

# w241: Experiments and Causality

## Unit 3

---

David Reiley, David Broockman, D. Alex Hughes

UC Berkeley, School of Information

Updated: 2021-05-12

# Sampling Distribution and Randomization Inference

# Standard Errors

- Standard deviation of the sampling distribution
- How spread out is the sampling distribution?
- How large are the typical chance differences?
- Later, we'll examine statistical power
  - The spread of the sampling distribution is the standard error
  - In what kinds of experiments are large and small differences likely to arise by chance?

# Sampling Distributions and R

- Groups may differ by chance, even if the treatment has no effect.
  - How much would the groups differ if the treatment had no effect?
  - How large of an "effect estimate" would we reach by chance?
- Distribution of estimates one would reach if treatment had no effect.
  - How likely is this estimate to have just arisen by chance?
- Similar to observational studies, but:
  - Intuition easy to see in experiments.
  - Testing a hypothesis about our sample, not a population.
  - Example code to walk through intuition on slides to follow

# Example: An Experiment with no Effect

- Does eating soybeans affect estrogen levels?
- 40 individuals: 20 men, 20 women.
- Simulate the potential outcomes of the control group.
- Simulate the potential outcomes of the treatment group.
- A simulated experiment with no effect.

```
group <- c(rep("Man",20),rep("Woman",20))

po_control <- c(
  seq(from = 1, to = 20),
  seq(from = 51, to = 70)
)

## Suppose there is no effect.
## Then, the potential outcomes to control are equal
## to the potential outcomes to treatment.

po_treatment <- po_control + 0

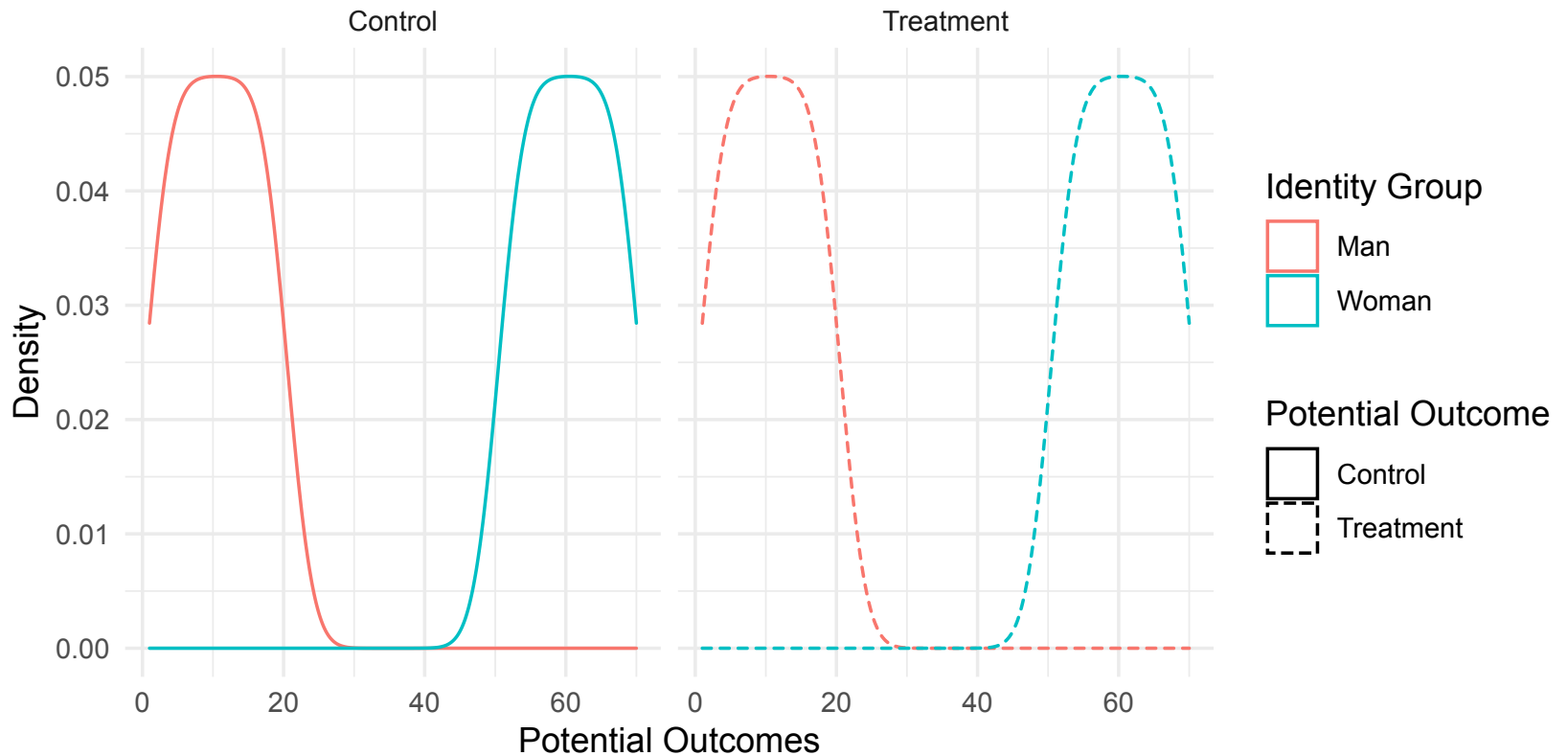
d <- data.frame(
  'Control' = po_control,
  'Treatment' = po_treatment,
  'group' = group
)
```

```
d %>%  
  head()
```

```
##   Control Treatment group  
## 1         1           1   Man  
## 2         2           2   Man  
## 3         3           3   Man  
## 4         4           4   Man  
## 5         5           5   Man  
## 6         6           6   Man
```

