AE 15: Hypothesis testing

Dav King

2/24/22

```
library(tidyverse)
library(tidymodels)
```

```
manhattan <- read_csv("manhattan.csv")
```

Learning goals

- Use simulation-based methods to test a claim about a population parameter
- Use simulation-based methods to generate the null distribution
- Calculate and interpret the p-value
- Use the p-value to draw conclusions in the context of the data

Recall

Terminology

Population: a group of individuals or objects we are interested in studying

Parameter: a numerical quantity derived from the population (almost always unknown)

Sample: a subset of our population of interest

Statistic: a numerical quantity derived from a sample

Common population parameters of interest and their corresponding sample statistic:

Quantity	Parameter	Statistic
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s
Median	M	$\tilde{x}$
Proportion	p	$\hat{p}$

Statistical inference

Statistical inference is the process of using sample data to make conclusions about the underlying population the sample came from.

Estimation: estimating an unknown parameter based on values from the sample at hand

Testing: evaluating whether our observed sample provides evidence for or against some claim about the population

We will now move to testing hypotheses.

Testing

How can we answer research questions using statistics?

Statistical hypothesis testing is the procedure that assesses evidence provided by the data in favor of or against some claim about the population (often about a population parameter or potential associations).

Example:

The state of North Carolina claims that students in 8th grade spent, on average, 200 minutes on Zoom each day in Spring 2021. *What do you make of this statement? How would you evaluate the veracity of the claim?* This claim actually seems a little low - that would only amount to 3 hours and 20 minutes on zoom per day (way more than anyone should be subject to, but probably less than many people were). This claim is, however, averaging over both students whose schools were entirely online that spring and students who were entirely in person (as well as anything in between) - thus, who knows what those data would look like?

We can - simply draw a sample, bootstrap it, take the confidence interval and determine whether 200 minutes lies within it or not.

The hypothesis testing framework

1. Start with two hypotheses about the population: the null hypothesis and the alternative hypothesis.
2. Choose a (representative) sample, collect data, and analyze the data.
3. Figure out how likely it is to see data like what we observed, **IF** the null hypothesis were in fact true.
4. If our data would have been extremely unlikely if the null claim were true, then we reject it and deem the alternative claim worthy of further study. Otherwise, we cannot reject the null claim.

Two competing hypotheses

The **null hypothesis** (often denoted H_0) states that "nothing unusual is happening" or "there is no relationship," etc.

On the other hand, the **alternative hypothesis** (often denoted H_1 or H_A) states the opposite: that there is some sort of relationship (usually this is what we want to check or really think is happening).

In statistical hypothesis testing we always first assume that the null hypothesis is true and then see whether we reject or fail to reject this claim.

1. Defining the hypotheses

The null and alternative hypotheses are defined for **parameters,** not statistics.

What will our null and alternative hypotheses be for this example?

- H_0: the true mean time spent on Zoom per day for 8th grade students is 200 minutes

- H_1: the true mean time spent on Zoom per day for 8th grade students is not 200 minutes

Expressed in symbols:

- $H_0 : \mu = 200$
- $H_1 : \mu \neq 200$,

where μ is the true population mean time spent on Zoom per day by 8th grade North Carolina students.

2. Collecting and summarizing data

With these two hypotheses, we now take our sample and summarize the data.

```
zoom_time <- c(299, 192, 196, 218, 194, 250, 183, 218, 207,
               209, 191, 189, 244, 233, 208, 216, 178, 209,
               201, 173, 186, 209, 188, 231, 195, 200, 190,
               199, 226, 238)
```

```
mean(zoom_time)
```

```
## [1] 209
```

The choice of summary statistic calculated depends on the type of data. In our example, we use the sample mean: $\bar{x} = 209$.

Do you think this is enough evidence to conclude that the mean time is not 200 minutes?

3. Assessing the evidence observed

Next, we calculate the probability of getting data like ours, *or more extreme*, if H_0 were in fact actually true.

This is a conditional probability: Given that H_0 is true (i.e., if μ were *actually* 200), what would be the probability of observing $\bar{x} = 209$?" This probability is known as the **p-value**.

