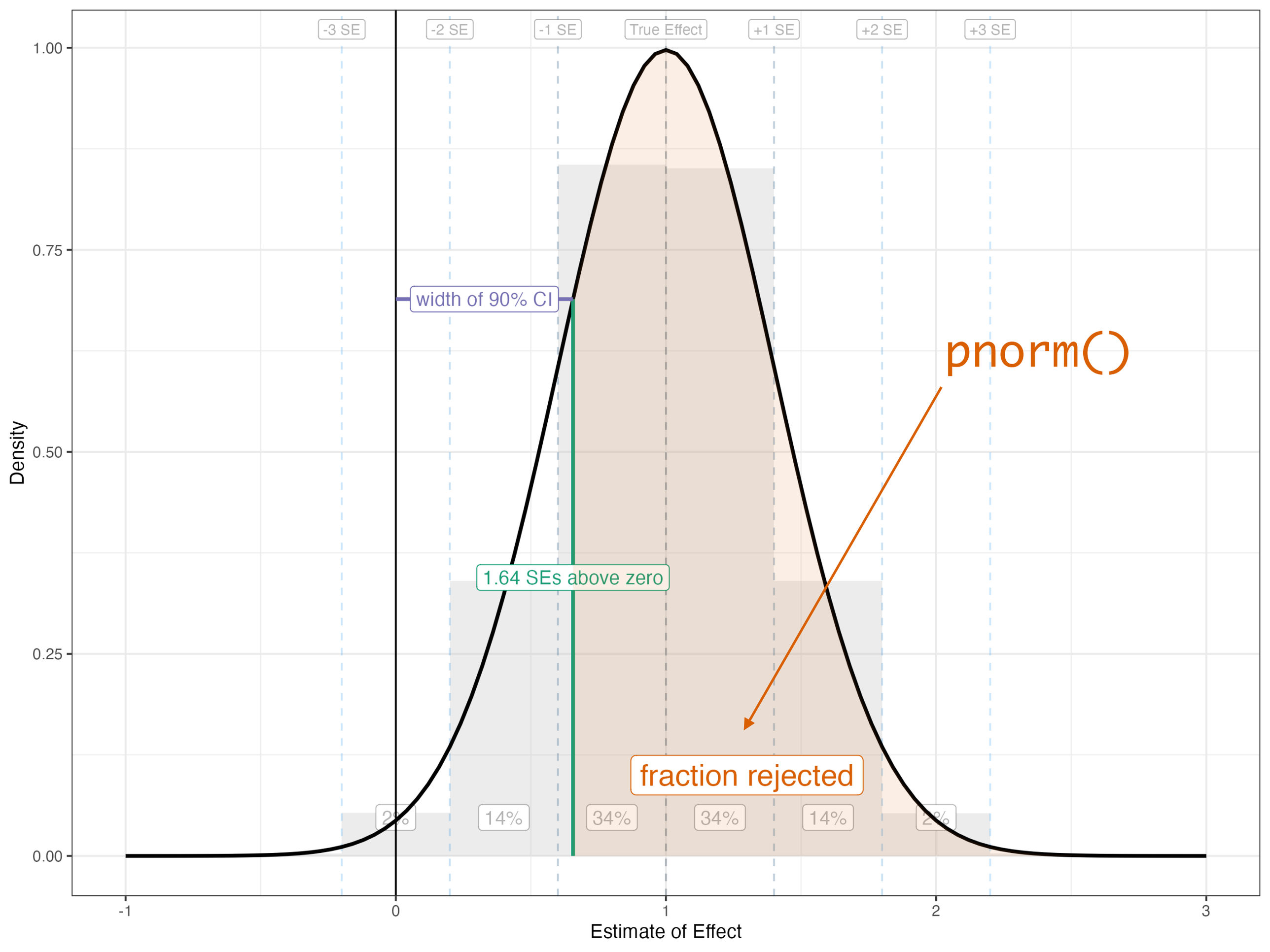
So what is power?