## Distribution of Potential Outcomes

Man identifiers have lower estrogen





# Random Assignment

- Define function to randomly assign units to treatment and control.
- Randomly pick 20 for treatment and 20 for control.
- Concatenate the two vectors.
- Get a different vector when you run it again.

```
randomize ← function(units_per_group) {  
  ## an (unnecessary) function to randomize units into  
  ## treatment and control  
  ## ---  
  ## args:  
  ## - units_per_group: how many zero and one should be returned  
  
  assignment_vector ← rep(c('Control', 'Treatment'), each = units_per_group)  
  sample.assignment_vector)  
}
```

# Random Assignment

```
randomize(units_per_group = 4)
```

```
## [1] "Treatment" "Treatment" "Control"    "Treatment" "Treatment" "Control"    "Control"
```

```
randomize(units_per_group = 4)
```

```
## [1] "Control"    "Treatment" "Treatment" "Treatment" "Treatment" "Control"    "Control"
```

# Realized Outcomes

- Treatment outcome for those randomized to treatment and control outcome for those randomized to control.
- Assign for each person in the vector.
- Same because we had an experiment with no effect.
- R code is often written in a compact manner; could also have been done separately for each group.
- Why are we doing this when there is no treatment effect?
- Because it should also work when there is one. We're looking at what happens when we randomly assign people to control and treatment groups.

# Realized Outcomes

```
treatment_assigned ← randomize()
```

```
outcomes ← po_treatment * I(treatment_assigned == "Treatment") +  
  po_control * I(treatment_assigned == "Control")
```

```
outcomes
```

```
## [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 51 52 53 54 55 56 57 58  
## [39] 69 70
```

# Function to Estimate the Average

- Subtract the mean outcome for the control group from the mean outcome of the treatment group.
- How much higher is the average in the treatment group versus the control group?
- We may have randomly selected someone with a higher or lower level of estrogen.
- Even though we know the effect is 0, we see chance differences.