4. Making a conclusion

We reject the null hypothesis if this conditional probability is small enough.

If it is very unlikely to observe our data (or more extreme) if H_0 were actually true, then that might give us enough evidence to suggest that it is actually false (and that H_1 is true).

What is "small enough"?

- We often consider a numeric cutpoint **significance level** defined *prior* to conducting the analysis.
- Many analyses use $\alpha = 0.05$. This means that if H_0 were in fact true, we would expect to make the wrong decision only 5% of the time.

What can we conclude?

Case 1: p-value $\geq \alpha$:

If the p-value is α or greater, we say the results are not statistically significant and we **fail to reject** H_0.

Importantly, **we never "accept" the null hypothesis** – we performed the analysis assuming that H_0 was true to begin with and assessed the probability of seeing our observed data or more extreme under this assumption.

Case 2: p-value $< \alpha$

If the p-value is less than α, we say the results are **statistically significant**. In this case, we would make the decision to **reject the null hypothesis**.

Similarly, **we never "accept" the alternative hypothesis**.

What is a p-value?

"The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true. We typically use a summary statistic of the data, in this section the sample proportion, to help compute the p-value and evaluate the hypotheses." (Open Intro Stats, pg. 194)

What isn't a p-value?

- "A p-value of 0.05 means the null hypothesis has a probability of only 5% of* being true.
- "A p-value of 0.05 means there is a 95% chance or greater that the null *hypothesis is incorrect"*

p-values do **not** provide information on the probability that the null hypothesis is true given our observed data.

Again, a p-value is calculated *assuming* that H_0 is true. It cannot be used to tell us how likely that assumption is correct. When we fail to reject the null hypothesis, we are stating that there is **insufficient evidence** to assert that it is false. This could be because...

- ... H_0 actually *is* true!
- ... H_0 is false, but we got unlucky and happened to get a sample that didn't give us enough reason to say that H_0 was false

Even more bad news, hypothesis testing does NOT give us the tools to determine which one of the two scenarios occurred.

What can go wrong?

Suppose we test a certain null hypothesis, which can be either true or false (we never know for sure!). We make one of two decisions given our data: either reject or fail to reject H_0.

We have the following four scenarios:

Decision	H_0 is true	H_0 is false
Fail to reject H_0	Correct decision	*Type II Error*

Decision	H_0 is true	H_0 is false
Reject H_0	*Type I Error*	Correct decision

It is important to weigh the consequences of making each type of error.

In fact, α is precisely the probability of making a Type I error. We will talk about this (and the associated probability of making a Type II error) in future lectures.

Let's conduct some hypothesis tests

Data

We'll continue to work with the sample of Zoom screen-time data we obtained. To make things easier with the `infer` functions, we'll create a tibble with `time` as a single variable.

```
zoom <- tibble(
  time = c(299, 192, 196, 218, 194, 250, 183, 218, 207,
           209, 191, 189, 244, 233, 208, 216, 178, 209,
           201, 173, 186, 209, 188, 231, 195, 200, 190,
           199, 226, 238))
```

```
zoom
```

```
## # A tibble: 30 x 1
##     time
##    <dbl>
##  1   299
##  2   192
##  3   196
##  4   218
##  5   194
##  6   250
##  7   183
##  8   218
##  9   207
## 10   209
## # ... with 20 more rows
```

Set seed

To obtain reproducible results, set the seed for the random number generation.

```
set.seed(1421)
```

Notes

Recall our hypothesis testing framework:

1. Start with two hypotheses about the population: the null hypothesis and the alternative hypothesis.
2. Choose a (representative) sample, collect data, and analyze the data.
3. Figure out how likely it is to see data like what we observed, **IF** the null hypothesis were in fact true.
4. If our data would have been extremely unlikely if the null claim were true, then we reject it and deem the alternative claim worthy of further study. Otherwise, we cannot reject the null claim.

