Advanced Quantitative Methods

Randomization inference & complier average causal effects

Instructor: Gregory Eady
Office: 18.2.10
Office hours: Fridays 13-15

Today

- Randomization inference
- Complier Average Causal Effects (CACE)
- Other considerations in experiments

Randomization can result in a massive number of possible assignments

Individual	Rand. 1	Rand. 2	Rand. 3	Rand. 4	Rand. 5	
1	1	0	0	1	1	$\cdots$
2	0	1	0	1	0	$\cdots$
3	1	1	0	1	0	$\cdots$
4	0	1	1	0	1	$\cdots$
5	1	0	0	0	0	$\cdots$
6	0	0	0	0	0	$\cdots$
7	1	0	1	1	1	$\cdots$
8	0	0	1	0	1	$\cdots$
9	0	1	1	0	1	$\cdots$
10	1	1	1	1	0	$\cdots$

Where does statistical uncertainty come from?

- Each possible randomization would lead to a different estimate of the average treatment effect (ATE)
  - Why? Because it is unlikely that effects are the same across all individuals
- Imagine that we could observe ATE estimates from all possible randomizations of a treatment
- This *distribution* of estimates represents our uncertainty about the sample ATE

How can we use this distribution to learn about our uncertainty about an effect estimate?

- We can assume that this distribution takes a certain shape as $N \to \infty$ (a normal distribution)
- We can also *simulate* potential experiments to see what would have happened if we had randomized differently
  - Randomization inference

Thinking about uncertainty

- Standard errors indicate the range of estimates assuming some hypothesis is true
- e.g. assume that the null hypothesis, H_0 is true
  - i.e. That our treatment has no effect on average
- Imagine we assign that wholly ineffective treatment to half of our sample at random; the other half is in the control condition
- We then measure some outcome
- Our *estimate* of the effect won't be exactly zero, even if the true effect is zero
- But if we ran our experiment infinite times, the distribution of those effect estimates would be normally distributed

If our sample size is large enough, then the normal distribution applies

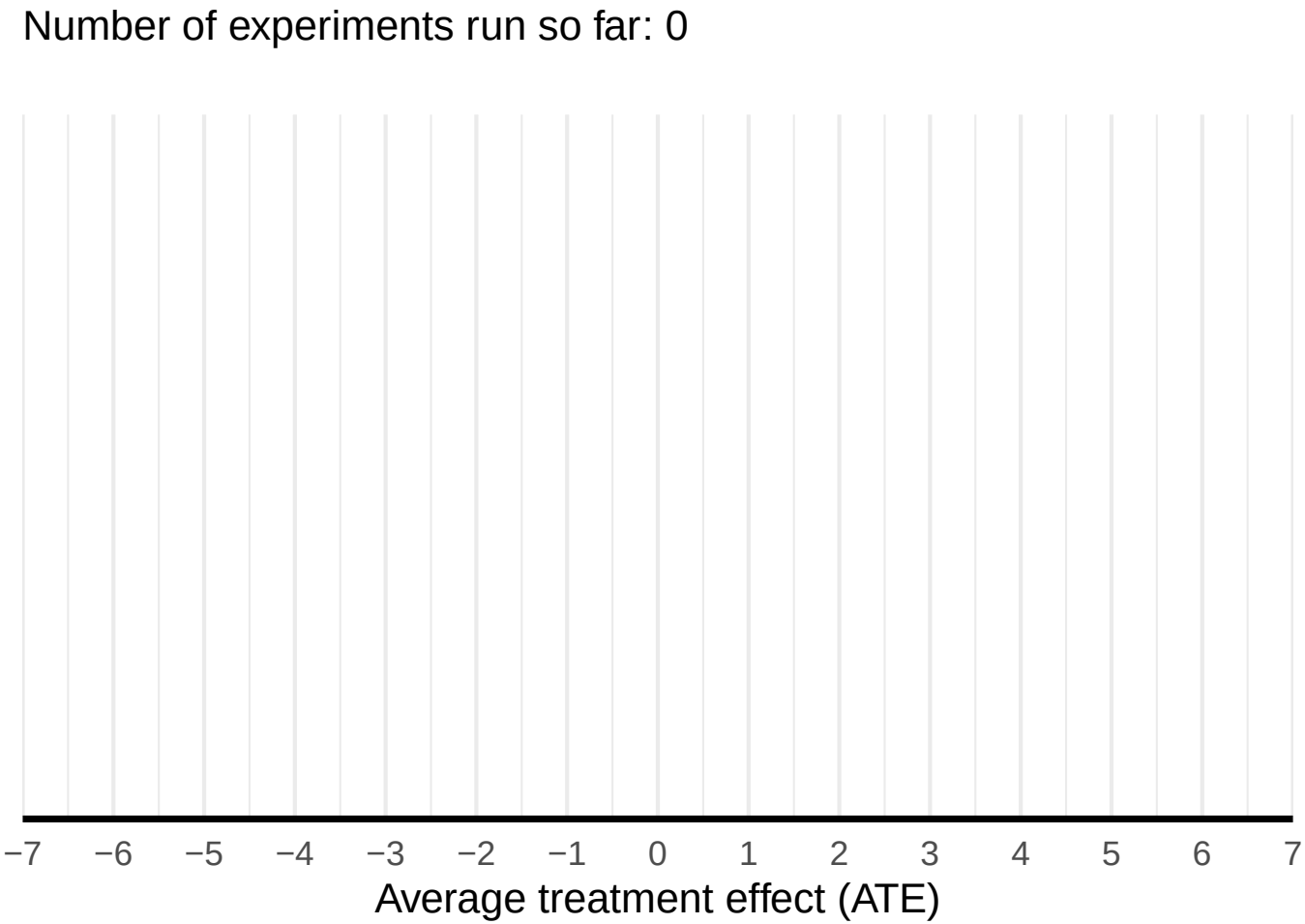
- ± 1.645 standard deviations $= 90\%$
- ± 1.96 standard deviations $= 95\%$

The standard error is a standard deviation... It's the standard deviation of the sampling distribution of a statistic of interest...

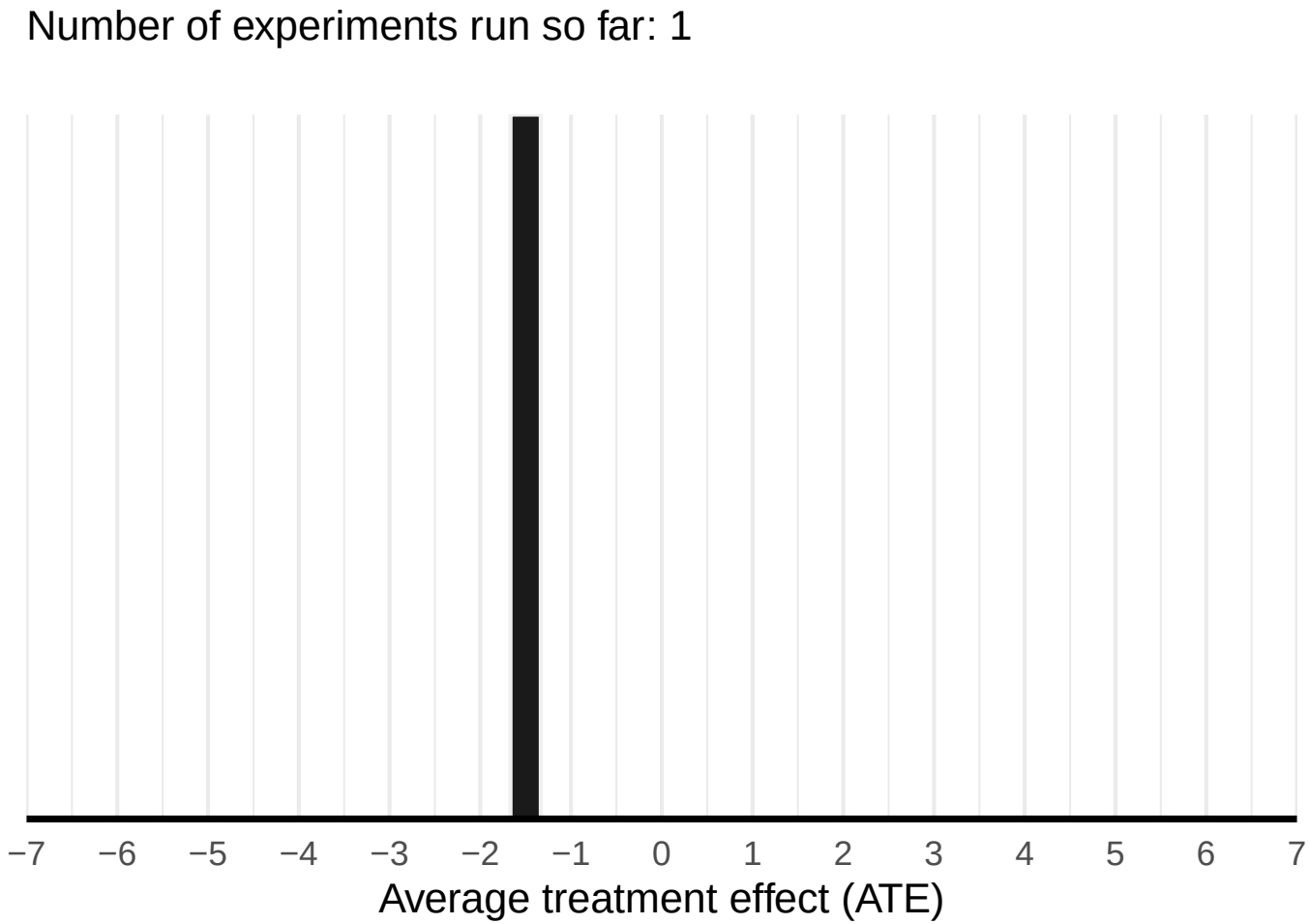
Example

- Say we want to know the effect of watching the final debate between Kamala Harris and Donald Trump on voters' approval of Donald Trump
- We measure approval on 0 to 100 scale
- We take 1,000 voters, and assign 500 to watch the debate and 500 to do something else
- We measure the effect of watching the debate by taking the difference in the average approval of Trump among those who we assigned to watch the debate and those who did not

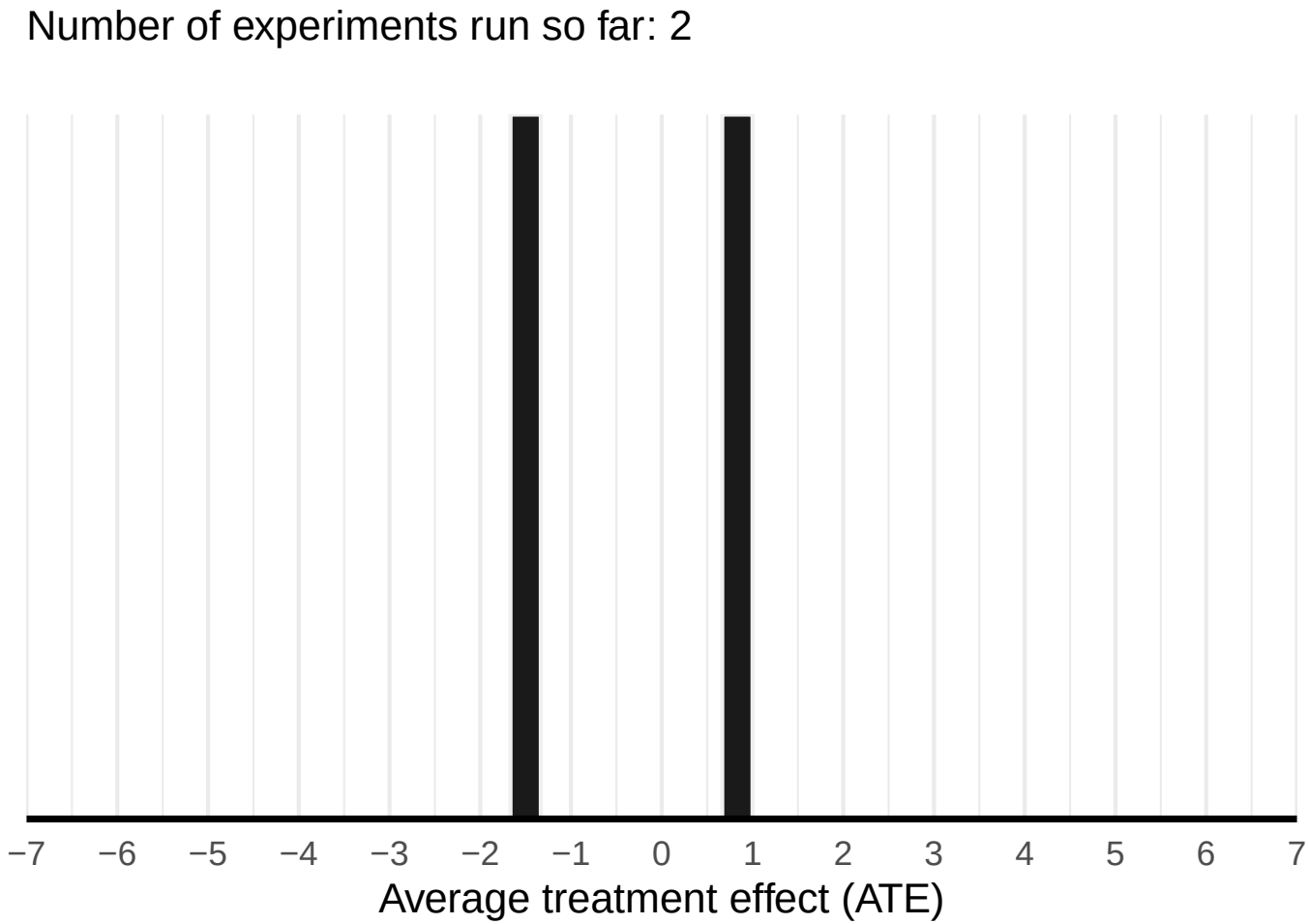
Let's say watching the debate truly has no effect. What might we see?



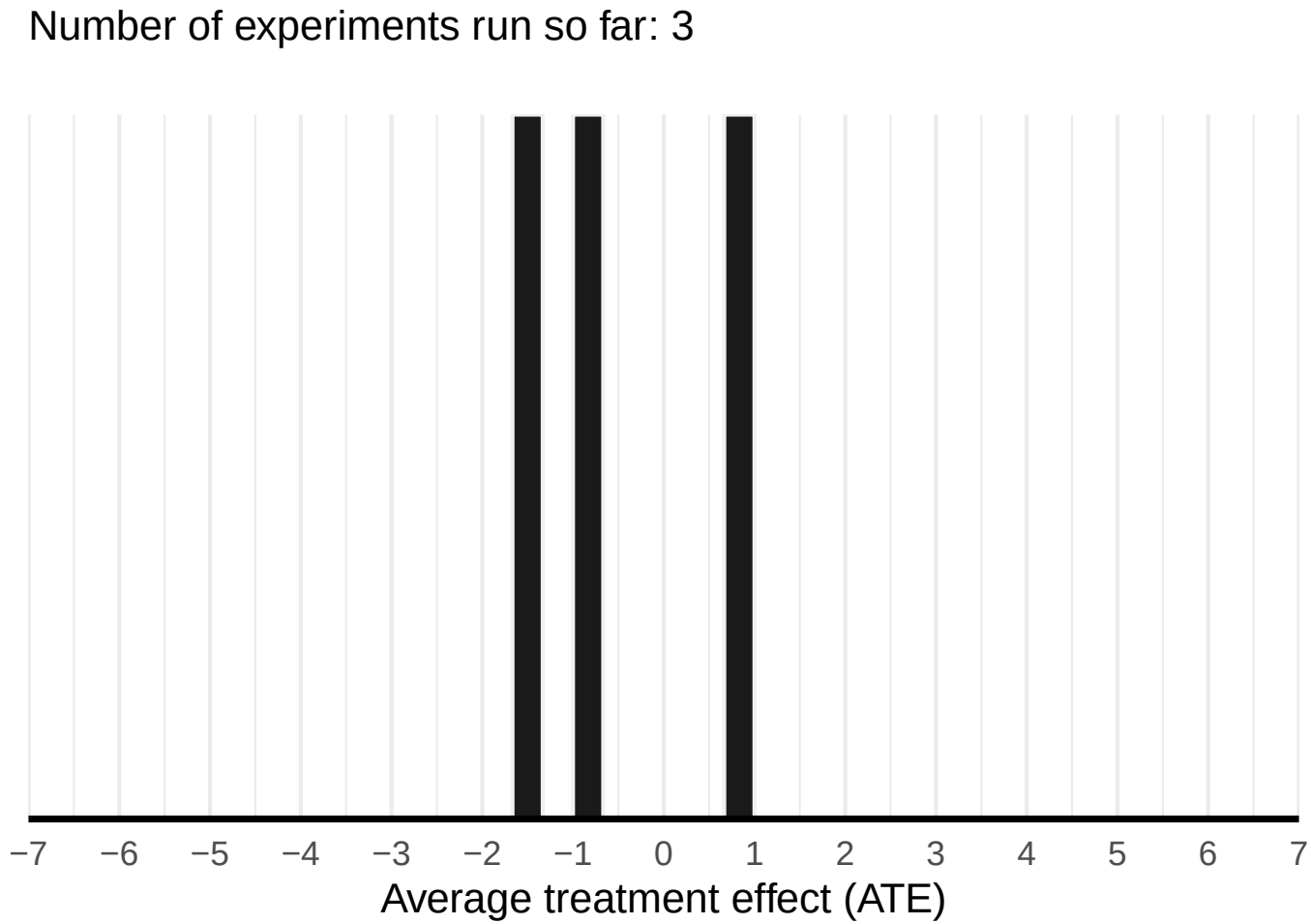
Let's say watching the debate truly has no effect. What might we see?



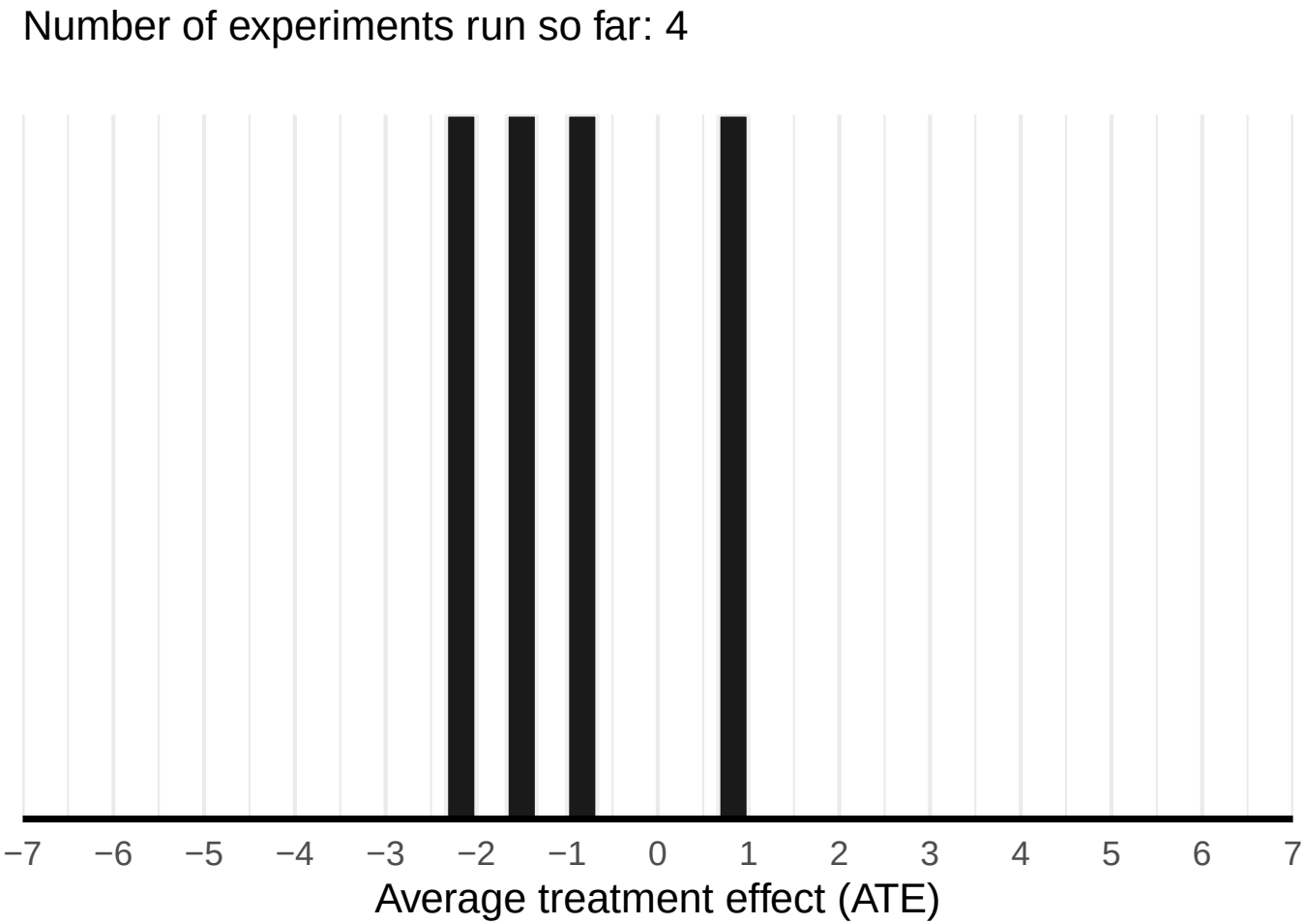
Let's say watching the debate truly has no effect. What might we see?



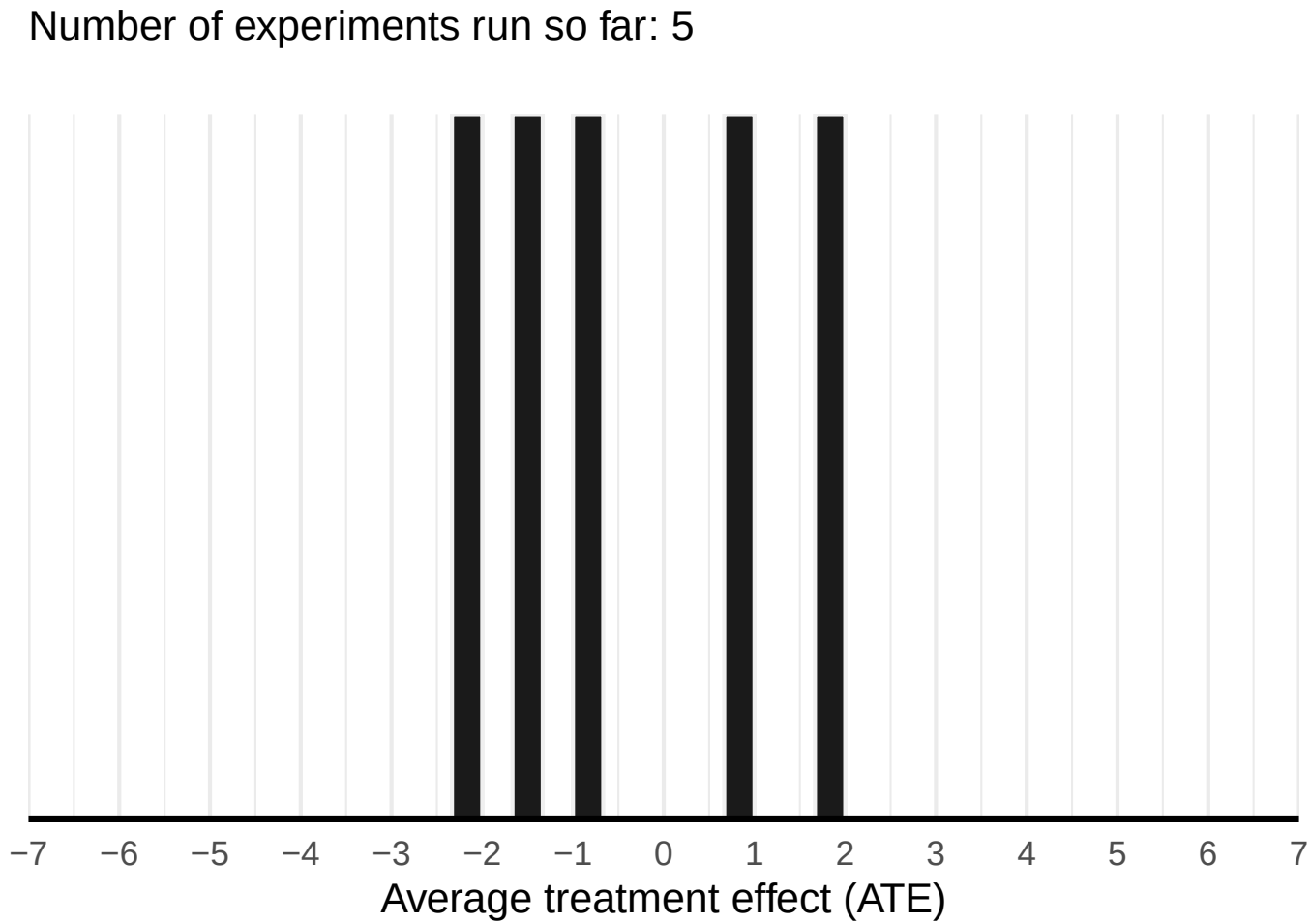
Let's say watching the debate truly has no effect. What might we see?



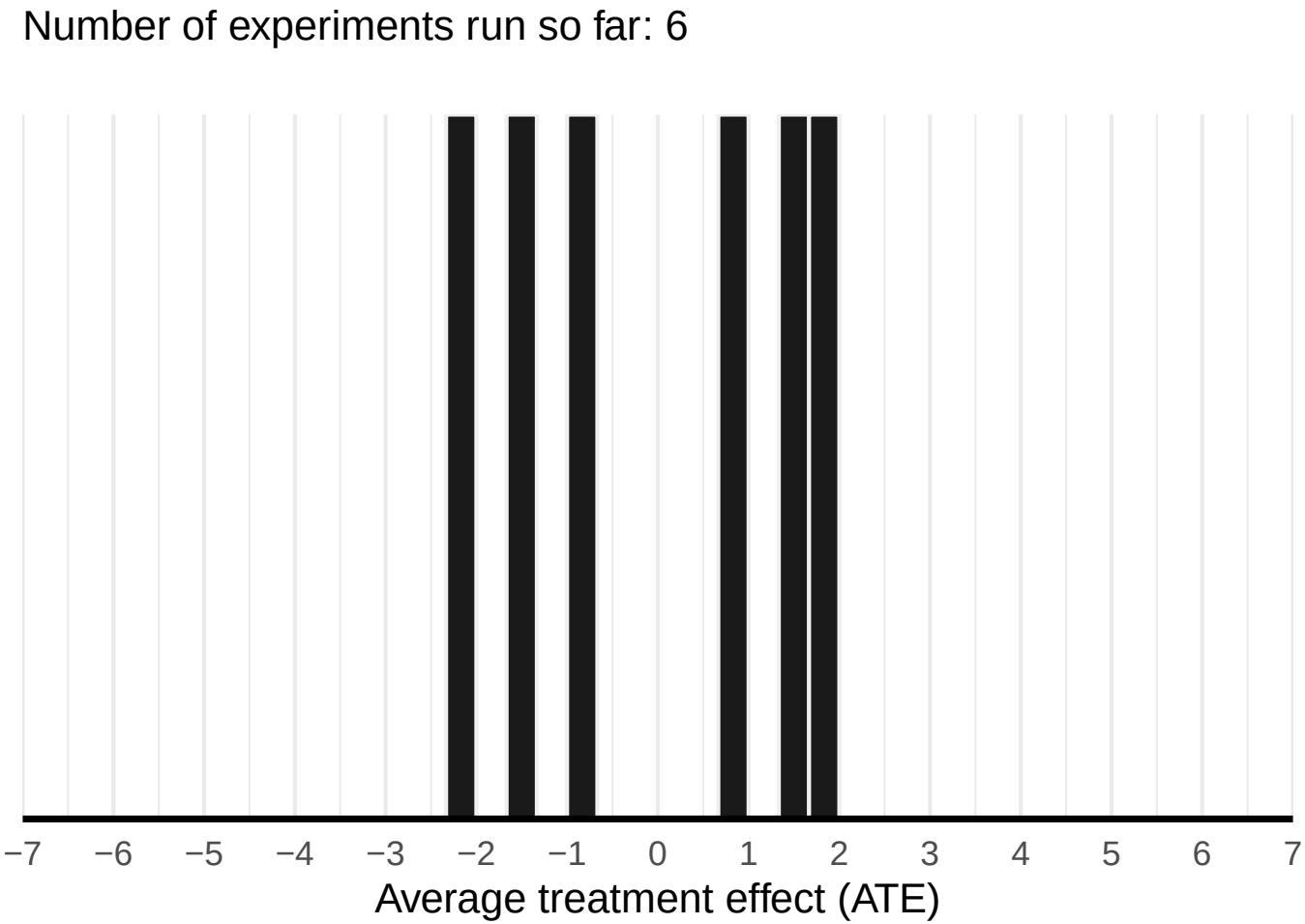
Let's say watching the debate truly has no effect. What might we see?