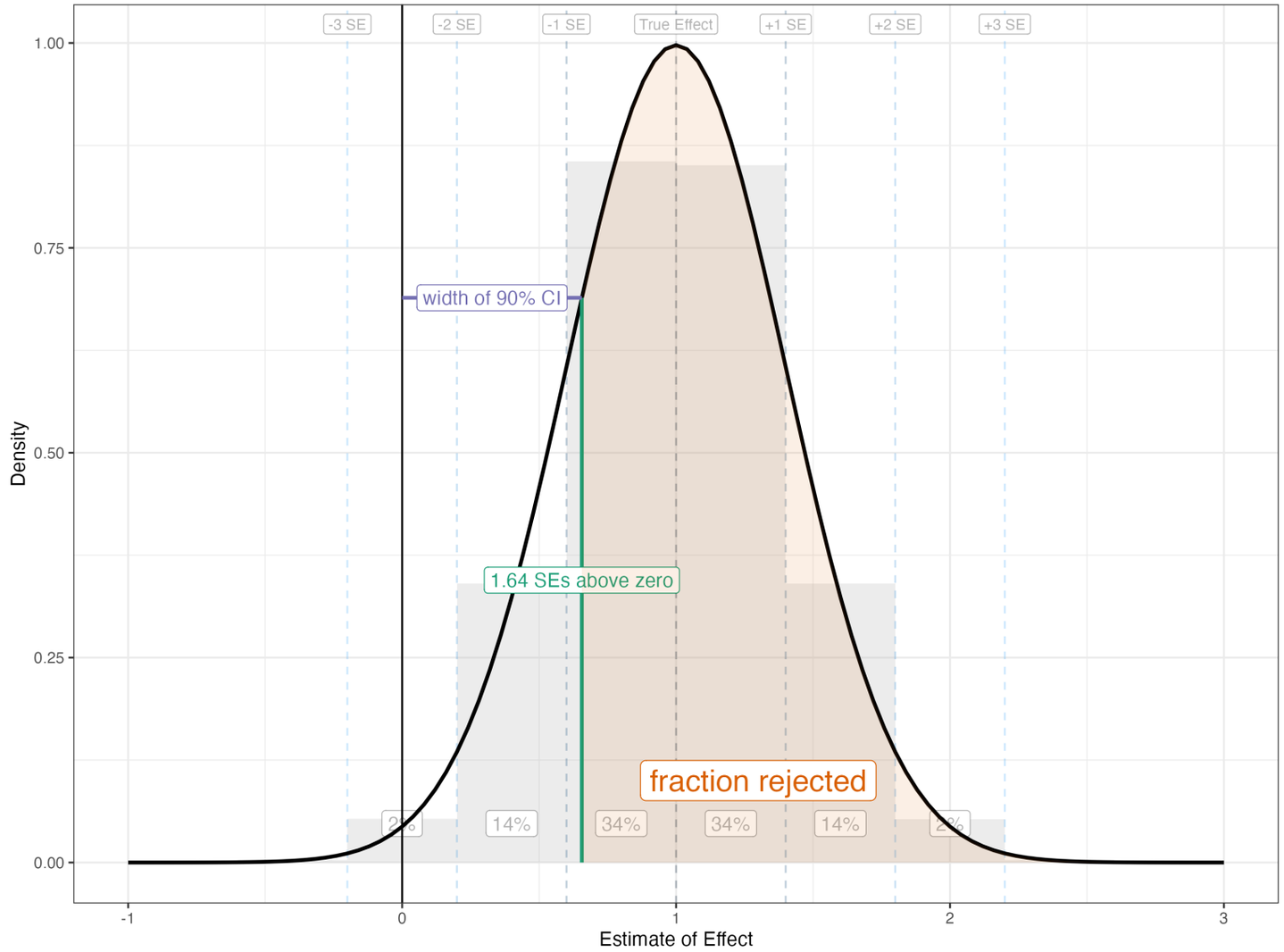

-3 SE
-2 SE
-1 SE
True Effect
+1 SE
+2 SE
+3 SE
1.00
0.75
0.50
0.25
0.00
Density
2%
14%
34%
34%
14%
2%
-1
0
1
2
3
Estimate of Effect


-3 SE
-2 SE
-1 SE
True Effect
+1 SE
+2 SE
+3 SE
1.00
0.75
0.50
0.25
0.00
Density
2%
14%
34%
34%
14%
2%
-1
0
1
2
3
Estimate of Effect


-3 SE
-2 SE
-1 SE
True Effect
+1 SE
+2 SE
+3 SE
1.00
0.75
0.50
0.25
0.00
Density
1.64 SEs above zero
2%
14%
34%
34%
14%
2%
-1
0
1
2
3
Estimate of Effect


-3 SE
-2 SE
-1 SE
True Effect
+1 SE
+2 SE
+3 SE
1.00
0.75
0.50
0.25
0.00
Density
width of 90% CI
1.64 SEs above zero
2%
14%
34%
34%
14%
2%
-1
0
1
2
3
Estimate of Effect