Example: testing population mean - μ

We've already done items 1 and 2, where

$$H_0 : \mu = 200$$
$$H_1 : \mu \neq 200$$

For this study, let $\alpha = 0.05$.

To tackle items 3 and 4, we'll use a simulation-based approach with functions from `infer`.

Simulate the null distribution

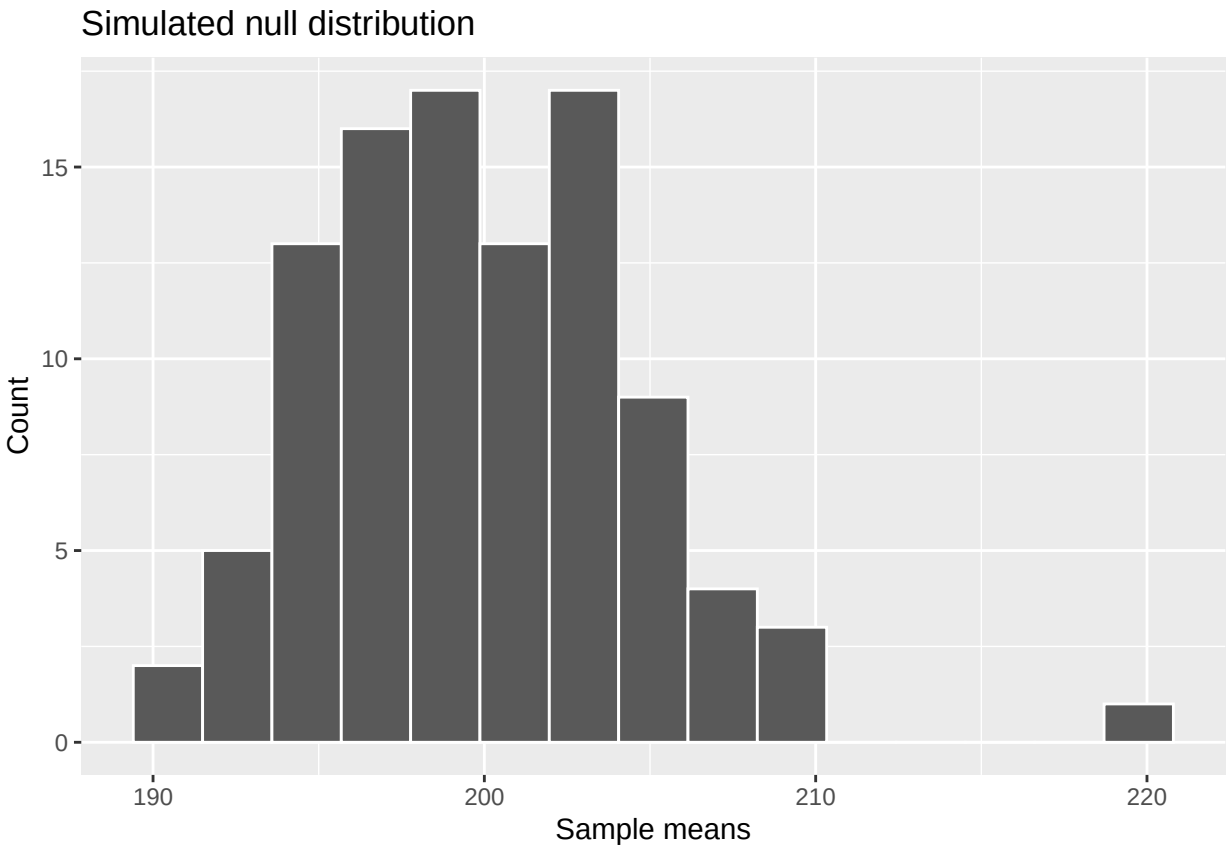
Recall that there is variability in the sampling distribution of the sample mean. We need to account for this in our statistical study. Just as we did for confidence intervals, we'll use a bootstrap procedure here.