# Function to Estimate the Average

```
estimate_ate ← function(y_values, treatment) {  
  
  treatment_group_mean ← mean(y_values[treatment == 'Treatment'])  
  control_group_mean   ← mean(y_values[treatment == 'Control'])  
  
  ate ← treatment_group_mean - control_group_mean  
  
  return(  
    list(  
      "tg_mean" = treatment_group_mean,  
      "cg_mean" = control_group_mean,  
      "ate"     = ate)  
    )  
  }  
}
```

*## In fact, there is \_no\_ effect, but... sampling!*

```
estimate_ate(y_values = outcomes, treatment = treatment_assigned)
```

```
## $tg_mean
```

```
## [1] 36
```

```
##
```

```
## $cg_mean
```

```
## [1] 35
```

```
##
```

```
## $ate
```

```
## [1] 1
```

*## To pull a single part of this, because it is a list, R indexes with .[[*

```
estimate_ate(y_values = outcomes, treatment = treatment_assigned)[['ate']]
```

```
## [1] 1
```

# The Null Hypothesis



# Rhetorical Posture of the Null

- You want to argue against a skeptic that a treatment has an effect.
- Assume the skeptic is right.

| Treatment has no effect.

- What is the chance that we would see this estimate by chance in that scenario?

| This is p-value.

- We'll see where it comes from visually.

# Average Size of the Difference

- Simulate this a few times to get a sense of how much our treatment effect estimate would vary by chance.
- We created an estimate function with the outcomes and the treatment group.
- Outcome vector will look the same regardless of the treatment vector.