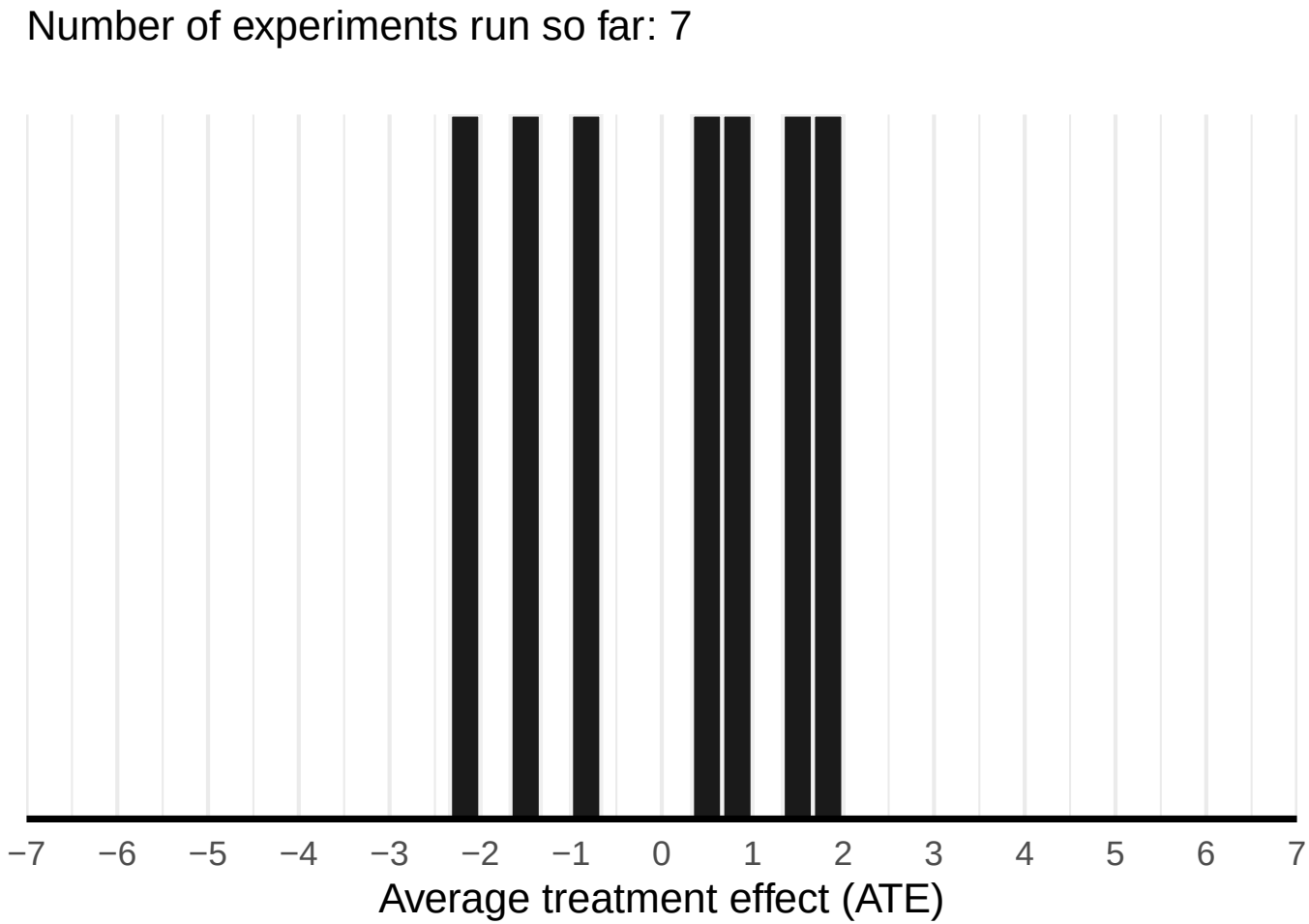
Let's say watching the debate truly has no effect. What might we see?



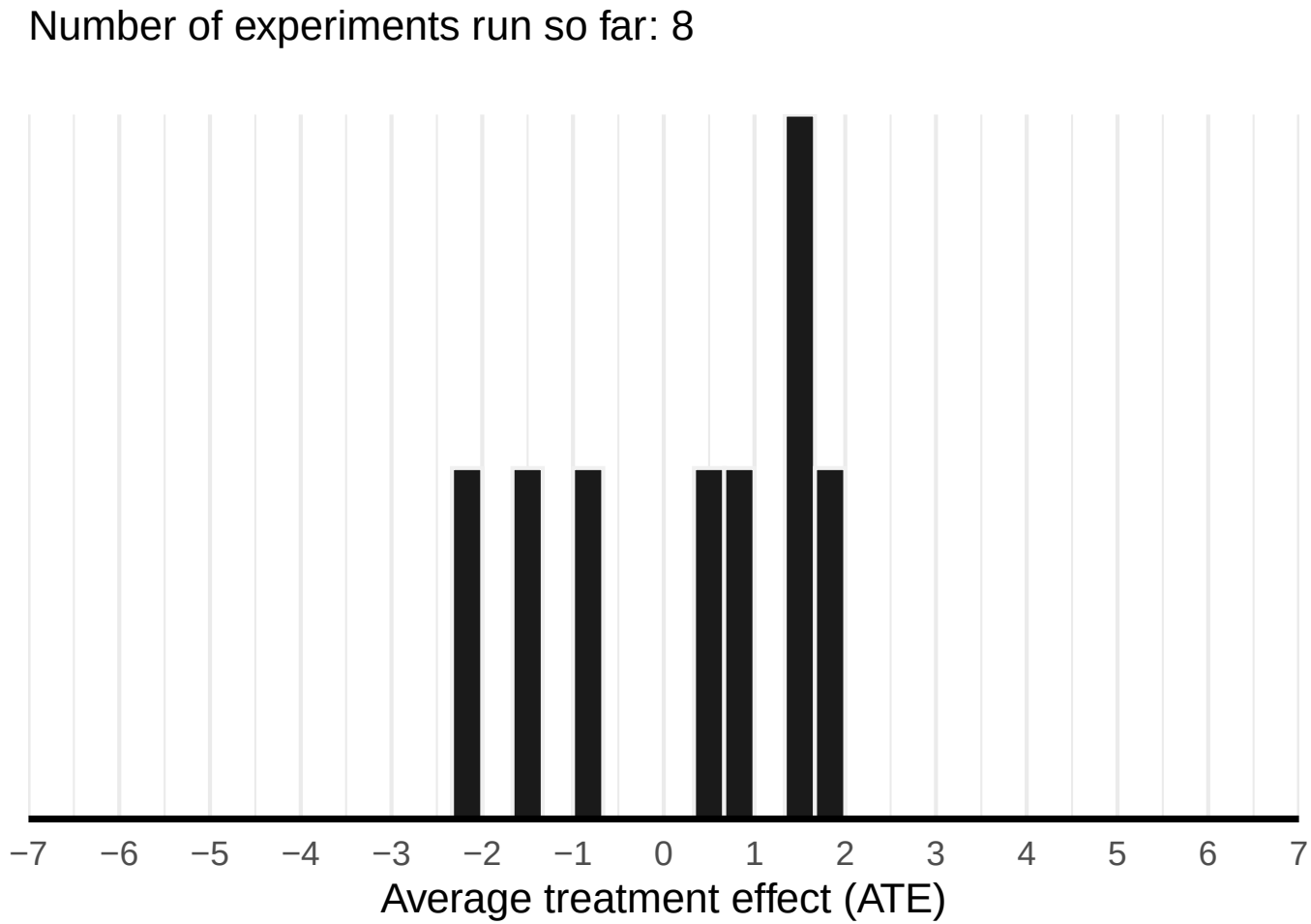
Let's say watching the debate truly has no effect. What might we see?



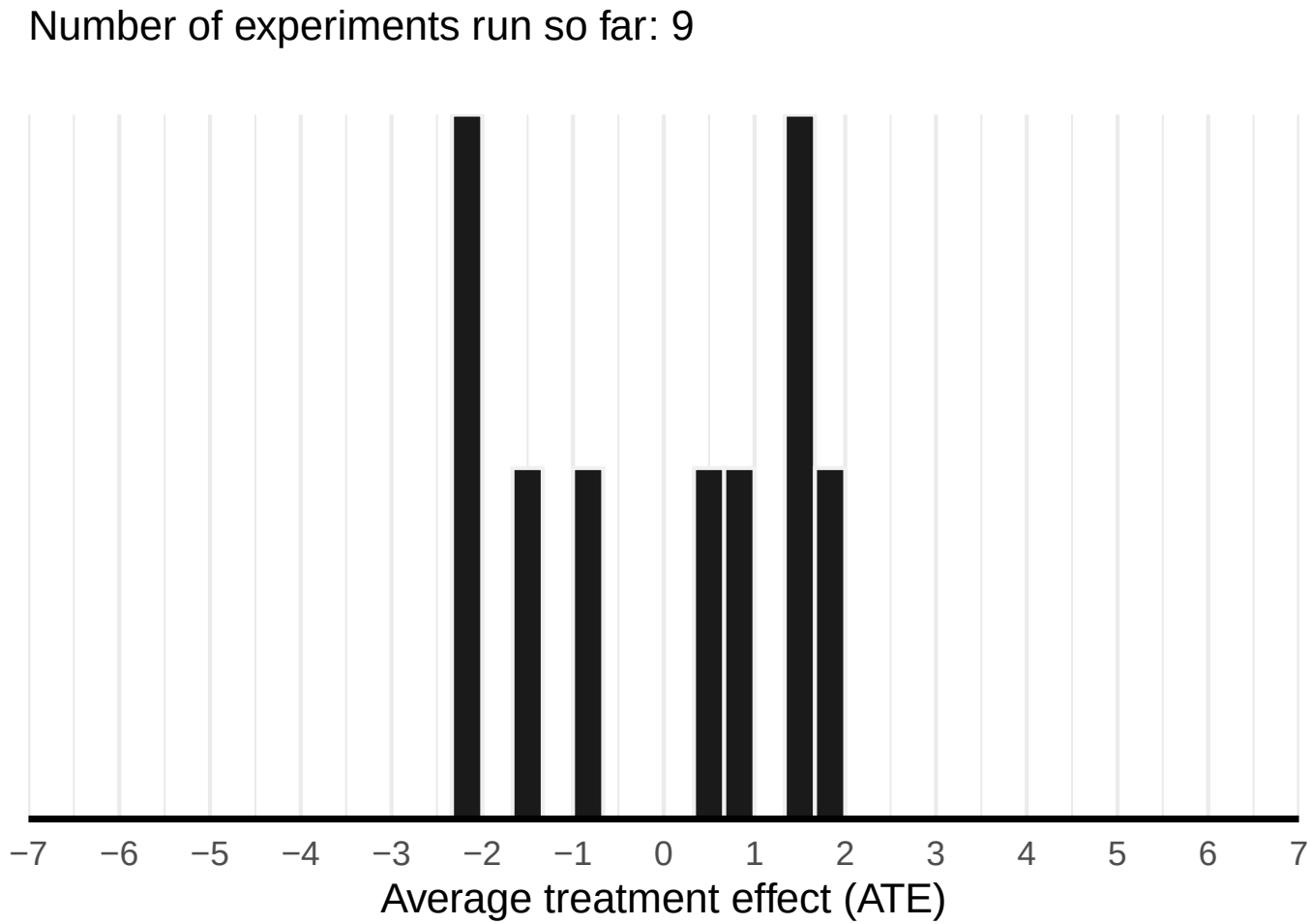
Let's say watching the debate truly has no effect. What might we see?



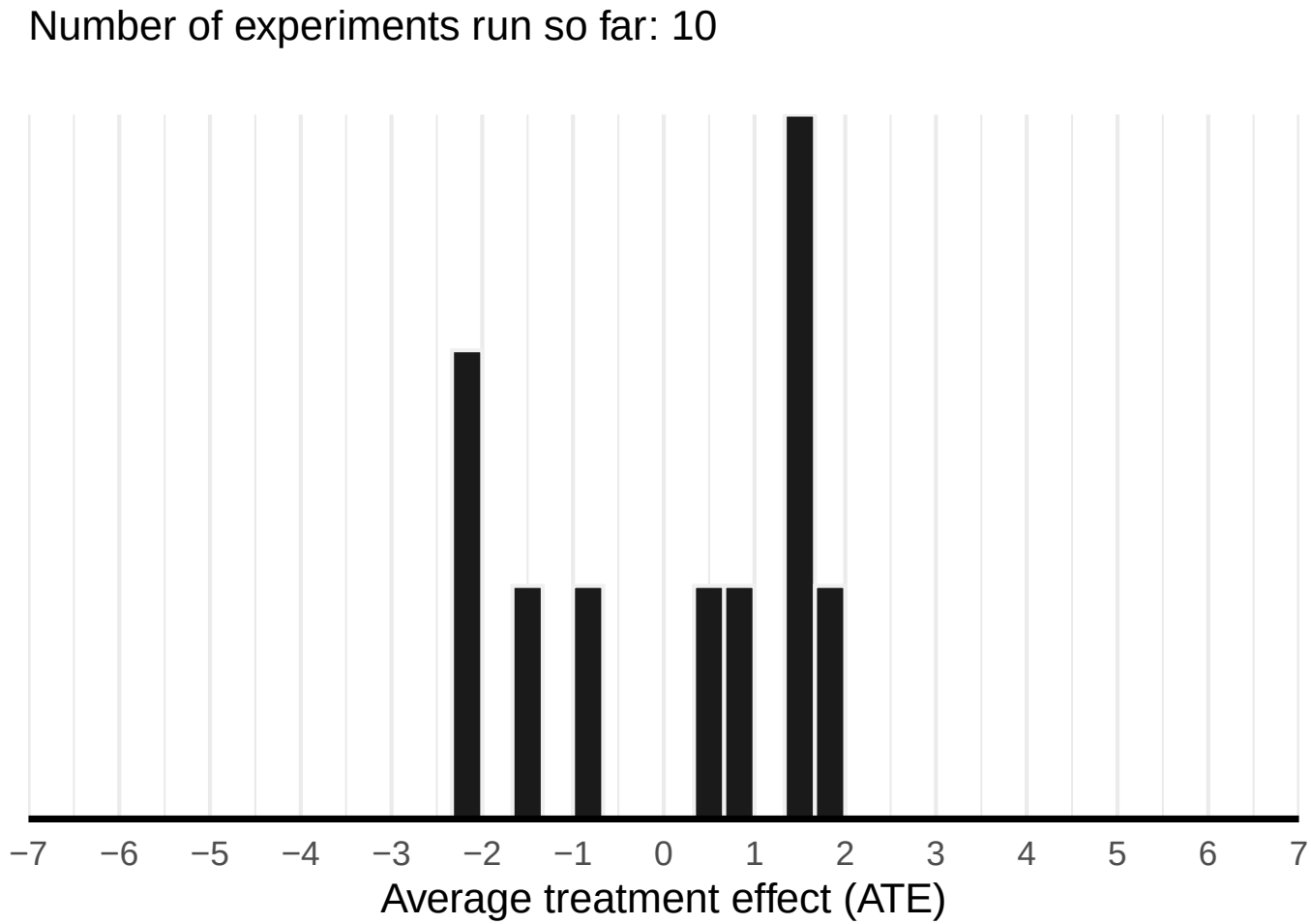
Let's say watching the debate truly has no effect. What might we see?



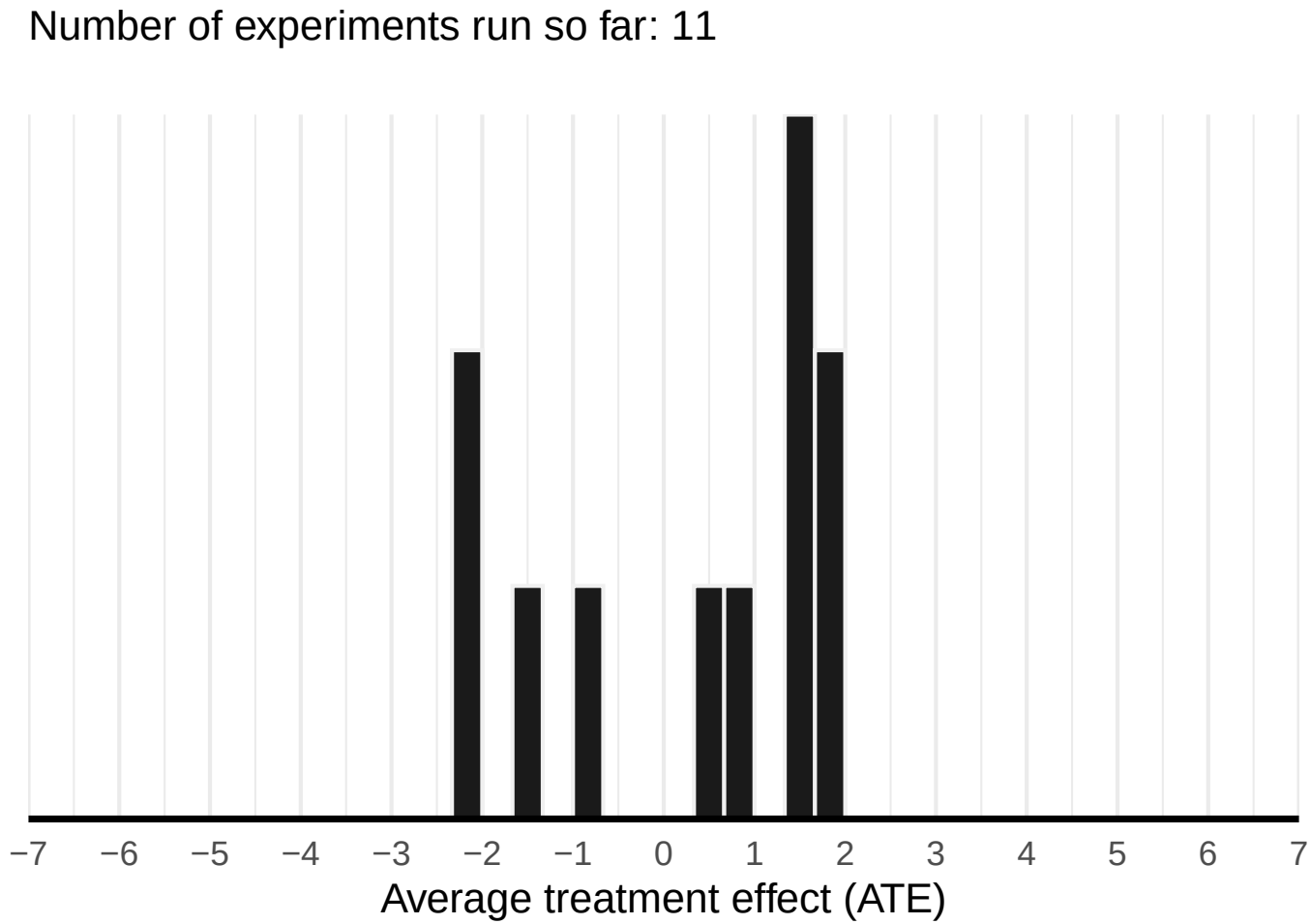
Let's say watching the debate truly has no effect. What might we see?



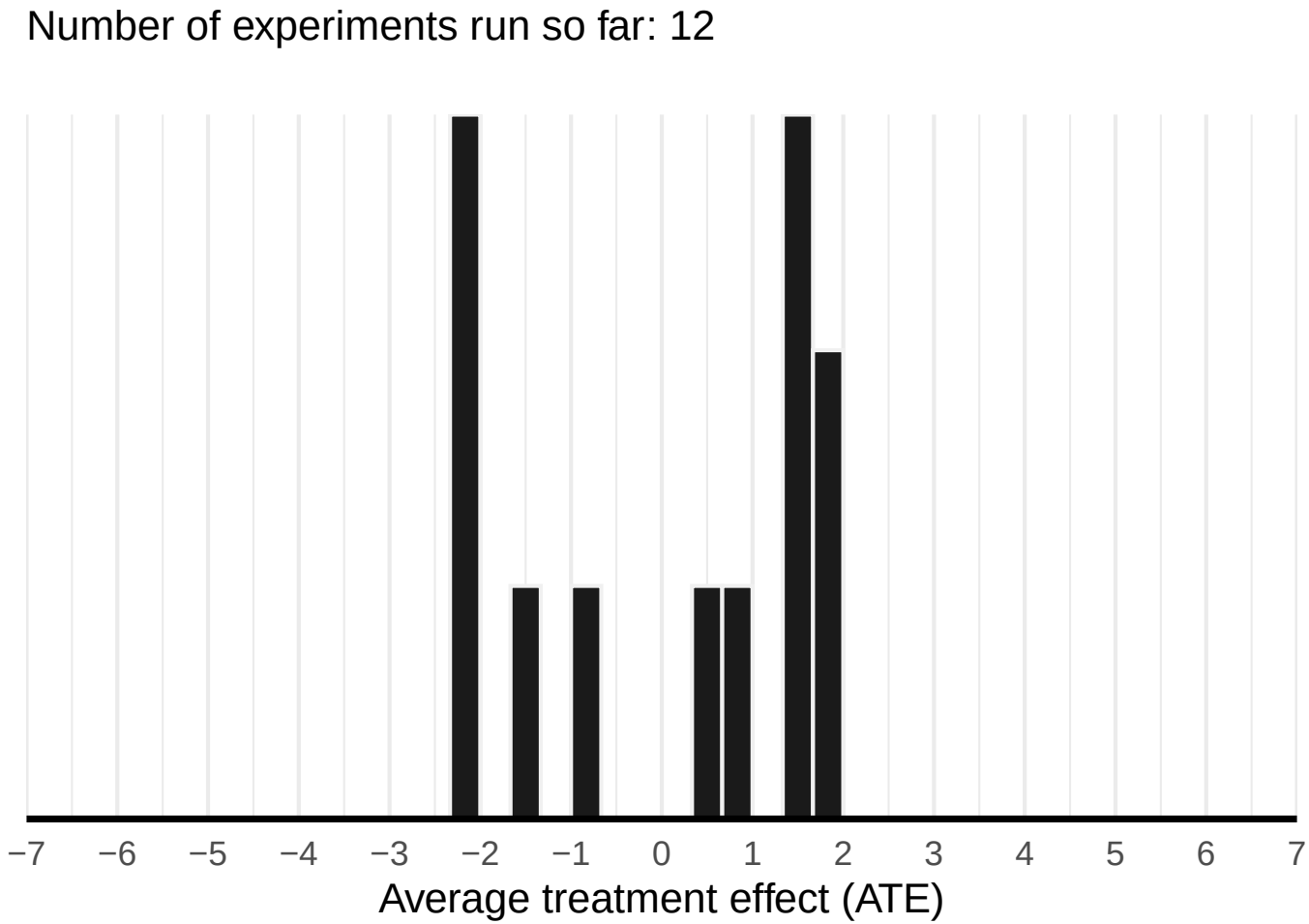
Let's say watching the debate truly has no effect. What might we see?



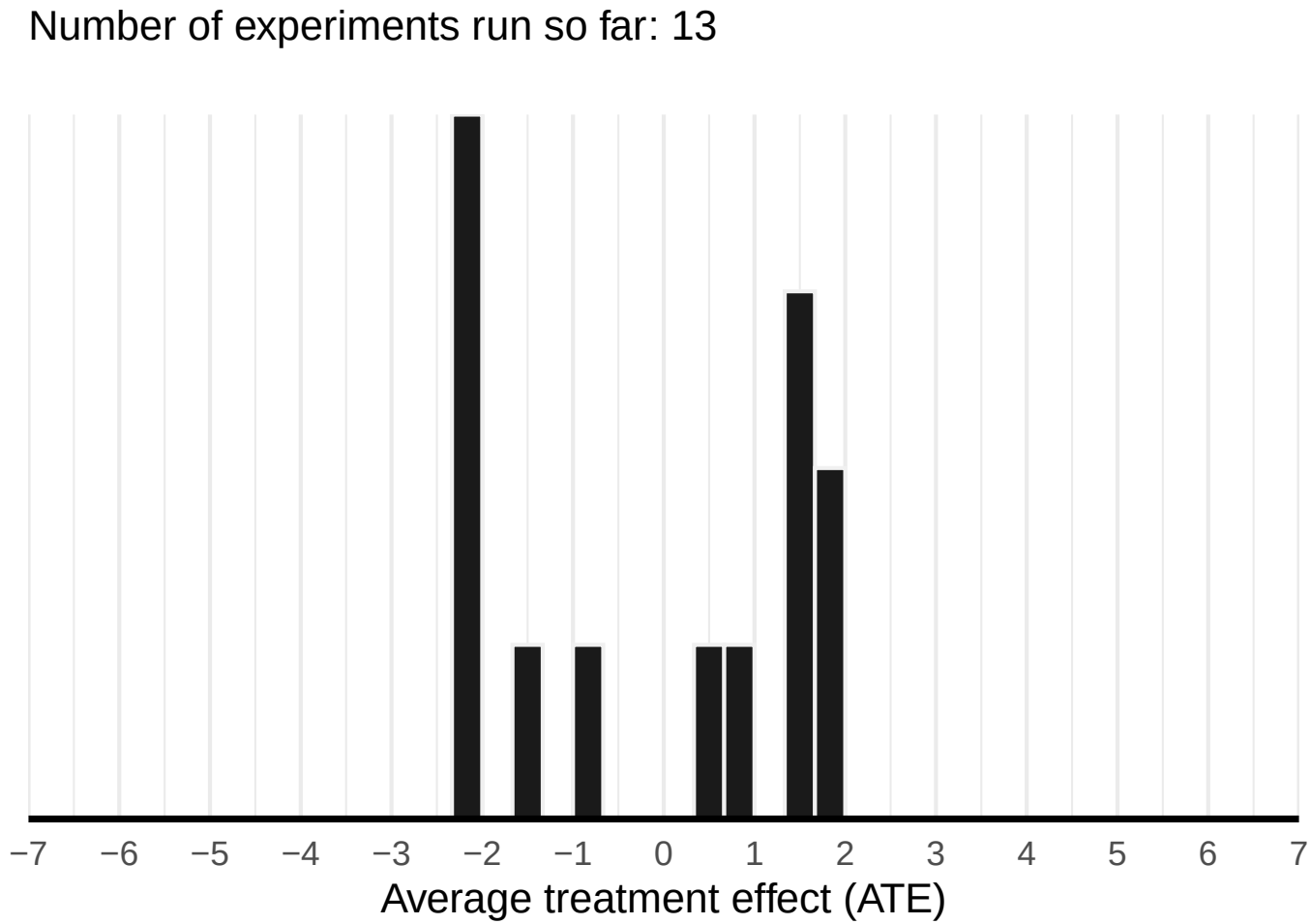
Let's say watching the debate truly has no effect. What might we see?



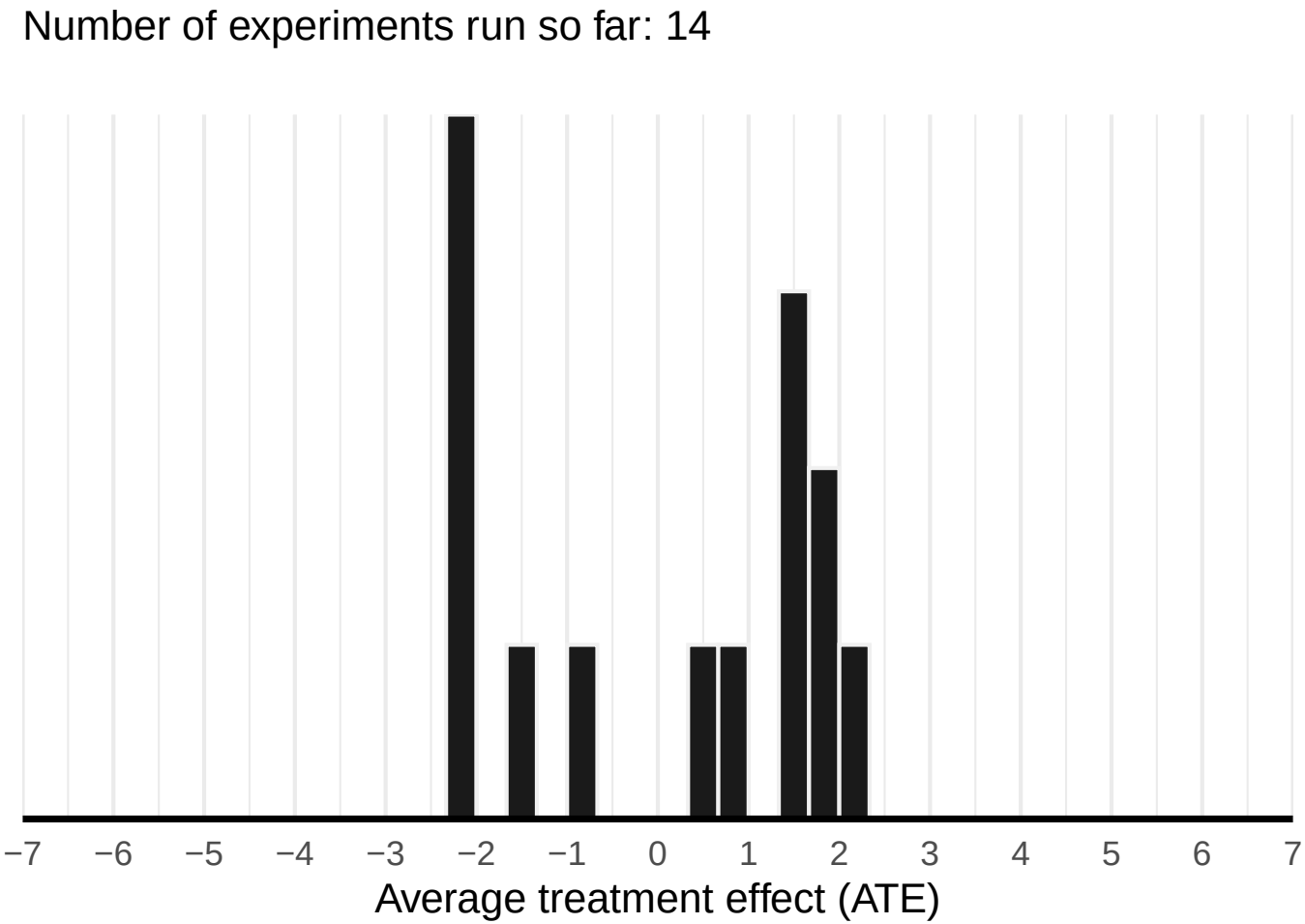
Let's say watching the debate truly has no effect. What might we see?