1. `specify()` the variable of interest
2. set the null hypothesis with `hypothesize()`
3. `generate()` the bootstrap samples
4. `calculate()` the statistic of interest

```
null_dist <- zoom %>%
  specify(response = time) %>%
  hypothesize(null = "point", mu = 200) %>%
  generate(reps = 100, type = "bootstrap") %>%
  calculate(stat = "mean")
```

Visualize the null distribution

```
visualise(null_dist) +
  labs(x = "Sample means", y = "Count", title = "Simulated null distribution")
```



What do you notice?

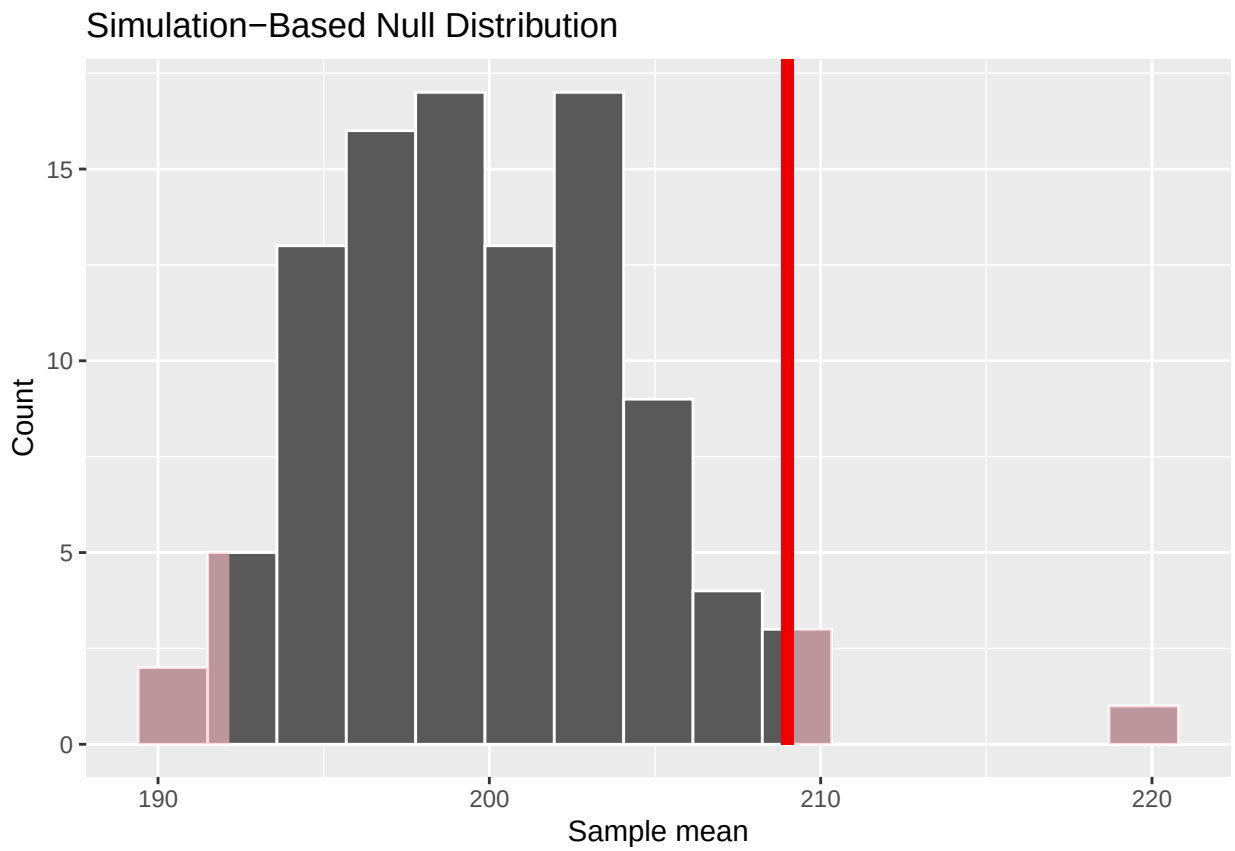
Compute p-value

Next, we calculate the probability of getting data like ours, *or more extreme*,if H_0 were in fact actually true. Our observed sample mean is 209 minutes.

```
x_bar <- zoom %>%
  summarise(mean_time = mean(time))
x_bar
```

```
## # A tibble: 1 x 1
##   mean_time
##       <dbl>
## 1       209
```

```
visualise(null_dist) +
  shade_p_value(obs_stat = x_bar, direction = "two-sided") +
  labs(x = "Sample mean", y = "Count")
```