-3 SE
-2 SE
-1 SE
True Effect
+1 SE
+2 SE
+3 SE
1.00
0.75
0.50
0.25
0.00
Density
width of 90% CI
1.64 SEs above zero
pnorm()
fraction rejected
2%
14%
34%
34%
14%
2%
-1
0
1
2
3
Estimate of Effect

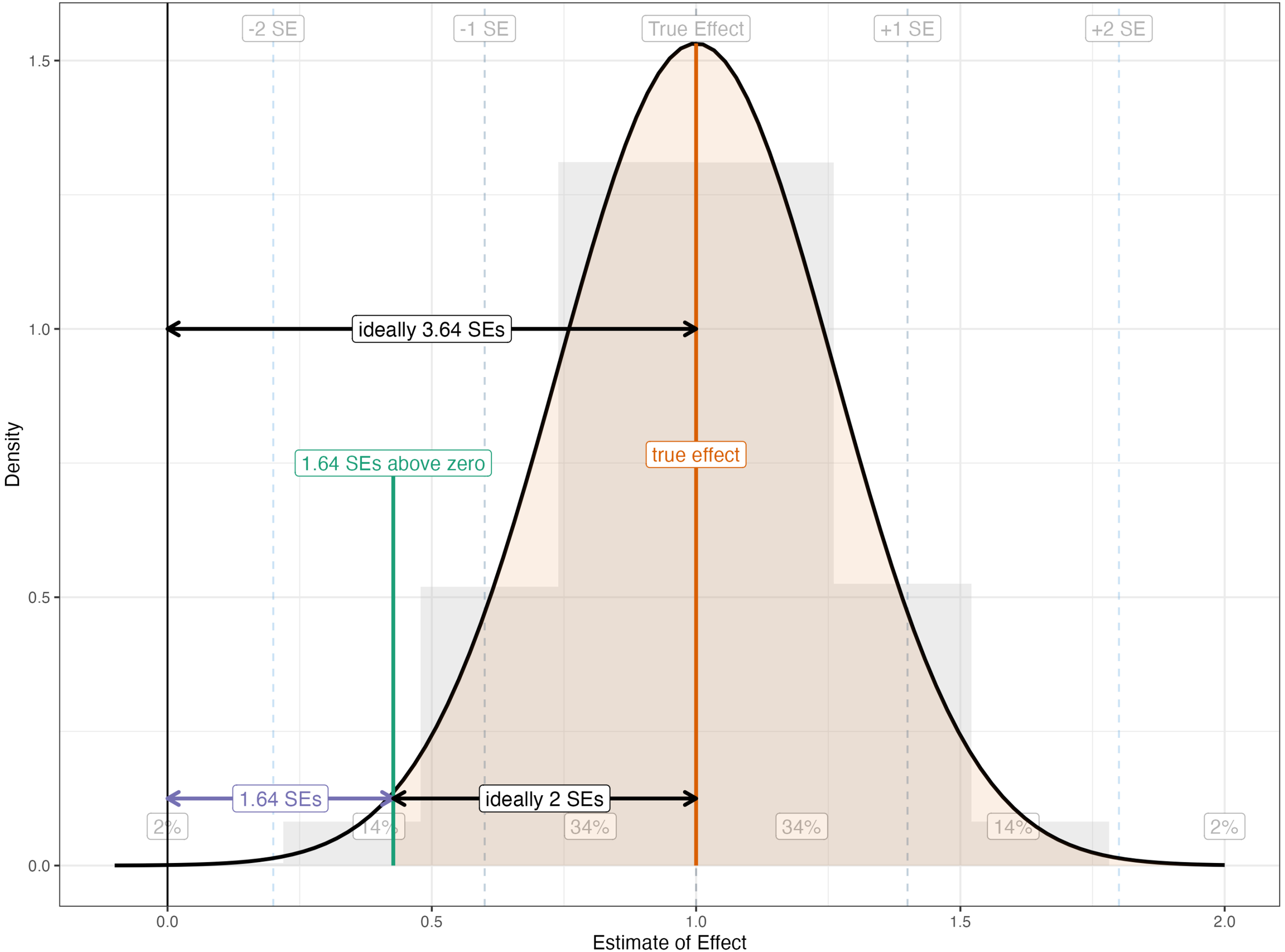



```
# mean and sd
true_effect <- 1.00
se <- 0.4

# compute power
pnorm(1.64*se,                  # want fraction above* 1.64 SE
      mean = true_effect,  # mean of sampling distribution
      sd = se,                  # sd of sampling distribution
      lower.tail = FALSE)  # fraction above, not below
```

Make sure that our result is informative.

Get the distribution above 1.64 SEs!