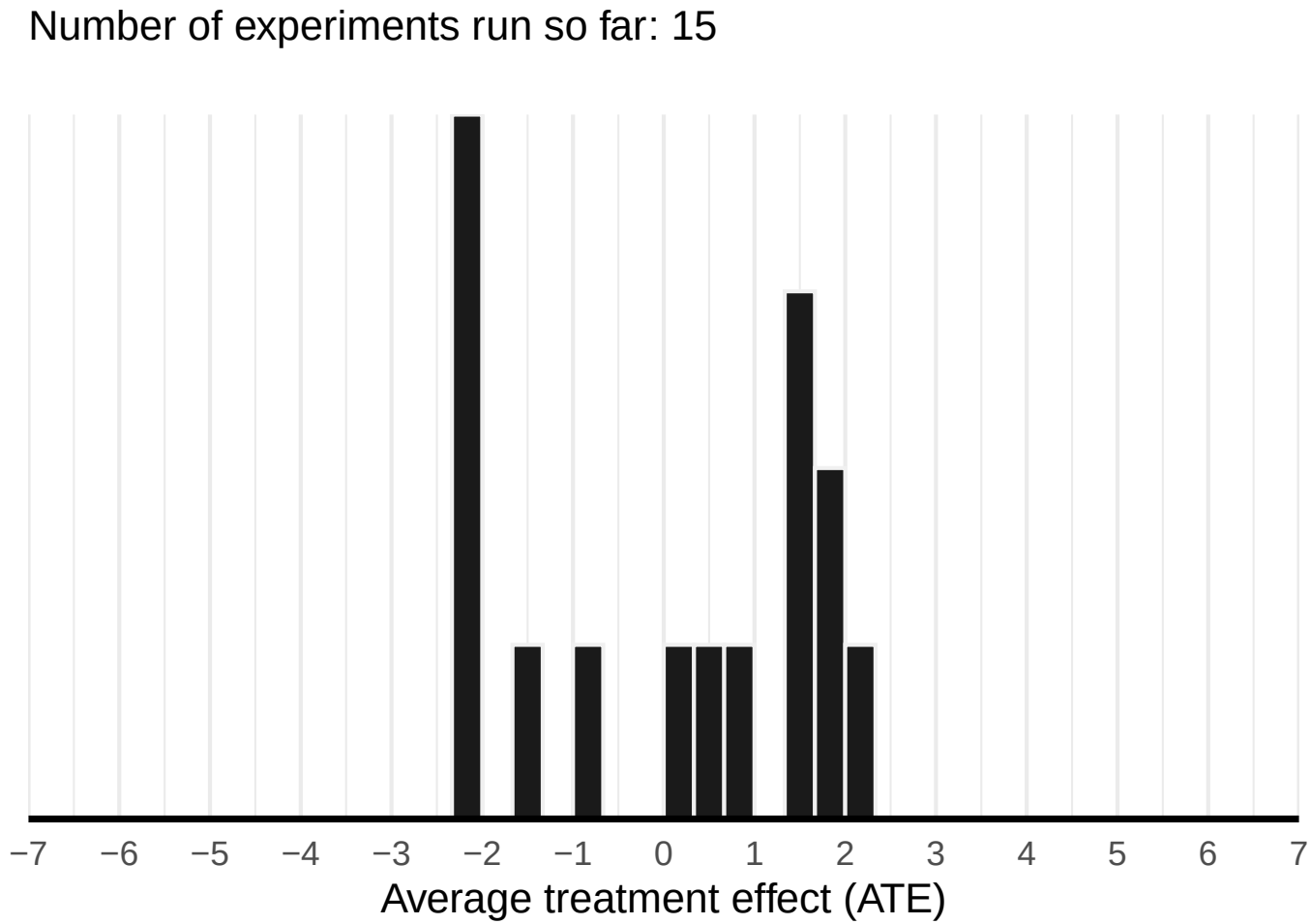
Let's say watching the debate truly has no effect. What might we see?



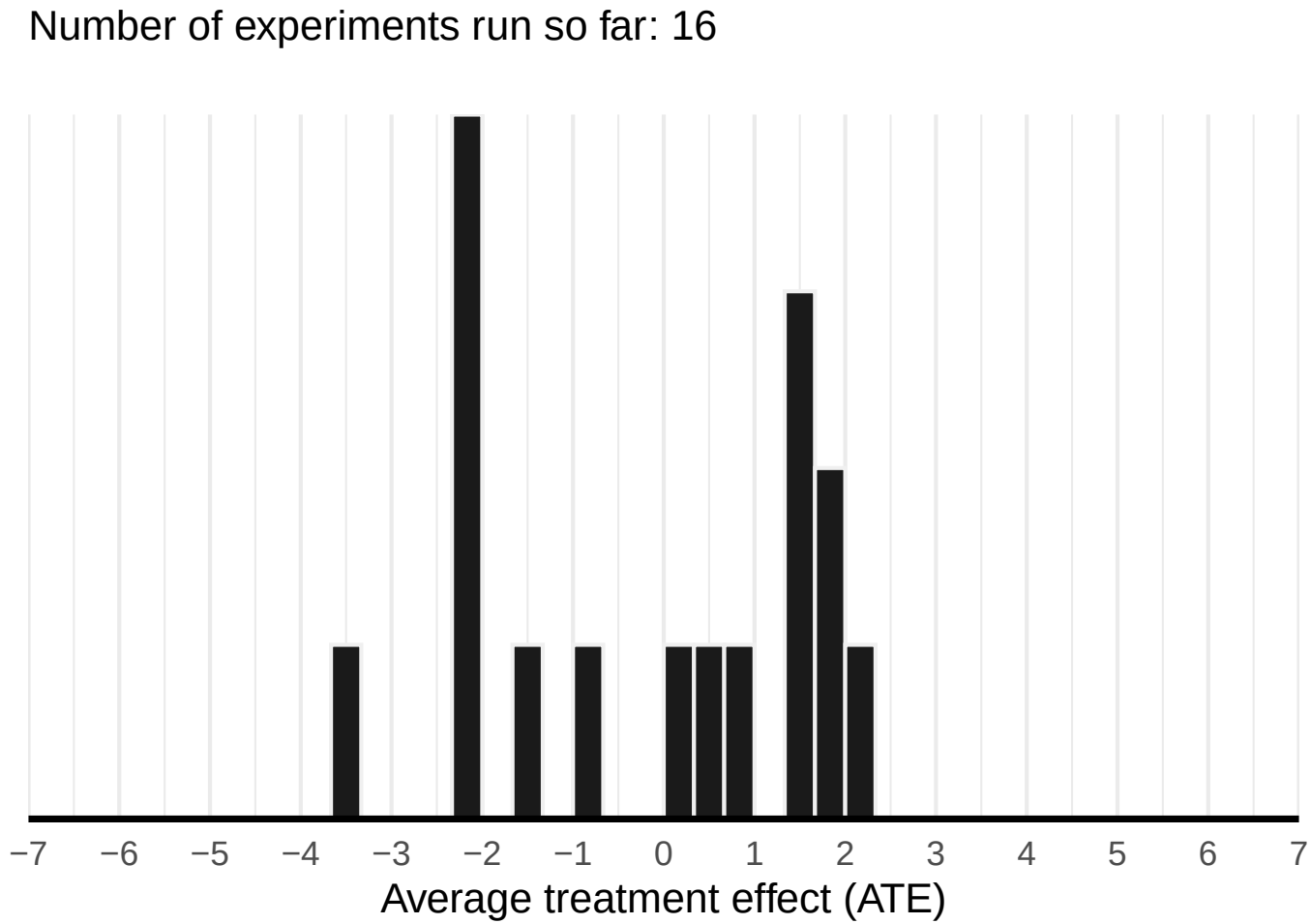
Let's say watching the debate truly has no effect. What might we see?



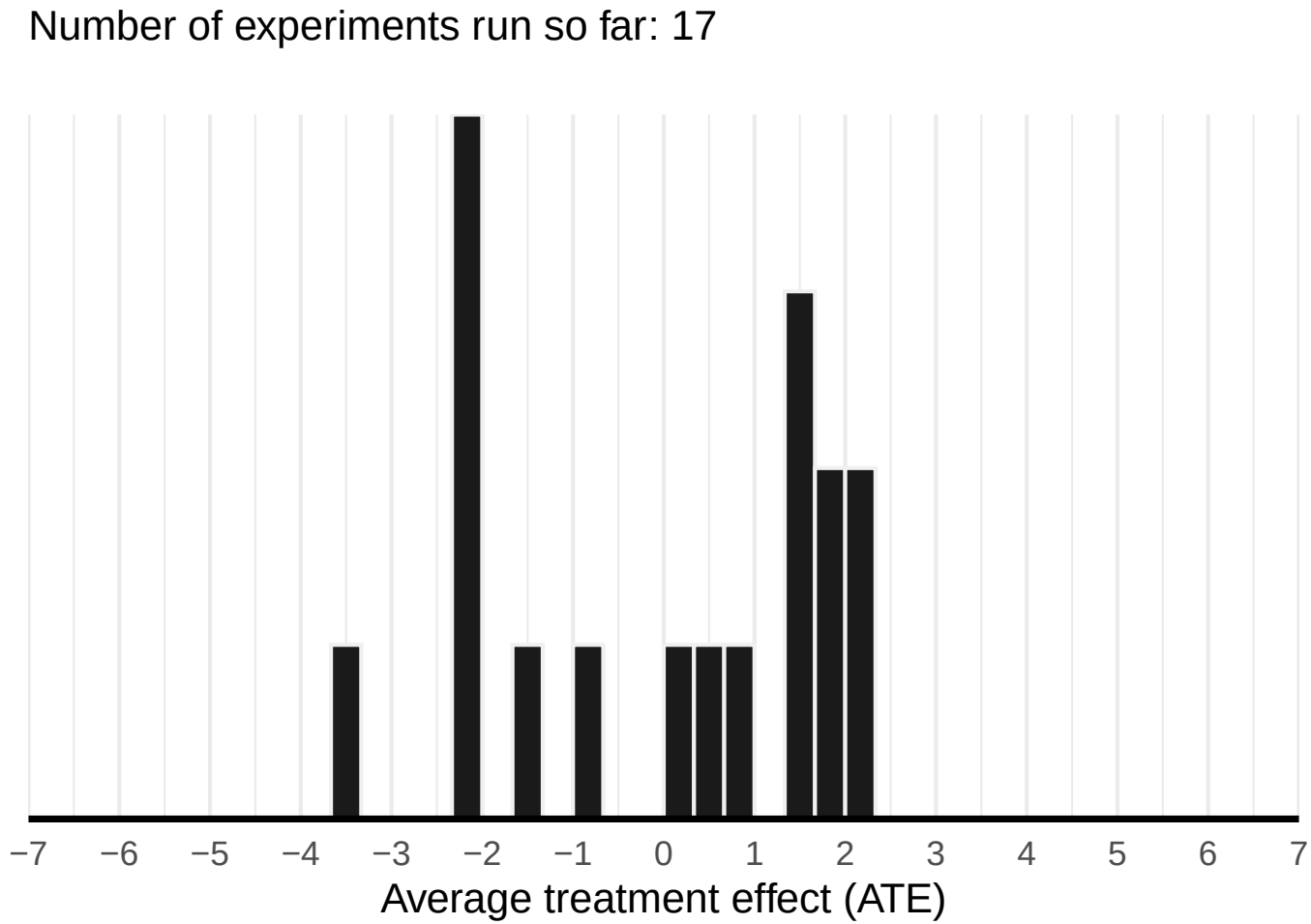
Let's say watching the debate truly has no effect. What might we see?



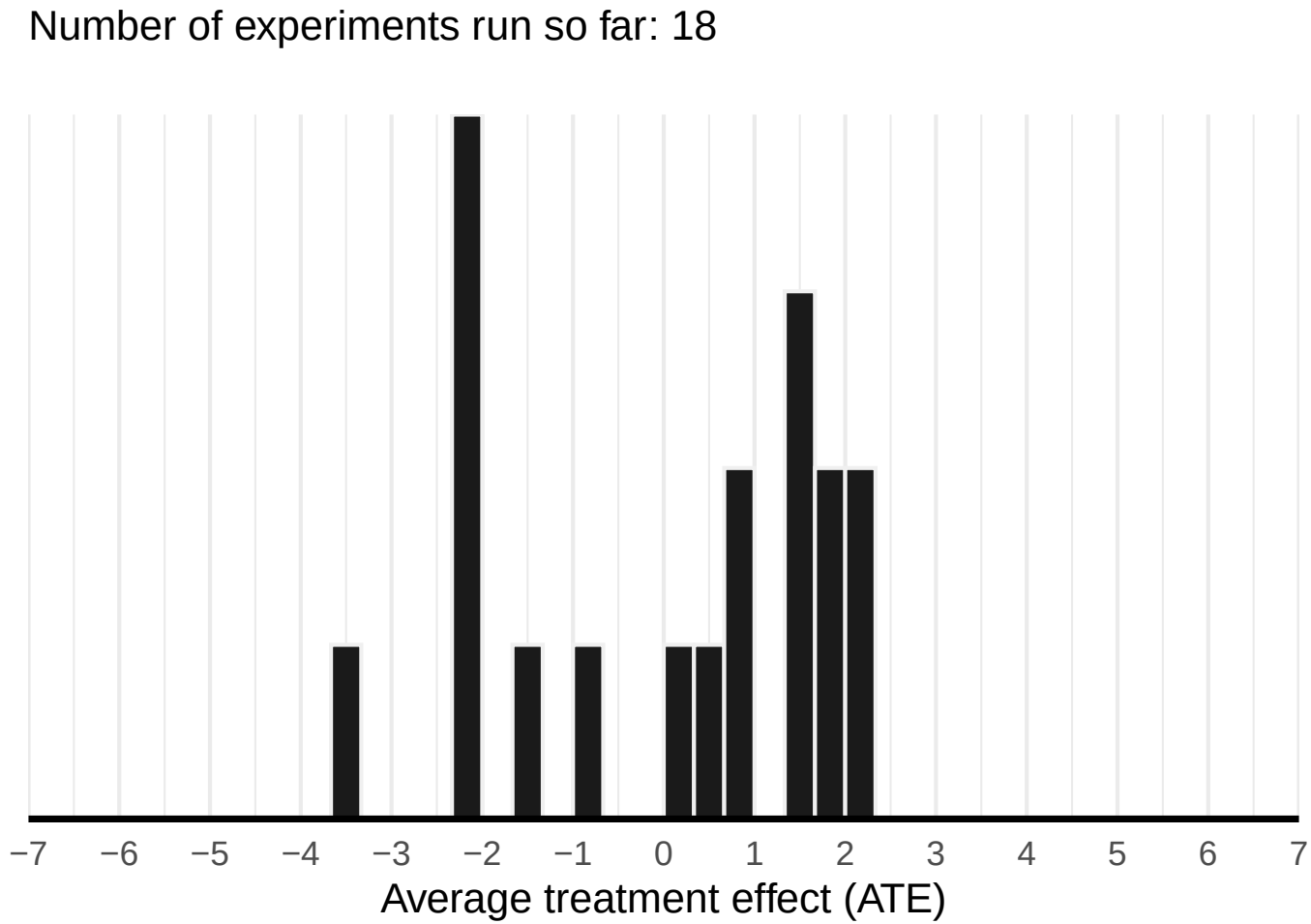
Let's say watching the debate truly has no effect. What might we see?



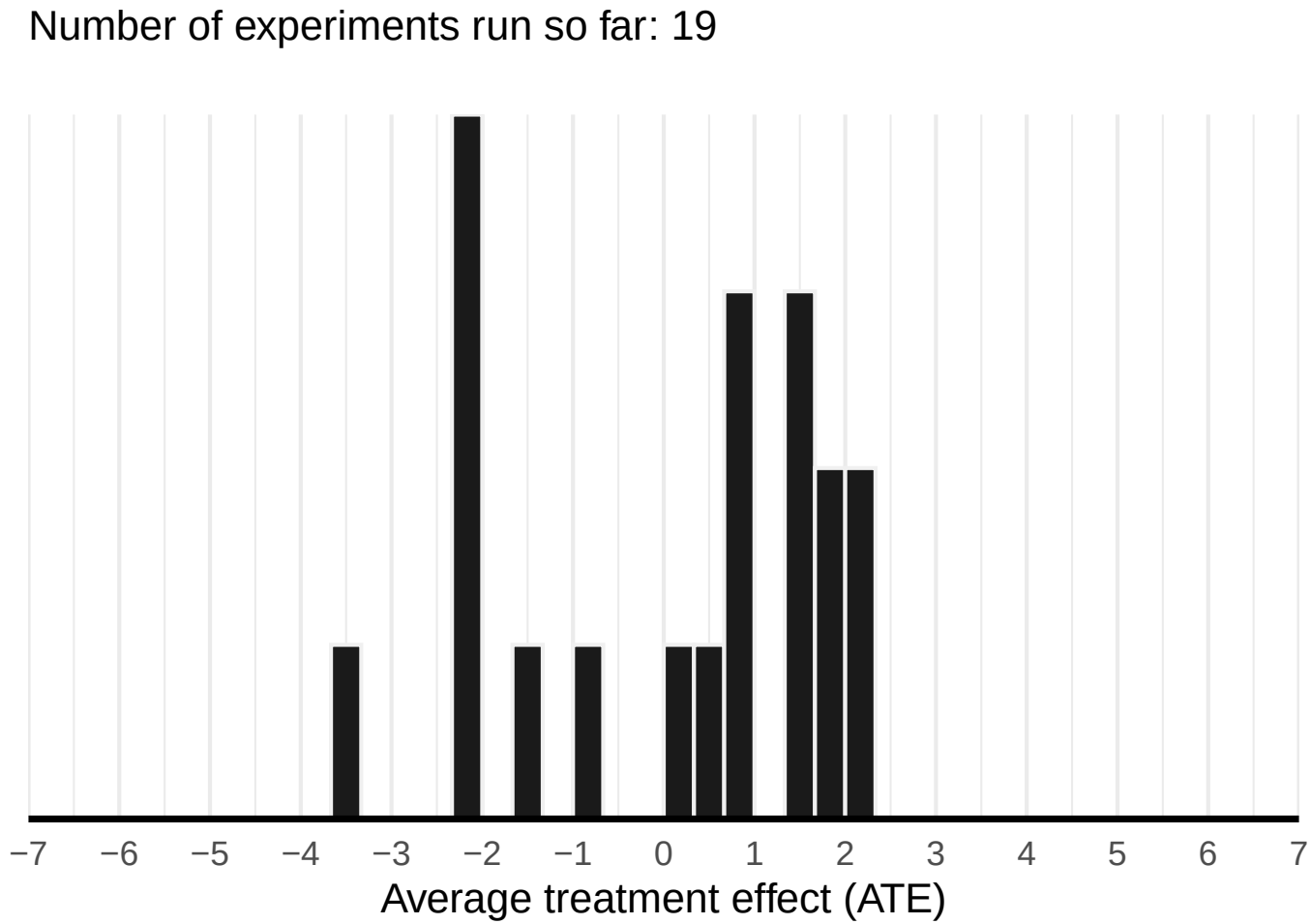
Let's say watching the debate truly has no effect. What might we see?



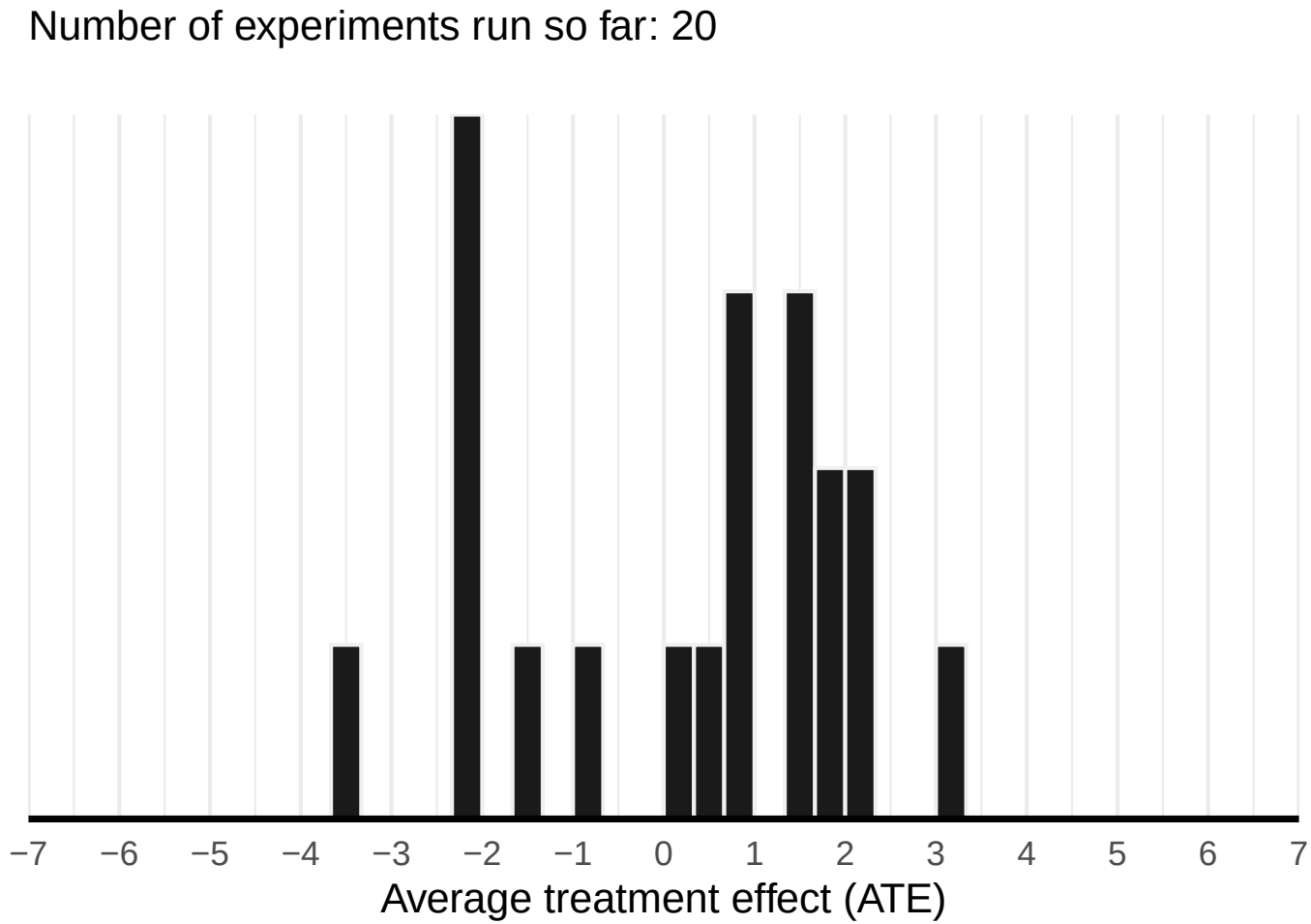
Let's say watching the debate truly has no effect. What might we see?



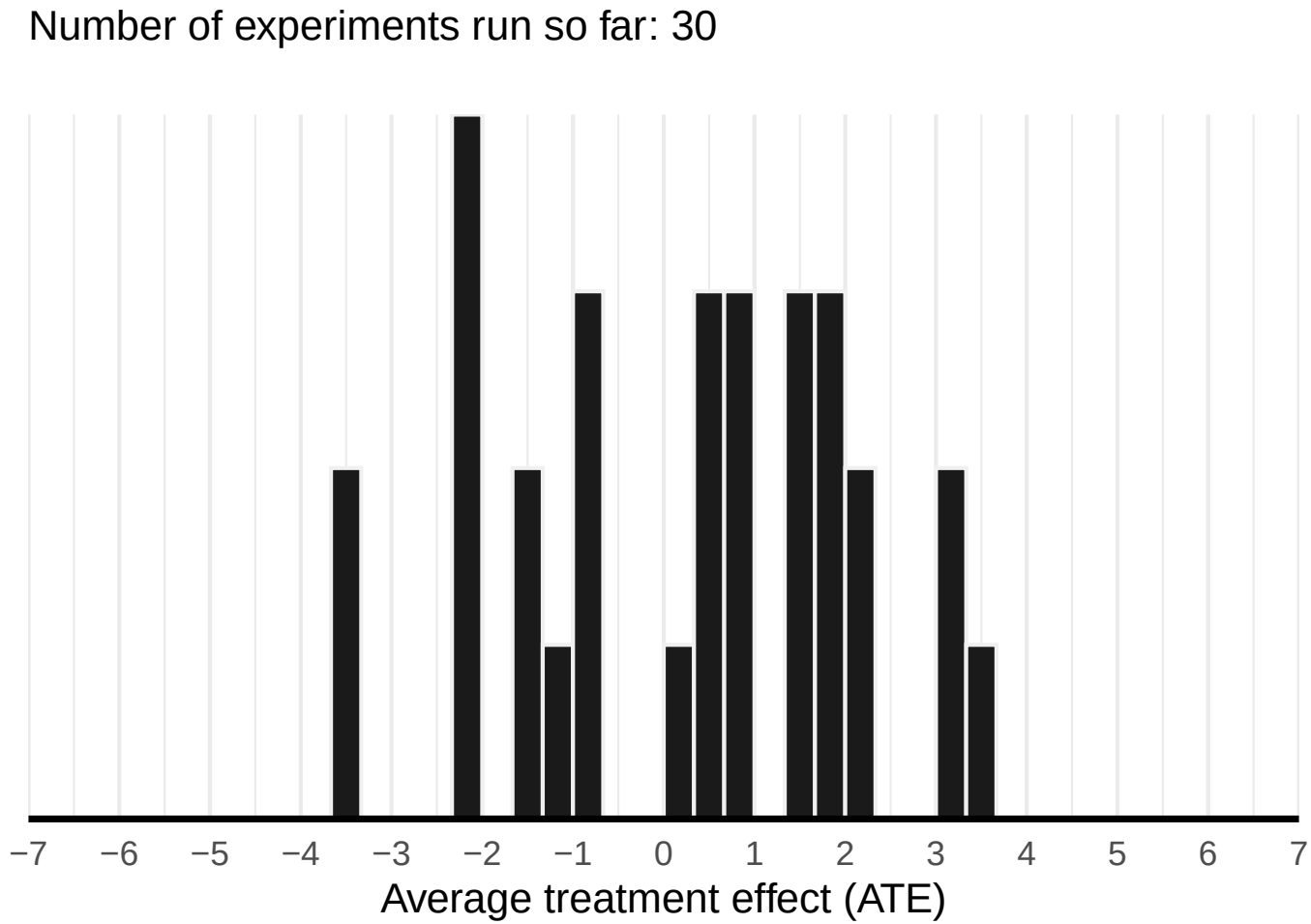
Let's say watching the debate truly has no effect. What might we see?



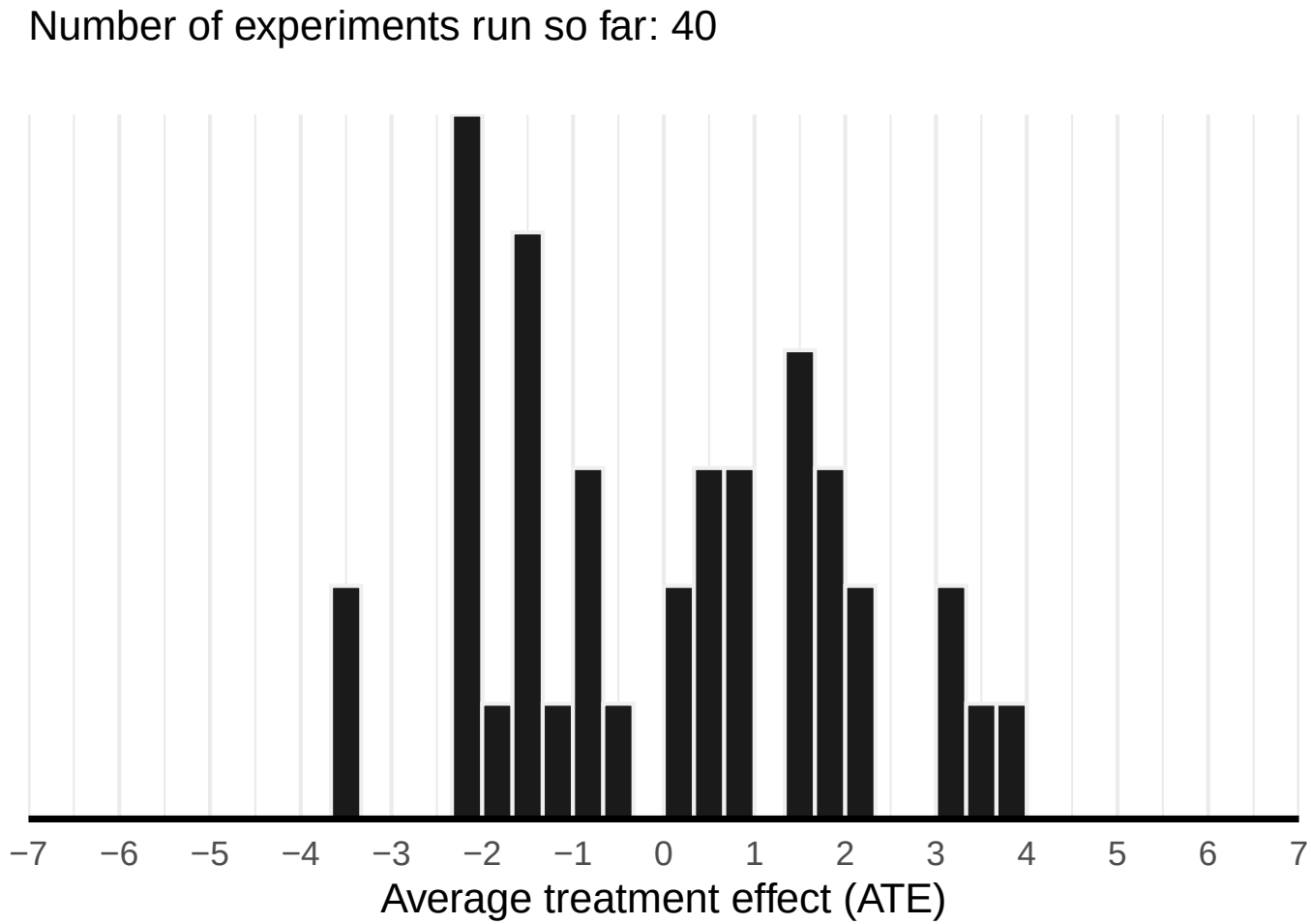
Let's say watching the debate truly has no effect. What might we see?