In the context of this simulation-based approach, the p-value is the proportion of observations shaded light-red. To compute this, `infer` provides a convenient function – `get_p_value()`.

```
null_dist %>%
  get_p_value(obs_stat = x_bar, direction = "two-sided")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1    0.06
```

Conclusion

Given the calculated p-value and the specified α, what conclusion do you make? Given that our p-value of 0.06 is greater than the designated $\alpha = 0.05$, we fail to reject the null hypothesis. We do not have enough evidence to say that our calculated statistic is significantly different from the null hypothesis' estimate of it.

Practice: Rent in Manhattan

On a given day in 2018, twenty one-bedroom apartments were randomly selected on Craigslist Manhattan from apartments listed as "by owner". The data are in the `manhattan` data frame. We will use this sample to conduct inference on the typical rent of one-bedroom apartments in Manhattan.

Exercise 1

Suppose you are interested in whether the mean rent of one-bedroom apartments in Manhattan is actually less than $3000. Choose the correct null and alternative hypotheses.

a. $H_0 : \mu = 3000$ vs. $H_a : \mu \neq 3000$
b. $H_0 : \mu = 3000$ vs. $H_a : \mu < 3000$
c. $H_0 : \mu = 3000$ vs. $H_a : \mu > 3000$
d. $H_0 : \bar{x} = 3000$ vs. $H_a : \bar{x} \neq 3000$
e. $H_0 : \bar{x} = 3000$ vs. $H_a : \bar{x} < 3000$
f. $H_0 : \bar{x} = 3000$ vs. $H_a : \bar{x} > 3000$

B

Exercise 2

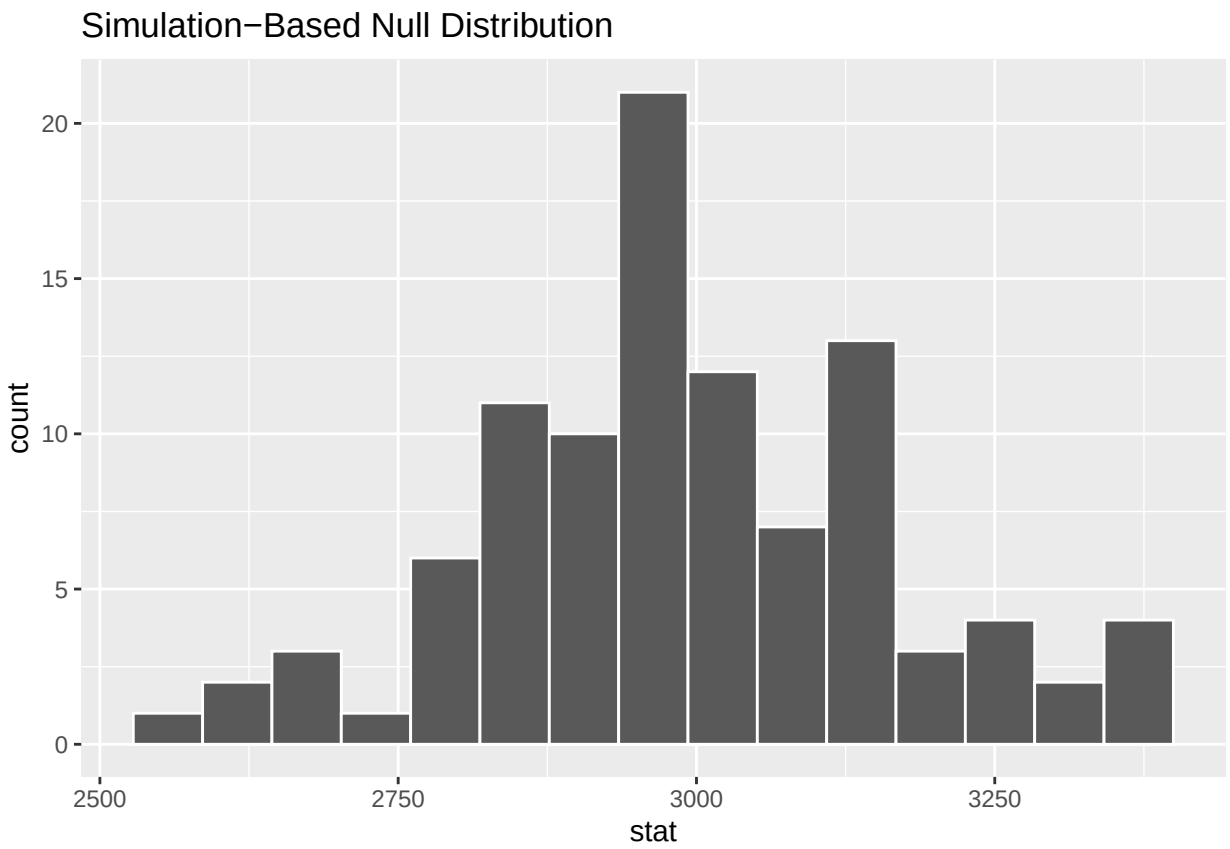
Let's use simulation-based methods to conduct the hypothesis test specified in Exercise 1. We'll start by generating the null distribution.