```
treatment_assigned_one ← randomize(units_per_group = 20)  
estimate_ate(y_values = outcomes, treatment = treatment_assigned_one)[['ate']]
```

```
## [1] 1.3
```

```
treatment_assigned_two ← randomize(units_per_group = 20)  
estimate_ate(y_values = outcomes, treatment = treatment_assigned_two)[['ate']]
```

```
## [1] 8
```

```
treatment_assigned_three ← randomize(units_per_group = 20)  
estimate_ate(y_values = outcomes, treatment = treatment_assigned_three)[['ate']]
```

```
## [1] -3.1
```

# Outcome With Different Assignments

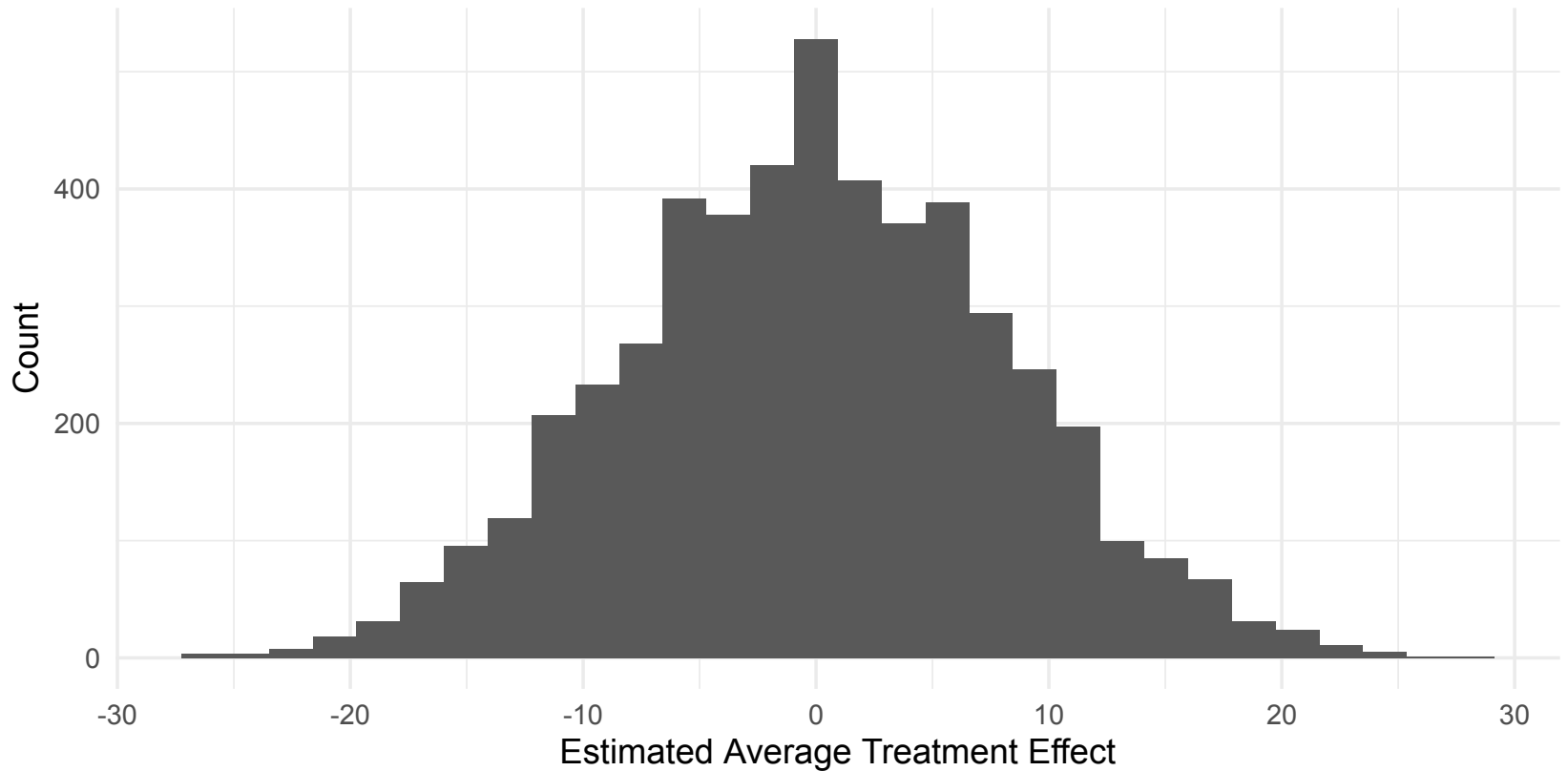
- Similar to re-sampling from a population.
- Re-randomizing from within the original population. Testing the null hypothesis from within the sample we already have.
- Re-shuffle the 40 people between treatment and control. Assuming the treatment effect for everyone is zero.
- **Sharp null hypothesis:** For every unit, there is no effect.
- Repeat this process to generate a synthetic distribution of effects if the sharp null hypothesis *were true*.
- Randomly sample the vector of assignments 5,000 times to generate an unbiased sample of all the effects.
- Literally, replicate 5,000 times, and save to a vector.

```
## going to move the randomization inside the `estimate_ate` function  
## for compactness
```

```
sharp_null ← replicate(  
  n = 5000,  
  expr = estimate_ate(  
    y_value = outcomes,  
    treatment = randomize(units_per_group = 20))[['ate']]  
)
```