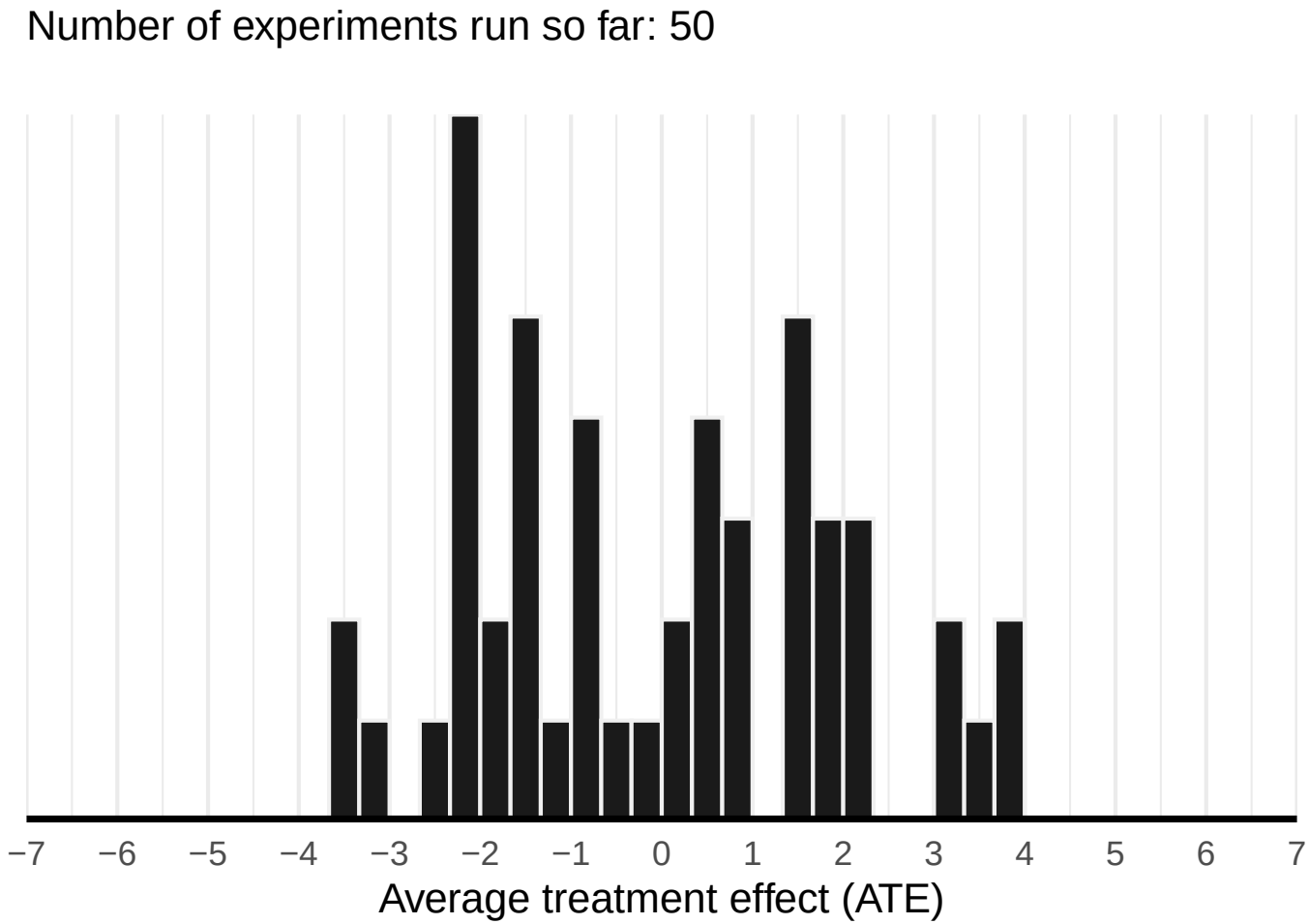
Let's say watching the debate truly has no effect. What might we see?



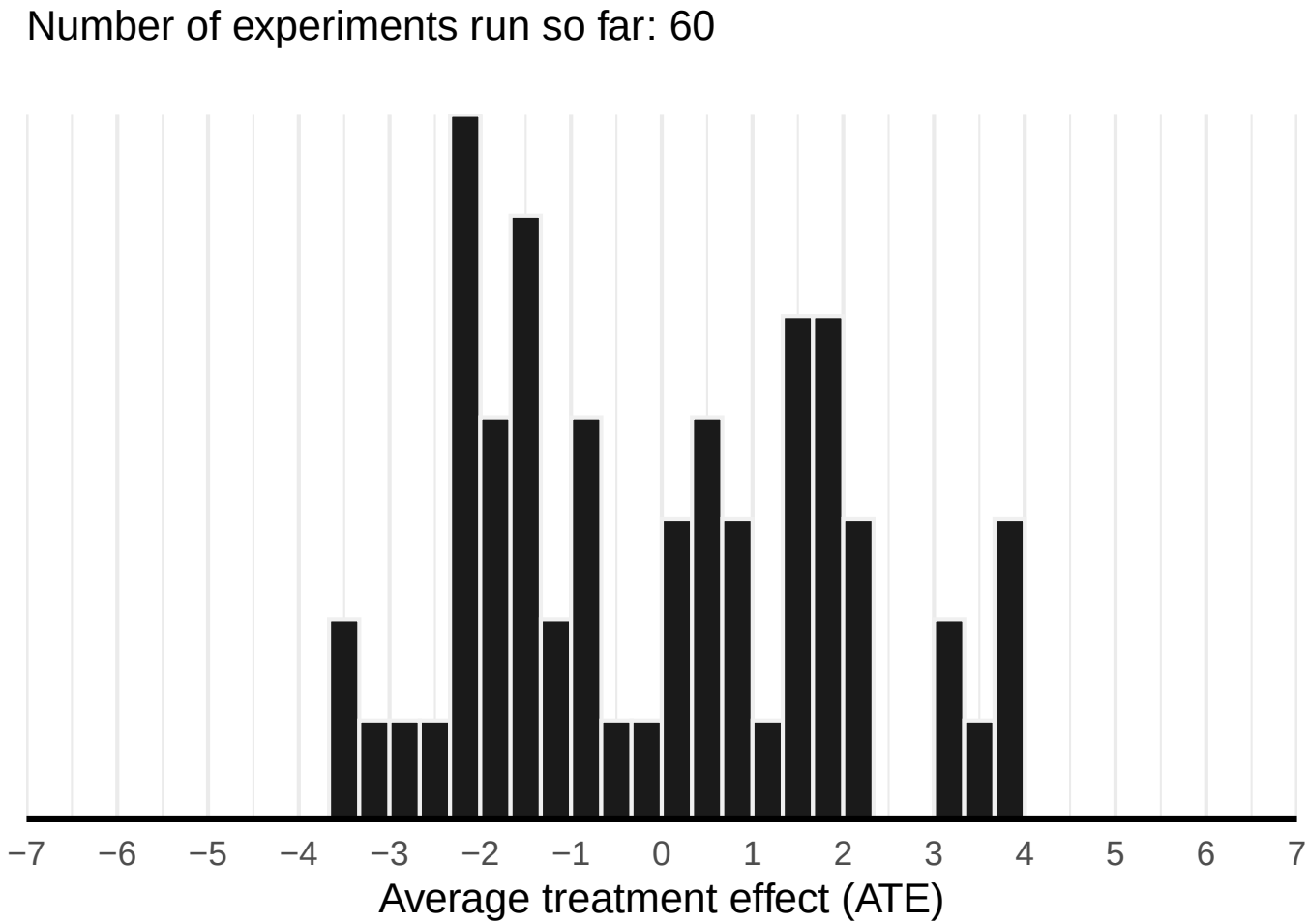
Let's say watching the debate truly has no effect. What might we see?



Let's say watching the debate truly has no effect. What might we see?

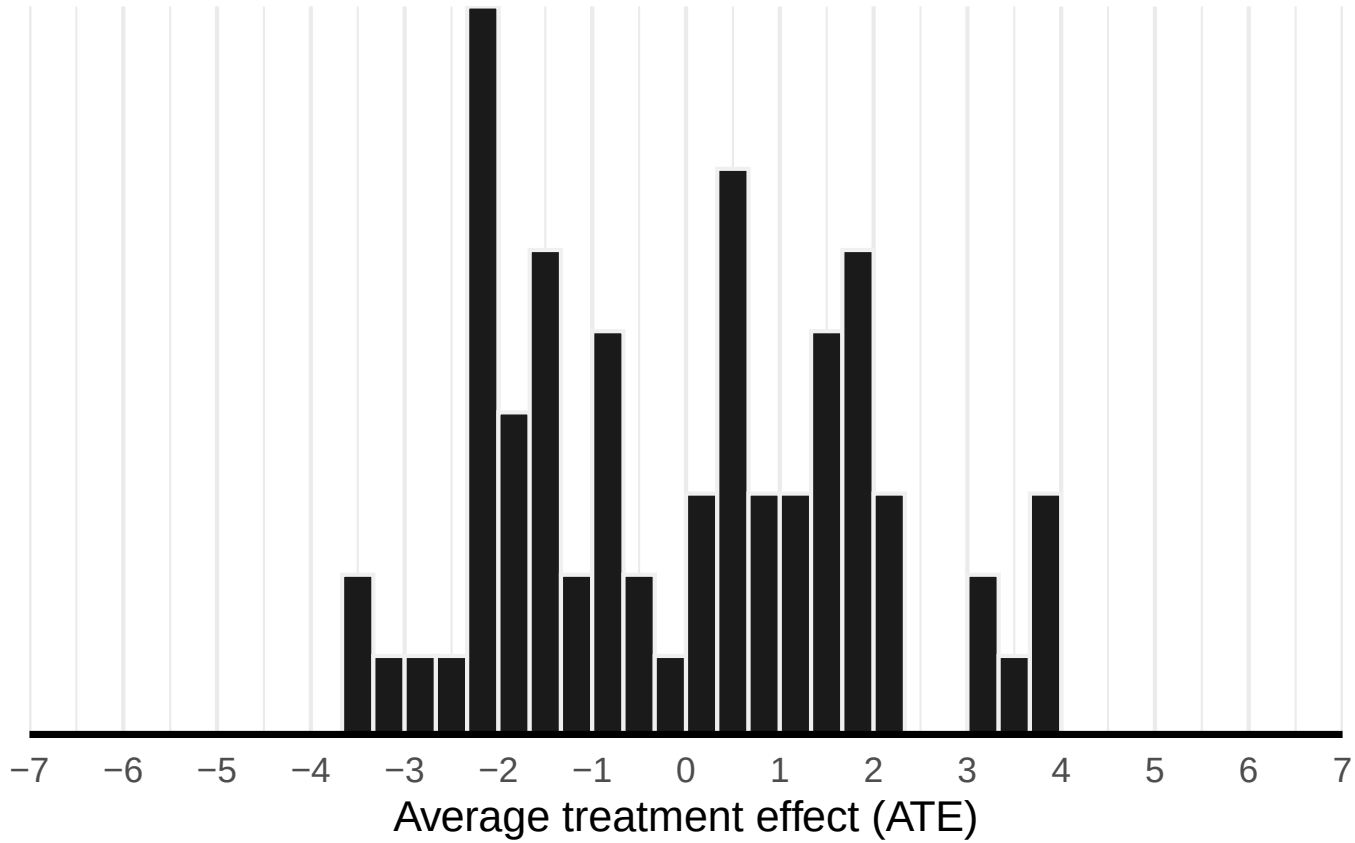


Let's say watching the debate truly has no effect. What might we see?

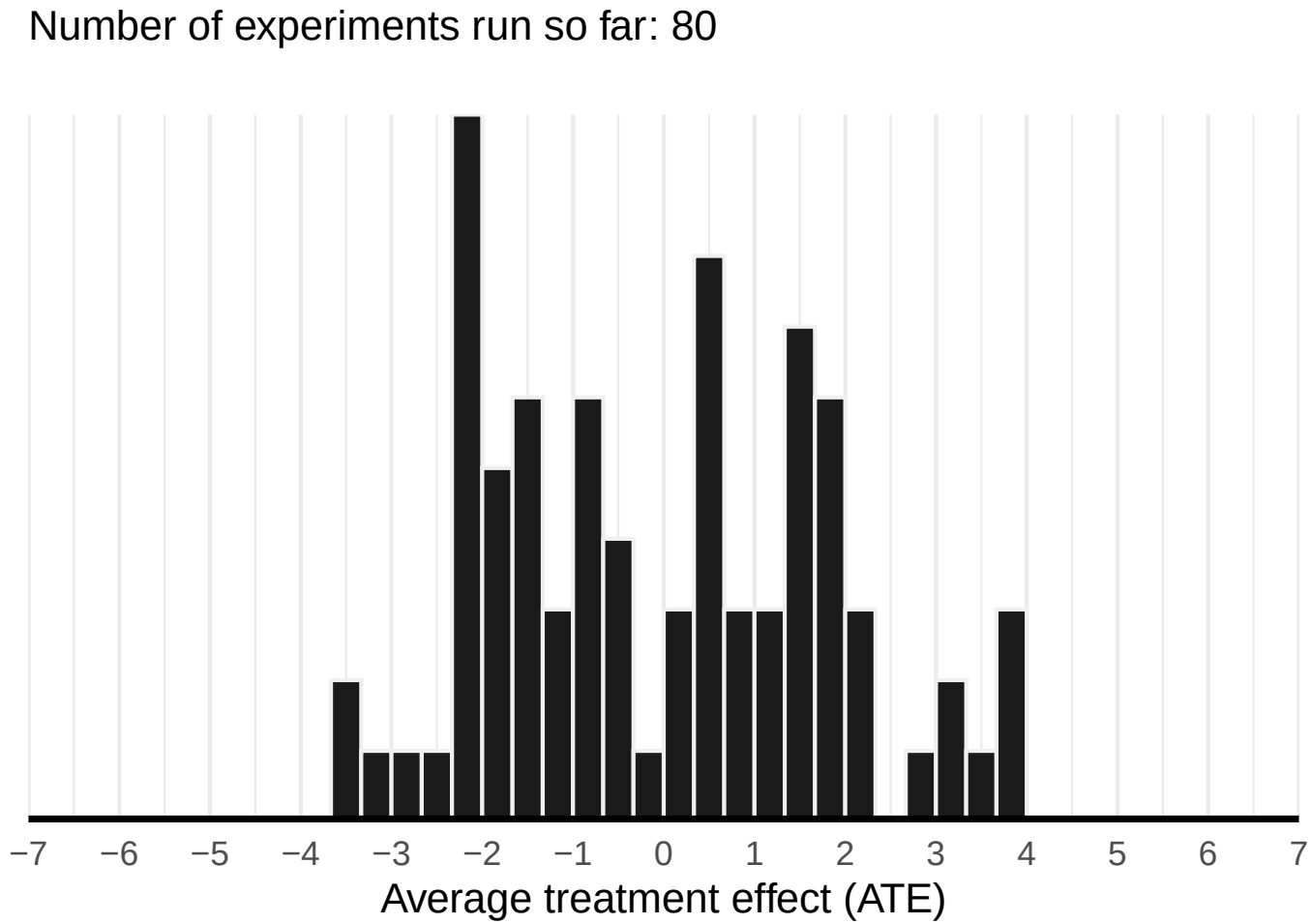


Let's say watching the debate truly has no effect. What might we see?

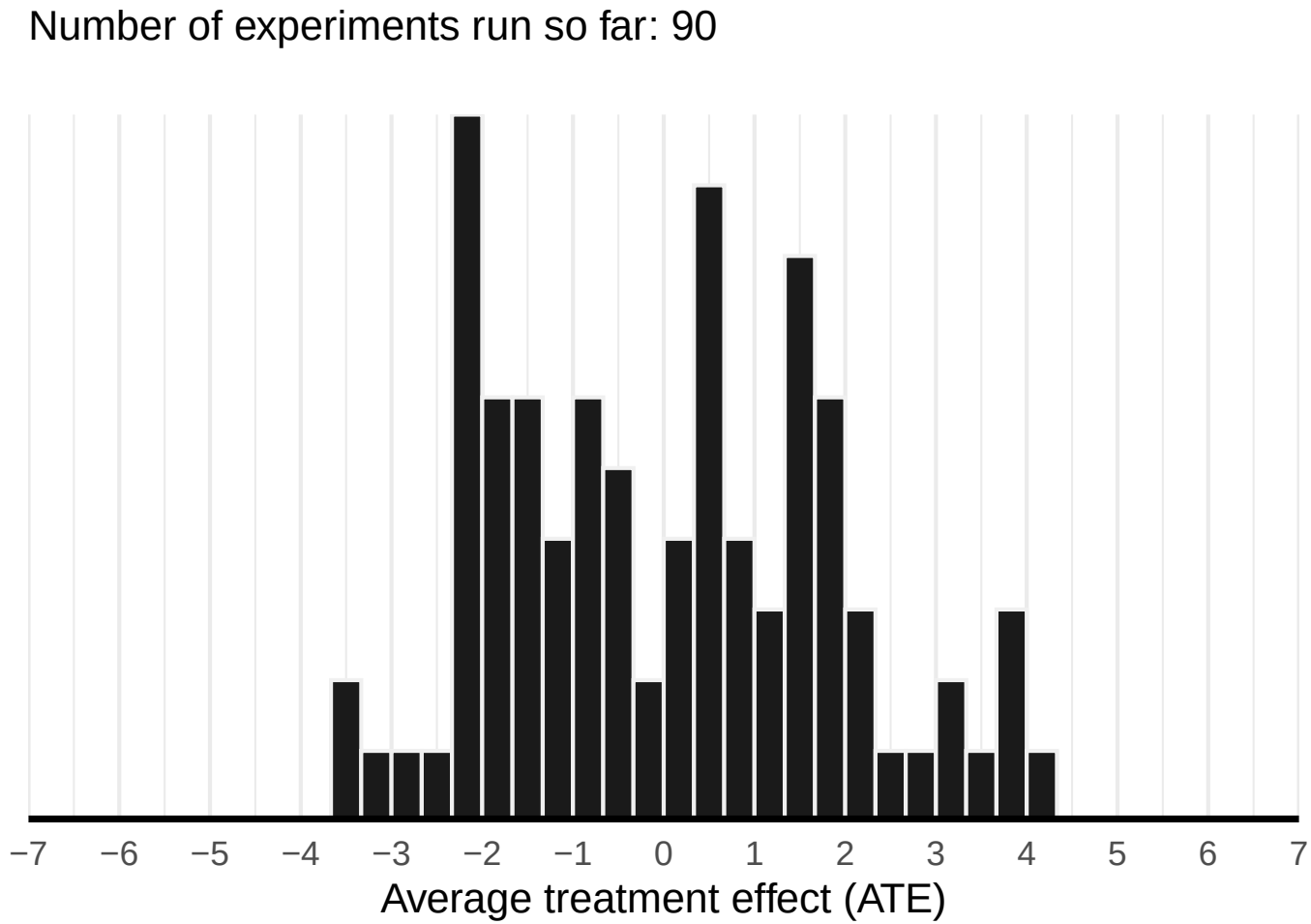
Number of experiments run so far: 70



Let's say watching the debate truly has no effect. What might we see?

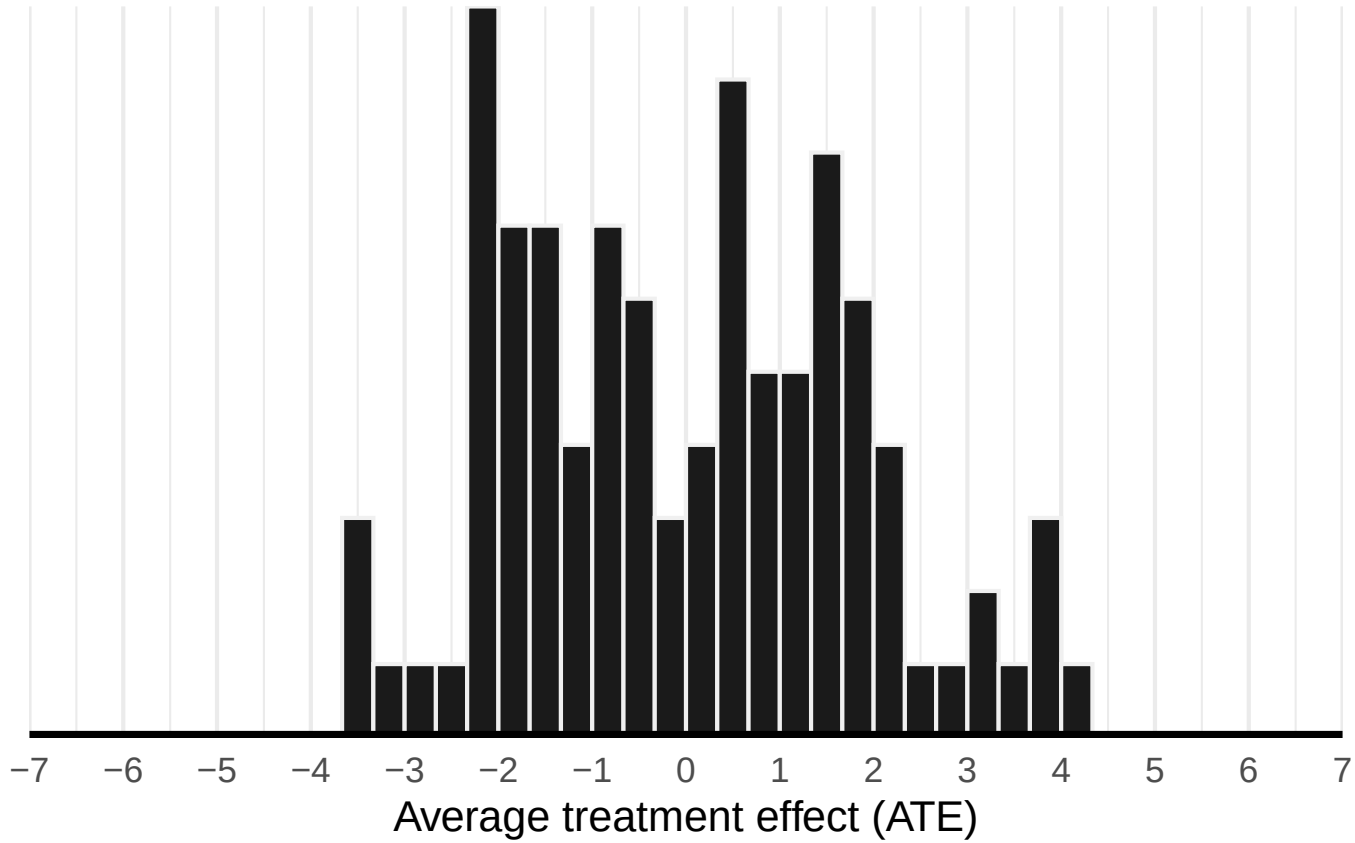


Let's say watching the debate truly has no effect. What might we see?

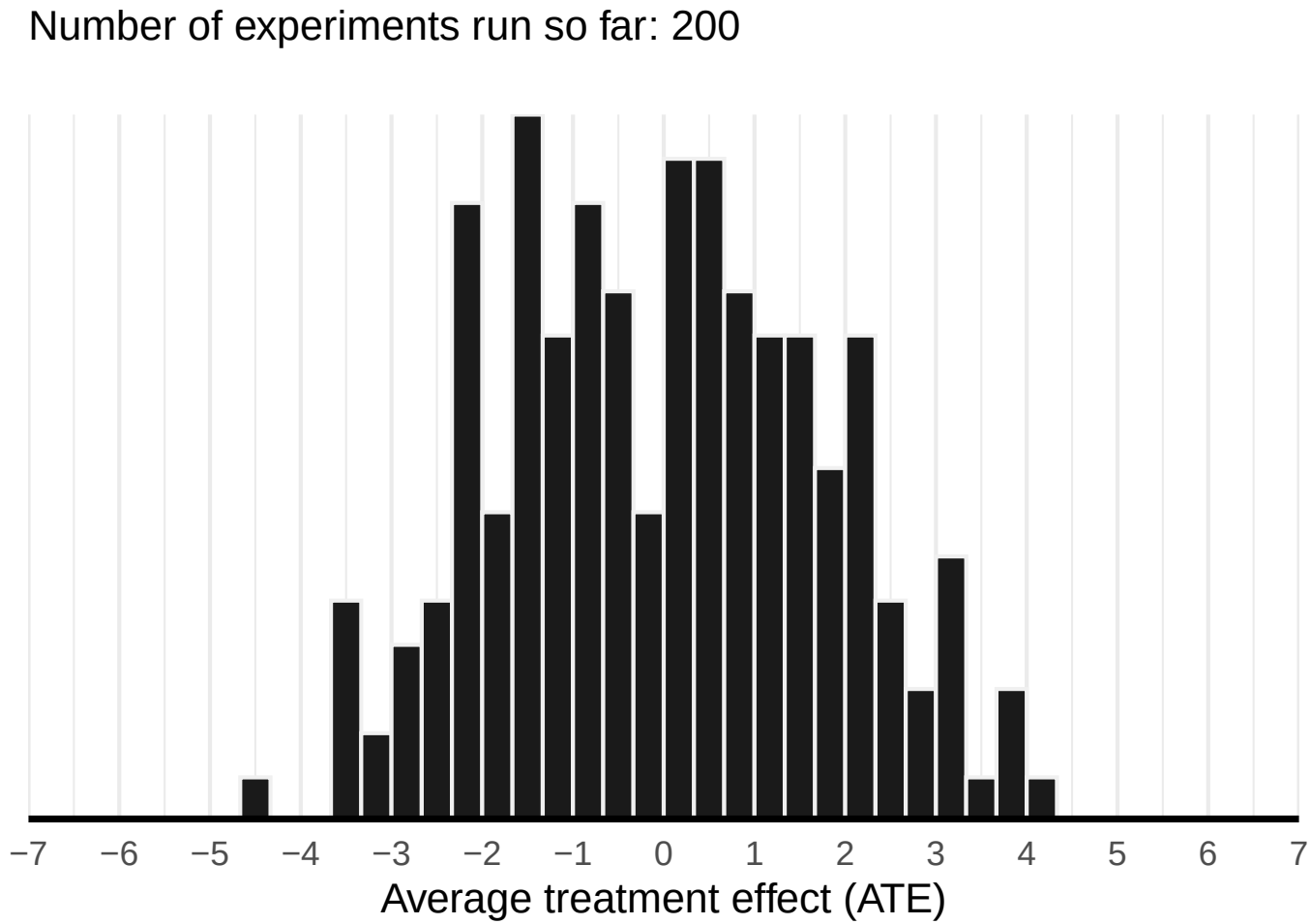


Let's say watching the debate truly has no effect. What might we see?

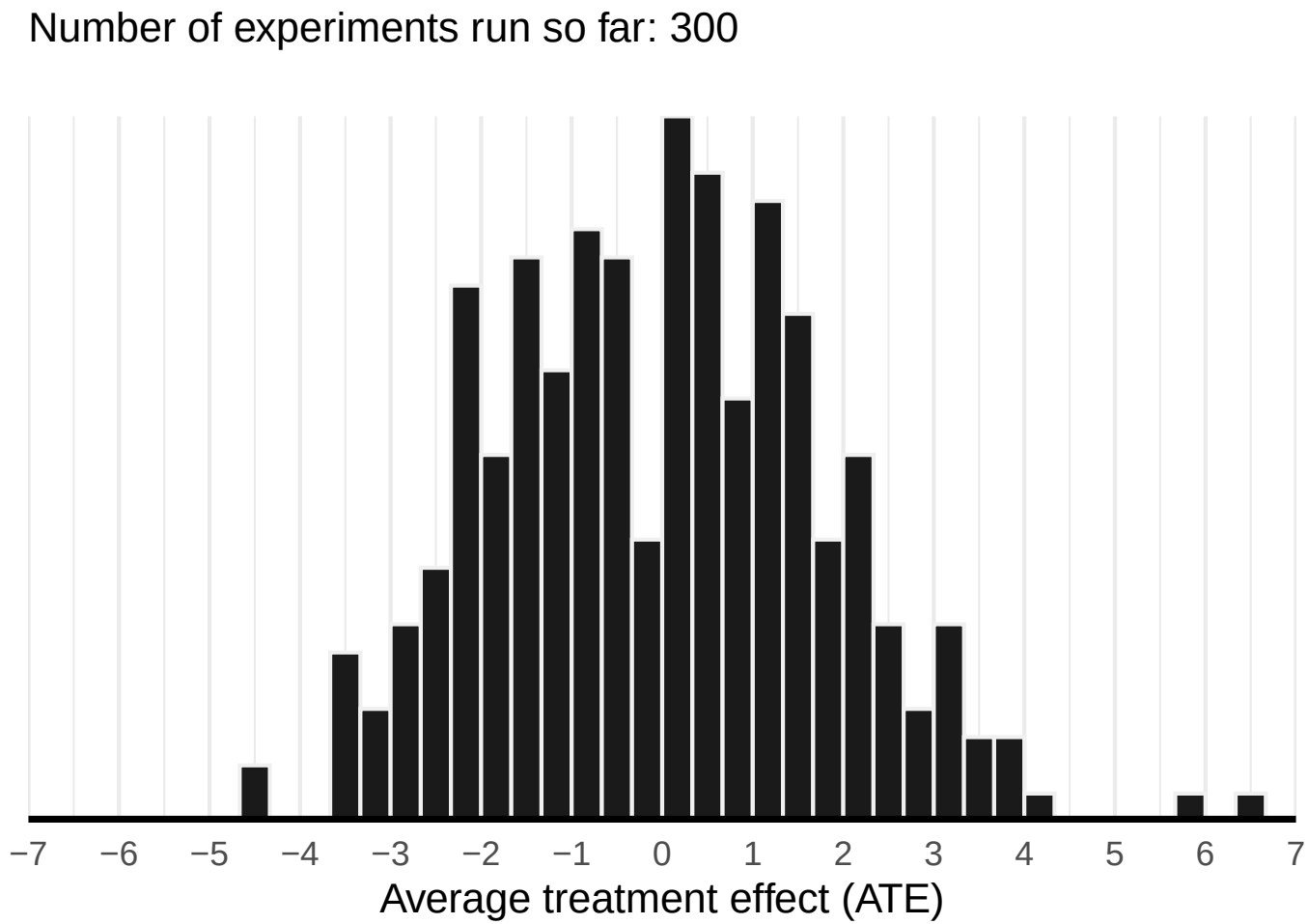
Number of experiments run so far: 100



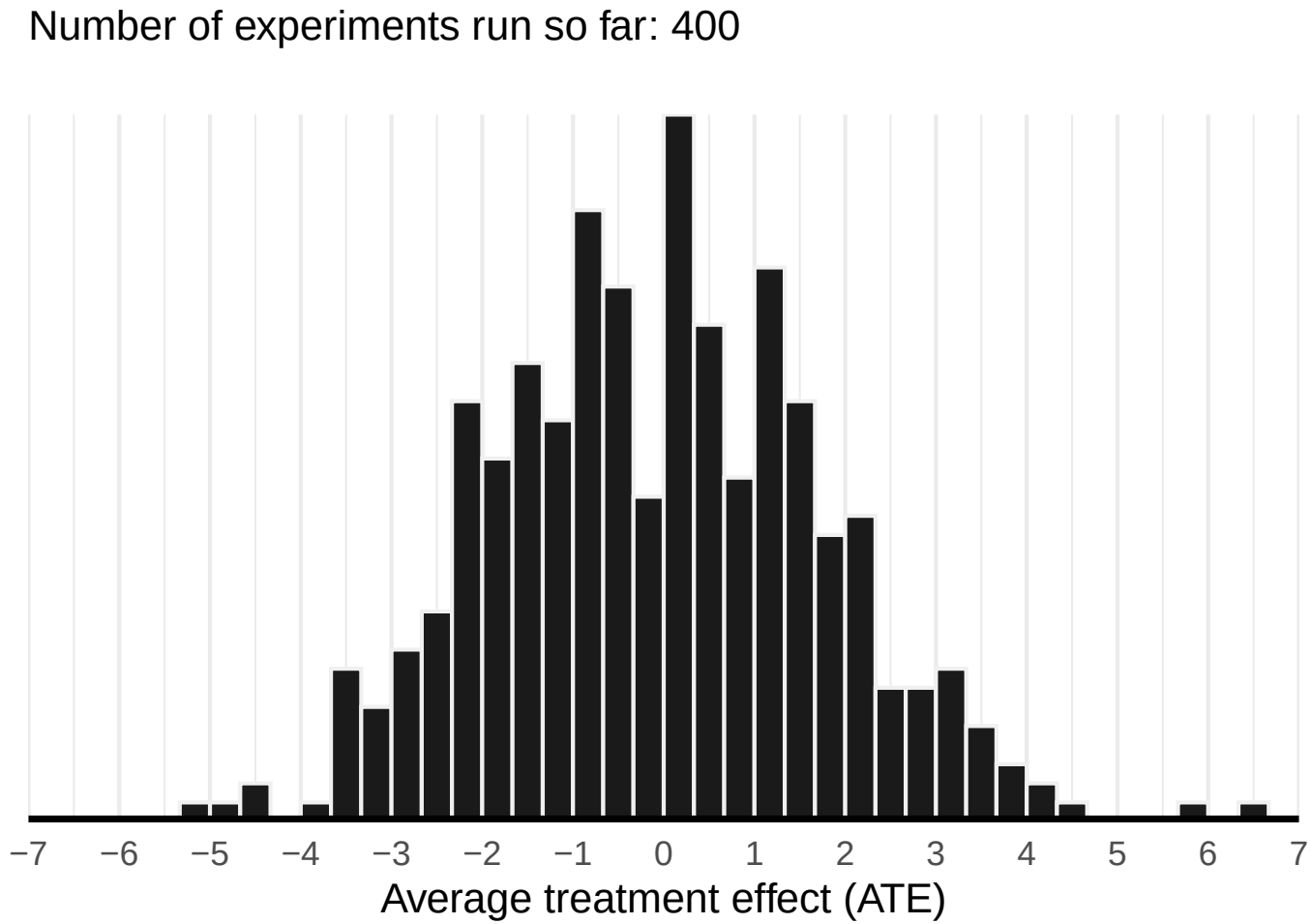
Let's say watching the debate truly has no effect. What might we see?



