Notes and Exercises Chapter 16 – Inference for a Single Proportion
Math 217
Saidi

Chapter 16 – Bootstrapping for Proportions

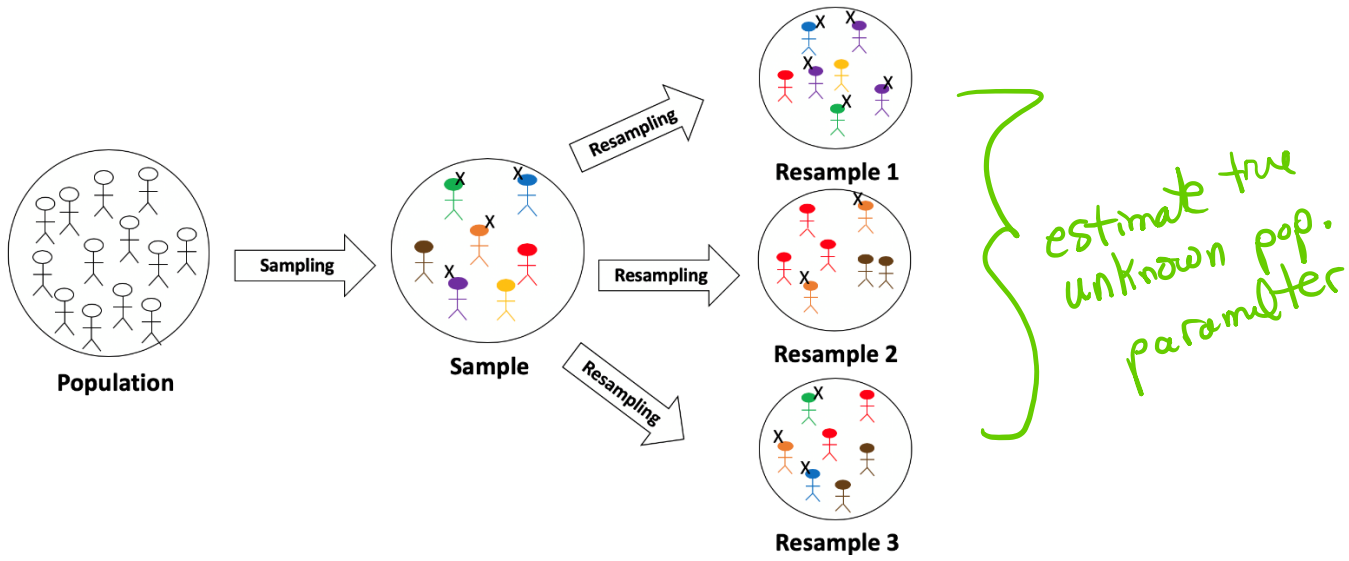
Bootstrapping

With hypothesis testing, it is important to understand how samples from a null population vary. We repeatedly sample from a null population, which gives a sense for the variability of the statistic under the random chance model.

Note that the *statistic* at hand is $\hat{p}$, which is the proportion of successes in the *sample*. The *parameter*, on the other hand, is the proportion of successes in the *population*.

In contrast, with confidence intervals, there is no null population. Instead, we need to understand how samples from the population of interest vary.

We expect the sample statistic to vary around the parameter, but how far is the statistic from the parameter? Bootstrapping is a method that allows us to estimate the distance from a statistic to the parameter.



Bootstrapping repeatedly of samples from the sample is used to estimate the variability of the statistic. Each time we resample, the data are sampled from the original data with replacement. It turns out that the process of resampling from the original sample is an excellent approximation for sampling from a population.

We call the bootstrapped statistic $\hat{p}*$, which is the proportion of successes in the resample.

Use StatKey to explore bootstrapping.

Example 1: People providing an organ for donation sometimes seek the help of a special "medical consultant." These consultants assist the patient in all aspects of the surgery, with the goal of reducing the possibility of complications during the medical procedure and recovery.

One consultant tried to attract patients by noting the average complication rate for liver donor surgeries in the US is about 10%, but her clients have only had 3 complications in the 62 liver donor surgeries she has facilitated. She claims this is strong evidence that her work meaningfully contributes to reducing complications (and therefore she should be hired!).

Using the data, is it possible to assess the consultant's claim that her complication rate is less than 10%?