-2 SE
-1 SE
True Effect
+1 SE
+2 SE
1.5
1.0
0.5
0.0
Density
ideally 3.64 SEs
1.64 SEs above zero
true effect
1.64 SEs
ideally 2 SEs
2%
14%
34%
34%
14%
2%
0.0
0.5
1.0
1.5
2.0
Estimate of Effect

Rule of 3.64

ATE needs to be at least 3.64x larger than SE.

SE needs to be at most 27% of ATE.

2.5x → 80% power
3.3x → 95% power
4.0x → 99% power

Point #2

Minimize your risk!

SE should be 25% to 40%
of the ATE.

Goal #3

Discuss a few handy rules

Something you don’t want to hear

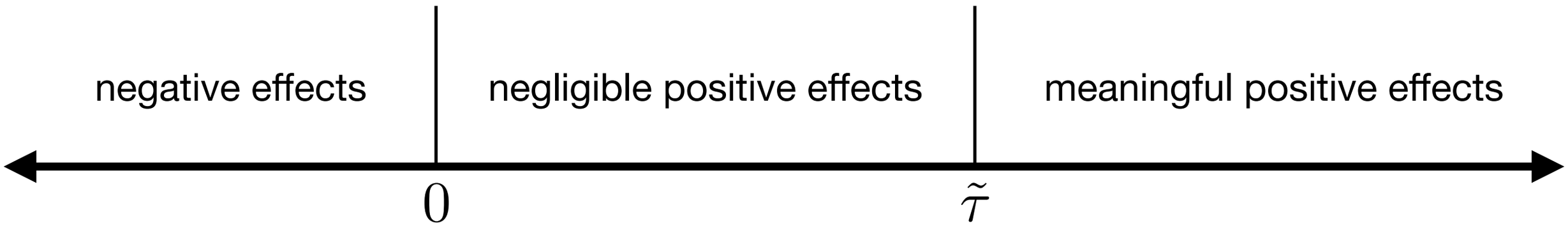
You must choose the smallest effect of substantive interest.

“smallest substantively meaningful effect”

“smallest effect size of interest” (SESOI)

“minimum effect of interest” (MEI)

“minimum meaningful effect” (MME)


negative effects
negligible positive effects
meaningful positive effects
0
$\tilde{\tau}$

	Features of a Population	Existing Study	Pilot Data
Minimum Detectable Effect			
Power			
Required Sample Size			

just algebra

(it's worthwhile to read through the paper and replicate the rules and examples with a paper and pencil)

	Features of a Population	Existing Study	Pilot Data
Minimum Detectable Effect			
Power			
Required Sample Size			

An experiment has

- 80% power to detect a treatment effect that is 2.5 times the standard error and
- 95% power to detect a treatment effect that is 3.3 times the standard error.

$$\text{The SE will be about } \frac{2 \cdot \widetilde{SD(Y)}}{\sqrt{2 \cdot n}}.$$

	Features of a Population	Existing Study	Pilot Data
Minimum Detectable Effect			
Power			
Required Sample Size			

An experiment has

- 80% power to detect a treatment effect that is 2.5 times the standard error and
- 95% power to detect a treatment effect that is 3.3 times the standard error.

$$\text{The SE will be about } \sqrt{\frac{n^{existing}}{n^{planned}}} \cdot SE_{\widehat{\tau}}^{existing}.$$

	Features of a Population	Existing Study	Pilot Data
Minimum Detectable Effect			
Power			
Required Sample Size			

An experiment has

- 80% power to detect a treatment effect that is 2.5 times the standard error and
- 95% power to detect a treatment effect that is 3.3 times the standard error.

Conservatively, the SE will be about $\sqrt{\frac{n^{pilot}}{n^{planned}}}\left[\left(\sqrt{\frac{1}{n^{pilot}}}+1\right)\cdot\widehat{SE}_{\widehat{\tau}}^{pilot}\right]$.

	Features of a Population	Existing Study	Pilot Data
Minimum Detectable Effect			
Power			
Required Sample Size			

Conservatively, the SE will be about $\sqrt{\frac{n^{pilot}}{n^{planned}}}\left[\left(\sqrt{\frac{1}{n^{pilot}}}+1\right)\cdot\widehat{SE}_{\widehat{\tau}}^{pilot}\right]$.

$$\text{Power equals } 1-\Phi(1.64\cdot SE;\mu=\tau,\sigma=SE).$$

$$\text{Alternatively, power equals } 1-\Phi_{std}\left(1.64-\frac{\tau}{SE}\right).$$

	Features of a Population	Existing Study	Pilot Data
Minimum Detectable Effect			
Power			
Required Sample Size			

For 80% power, the sample size per condition will need to be (conservatively) about $n^{pilot} \cdot \left[\frac{2.5}{\widetilde{\tau}} \cdot \left(\sqrt{\frac{1}{n^{pilot}}} + 1\right) \cdot \widehat{SE}_{\widehat{\tau}}^{pilot}\right]^2$.