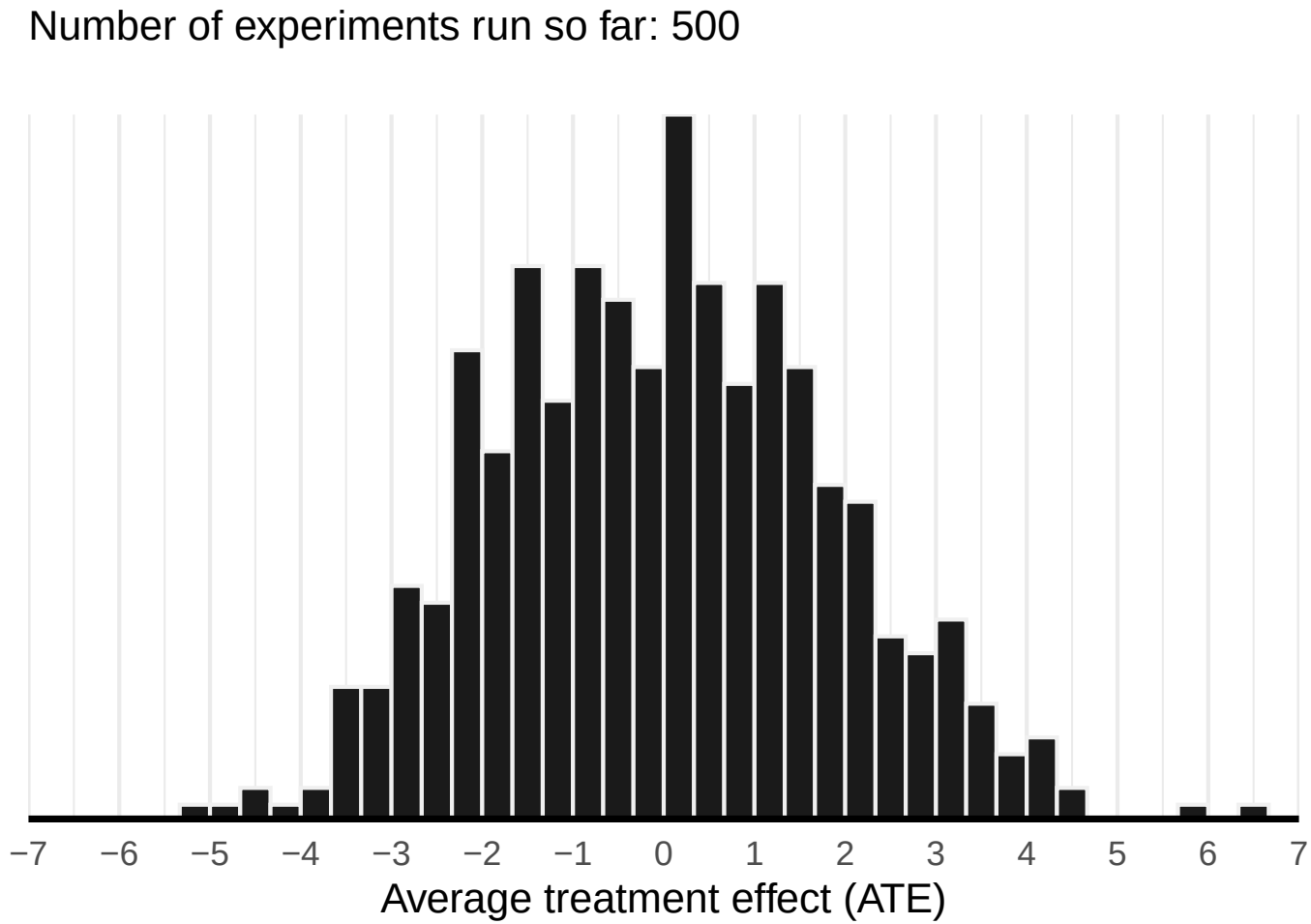
Let's say watching the debate truly has no effect. What might we see?



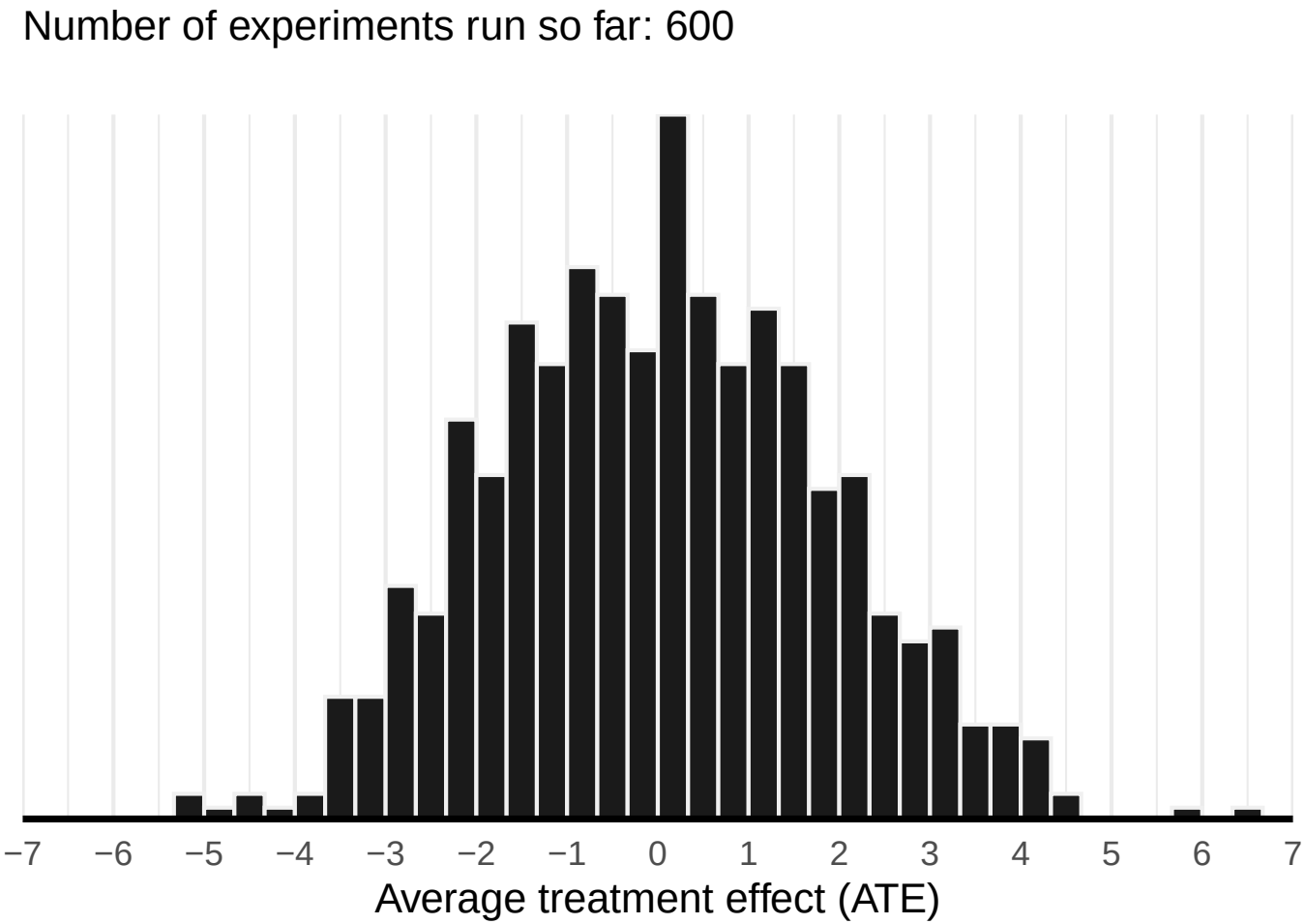
Let's say watching the debate truly has no effect. What might we see?



Let's say watching the debate truly has no effect. What might we see?

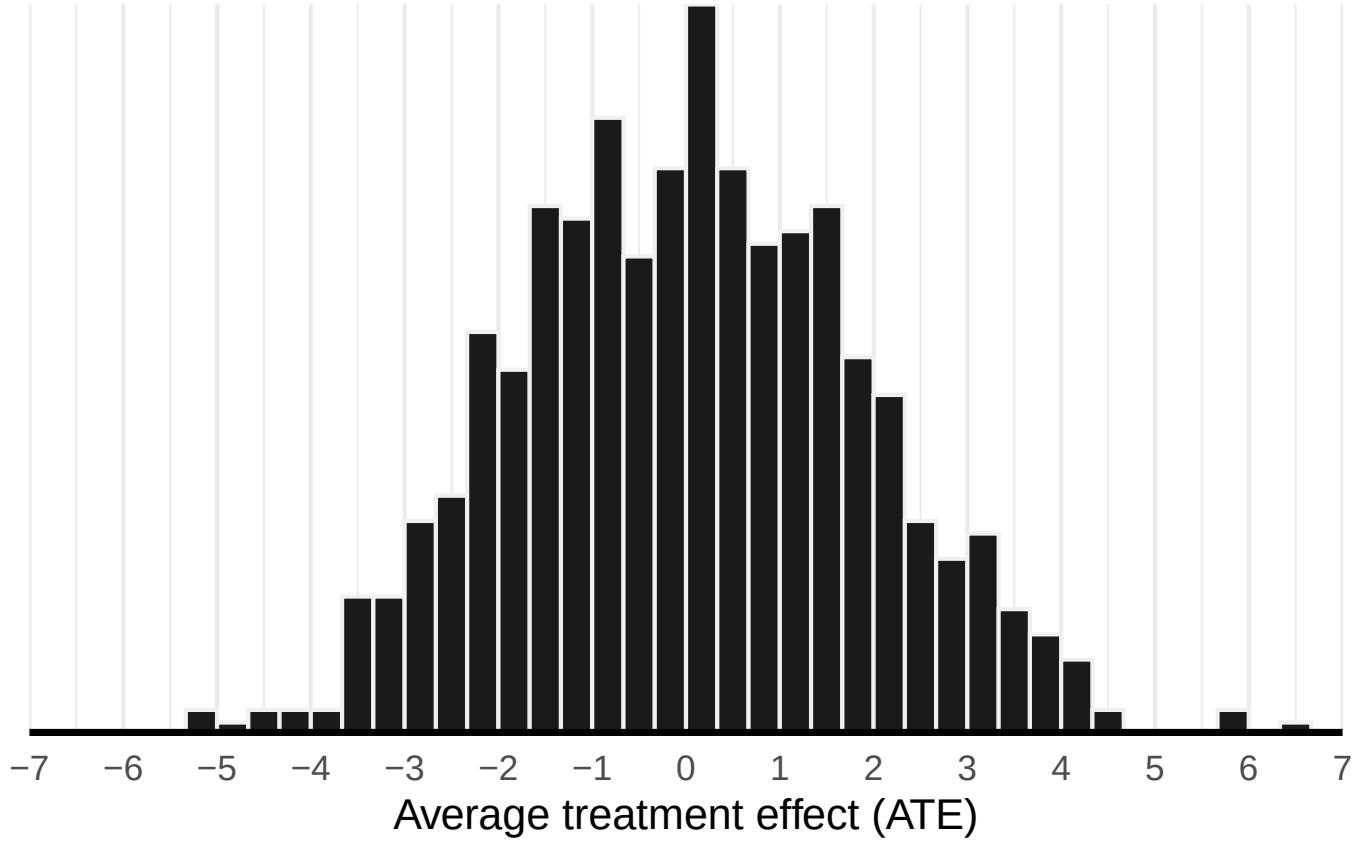


Let's say watching the debate truly has no effect. What might we see?

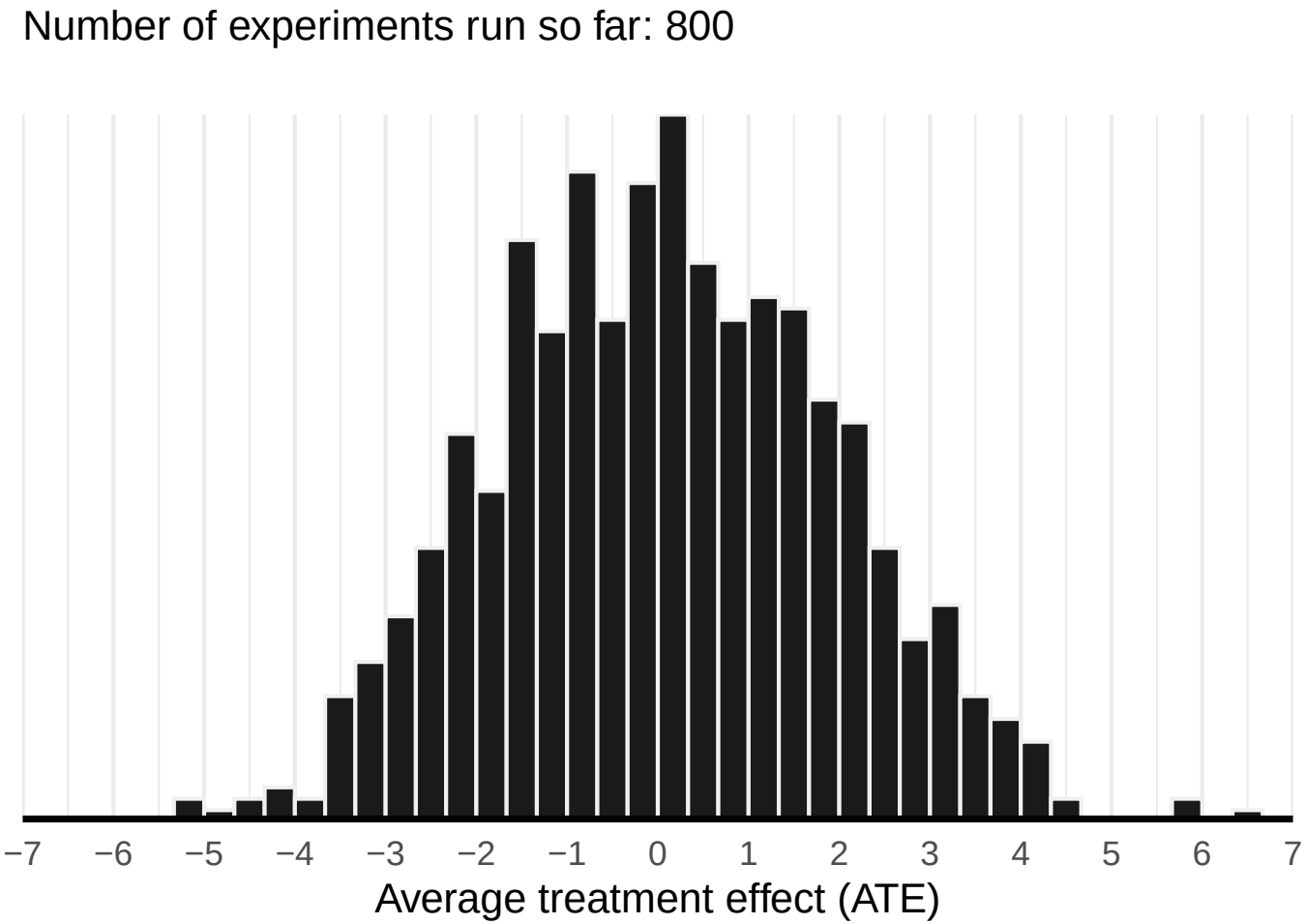


Let's say watching the debate truly has no effect. What might we see?

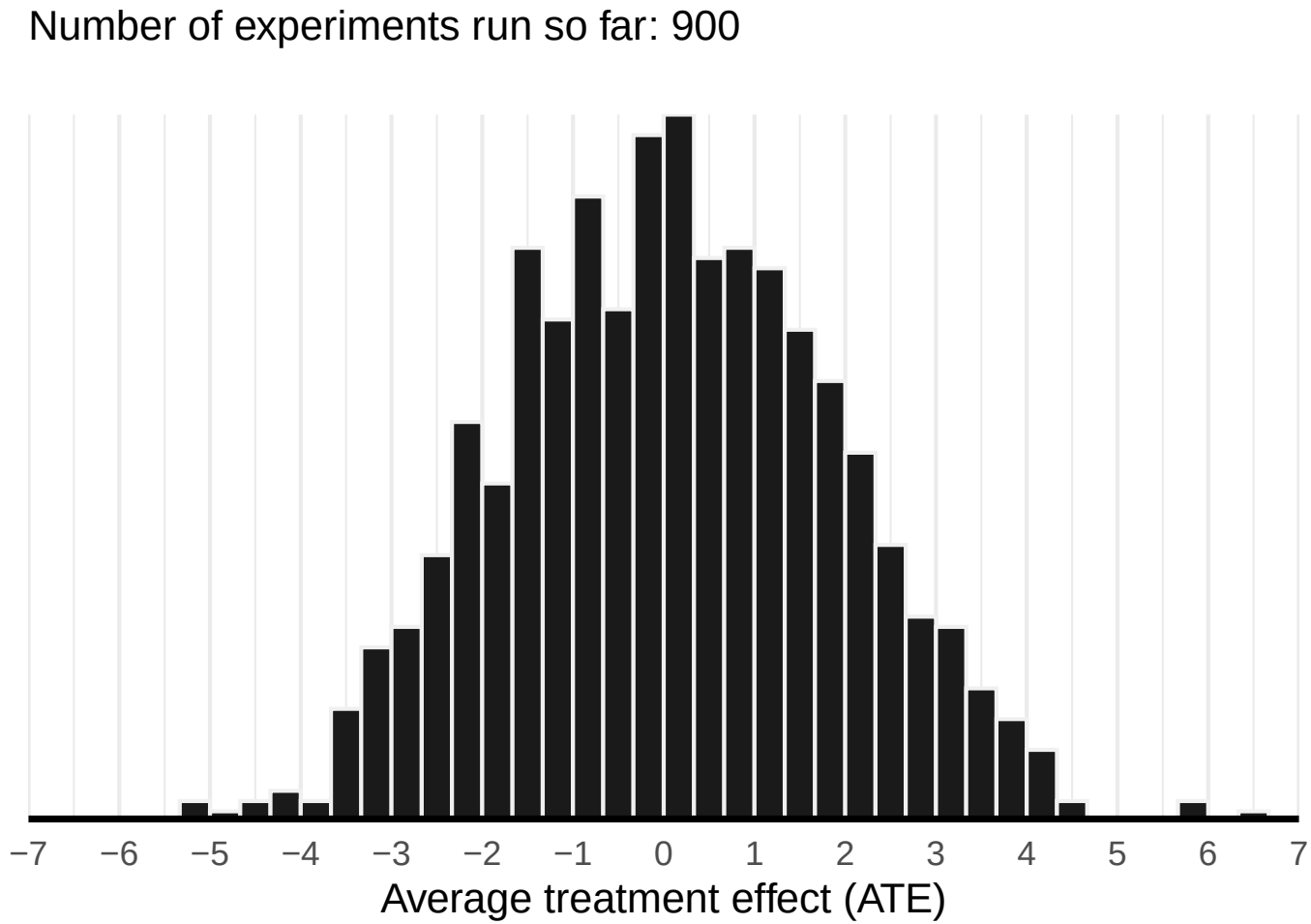
Number of experiments run so far: 700



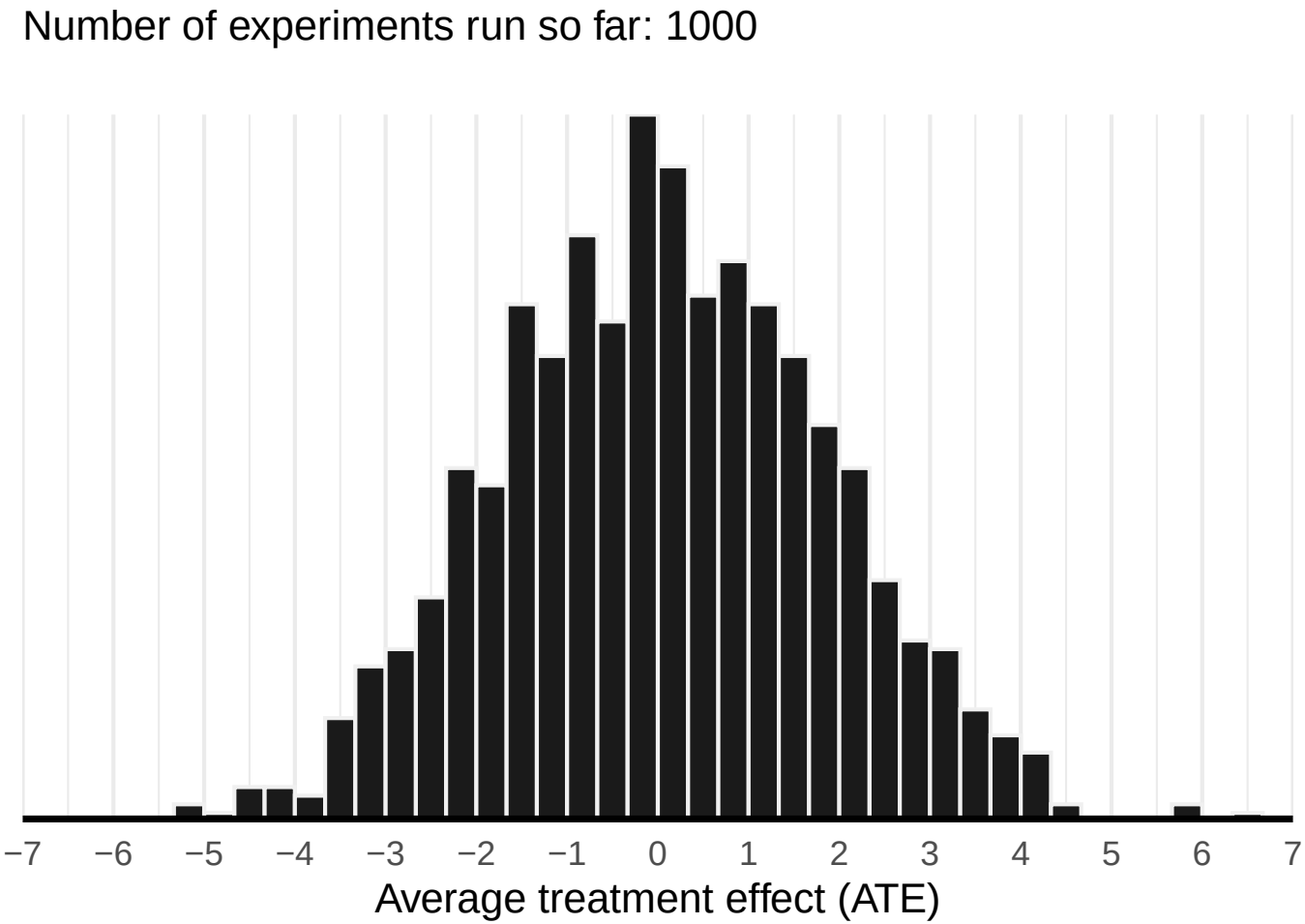
Let's say watching the debate truly has no effect. What might we see?



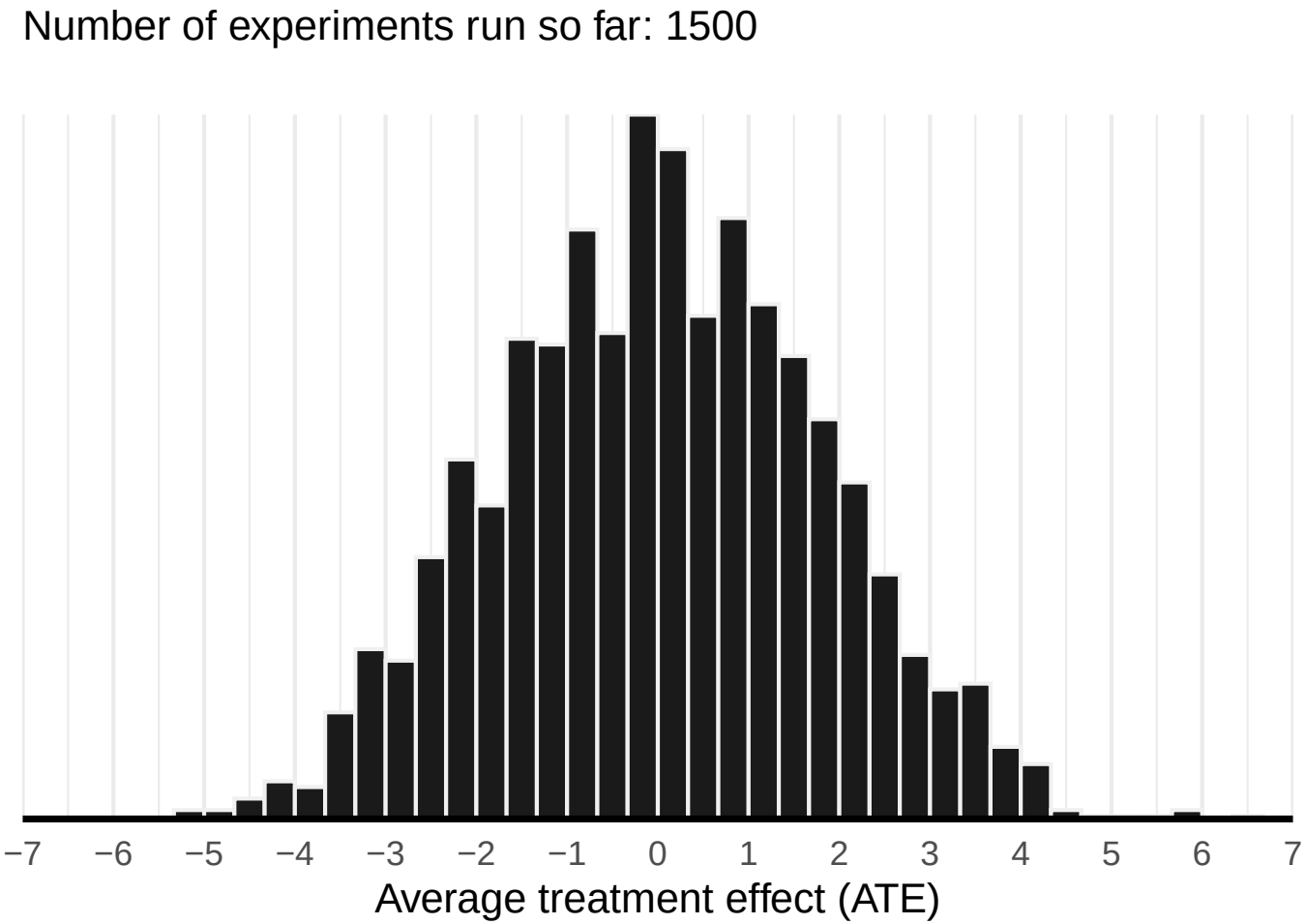
Let's say watching the debate truly has no effect. What might we see?



Let's say watching the debate truly has no effect. What might we see?

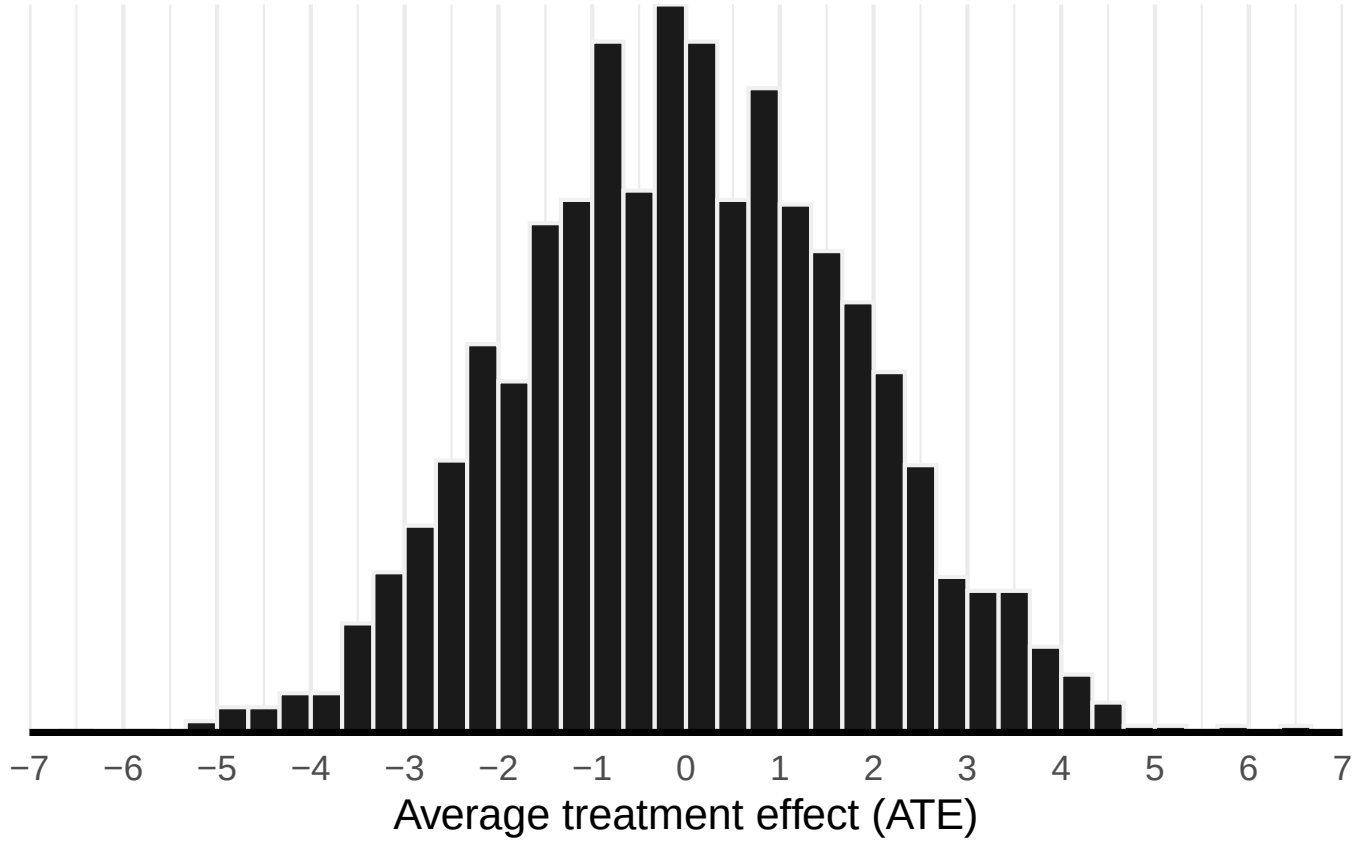


Let's say watching the debate truly has no effect. What might we see?