The answer is no because this is observational data, but we can test to see if there is an association between the complication rates for her clients. We ask, *"could the low complication rate of $\hat{p}$ =0.048 have simply occurred by chance, if her complication rate does not differ from the US standard rate?"*

1. First, check conditions to test proportions (from page 2 of chapter 13 notes), that at least 10 expected successes and 10 expected failures in the sample.

2. We want to see the variability we can expect from sample proportions if the null hypothesis was true.
3. Then we plan to use this information to decide whether there is enough evidence to reject the null hypothesis.
4. Finally, create a **bootstrap simulation** for the observations using a hypothesized null parameter.

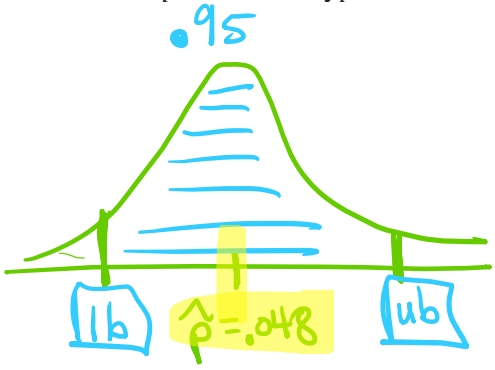
Under the null hypothesis, 10% of liver donors have complications during or after surgery. Suppose this rate was really no different for the consultant's clients (for *all* the consultant's clients, not just the 62 previously measured). If this was the case, we could *simulate* 62 clients to get a sample proportion for the complication rate from the null distribution.

> Similar to the process described in Chapter 12, each client can be simulated using a bag of marbles with 10% red marbles and 90% white marbles. Sampling a marble from the bag (with 10% red marbles) is one way of simulating whether a patient has a complication *if the true complication rate is 10%*. If we select 62 marbles and then compute the proportion of patients with complications in the simulation, $\widehat{p_{sim\_1}}$, then the resulting sample proportion is a sample from the null distribution.
>
> There were 5 simulated cases with a complication and 57 simulated cases without a complication, i.e., $\widehat{p_{sim\_1}}$=5/62=0.081. Repeat this process 10,000 or more times.

The proportions that are equal to or less than $\hat{p}$=0.048 are shaded. The shaded areas represent sample proportions under the null distribution that provide at least as much evidence as $\hat{p}$ favoring the alternative hypothesis. There were 420 simulated sample proportions with p_{sim}≤0.048. We use these to construct the null distribution's left-tail area and find the p-value:

$$left\ tail\ area = \frac{\#\ observed\ simulations\ with\ \widehat{p_{sim}} \leq 0.048}{10000}$$

Of the 10,000 simulated $\widehat{p_{sim}}$, 420 were equal to or smaller than $\hat{p}$. Since the hypothesis test is one-sided, the estimated p-value is equal to this tail area: 0.042.



Conditions for sampling distribution of $\hat{p}$

The sampling distribution for $\hat{p}$ based on a sample of size n from a population with a true proportion p is nearly normal when:

1. The sample's observations are independent, e.g., are from a simple random sample.
2. We expected to see at least 10 successes and 10 failures in the sample, i.e., $np \geq 10 \; and \; n(1-p) \geq 10$. This is called the **success-failure condition**.

When these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error of $\hat{p}$ as $se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Standard error for $\hat{p}$

$se_{\hat{p}} = \sqrt{\frac{\text{p}(1-\text{p})}{n}}$, but since we almost never know the true value of p, $se = \sqrt{\frac{\text{"guess for p"}(1-\text{"guess for p"})}{n}}$, where that "guess for p" is often the null value, p_0.

Margin of error for $\hat{p}$

The margin of error is $z\star\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where z⋆ is calculated from a specified percentile on the normal distribution (explained in Ch 13 notes).