## Distribution of Treatment Effects Under Sharp Null

Distribution is Centered at Zero, And Symmetric



# Size of the Observed Difference

- The p-value.
- How often did I get a randomization under the sharp null where the estimate was larger than my actual estimate?
- For each, is it larger than the average treatment effect estimate?
- This is a sampling distribution.
- How big is my estimate relative to the distribution of estimates?

# Size of the Observed Difference

In this particular case,

```
experimental_randomization <- randomize(units_per_group = 20)
experimental_ate <- estimate_ate(
  y_values = outcomes,
  treatment = experimental_randomization)[['ate']]

sharp_null <- replicate(
  n = 5000,
  expr = estimate_ate(
    y_value = outcomes,
    treatment = randomize(units_per_group = 20))[['ate']]
)

mean(abs(sharp_null) > abs(experimental_ate))

## [1] 0.0686
```

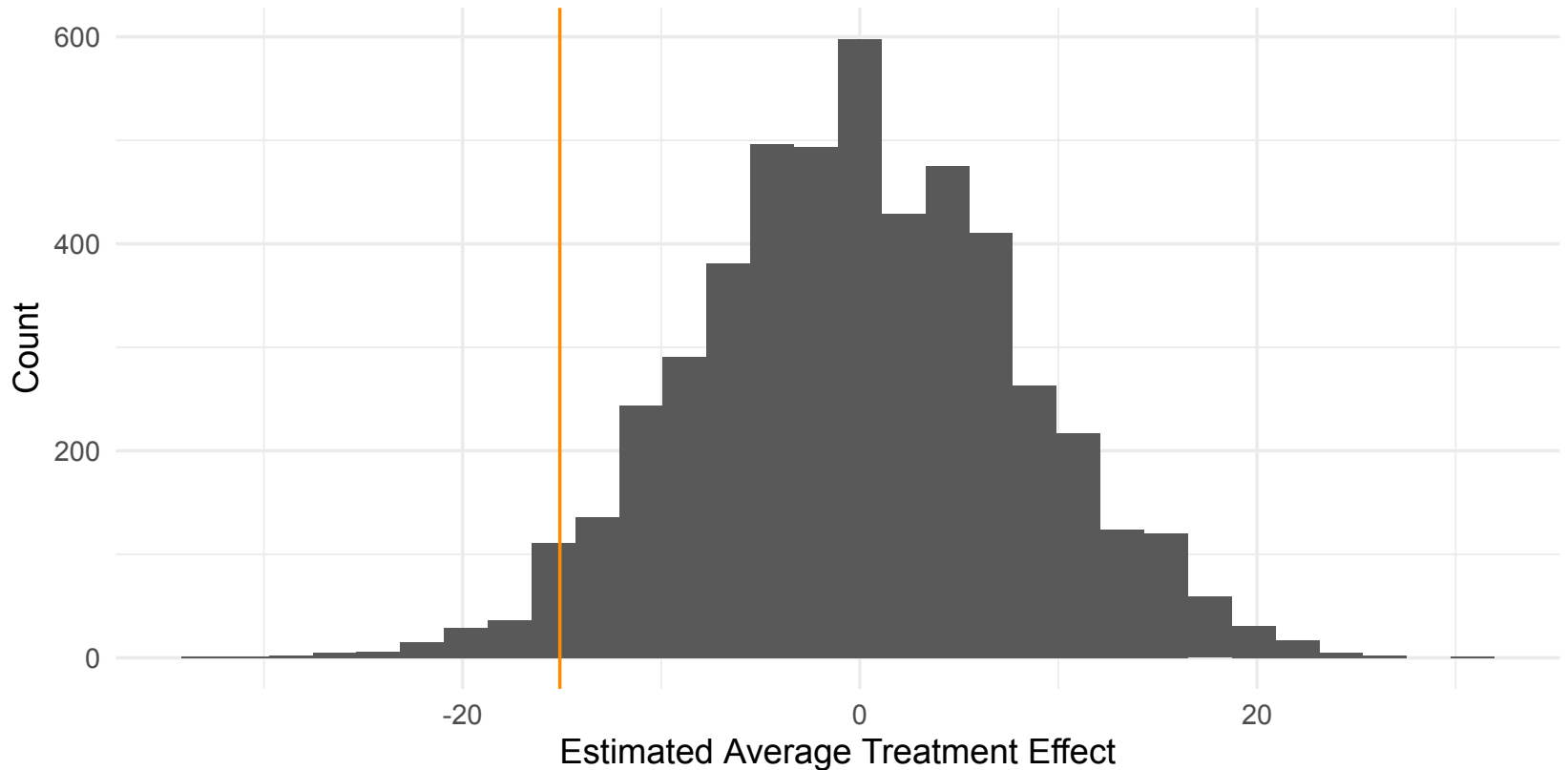
# Size of the Observed Difference

```
histogram_no_effect <- ggplot() +  
  aes(x = sharp_null) +  
  geom_histogram() +  
  geom_vline(xintercept = experimental_ate, color = 'darkorange') +  
  labs(  
    title = "No Effect: Distribution of Treatment Effects Under Sharp Null",  
    subtitle = "Distribution is Centered at Zero, And Symmetric",  
    x = "Estimated Average Treatment Effect",  
    y = "Count"  
  )
```



# Size of the Observed Difference

No Effect: Distribution of Treatment Effects Under Sharp Null  
Distribution is Centered at Zero, And Symmetric



# P-Values and Hypothesis Tests

# P-Values

- If the treatment had no effect, how likely is it that the data would generate a difference this extreme, *just by chance*?
- What is the difference between the mean in the control and treatment groups?
- Different from how likely it is the treatment has an effect
- Convention is to reject the null with p-value under 0.05.
- p-values don't tell you for sure that the treatment has an effect.
- They just tell you how likely it is you would have gotten that result by chance.
- The sampling distribution tells us how large the differences are we find by chance.
- Can find p-values  $< 0.05$  even when the null hypothesis is correct.