Fill in the code and uncomment the lines below to generate then visualize null distribution.

```
set.seed(101321)
```

```
null_dist_2 <- manhattan %>%
  specify(response = rent) %>%
  hypothesize(null = "point", mu = 3000) %>%
  generate(reps = 100, type = "bootstrap") %>%
  calculate(stat = "mean")
```

```
visualize(null_dist_2)
```



Exercise 3

Fill in the code and uncomment the lines below to calculate the p-value using the null distribution from Exercise 2.

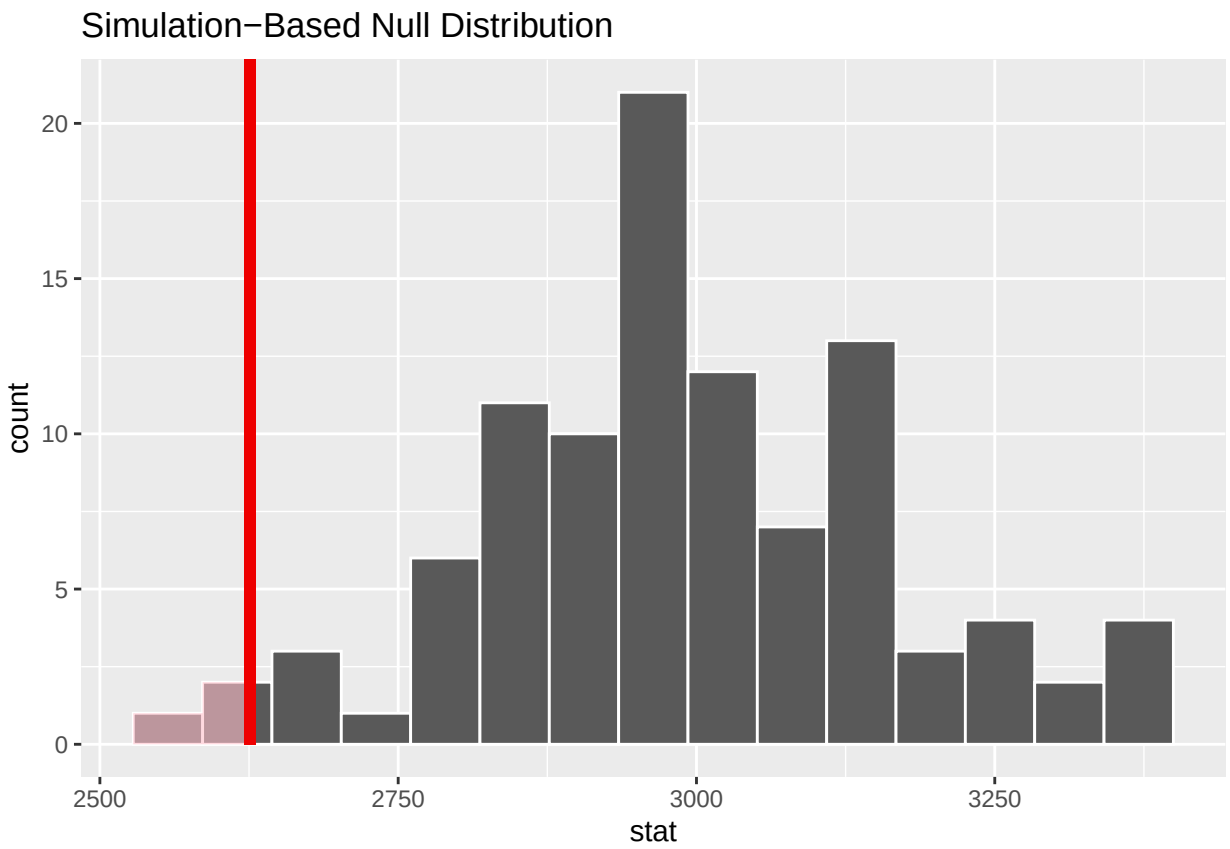
```
mean_rent <- manhattan %>%
  summarise(mean_rent = mean(rent)) %>%
  pull()
```

```
null_dist_2 %>%
  get_p_value(obs_stat = mean_rent, direction = "less")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1    0.02
```