Point #3

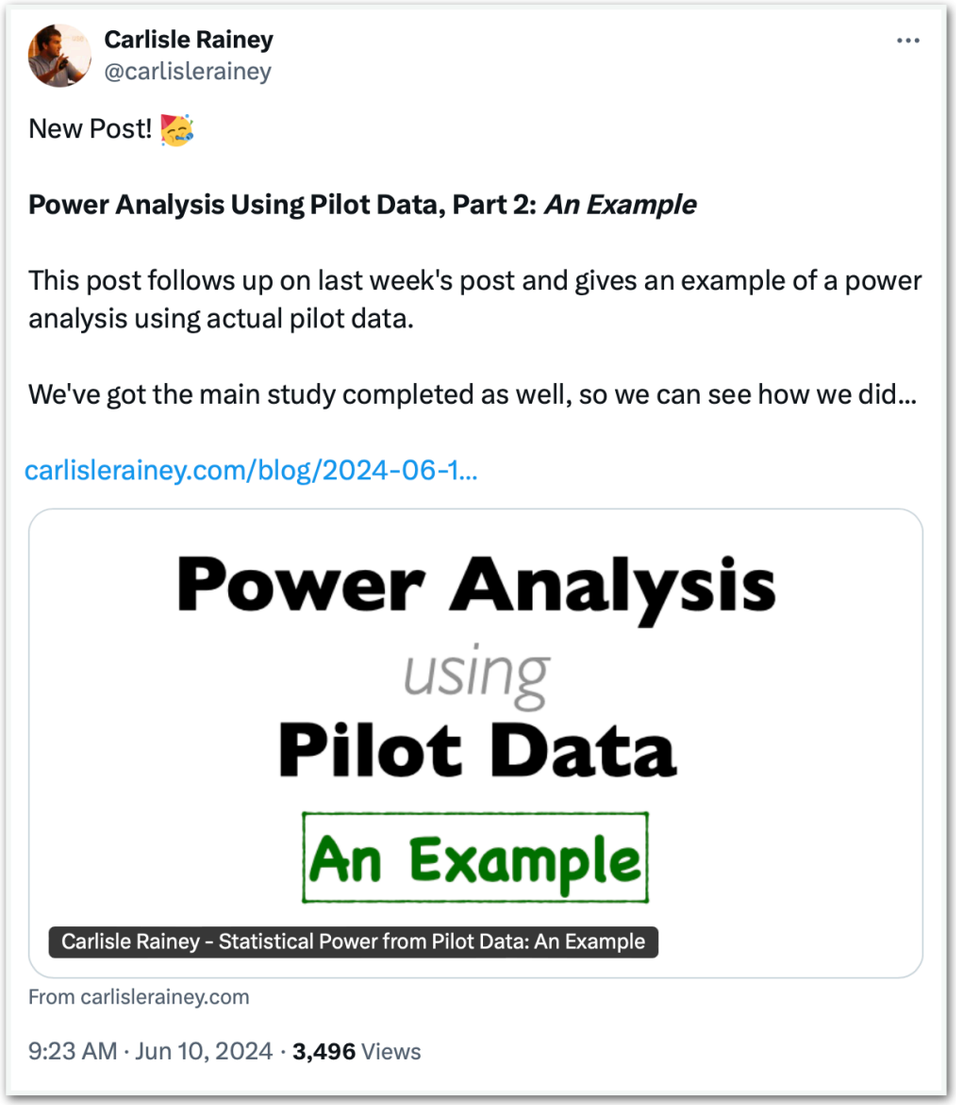
There are easy rules!