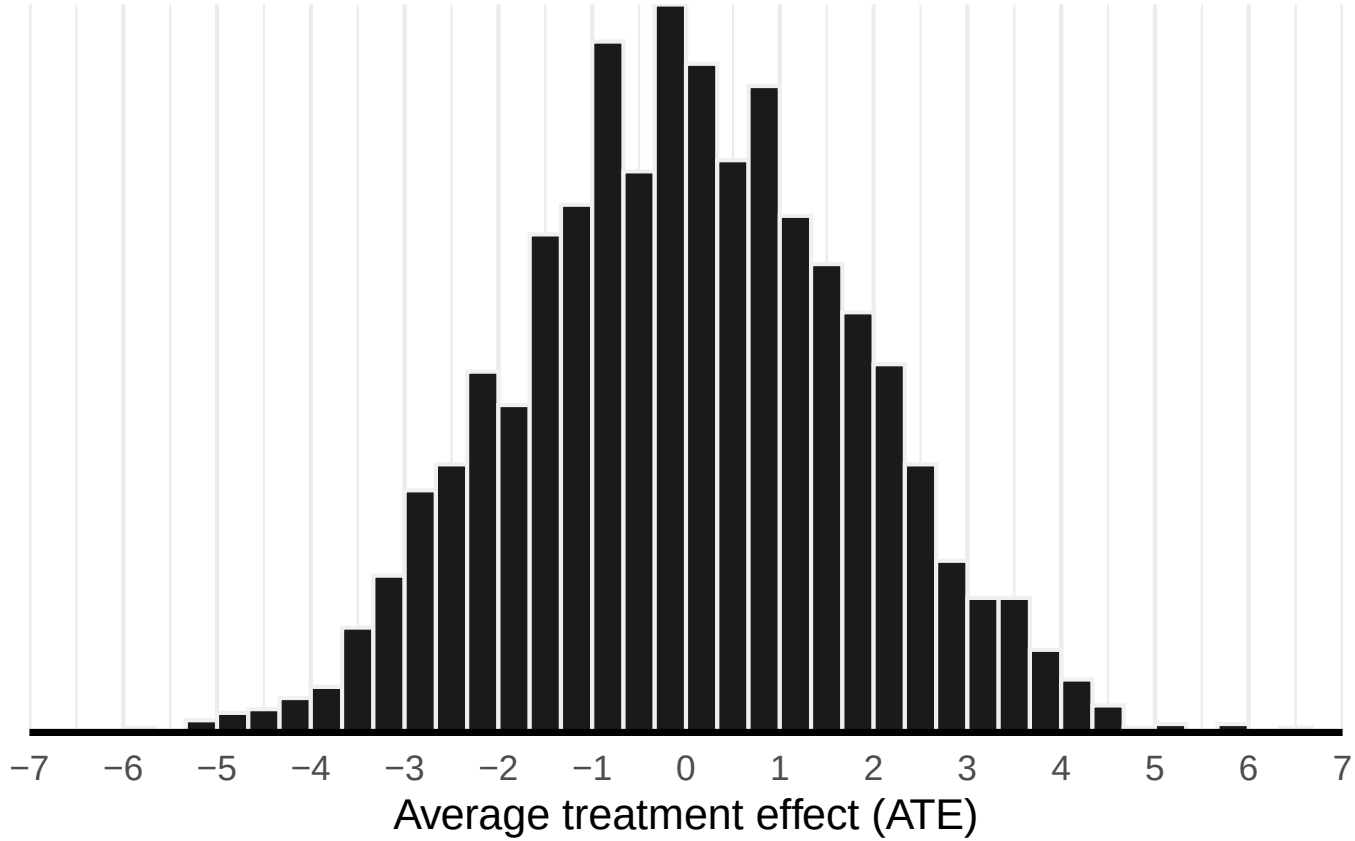
Let's say watching the debate truly has no effect. What might we see?

Number of experiments run so far: 2000

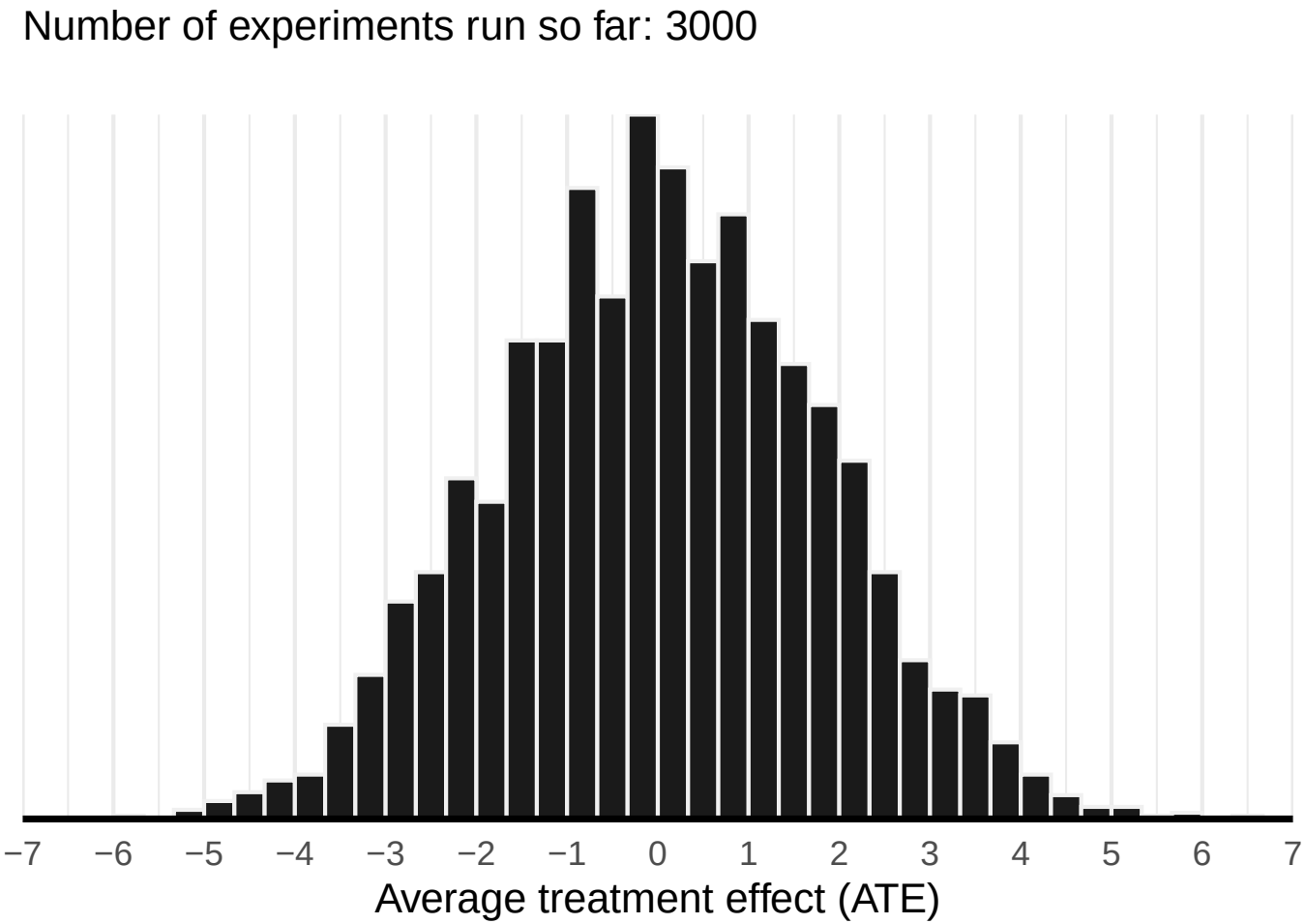


Let's say watching the debate truly has no effect. What might we see?

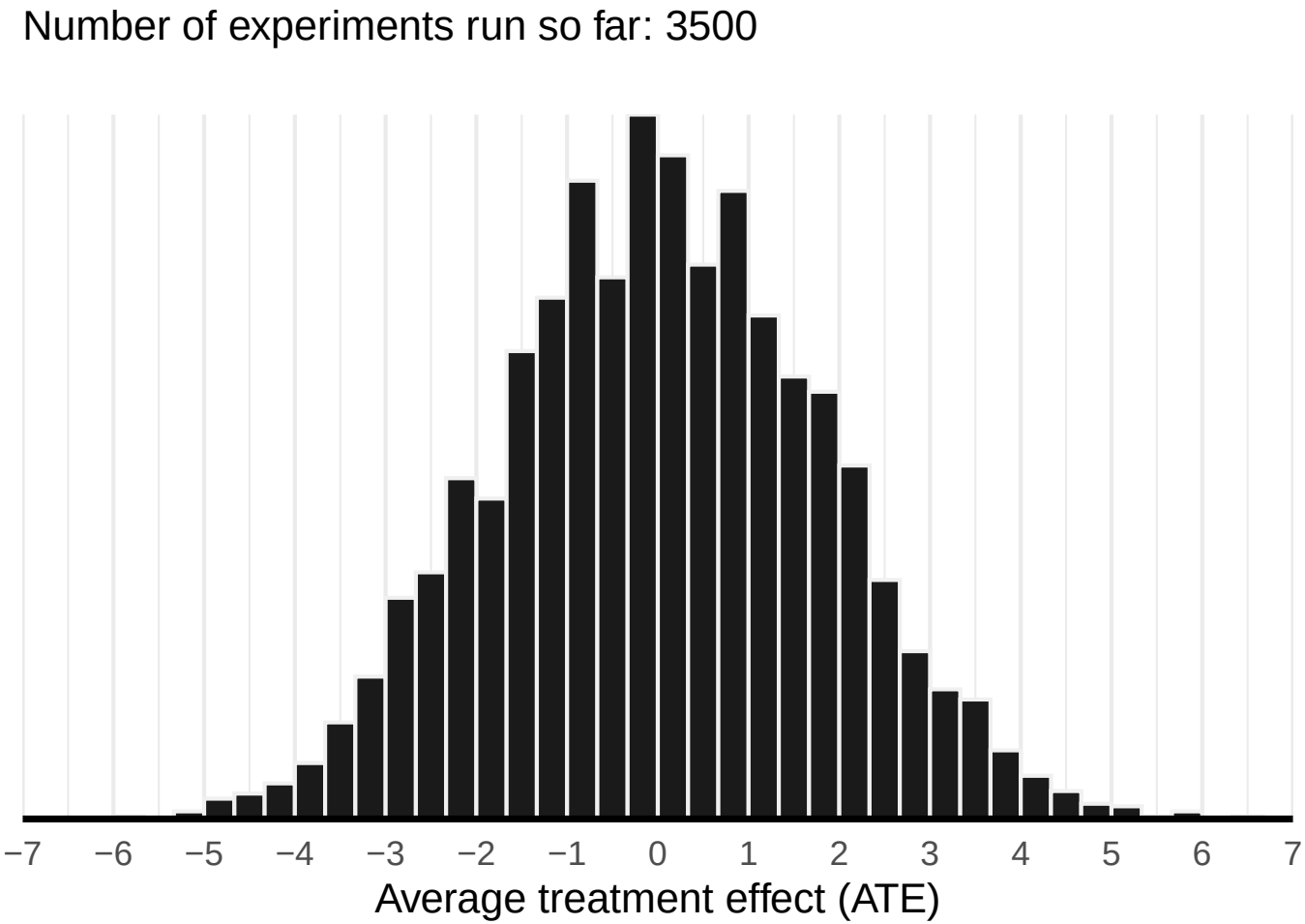
Number of experiments run so far: 2500



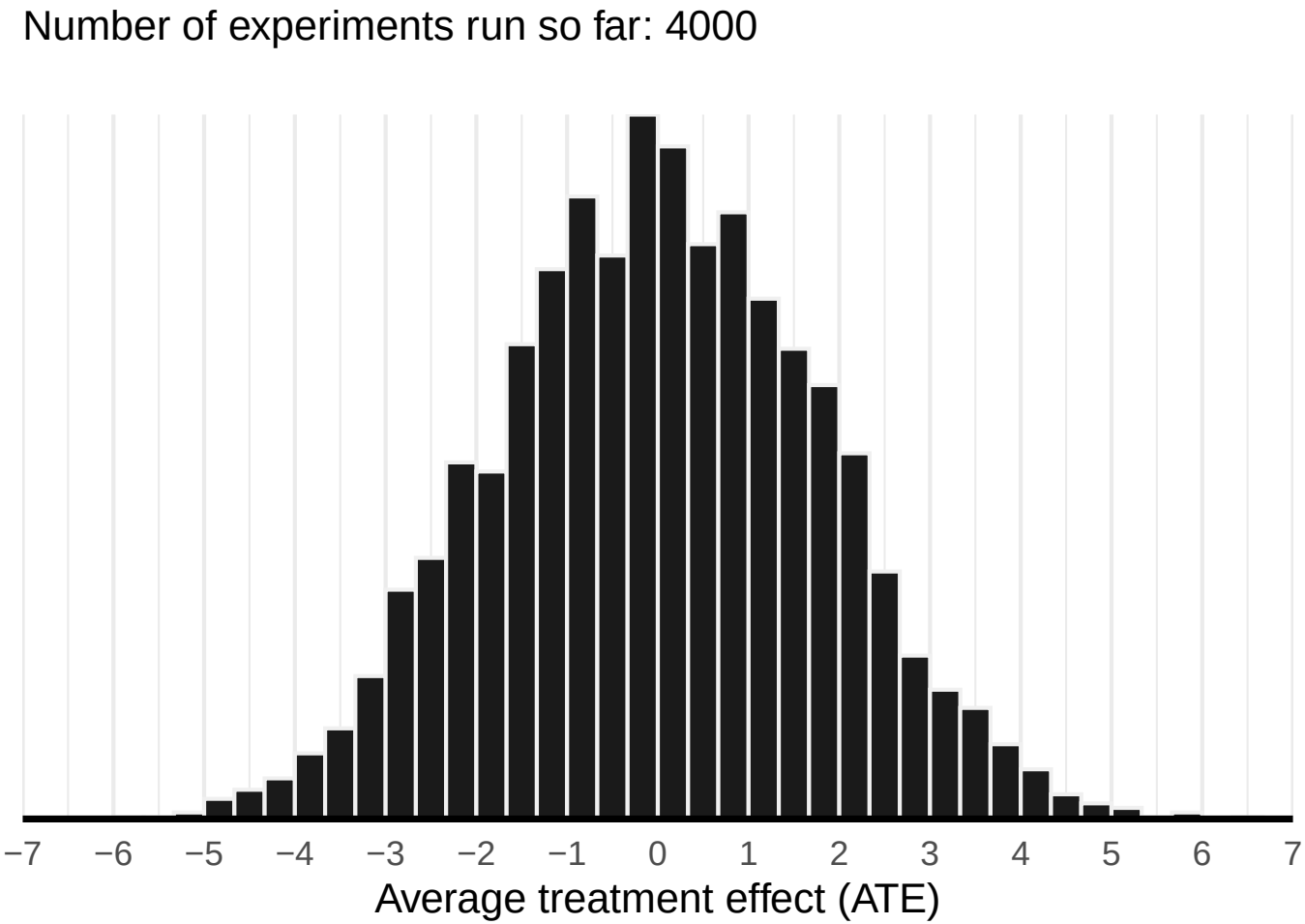
Let's say watching the debate truly has no effect. What might we see?



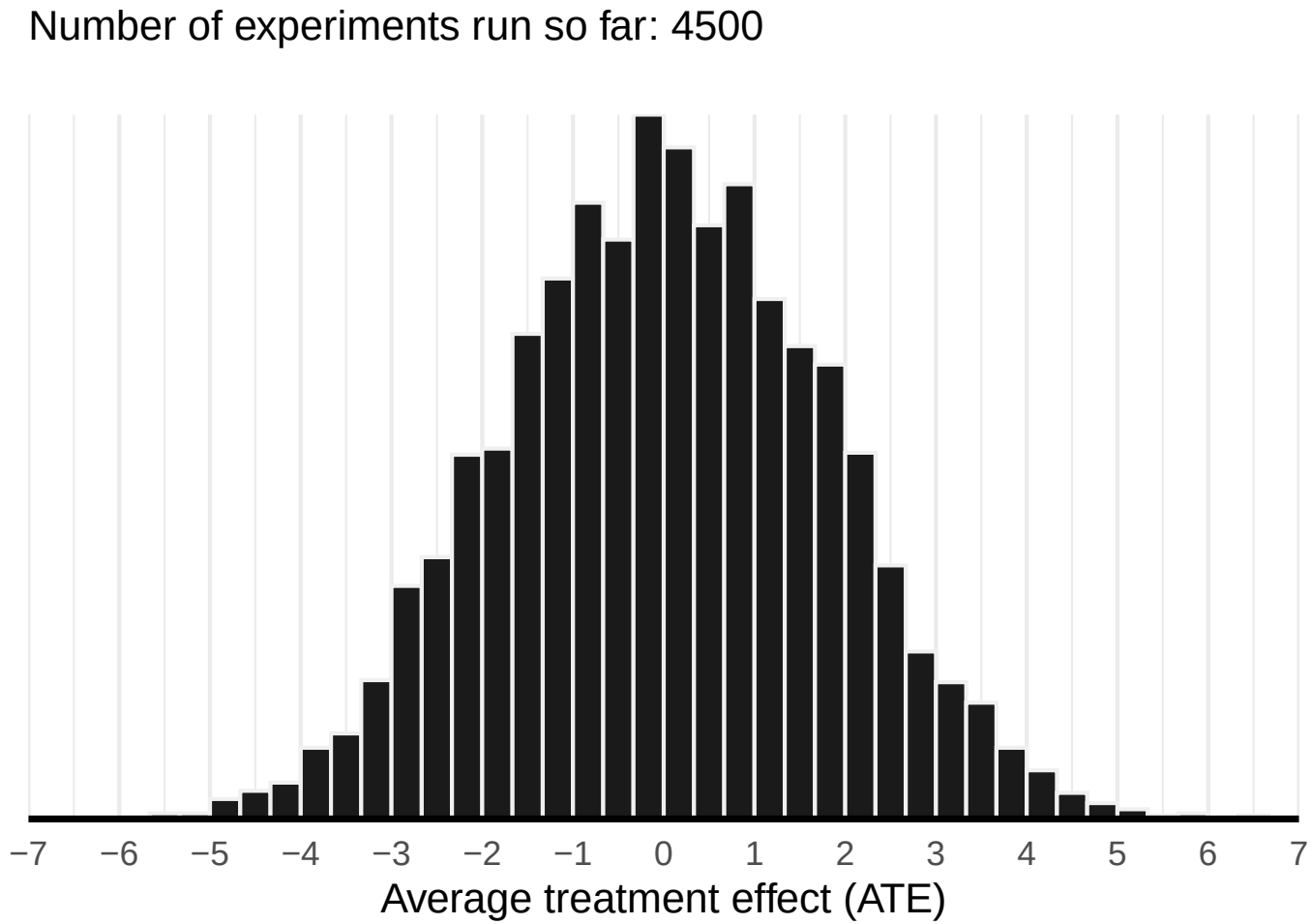
Let's say watching the debate truly has no effect. What might we see?



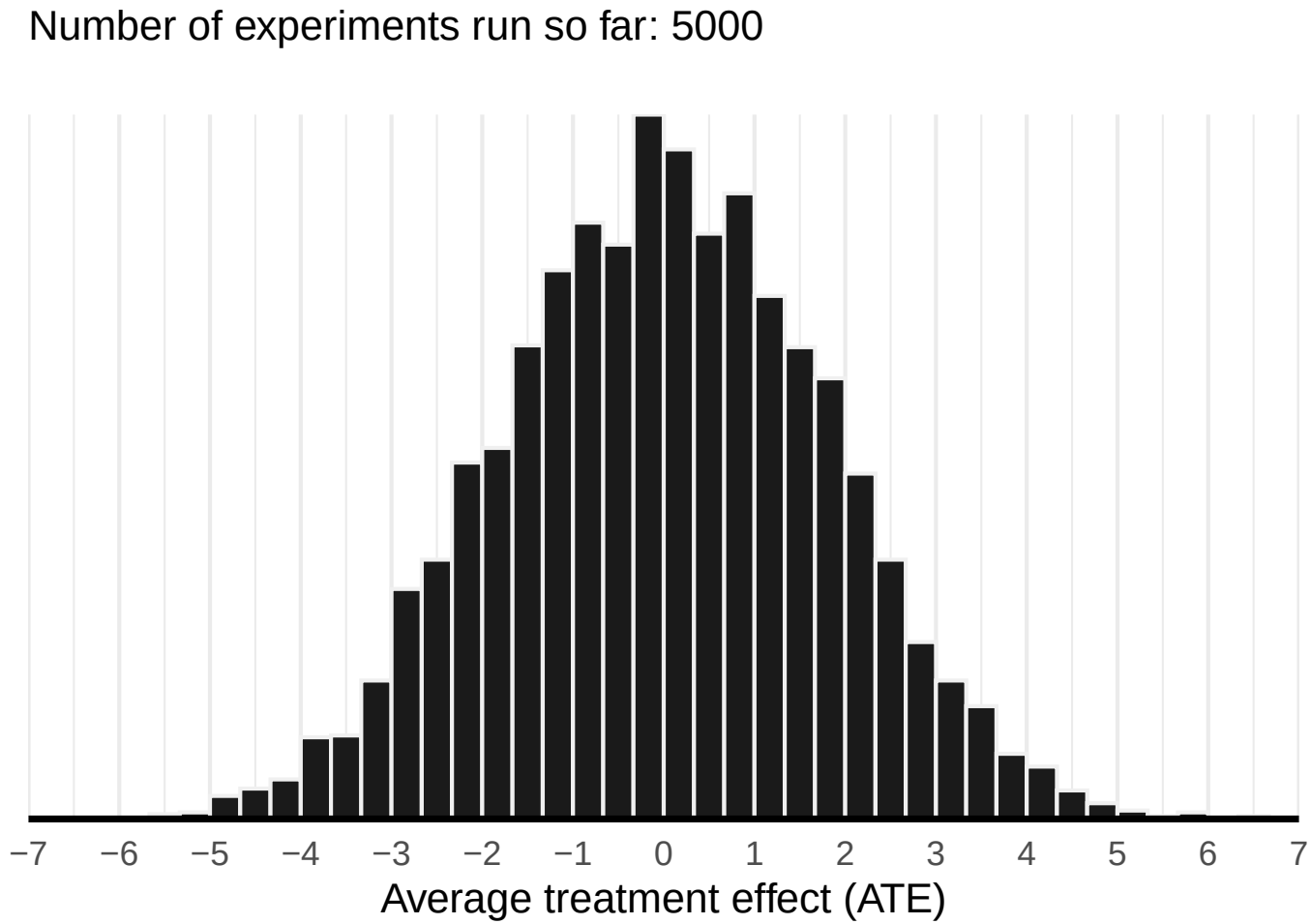
Let's say watching the debate truly has no effect. What might we see?



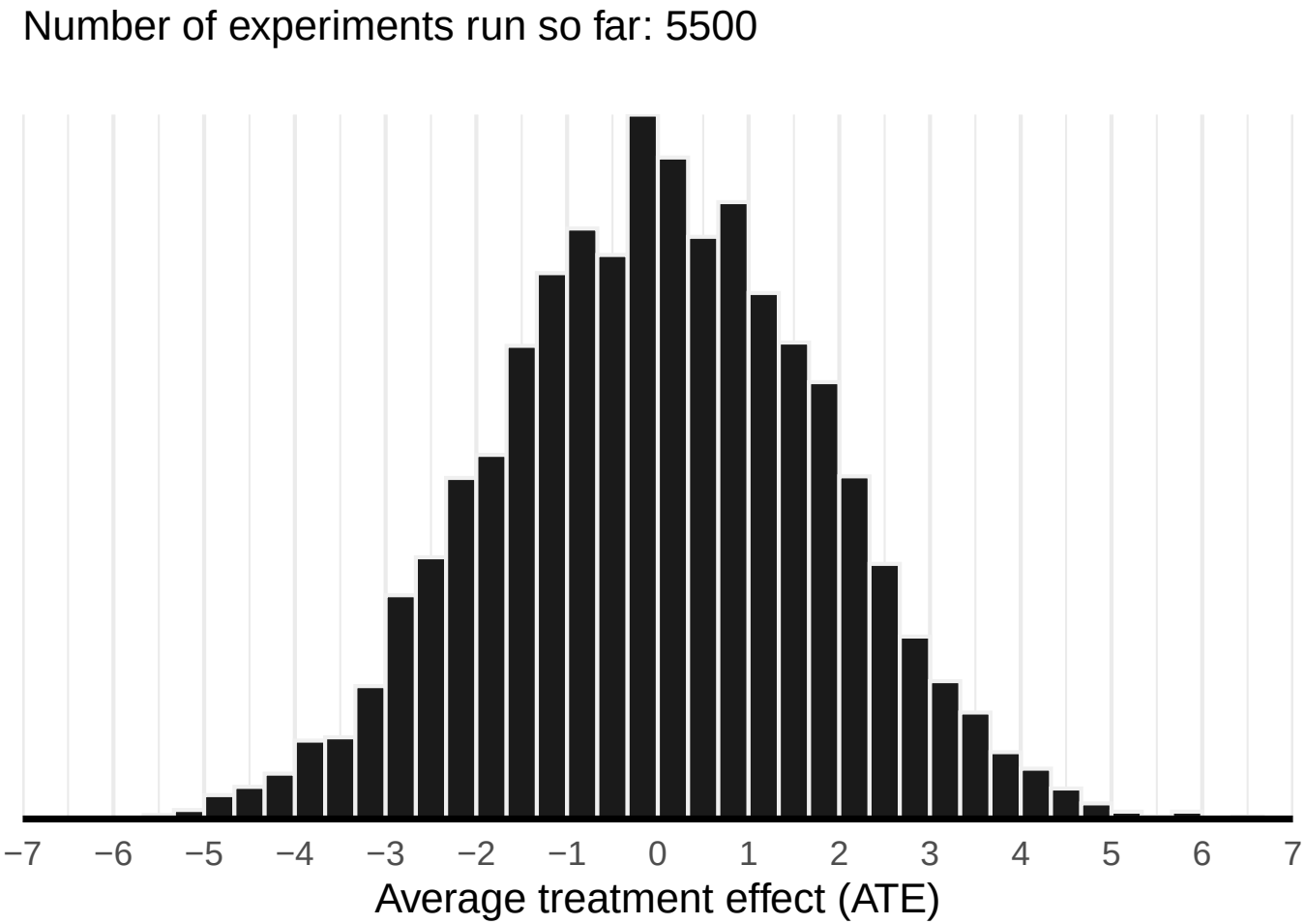
Let's say watching the debate truly has no effect. What might we see?



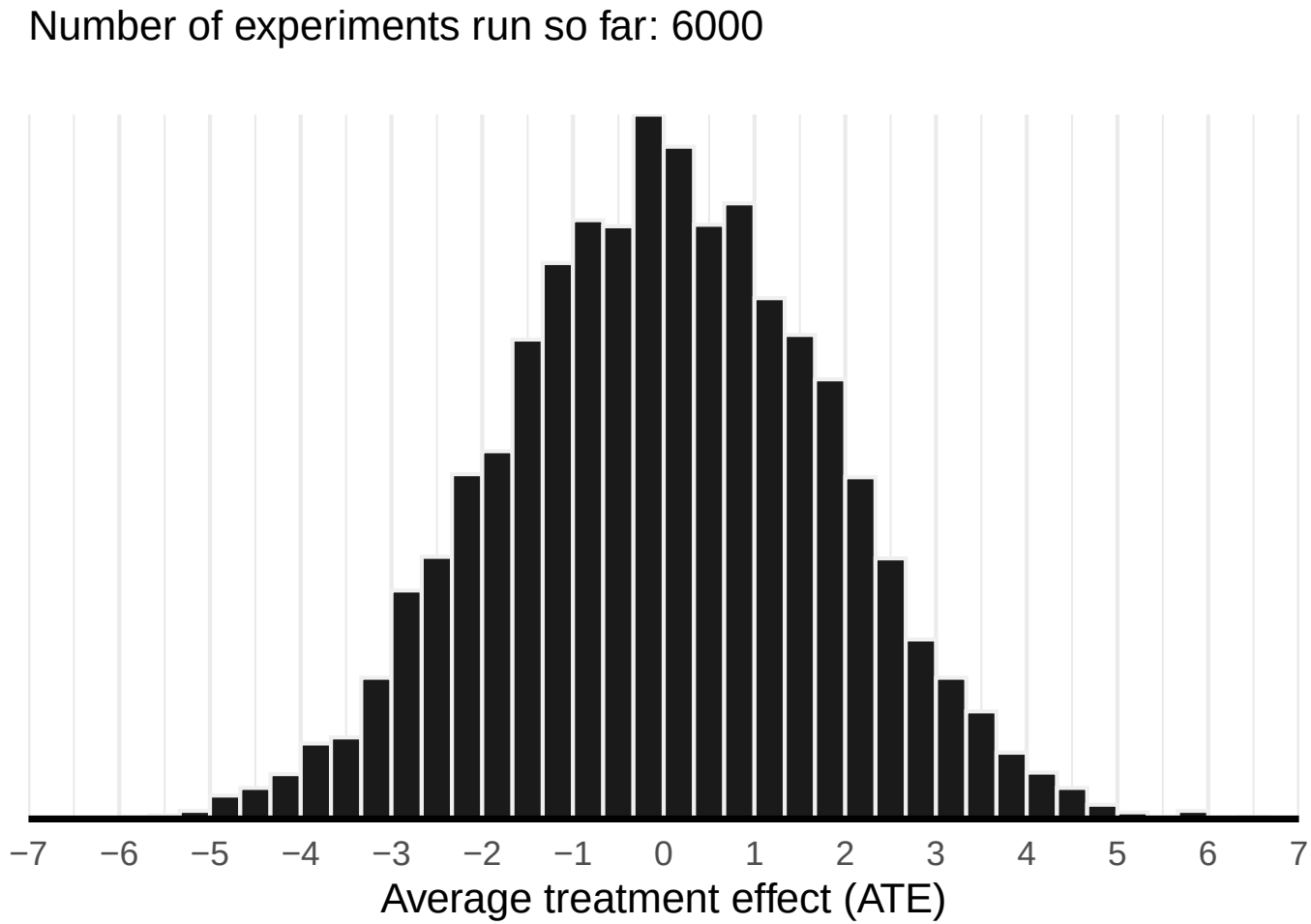
Let's say watching the debate truly has no effect. What might we see?