Confidence interval for a single proportion

$$\hat{p} \pm z^*(se) = \hat{p} \pm z^* \sqrt{\frac{\text{p}(1-\text{p})}{n}}$$

Confidence interval conclusion

We are 95% confident that the true population _____ (mean/proportion) for

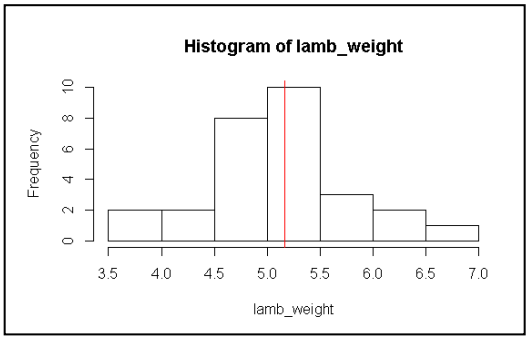
_____(context of problem) is between _____ and _____.

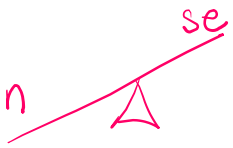
What is se?

- SD describes the **dispersion** of the data
- SE describes the **precision** using the sample statistic to predict the true population parameter.

Example: Lamb birthweights

Single sample





Sampling distributions of various sizes

	$n = 28$	$n = 280$	$n = 2{,}800$	$n \to \infty$
$\bar{y}$	5.17	5.19	5.14	$\bar{y} \to \mu$
s	0.65	0.67	0.65	$s \to \sigma$
SE	0.12	0.040	0.012	SE $\to 0$
Sample distribution				

Notice how the distribution shape changes with larger sample sizes and becomes more **normal**…..

Example 1 revisited: Check conditions to calculate a 95% confidence interval to estimate the true proportion of complications.

1. Independence of sample observations seems valid, as long as each surgery is from a different surgical team.
2. Success-failure condition fails: $np = 62 * .10 = 6.2$ – this is **not** ≥ 10.

Because the conditions are not met, we should not use this method to estimate the CI for the true proportion.

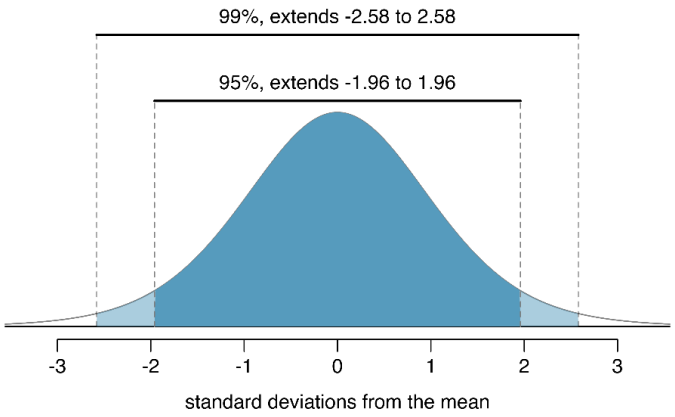


Figure 16.2: The area between -z⋆ and z⋆ increases as |z⋆| becomes larger. If the confidence level is 99%, we choose z⋆ such that 99% of the normal curve is between -z⋆ and z⋆, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: z⋆=2.58.

16.2.5 Hypothesis test for a proportion

Conditions:

- independent observations
- large samples ($np_0 \geq 10$ and $n(1 - p_0) \geq 10$

The p-value is always derived by analyzing the null distribution of the test statistic. The normal model poorly approximates the null distribution for $\hat{p}$ when the success-failure condition is not satisfied. As a substitute, we can generate the null distribution using simulated sample proportions and use this distribution to compute the tail area, i.e., the p-value.

Example 4 : A study of 100 patients who had stents implanted at a particular hospital estimated a 9% increase in the number of patients who had a stroke after implanting a stent. If the CDC reports that the national rate for strokes after stent implants is 6%, is this hospital's rate different than the national rate?

Hypothesis test conclusion

Reject the null/fail to reject the null. There is _____ compelling/convincing evidence that