# Simulating an Experiment with a Large

- Vector of outcomes and control
- 40-row table with potential outcomes in control and treatment.
- This time, with a difference of 25

# Simulating an Experiment with a Large

```
po_control    ← c(1:20, 51:70)
po_treatment  ← po_control + 25

treatment_assigned ← randomize(units_per_group = 20)

outcomes ← po_treatment * I(treatment_assigned == "Treatment") +
  po_control * I(treatment_assigned == "Control")

experimental_ate_big_effect ← estimate_ate(
  y_values = outcomes,
  treatment = treatment_assigned
)[['ate']]
```

# Simulating an Experiment with a Large

```
sharp_null_big_effect <- replicate(  
  n = 5000,  
  expr = estimate_ate(  
    y_values = outcomes,  
    treatment = randomize(units_per_group = 20))['ate']  
)
```

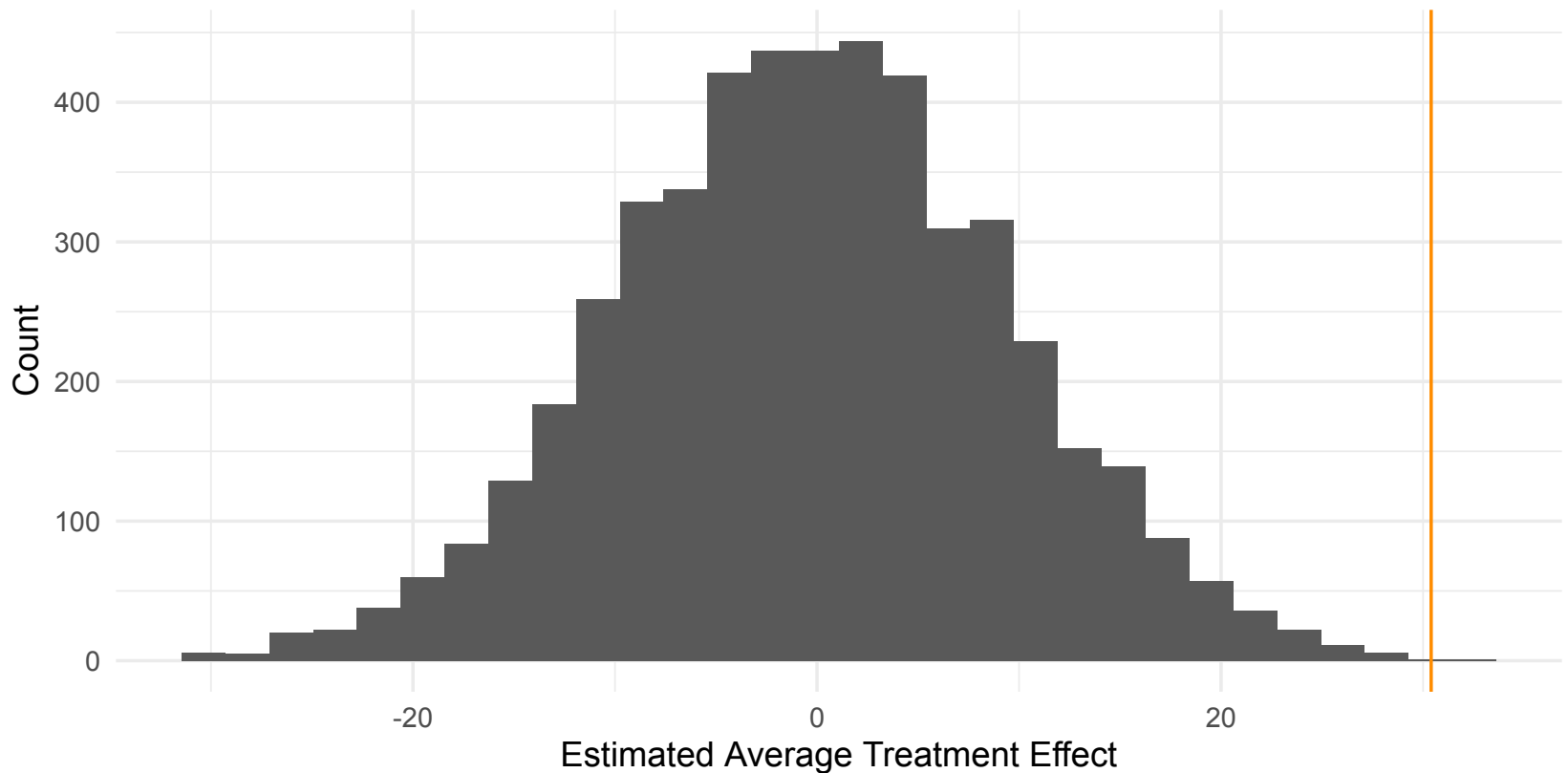
# Simulating an Experiment with a Large

```
histogram_big_effect <- ggplot() +  
  aes(x = sharp_null_big_effect) +  
  geom_histogram() +  
  geom_vline(xintercept = experimental_ate_big_effect, color = 'darkorange') +  
  labs(  
    title = "Big Effect: Distribution of Treatment Effects Under Sharp Null",  
    subtitle = "Distribution is Centered at Zero, And Symmetric",  
    x = "Estimated Average Treatment Effect",  
    y = "Count"  
  )
```