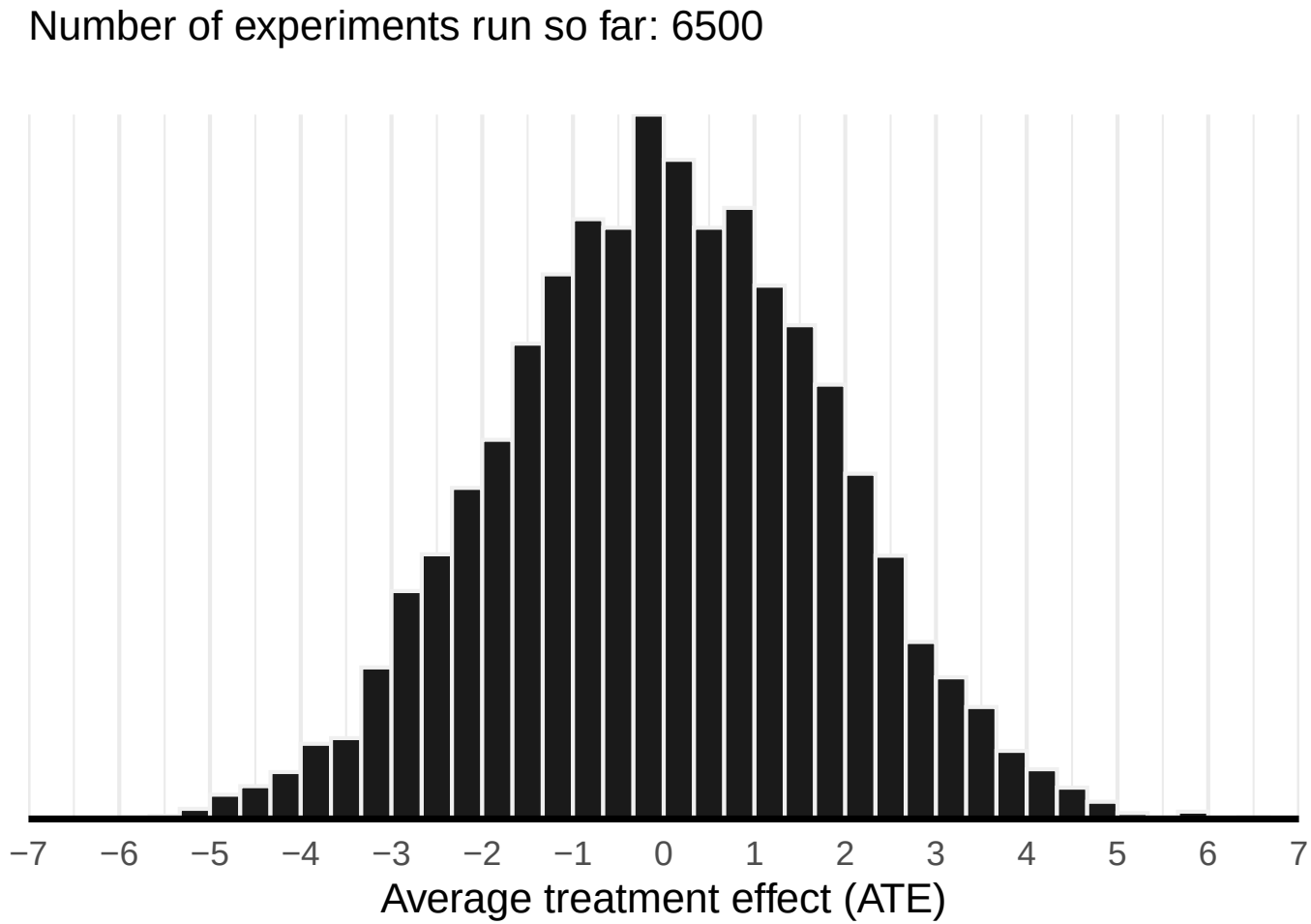
Let's say watching the debate truly has no effect. What might we see?



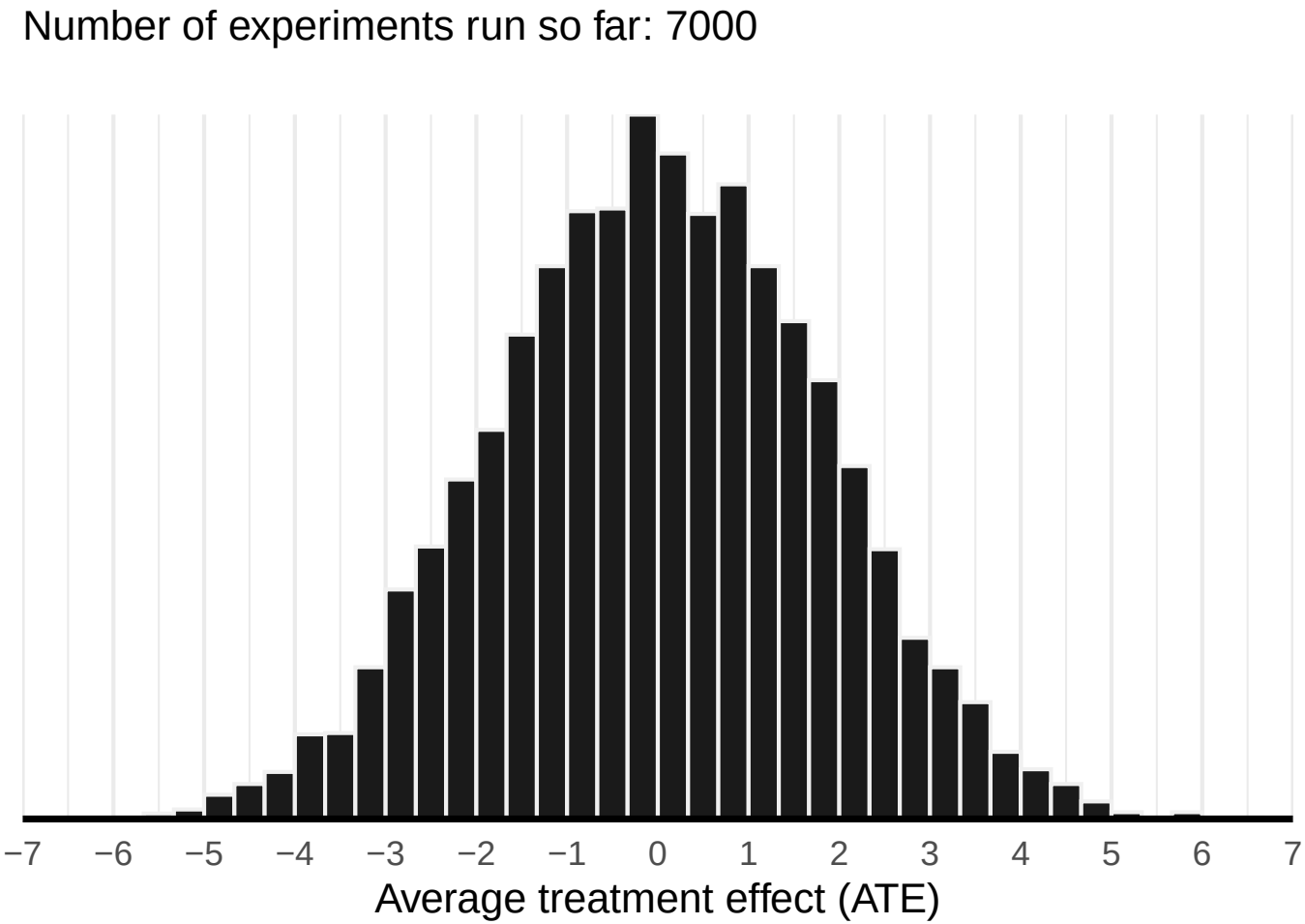
Let's say watching the debate truly has no effect. What might we see?



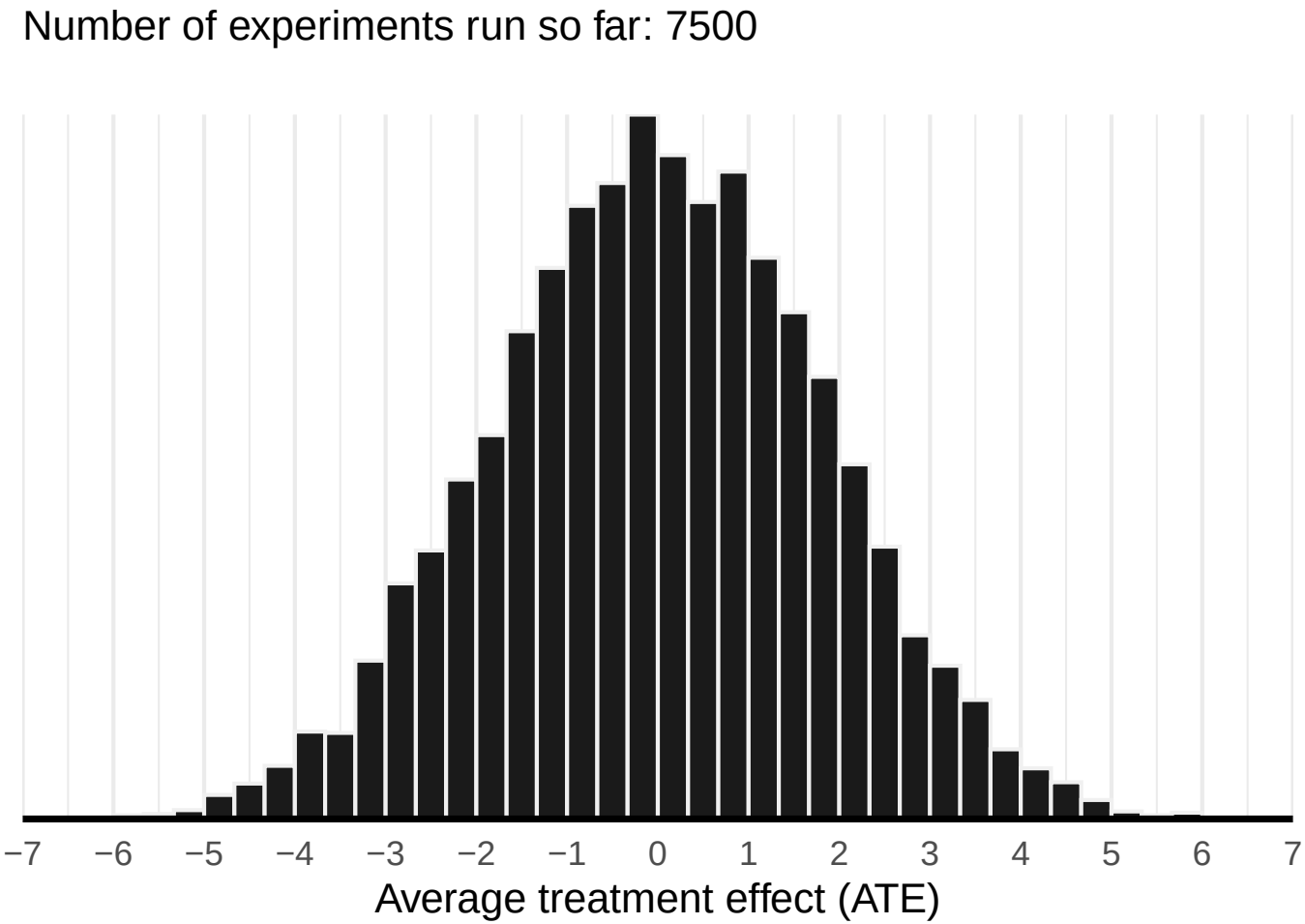
Let's say watching the debate truly has no effect. What might we see?



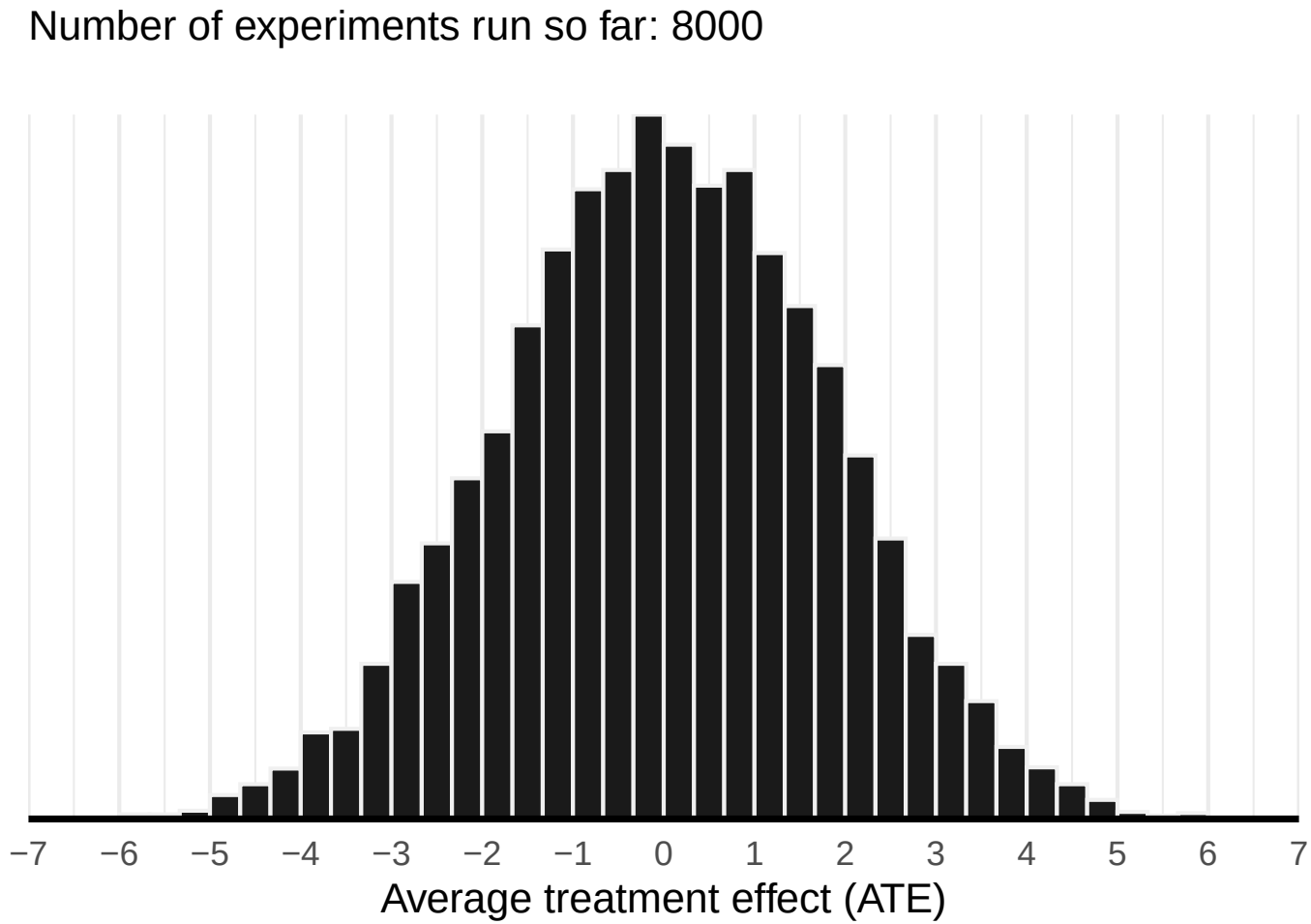
Let's say watching the debate truly has no effect. What might we see?



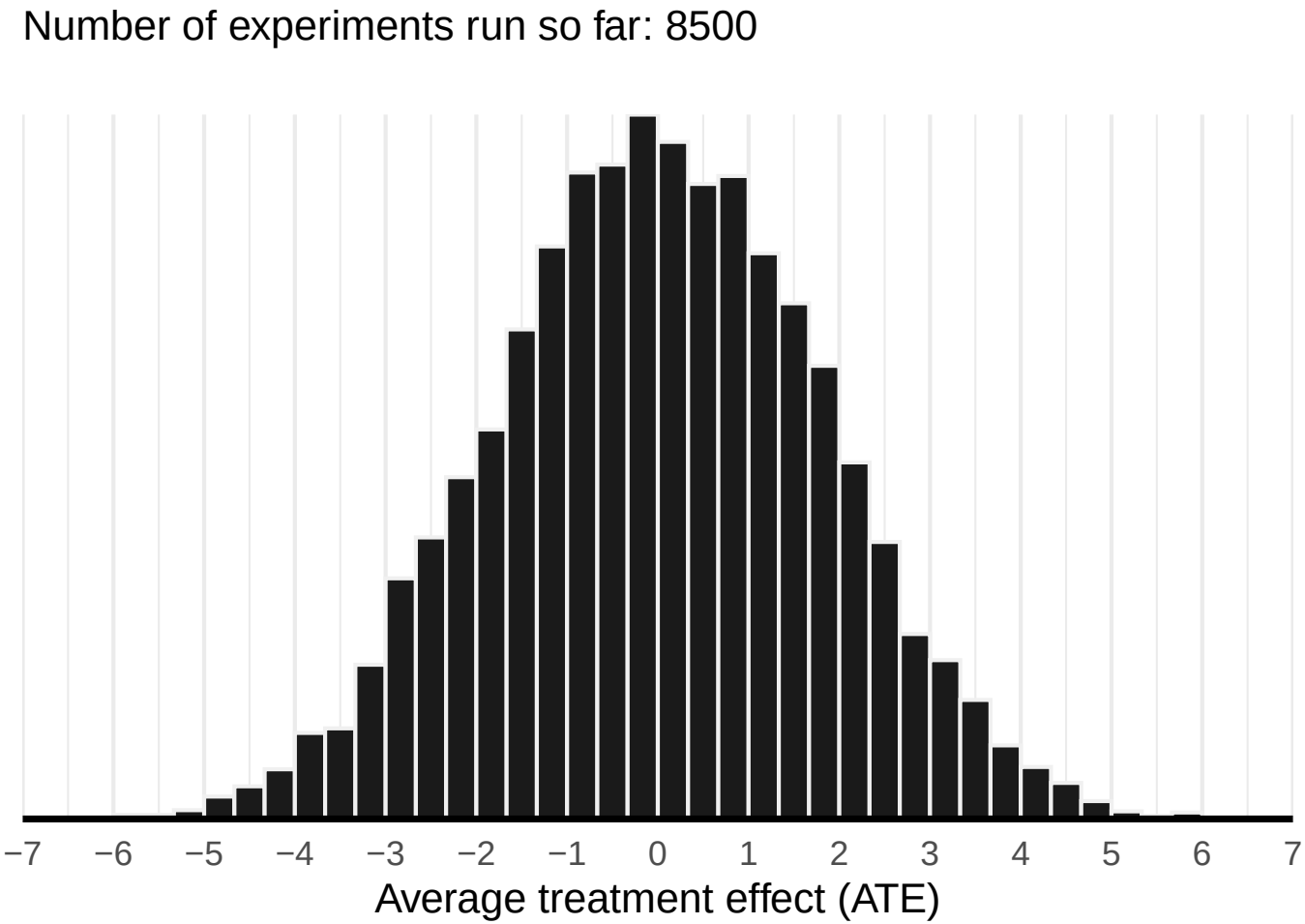
Let's say watching the debate truly has no effect. What might we see?



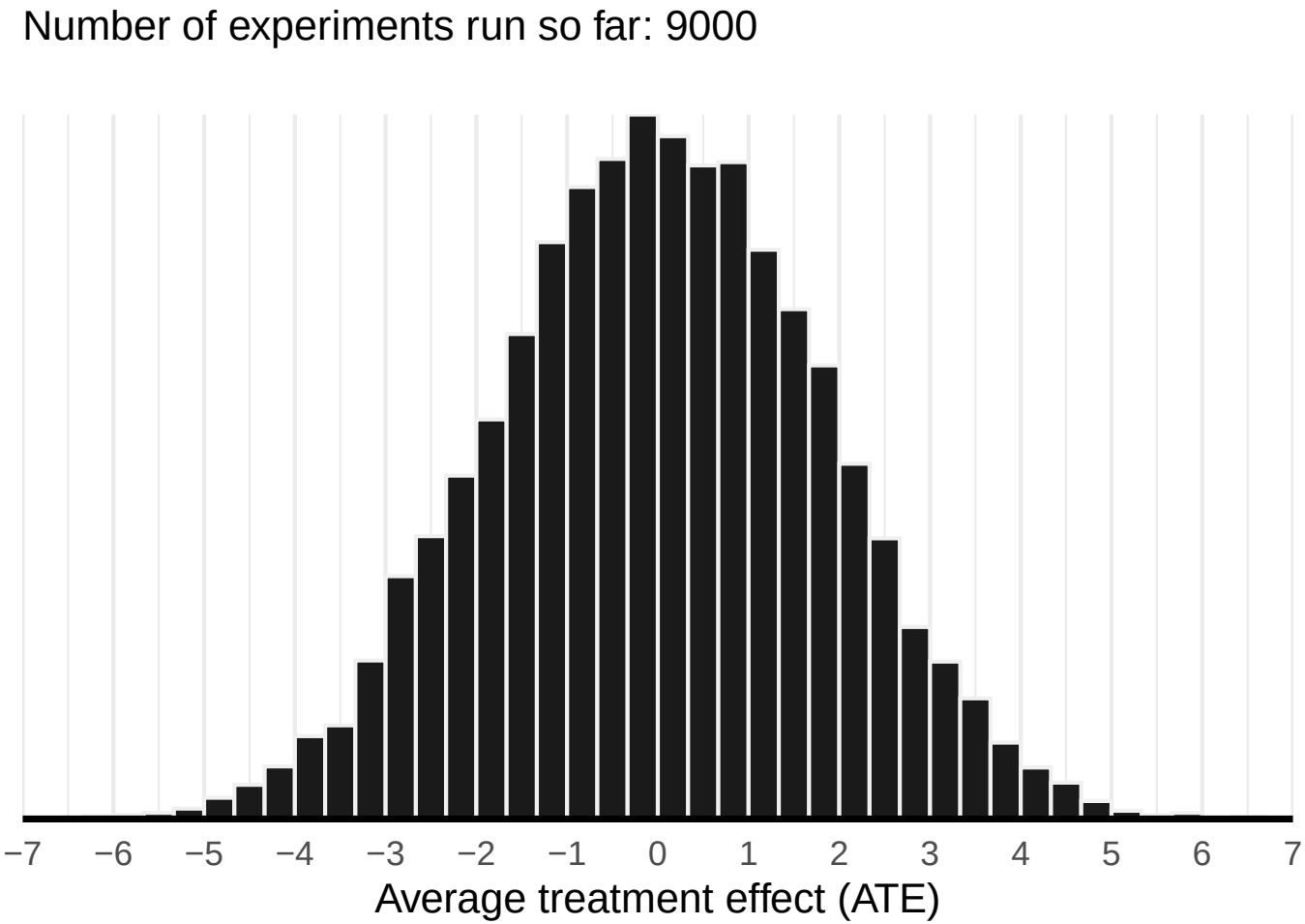
Let's say watching the debate truly has no effect. What might we see?



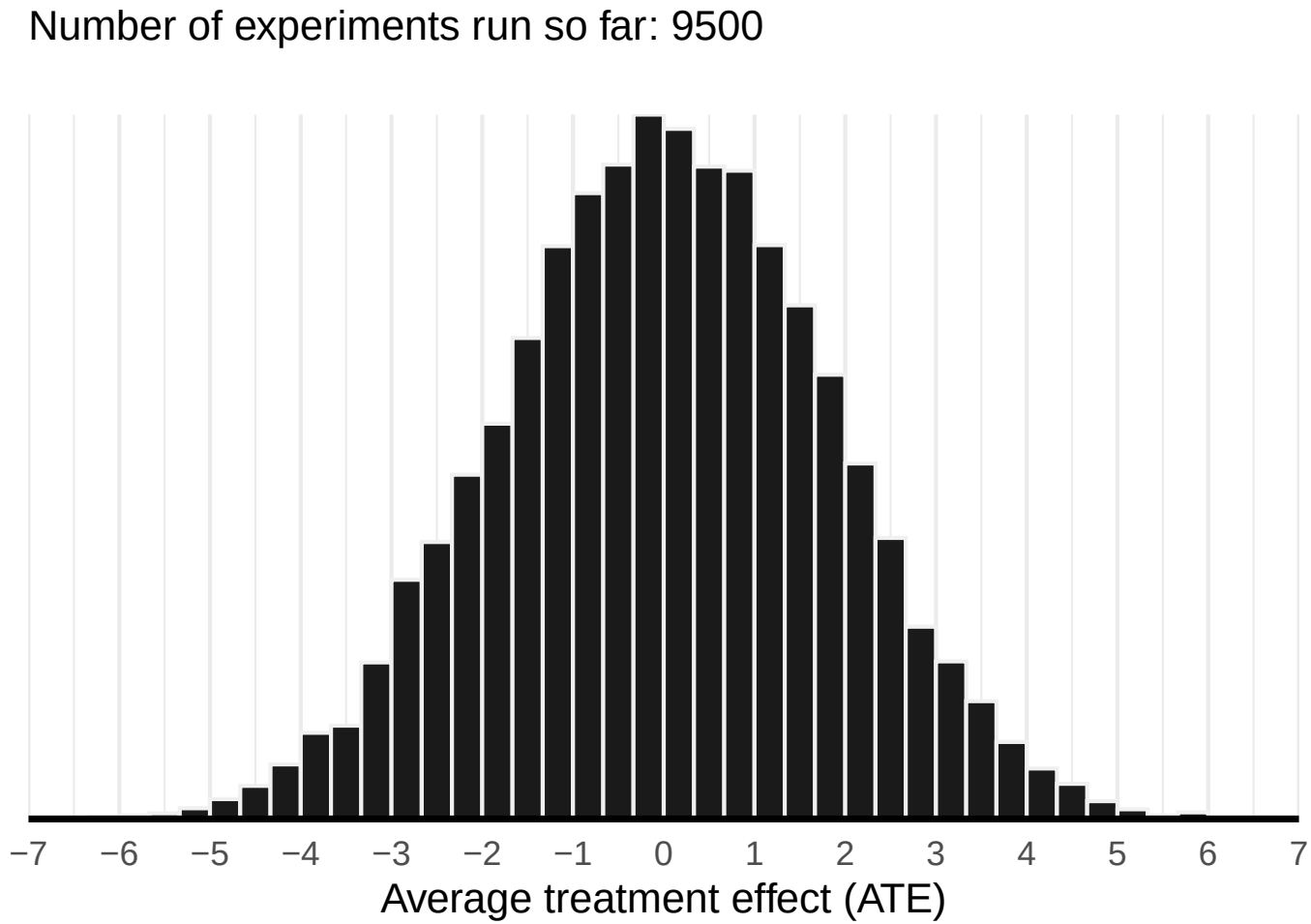
Let's say watching the debate truly has no effect. What might we see?



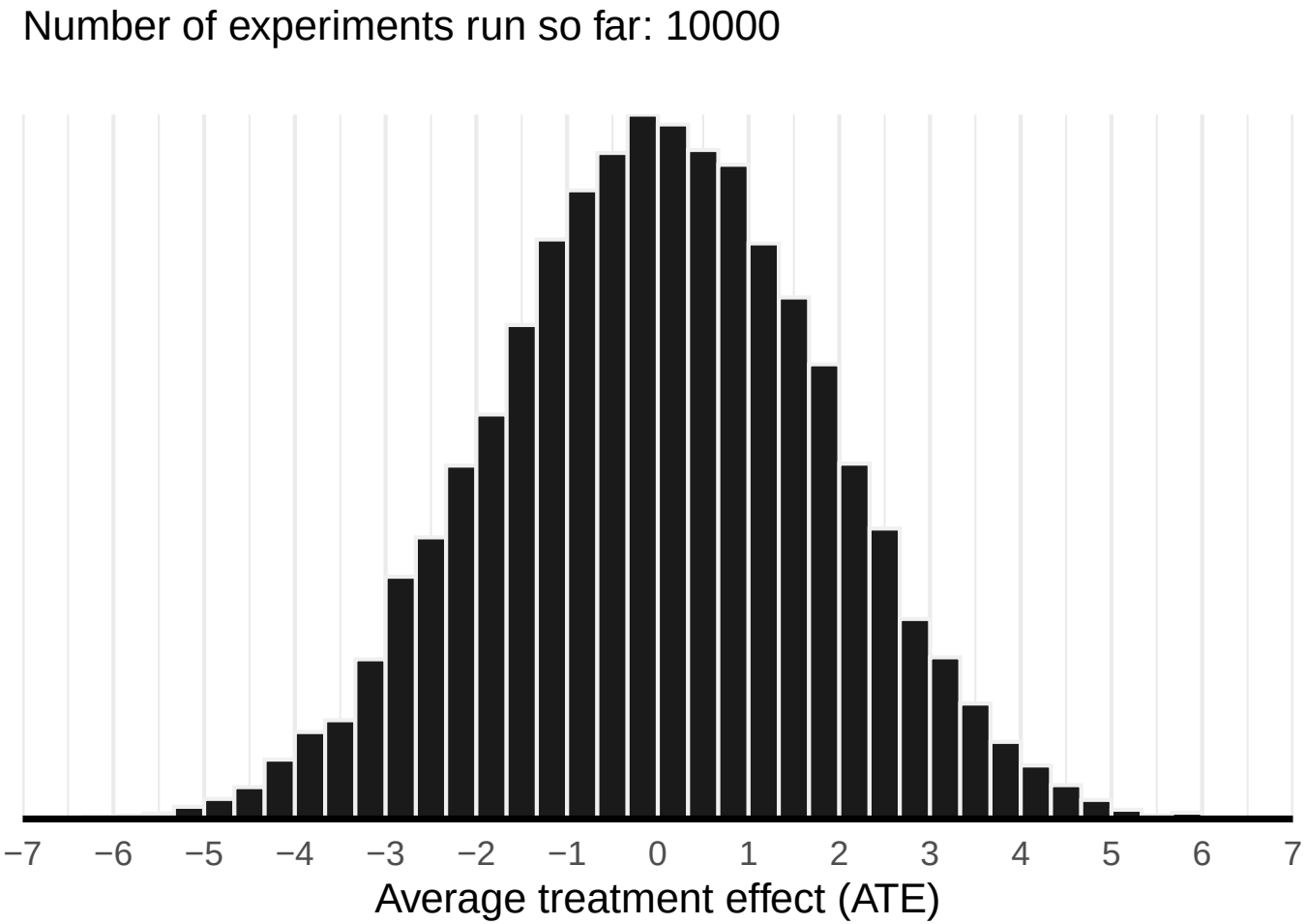
Let's say watching the debate truly has no effect. What might we see?