Fill in the direction in the code below and uncomment to visualize the shaded area used to calculate the p-value.

```
visualize(null_dist_2) +
  shade_p_value(obs_stat = mean_rent, direction = "less")
```



Let's think about what's happening when we run `get_p_value`. Fill in the code below to calculate the p-value "manually" using some of the `dplyr` functions we've learned.

```
null_dist_2 %>%
  filter(stat < mean_rent) %>%
  summarise(p_value = n()/100)
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1    0.02
```

Exercise 4

Use the p-value to make your conclusion using a significance level of 0.05. Remember, the conclusion has 3 components

- How the p-value compares to the significance level
- The decision you make with respect to the hypotheses (reject H_0 /fail to reject H_0).
- The conclusion in the context of the analysis question.

Our P-value obtained was 0.02, which is less than the given significance level $\alpha = 0.05$. Thus, we reject H_0. It is significantly unlikely for us to calculate a mean like this from the sample, given that the population mean is $3000.

Exercise 5

Suppose instead you wanted to test the claim that the mean price of rent is not equal to $3000. Which of the following would change? *Select all that apply.*

a. Null hypothesis
b. Alternative hypothesis
c. Null distribution
d. p-value

B and D

Exercise 6

Let's test the claim in Exercise 5. Conduct the hypothesis test, then state your conclusion in the context of the data.

```
null_dist_2 %>%
  get_p_value(obs_stat = mean_rent, direction = "both")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1    0.04
```

Our p-value of 0.04 is less than the alpha level 0.05. Thus, we reject H_0. We have significant evidence to say that the true population mean is not equal to $3000.

Exercise 7

Create a new variable `over2500` that indicates whether or not the rent is greater than $2500.

```
manhattan2 <- manhattan %>%
  mutate(over2500 = if_else(rent > 2500, "1", "0"))
```