# Simulating an Experiment with a Large

## Big Effect: Distribution of Treatment Effects Under Sharp Null

Distribution is Centered at Zero, And Symmetric





# Simulating an Experiment with a Large

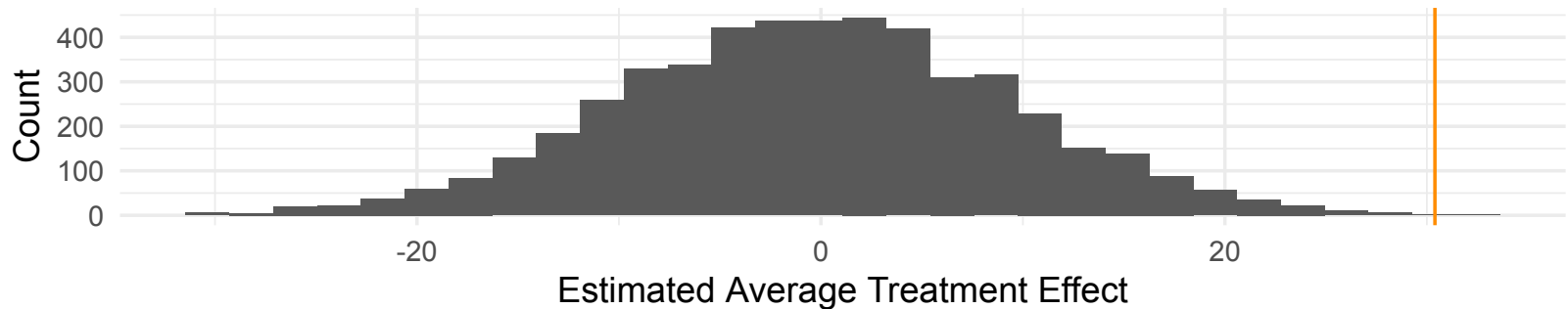
```
mean(abs(sharp_null_big_effect) > abs(experimental_ate_big_effect))
```

```
## [1] 0.0012
```