_____(context of problem in terms of the alternative hypothesis wording).

Coding (from Tutorial)

https://rpubs.com/rsaidi/1102472

Chapter 16 Interactive Notes

Load the libraries and data

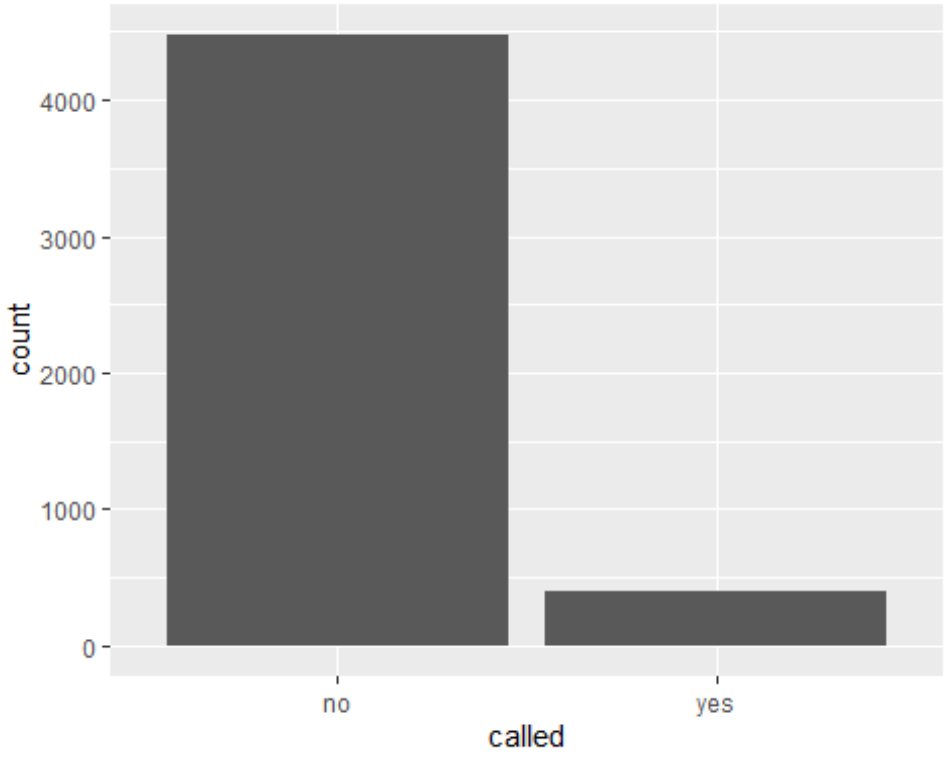
```
library(tidyverse)
library(openintro)
library(tidymodels)
data("resume")
```

Rename the factor levels for callbacks

```
resume1 <- resume |>
  mutate(called = ifelse(received_callback == 1, "yes", "no"))
```

View the counts for applicants receiving callbacks

```
ggplot(resume1, aes(x = called)) +
  geom_bar()
```



calculate exact proportion of the sample that responded this way: p-hat

```
p_hat_call <- resume1 |>
  summarize(prop_call = mean(called == "yes")) |>
  pull()
p_hat_call
```

```
## [1] 0.08049281
```

The proportion who received a callback in the dataset was 0.080.

Test Basic Assumptions

- observations are independent

- np_hat >= 10 n(1-p_hat) >= 10

```
resume1 |>
  group_by(called)|>
  count()
```

```
## # A tibble: 2 × 2
## # Groups:   called [2]
##   called     n
##   <chr>  <int>
## 1 no      4478
## 2 yes      392
```

The 95% confidence interval can be calculated as the sample proportion plus or minus two standard errors of the sample proportion

The bootstrap is done with the function: **specify()**

We do this many times to create many bootstrap replicate data sets.

Do this with the function **generate()**

Next, for each replicate, we calculate the sample statistic, in this case: the proportion of respondents that said "yes" to receiving callbacks.