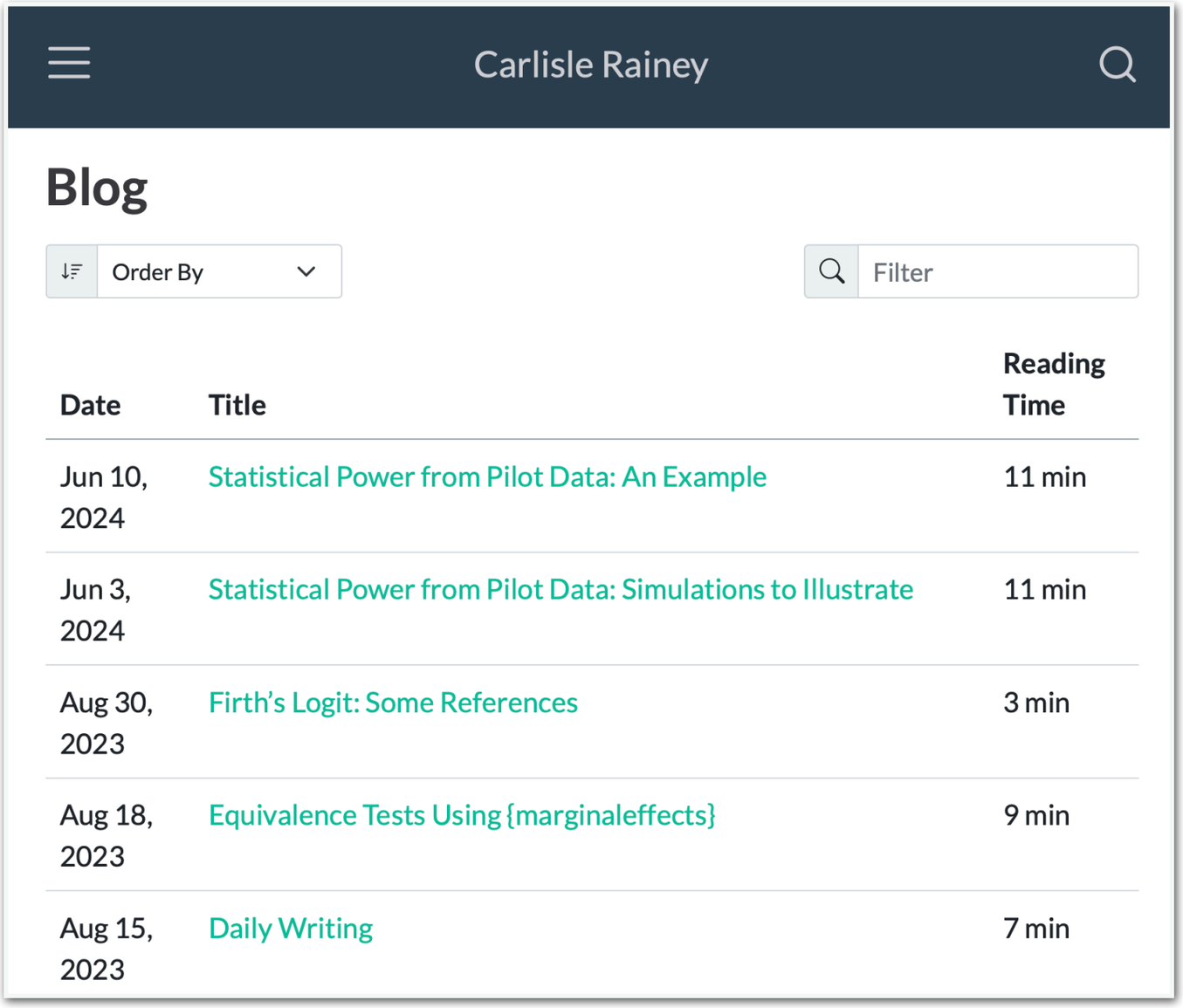
Table 1: This table shows how we can use the rules to answer questions about the experiment using features of a reference population, existing studies, and pilot data.

Goal	Source	Estimation Strategy	Required Information	Rule Sequence
I know my sample size and the statistical power I want. What is my minimum detectable effect?	features of reference population	no control variables	$SD(Y)$ in reference population	Rule 3 → Rule 5
		control variables	$SD(Y)$ and R^2 of control variables in reference population	Rule 3 → Rule 4 → Rule 5
	existing similar study	—	estimated standard error from existing study	Rule 7 → Rule 5
	pilot data	—	estimated standard error from pilot data	Rule 9 → Rule 5
I know my sample size and the effect of interest. What is my statistical power?	features of reference population	no control variables	$SD(Y)$ in reference population	Rule 3 → Rule 2
		control variables	$SD(Y)$ and R^2 of control variables in reference population	Rule 3 → Rule 4 → Rule 2
	existing similar study	—	estimated standard error from existing study	Rule 7 → Rule 2
	pilot data	—	estimated standard error from pilot data	Rule 9 → Rule 2
I know the statistical power I want and the effect of interest. What sample size do I need?	features of reference population	no control variables	$SD(Y)$ in reference population	Rule 6
		control variables	$SD(Y)$ and R^2 of control variables in reference population	Rule 6
	existing similar study	—	estimated standard error from existing study	Rule 8
	pilot data	—	estimated standard error from pilot data	Rule 10