# Compare Big Effect and No Effect Sharp

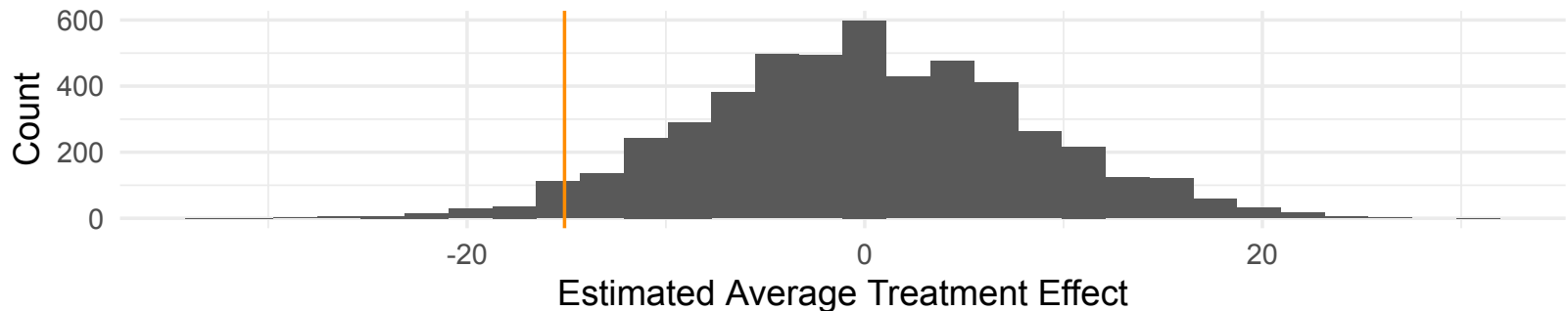
## Big Effect: Distribution of Treatment Effects Under Sharp Null

Distribution is Centered at Zero, And Symmetric



## No Effect: Distribution of Treatment Effects Under Sharp Null

Distribution is Centered at Zero, And Symmetric



# Statistical Power

# Detecting Non-Zero Treatment Effects

Suppose the treatment effect is 10.

# Create Whole Study Function

```
simulate_study ← function(effect_size) {  
  # requires all functions to date  
  
  # generate world  
  po_control   ← c(1:20, 51:70)  
  po_treatment ← po_control + effect_size  
  
  # assign treatment and measure outcomes  
  treatment_assigned ← randomize(20)  
  outcomes ← po_treatment * I(treatment_assigned == "Treatment") +  
    po_control * I(treatment_assigned == "Control")  
  
  # estimate ate  
  estimated_ate ← estimate_ate(y_values = outcomes, treatment = treatment_assigned)[1]  
  
  # generate sharp null distribution  
  sharp_null ← replicate(  
    n = 100,  
    expr = estimate_ate(y_values = outcomes, treatment = randomize(20))  
      [['ate']]  
  )  
  
  p_value ← mean(abs(sharp_null) > abs(estimated_ate))  
  return(list(  
    'estimated_ate' = estimated_ate,
```

# Generate P-Values, Effect Size: 0

```
## notice: we now have two loops:  
## - We're running 1,000 simulations;  
## - In each simulation, there are 1,000 sharp nulls drawn out  
## - So get some coffee if you're running this at home  
  
distribution_of_p_values_0 ← replicate(  
  n = 200,  
  expr = simulate_study(effect_size = 0)[['p_value']]  
)
```

# Plot Distribution of P-Values, Effect Size:

```
histogram_pvalues_0 <- ggplot() +  
  aes(distribution_of_p_values_0) +  
  geom_histogram() +  
  labs(  
    title = "Effect Size: 0",  
    subtitle = paste0("Percent P-Values < 0.05: ", round(mean(distribution_of_p_values  
x = "P-Value"))
```

```
distribution_of_p_values_10 <- replicate(  
  n = 200,  
  expr = simulate_study(effect_size = 10)[['p_value']]  
)
```

# Plot Distribution of P-Values Effect Size:

```
histogram_pvalues_10 <- ggplot() +  
  aes(x = distribution_of_p_values_10) +  
  geom_histogram() +  
  labs(  
    title = "Effect Size: 10",  
    subtitle = paste0("Percent P-Values < 0.05: ", round(mean(distribution_of_p_values  
    x = 'P-Value'  
  )
```



# Generate P-Value, Effect Size 20

```
distribution_of_p_values_20 ← replicate(  
  n = 200,  
  expr = simulate_study(effect_size = 20)[['p_value']]  
)
```

```
histogram_pvalues_20 ← ggplot() +  
  aes(x = distribution_of_p_values_20) +  
  geom_histogram() +  
  labs(  
    title = "Effect Size: 20",  
    subtitle = paste0("Percent P-Values < 0.05: ", round(mean(distribution_of_p_values_20 < 0.05), 2)),  
    x = 'P-Value')
```