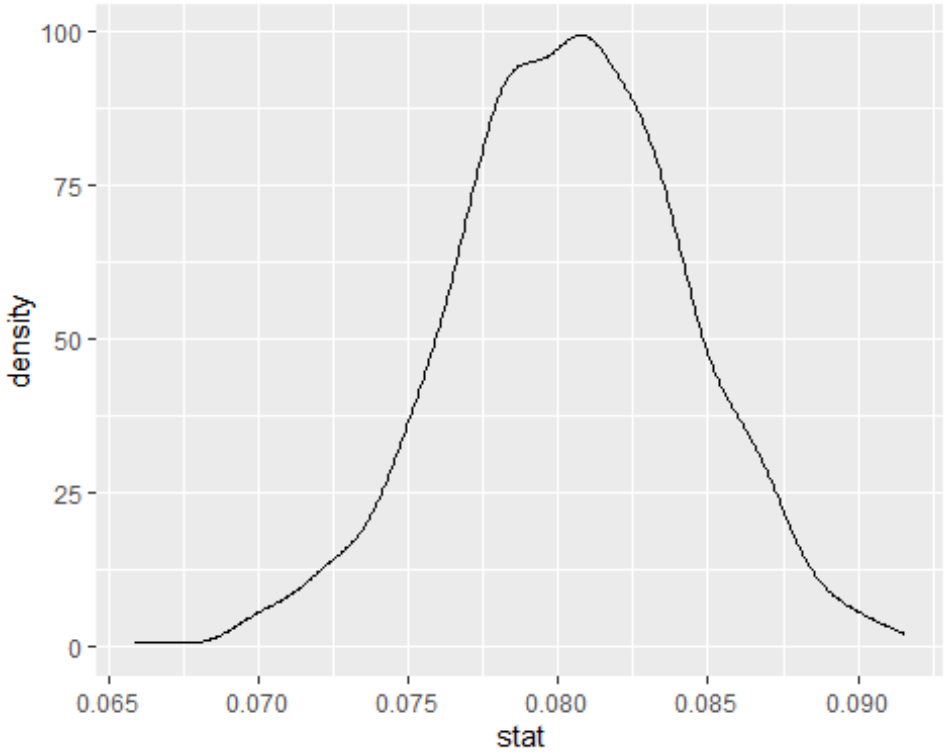
Do this with the function **calculate()**

```
boot_dist_call <- resume1 |>
  specify(response = called, success = "yes") |> # single group
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "prop")
boot_dist_call
```

```
## Response: called (factor)
## # A tibble: 1,000 × 2
##    replicate   stat
##        <int>  <dbl>
##  1         1 0.0881
##  2         2 0.0809
##  3         3 0.0856
##  4         4 0.0823
##  5         5 0.0786
##  6         6 0.0908
##  7         7 0.0782
##  8         8 0.0825
##  9         9 0.0840
## 10        10 0.0782
## # ℹ 990 more rows
```

Plot the density curve of this distribution

```
ggplot(boot_dist_call, aes(x = stat)) +
  geom_density()
```



The standard deviation of the stat variable in this data frame (the bootstrap distribution) is the bootstrap standard error and it can be calculated using the summarize() function.

```
SE_call <- boot_dist_call |>
  summarize(se = sd(stat)) |>
  pull()
SE_call
```

```
## [1] 0.003929482
```

We can use this value, along with our point estimate, to roughly calculate a 95% confidence interval:

$$\hat{p} \pm z^* se$$

```
c(p_hat_call - 1.96 * SE_call, p_hat_call + 1.96 * SE_call)
```

```
## [1] 0.07279103 0.08819460
```

We are 95% confident that the true proportion of applicants receiving callbacks is between 7.28% and 8.81%.

A more efficient way to get the CI

Use get_ci

```
ci <- get_ci(boot_dist_call, level=.95)
ci
```

```
## # A tibble: 1 × 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   0.0721   0.0877
```

We are 95% confident that the true proportion of applicants receiving callbacks is between 7.33% and 8.77%.

The normal distribution for confidence interval

Another option for calculating the CI is by estimating it by using the Normal Distribution (the bell curve)

If

1. observations are independent
2. n is large (S-F condition is met)

Then

$\hat{p}$ follows a normal distribution

Steps

Calculate proportion receiving callback

```
n <- nrow(resume1)
p_hat_call <- resume1|>
  summarize(prop_call = mean(called == "yes")) |>
  pull()
p_hat_call
```

```
## [1] 0.08049281
```

Check conditions (both should be >= 10)

```
n * p_hat_call
```

```
## [1] 392
```

```
n * (1 - p_hat_call)
```

```
## [1] 4478
```

Calculate SE

```
se_call_approx <- sqrt(p_hat_call * (1 - p_hat_call) / n)
```

Calculate

$$z^*$$

```
z_star <- qnorm(.975, m=0, sd = 1)
```

Form 95% CI

```
c(p_hat_call - z_star * se_call_approx, p_hat_call + z_star * se_call_approx)
```

```
## [1] 0.07285200 0.08813363
```