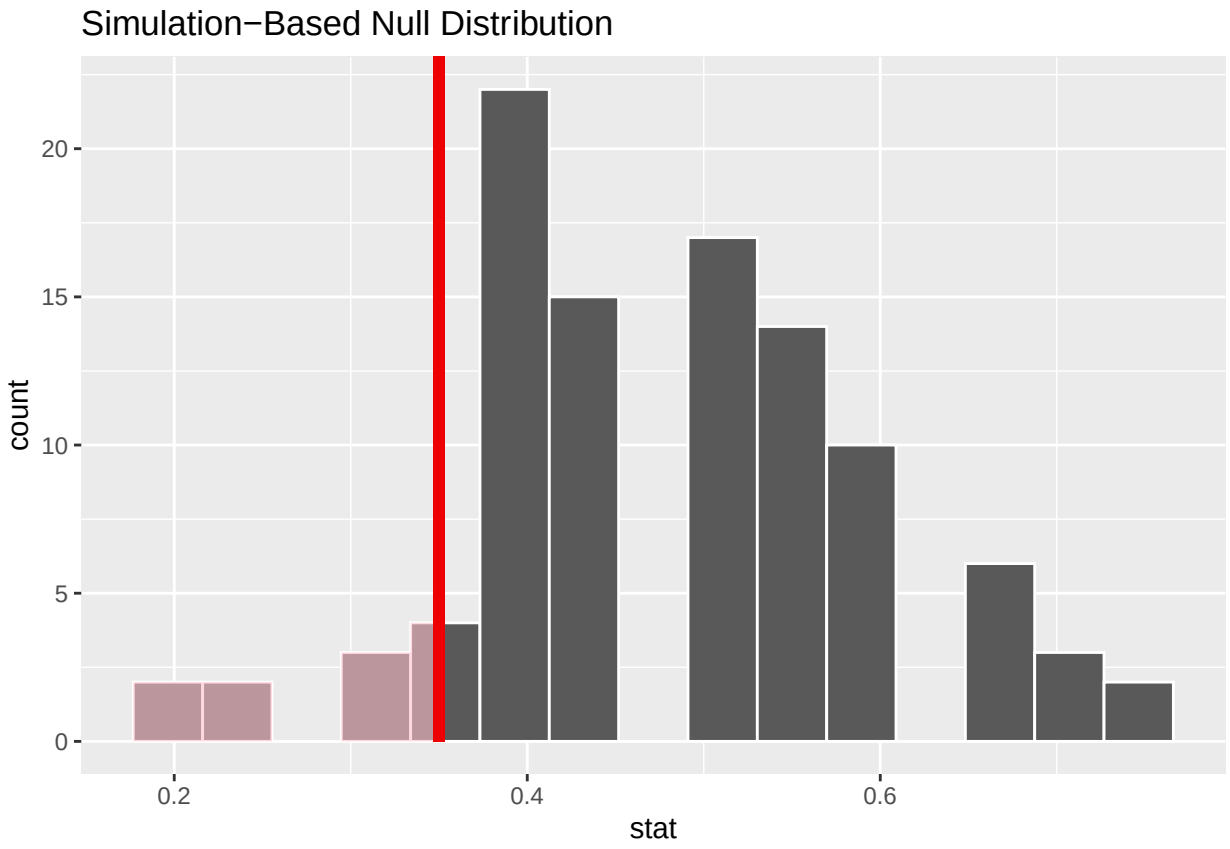
Suppose you are interested in testing whether a majority of one-bedroom apartments in Manhattan have rent greater than $2500.

- State the null and alternative hypotheses. $H_0 : p > 0.5$ $H_1 : p <= 0.5$
- Fill in the code to generate the null distribution.

```
null_dist_3 <- manhattan2 %>%
  specify(response = over2500, success = "1") %>%
  hypothesize(null = "point", p = 0.5) %>%
  generate(reps = 100, type = "simulate") %>%
  calculate(stat = "prop")
avg_p <- manhattan2 %>%
  summarize(avg_p = mean(as.integer(over2500))) %>%
  pull()
```

- Visualize the null distribution and shade in the area used to calculate the p-value.

```
visualize(null_dist_3) +
  shade_p_value(obs_stat = avg_p, direction = "less")
```



- Calculate your p-value. Then use the p-value to make your conclusion using a significance level of 0.05.

```
null_dist_3 %>%
  get_p_value(obs_stat = avg_p, direction = "less")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1    0.11
```

At $\alpha = 0.05$, our p-value of 0.11 means that we fail to reject the null hypothesis. We do not have significant evidence to say that the proportion is different from 0.5.