Goal #4

An Example





Date	Title	Reading Time
Jun 10, 2024	Statistical Power from Pilot Data: An Example	11 min
Jun 3, 2024	Statistical Power from Pilot Data: Simulations to Illustrate	11 min
Aug 30, 2023	Firth's Logit: Some References	3 min
Aug 18, 2023	Equivalence Tests Using {marginaleffects}	9 min
Aug 15, 2023	Daily Writing	7 min

```r
# load pilot data and keep only success condition
robbins2_pilot <- crdata::robbins_pilot |>
  subset(failure == "Success") |>
  glimpse()
```

```
Rows: 147
Columns: 5
$ cong_overall <dbl> 3, 1, -2, 0, -2, -1, 0, -1, 0, 0, 0, 2, -1, 3, -3, 0, 0, …
$ failure      <fct> Success, Success, Success, Success, Success, Success, Suc…
$ amplify      <fct> Ignore, Ignore, Ignore, Amplify, Ignore, Ignore, Ignore, …
$ pid7         <fct> Strong Democrat, Not very strong Republican, Strong Democ…
$ pid_strength <dbl> 3, 2, 3, 2, 2, 3, 2, 2, 2, 2, 2, 1, 0, 3, 3, 3, 3, 1, 1, …
```

```r
# t test
fit_pilot <- t.test(cong_overall ~ amplify, data = robbins2_pilot)
```

```r
# create a table showing the observations per condition
table(robbins2_pilot$amplify)
```

```
Ignore Amplify
    70      77
```

```r
# sample size per condition
n_pilot <- mean(table(robbins2_pilot$amplify))
n_pilot
```

```
[1] 73.5
```

```r
# get estimated standard error from pilot
se_hat_pilot <- fit_pilot$stderr
se_hat_pilot
```

```
[1] 0.2761011
```

```r
n_planned <- 250
```

```r
pred_se_cons <- sqrt(n_pilot/n_planned)*((sqrt(1/n_pilot) + 1)*se_hat_pilot)
pred_se_cons
```

```
[1] 0.1671691
```

```r
# compute conservative minimum detectable effect
2.5*pred_se_cons
```

```
[1] 0.4179227
```

```r
# compute power as a percent
1 - pnorm(1.64 - 0.35/pred_se_cons)
```

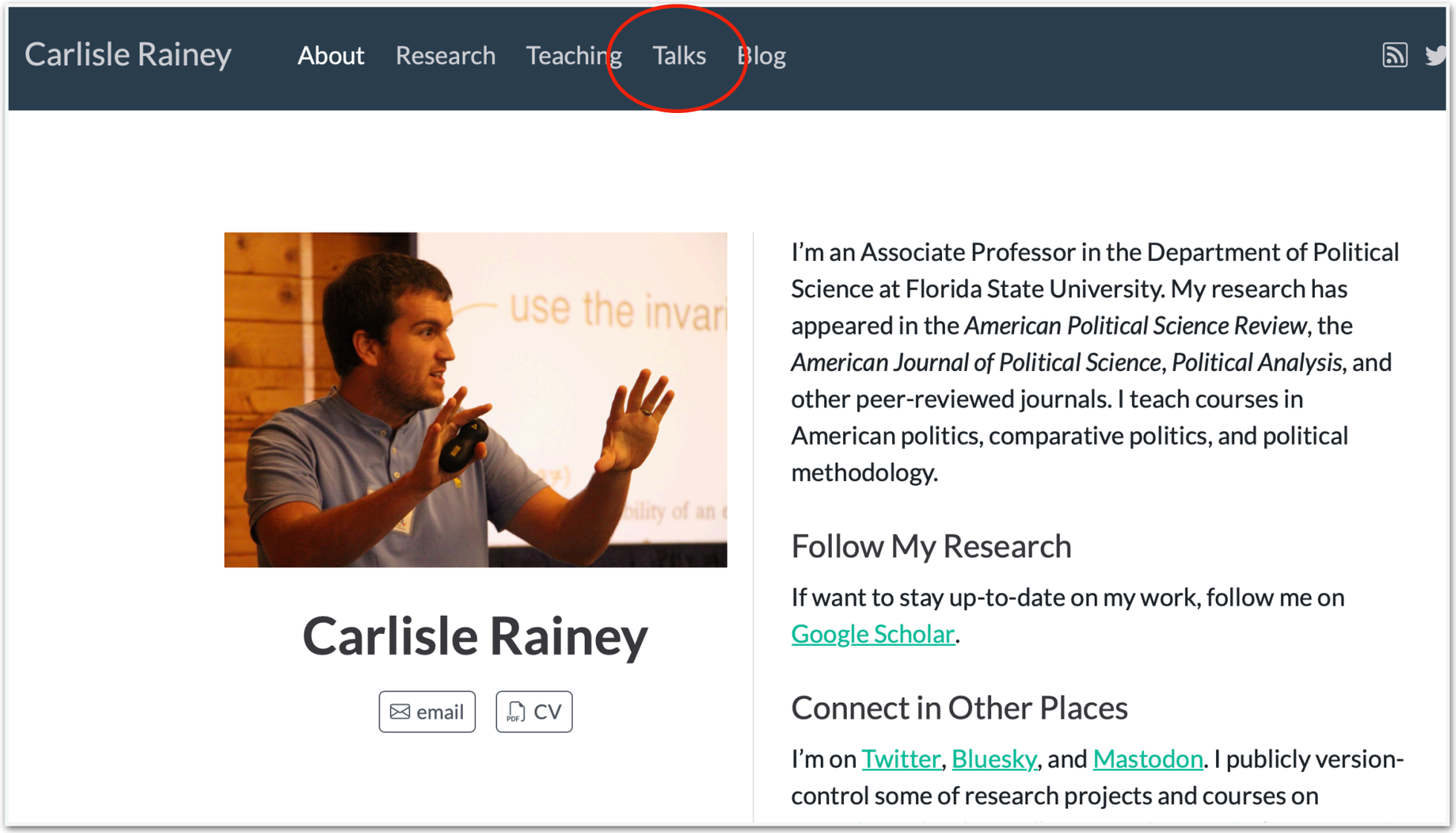
```
[1] 0.6749736
```

```r
n_pilot*((2.5/0.35)*((sqrt(1/n_pilot) + 1)*se_hat_pilot))^2
```

```
[1] 356.4477
```