Hypothesis Test

Use the General Social Survey (gss) data

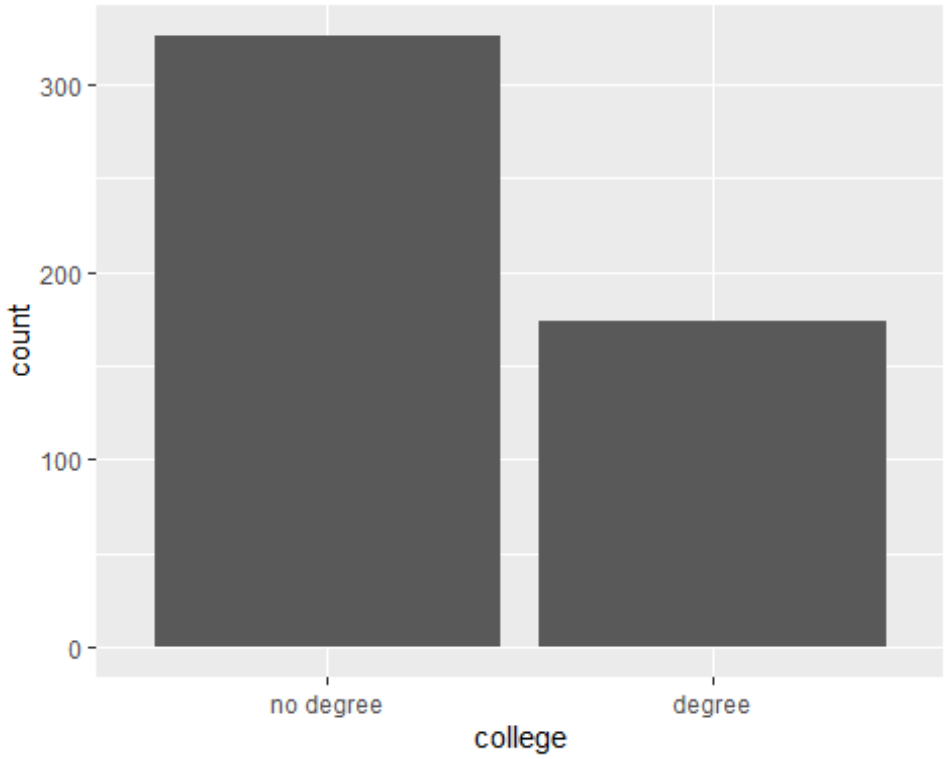
```
data("gss")
head(gss)
```

```
## # A tibble: 6 × 11
##    year   age sex    college   partyid hompop hours income  class finrela weight
##   <dbl> <dbl> <fct>  <fct>     <fct>    <dbl> <dbl> <ord>   <fct> <fct>    <dbl>
## 1  2014    36 male   degree    ind          3    50 $25000… midd… below …  0.896
## 2  1994    34 female no degree rep          4    31 $20000… work… below …  1.08
## 3  1998    24 male   degree    ind          1    40 $25000… work… below …  0.550
## 4  1996    42 male   no degree ind          4    40 $25000… work… above …  1.09
## 5  1994    31 male   degree    rep          2    40 $25000… midd… above …  1.08
## 6  1996    32 female no degree rep          4    53 $25000… midd… average  1.09
```

Do 1/3 of Americans have a college degree?

Use a hypothesis test to study a question on the gss. Respondents were asked if have a college degree. You can look at the distribution of answers by forming a bar chart. You see that of the 500 respondents, about 180 have a degree.

```
ggplot(gss, aes(x=college)) +
  geom_bar()
```



State the null and alternative hypotheses

$H_o: p = 1/3$ Ho: The proportion of US adults who have a college degree is 1/3.

$H_a: p \neq 1/3$ Ha: The proportion of US adults who have a college degree is different from 1/3.