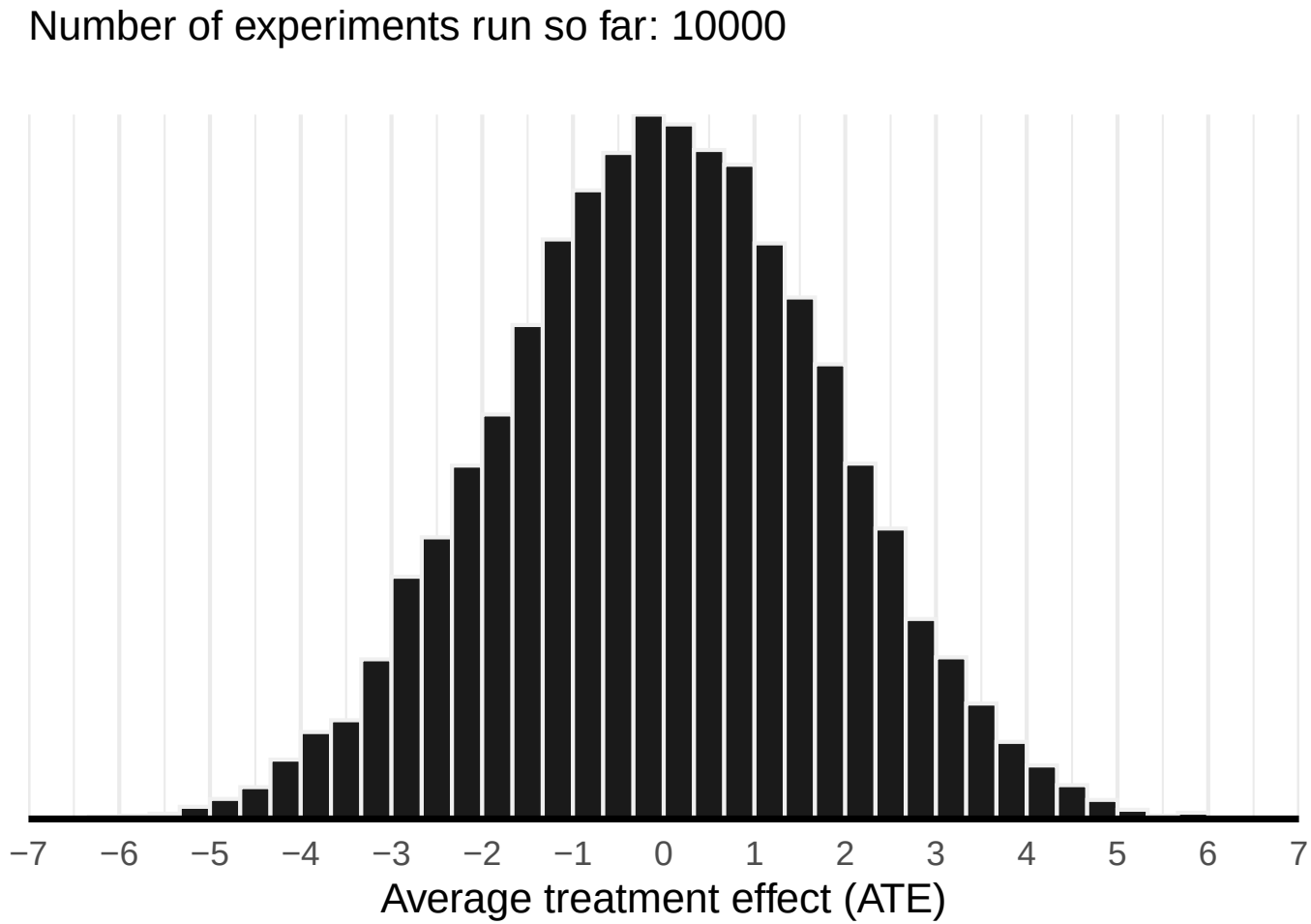
Let's say watching the debate truly has no effect. What might we see?



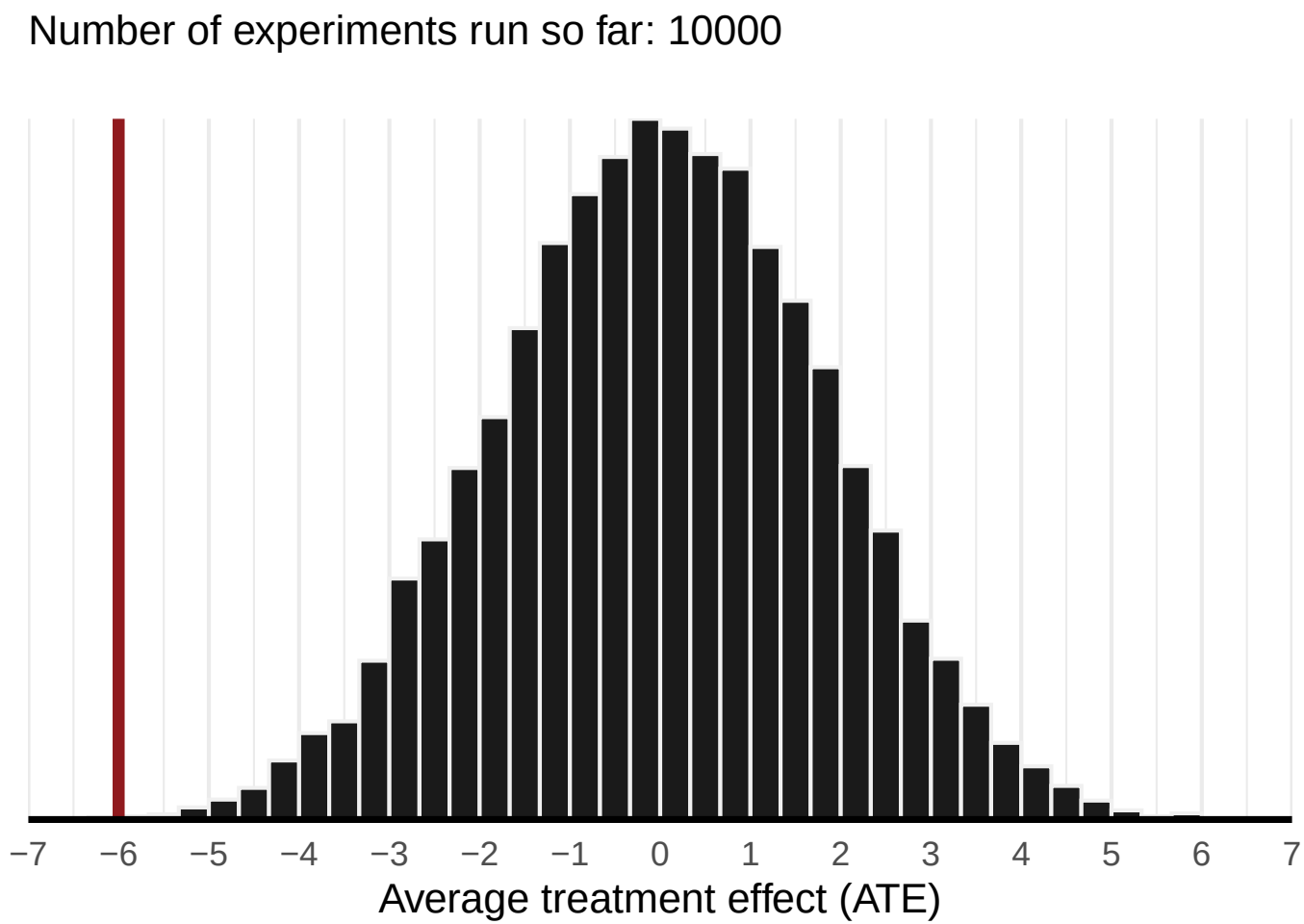
Let's say watching the debate truly has no effect. What might we see?



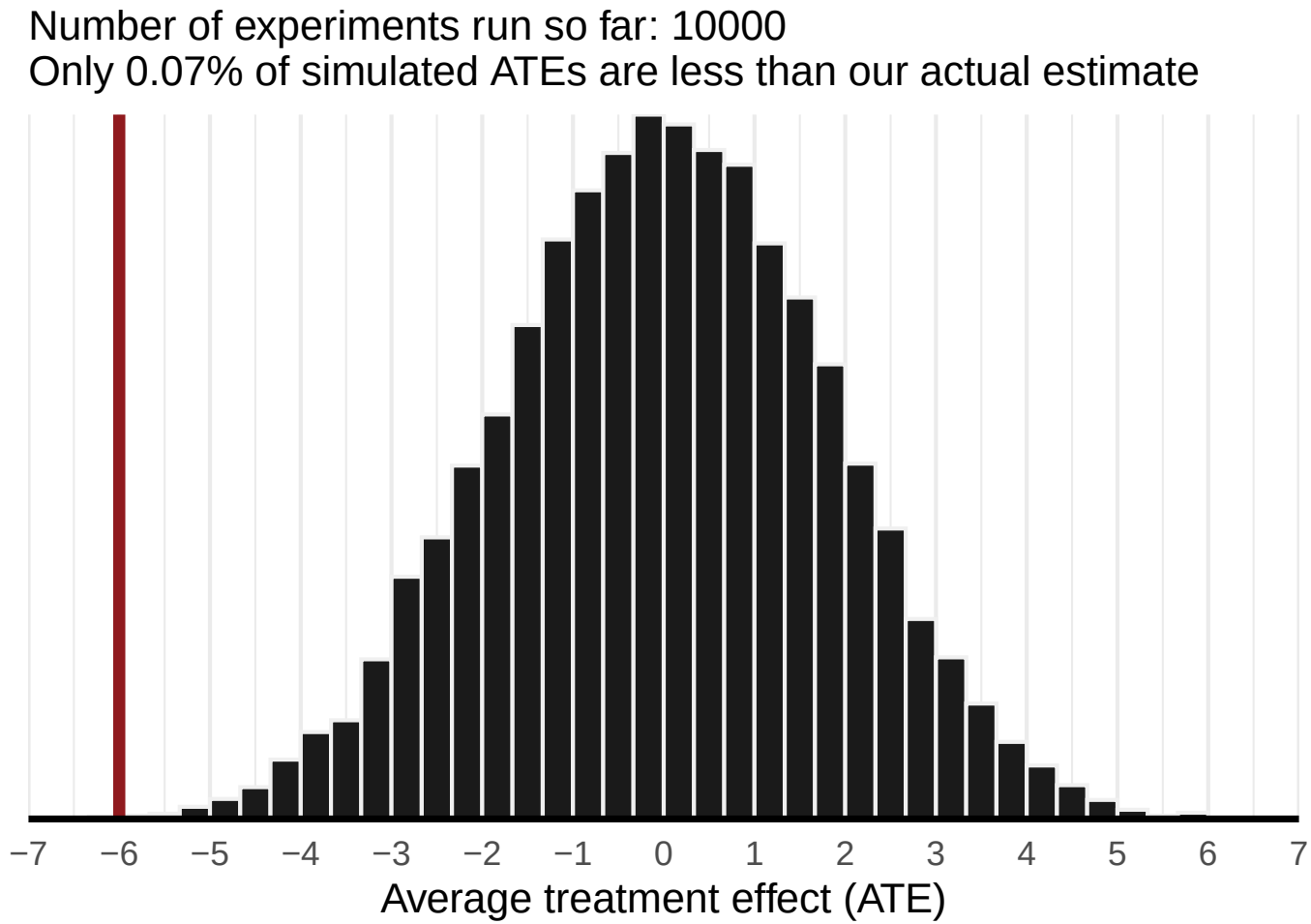
But let's say we actually ran this experiment. And we get the following...



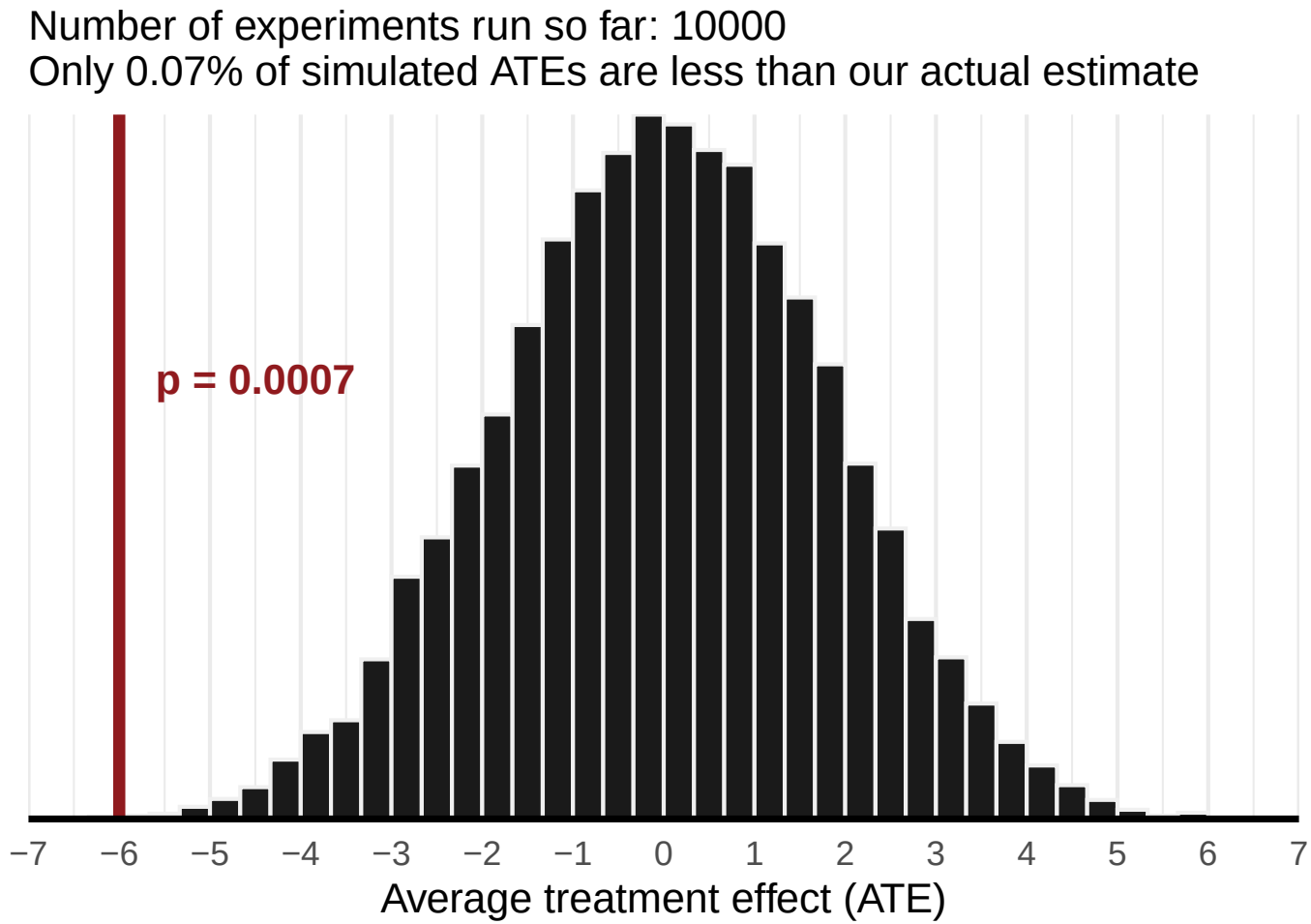
But let's say we actually ran this experiment. And we get the following...



But let's say we actually ran this experiment. And we get the following...



But let's say we actually ran this experiment. And we get the following...



We typically assume that the normal distribution is sufficient, but:

- This can requires a relatively large sample size
- Can fall apart if high-leverage outcomes i.e. outliers (Young 2019)
- Are also cases where it's unclear how to get a standard error or p-value

Fortunately, we can calculate a p-value more robustly using randomization inference

- Randomization inference assumes that our null hypothesis is "sharp"
  - A sharp null hypothesis is that the treatment effect is zero for everyone in the sample, not just on average
- If so, then the outcome variable is going to be the same value for each person regardless of whether they are assigned to the treatment or control condition

What does a sharp null hypothesis look like?

Village	Budget Share Leader = Female	Budget Share Leader = Male	Treatment effect
1	15	15	0
2	15	15	0
3	20	20	0
4	20	20	0
5	10	10	0
6	15	15	0
7	30	30	0
Average	17.9	17.9	0

If the effect is zero for every individual, then:

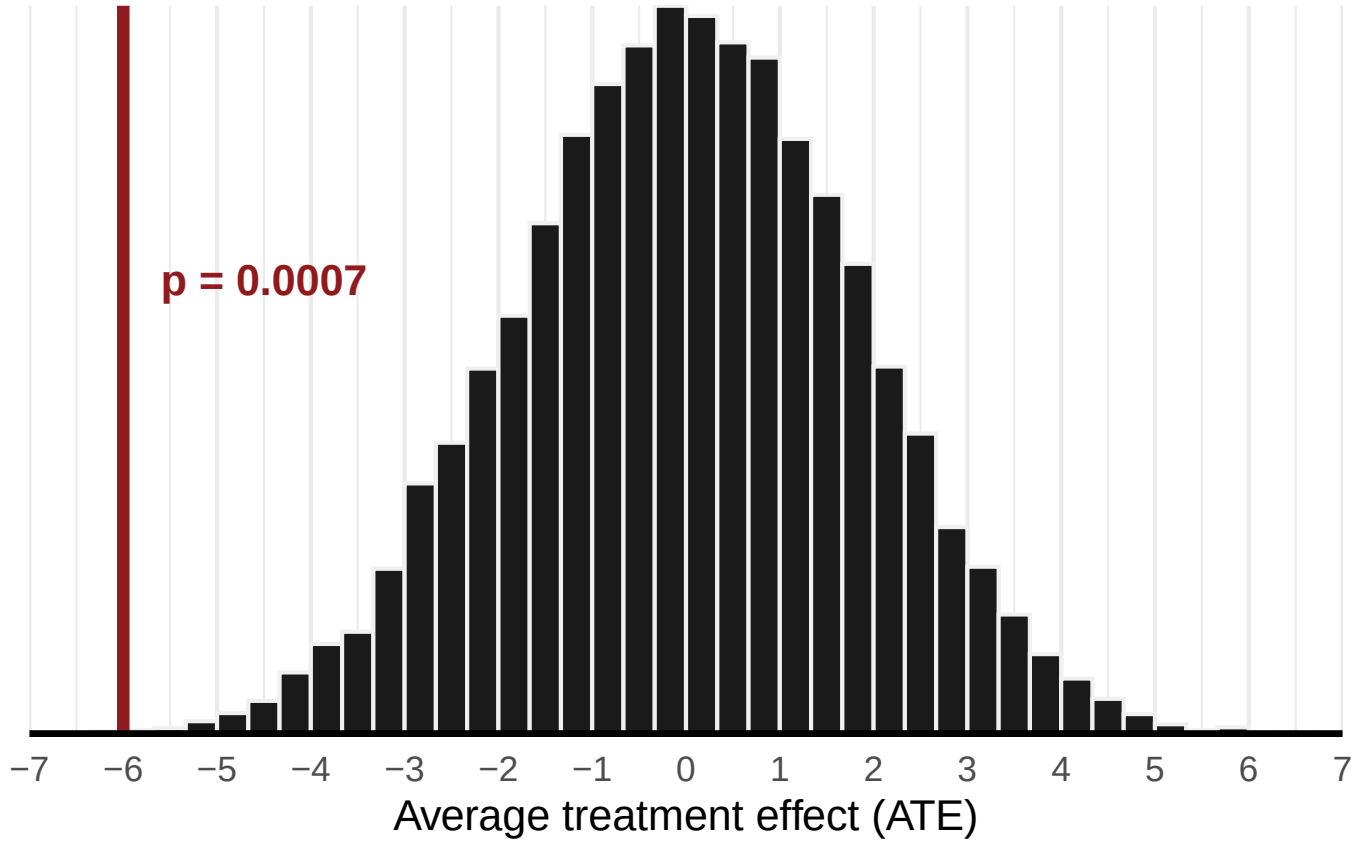
- We can randomly re-assign treatment and control as much as we want and the "effect" will always be the same
- Therefore, we can randomly assign treatment and control in our data and calculate an effect size estimate over and over to get the sampling distribution of the null effect distribution
- In fact, we can get the "exact" distribution because there is a finite number of possible treatment and control assignments
  - This is often impractical, so we just simulate instead

Applying to our simple 7-village case:

		Randomization				
Village	**Budget share**	**A**	**B**	**C**	**D**	$\cdots$
1	15	1	1	1	0	$\cdots$
2	15	1	0	0	0	$\cdots$
3	20	0	0	0	0	$\cdots$
4	20	1	1	0	1	$\cdots$
5	10	0	1	1	1	$\cdots$
6	15	0	0	1	1	$\cdots$
7	30	1	1	1	1	$\cdots$
τ		5	2.08	-0.83	2.08	$\cdots$

Randomization inference is the simulation that I showed earlier

Number of experiments run so far: 10000
Only 0.07% of simulated ATEs are less than our actual estimate



Randomization inference

- Flexible, non-parametric approach to calculating a p-value
  - It makes no assumptions about the sampling distribution of your experimental effect
- We cannot do an 'exact' test in practice, unless have a *very* small sample
- Our solution: Just simulate it instead
  - We can do this manually with some pretty simple coding in R, or can use an R library like `ri2`

Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*

Alwyn Young

The Quarterly Journal of Economics, Volume 134, Issue 2, May 2019, Pages 557–598,
https://doi.org/10.1093/qje/qjy029
Published: 21 November 2018

Abstract

I follow R. A. Fisher's,*The Design of Experiments* (1935), using randomization statistical inference to test the null hypothesis of no treatment effects in a comprehensive sample of 53 experimental papers drawn from the journals of the American Economic Association. In the average paper, randomization tests of the significance of individual treatment effects find 13% to 22% fewer significant results than are found using authors' methods. In joint tests of multiple treatment effects appearing together in tables, randomization tests yield 33% to 49% fewer statistically significant results than conventional tests. Bootstrap and jackknife methods support and confirm the randomization results.

Non-compliance in experiments

- What do we do if research subjects don't follow through with a treatment?
- Examples:
  - Effect of door-to-door get-out-the-vote (GOTV) campaign: not everyone answers the door
  - Effect of watching an election debate on vote choice: not everyone asked to watch will do so
  - Effect of providing a VPN on access to foreign news beyond China's Great Firewall: not everyone installs the VPN
  - Effect of giving tips about how to spot fake news on actually spotting fake news: not everyone reads the tips

Problem

- Without very strong assumptions, we can't estimate the average treatment effect (ATE)
  - i.e. that the effect of a treatment on *non*-compliers is the same as that on compliers
- Why? Because those who comply and those who don't are likely to be different and might respond to the treatment differently
- We thus need to change our estimand:
  1. Intent-to-treat effect (ITT): Calculated in the same way as the ATE, but interpretation is different: the effect on those we intended to treat (even if some weren't treated at all)
  2. Complier average causal effect (CACE): The effect on those who *actually* got treated

There are four classes of research subjects when there are non-compliers

- Compliers in the treatment group
- Compliers in the control group (unobserved)
- Never-takers in the treatment group
- Never-takers in the control group (unobserved)

What is the Intent-to-Treat (ITT) effect?

- $Z_i \in \{0, 1\}$ is treatment assignment
- $D_i \in \{0, 1\}$ is actual receipt of treatment

$$ITT = E[Y_i(z = 1, d(1))] - E[Y_i(z = 0, d(0))] \quad (1)$$

This is just the average effect on those in the treatment group, regardless of whether one actually receives the treatment

What is the Complier Average Causal Effect (CACE)?

- $Z_i \in \{0, 1\}$ is treatment assignment
- $D_i \in \{0, 1\}$ is actual receipt of treatment

$$CACE = E[Y_i(d = 1) - Y_i(d = 0)|d_i(1) = 1] \tag{2}$$

We can estimate this pretty simply in a randomized experiment:

$$\frac{ITT_Y}{ITT_D}, \tag{3}$$

Where ITT_D is the proportion of those in the treatment group who are compliers

Why ITT_Y/IIT_D?

Research subject	Treatment	Complier	Outcome‡
1	1	1	10
2	1	1	20
3	1	0	0†
4	1	0	0†
5	0	1*	0
6	0	1*	0
7	0	0*	0
8	0	0*	0
ITT			7.5 - 0 = 7.5
CACE			(7.5 - 0) / 0.5 = 15

‡ Relative to the control; † 0 because non-complier

* Compliers and non-compliers in the control group are unknown, however

BUT, if treatment assignment itself has an effect regardless of whether you receive treatment, the "exclusion restriction" is broken

- The exclusion restriction says that a treatment only operates through a specific channel and no other
- e.g. If encouraging someone to watch the debate affects how much post-debate coverage they read, even if they didn't end up watching the debate itself

If the exclusion restriction is violated, then our estimate of the CACE is biased

$$\frac{ITT_y}{ITT_D} = CACE + (\frac{1 - ITT_D}{ITT_D}) E[\underbrace{Y_i(z = 1, d = 0)}_{\text{Non-compliers under treatment}} - \underbrace{Y_i(z = 0, d = 0)}_{\text{Non-compliers under control}}].$$

- We will discuss this much more in class on instrumental variables

Can't we just remove the non-compliers from the data when we estimate the treatment effect? No.

- If we remove non-compliers then we are comparing the treatment group compliers to control group compliers *and* non-compliers

Clustered treatment assignment

- Households instead of individuals
- Classrooms instead of students
- Media markets instead of media consumers

Consequence:

- Less precision in treatment effect estimates (i.e. larger std. errors)
- Need to use clustered standard errors that take account of the level of the treatment

Blocking

- Can get more precision by random assignment within "blocks"
- e.g. If you think political ideology explains a lot of variation in your outcome variable, randomize treatment/control among those on the left, and those on the right
- Same proportion of those assigned to treatment and control within each block
- Higher precision (lower SEs) because less sampling error due to important covariates

Pre- & post-treatment outcome measures

- Measure your outcome before treatment, and then compare it to the measure after an individual receives the treatment or control
- Within-individual comparisons
- Lower SEs

Want to run an experiment? Here's a small check list:

1. Pre-register your hypotheses and analysis
    - e.g. https://osf.io
2. Conduct a power analysis to ensure that you have a large enough sample to detect the effect that you hypothesize
3. If possible, block treatment assignment on important covariates
4. If possible, measure the outcome before treatment assignment to do a pre-/post-comparison
5. If possible, create an index using multiple questions to reduce measurement error (if using a survey measure)
6. If possible, use a placebo if there are non-complier to allow direct complier versus complier comparison
7. If non-compliers, estimate both the ITT and CACE
8. If clustered treatment assignment, correct your standard errors

Exercise

Complete the exercise from the R file on the course website

Exercise solutions

```
# Estimate the Intent to treat (ITT) effect with standard OLS using lm()
# Recall that the ITT is just the effect of the treatment on the outcome
# regardless of whether people in the treatment group were compliers
# In this experiment, this means regardless of whether they could be reach by
# the door-to-door canvassers
model_itt <- lm(voted ~ assigned, data = GG)

# Save the value of the estimated ITT effect to a variable
# Use the coef() command which gives the coefficients in a model
# Save only the coefficient that is your estimate of the ITT
itt <- coef(model_itt)[2]


# Estimate the effect of treatment assignment on actually being treated
# i.e. the effect of being assigned to have a door-to-door cavasser knock on
# someone's door on that person actually answering it and being told to vote
model_ittd <- lm(treated ~ assigned, data = GG)

# Save the value of the estimated ITT_d effect to a variable
# Use the coef() command which gives the coefficients in a model
# Save only the coefficient that is your estimate of the ITT
ittd <- coef(model_ittd)[2]
```

Exercise solutions

```
# Recall that the CACE is the effect of the treatment on those people who
# actually complied (i.e. those assigned to treatment who also answered their
# door). To estimate this, we need to take the intent-to-treat (ITT) effect
# and divide is by the estimated proportion of compliers.
# Given that you saved the estimate of the ITT effect above, and you saved the
# estimate of the proportion of compliers as well, calcuate the CACE
# Hint: You're just using "itt" and "ittd" that you saved above
cace <- itt / ittd
```

Exercise solutions

```
# Now that we have all the simulated values in the data.frame "S", and the
# estimate from the actual experiment in "experiment_estimate", we can graph
# the simulated values with geom_histogram(), and then put a vertical line
# where our actual experimental estimate is.
# See if you can figure out how to do this with geom_histogram() and
# geom_vline(). Add some extras to make the graph look a bit better

ggplot(S, aes(x = sim_est)) +
  labs(x = "Estimate", y = "Density") +
  coord_cartesian(ylim = c(-10, 550), expand = FALSE) +
  scale_y_continuous(breaks = c()) +
  scale_x_continuous(breaks = c(-0.04, -0.02, 0, 0.02, 0.04),
                     labels = c("-.04", "-.02", "0", ".02", ".04")) +
  geom_histogram(binwidth = 0.003, color = "white", fill = "blue", size = 0.3) +
  geom_hline(yintercept = 0, size = 0.75) +
  geom_vline(xintercept = experiment_estimate,
             size = 0.5, linetype = 1, color = "red") +
  theme_minimal()
```