Resources

http://www.carlislerainey.com/talks/



Wrap

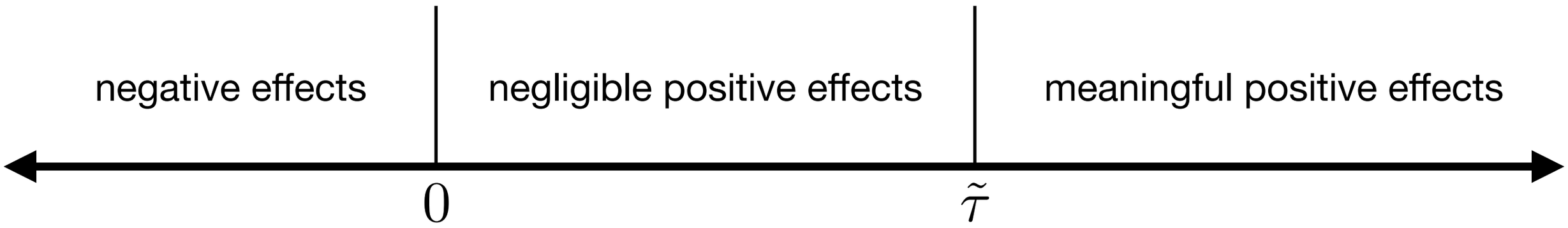
Experiments are noisy—you can easily waist your time.

You need your standard errors to be 25% to 40% of the ATE.

The paper provides some helpful rules, especially about pilot data.

Addendum

A Brief Argument for 95% Power


negative effects
negligible positive effects
meaningful positive effects
0
$\tilde{\tau}$

Another Addendum

A Question for the Disciple to Wrestle with

Carlisle Rainey
@carlislerainey

An author conducts an experiment and obtains a 95% CI that is inconsistent with a widely believed fact.

Reviewer 2 argues for rejection because the experiment did not have sufficient statistical power.

Should the editor consider statistical power when making the decision?

Should consider power	**66.8%**
Should NOT consider power	33.2%

208 votes · Final results

5:31 AM · May 21, 2024 · **3,853** Views

Carlisle Rainey @carlislerainey · Apr 27

Excellent 🏛️ 🏛️. I agree.

If one interprets tests correctly, then you should never make a Type II "error". Instead, the study simply isn't as informative as we would wish.

Underpowered studies aren't "misleading" or "unsafe", the CI is just too wide for super-strong claims.

2 · 2 · 189

Ryan Briggs
@ryancbriggs

They are misleading and unsafe! (Because of significance filtering at the publication stage, which absolutely exists)

1:15 PM · Apr 27, 2024 from Toronto, Ontario · **160** Views