Check Success-Failure Condition

np >=10 n(1-p)>=10

```
gss |>
  group_by(college)|>
  count()
```

```
## # A tibble: 2 × 2
## # Groups:   college [2]
##   college       n
##   <fct>     <int>
## 1 no degree   326
## 2 degree      174
```

Calculate the sample proportion

```
p_hat <- gss |>
  summarize(mean(college == "degree")) |>
  pull()
p_hat
```

```
## [1] 0.348
```

p-hat = 0.348

Now perform the hypothesis test

```
null_dist <- gss |>
  specify(response = college, success = "degree") |>
  hypothesize(null = "point", p = 1/3) |>
  generate(reps = 1000, type = "draw") |>
  calculate(stat = "prop")
null_dist
```

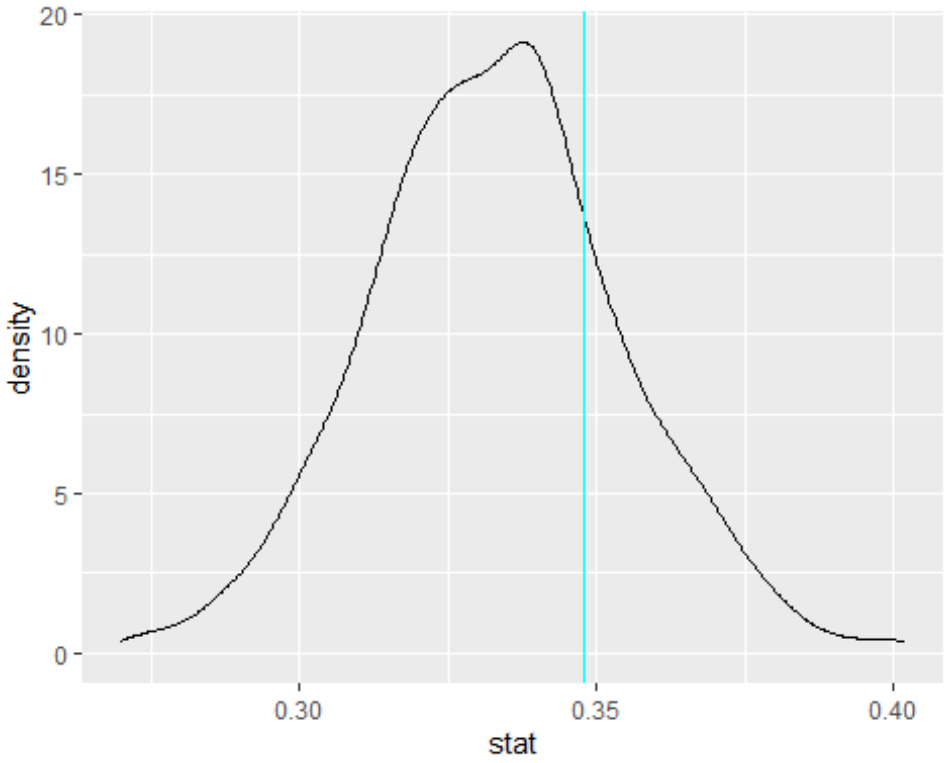
```
## Response: college (factor)
## Null Hypothesis: point
## # A tibble: 1,000 × 2
##    replicate  stat
##        <int> <dbl>
##  1         1 0.334
##  2         2 0.37
##  3         3 0.338
##  4         4 0.322
##  5         5 0.326
##  6         6 0.316
##  7         7 0.342
##  8         8 0.334
##  9         9 0.322
## 10        10 0.33
## # ℹ 990 more rows
```

Create a density curve of the null distribution

Add a vertical line (geom_vline) to indicate where p-hat lies on the curve.

```
ggplot(null_dist, aes(x = stat)) +
  geom_density() +
  geom_vline(
    xintercept = p_hat,
```

```
    color = "cyan"
  )
```



Calculate a two-tailed p-value (multiply by two)

```
pvalue <- null_dist |>
  summarize(mean(stat > p_hat)) |>
  pull() * 2
pvalue
```

```
## [1] 0.434
```

```
pvalue2 <- get_p_value(null_dist, p_hat, direction = "two-sided")
pvalue2
```

```
## # A tibble: 1 × 1
##   p_value
##     <dbl>
## 1   0.476
```

The p-value is very large, at 0.534. We fail to reject the null. There is no evidence that the proportion of US adults with a college degree is different than 1/3 the population.

Homework Chapter 16

1. Review section 16.3 (the chapter review)
2. **Suggested**: from textbook section 16.4 exercises: 1, 3, 8, 10, 14, 19, 25

3. **Suggested**: tutorials

1 - Inference for a single proportion

2 - Hypothesis tests to compare proportions