Lab6 - Inference with Categorical Variables

Keith Colella

2023-03-19

Setup

Config

```
library(tidyverse)
library(openintro)
library(infer)
library(cowplot)
set.seed(2012)
```

Data

```
data('yrbss', package='openintro')
no_helmet <- yrbss %>%
  filter(helmet_12m == "never")
```

Exercise 1

Question

What are the counts within each category for the amount of days these students have texted while driving within the past 30 days?

Response

```
table(yrbss$text_while_driving_30d)
```

```
##
##             0           1-2         10-19         20-29           3-5
##          4792           925           373           298           493
##            30           6-9 did not drive
##           827           311          4646
```

```
yrbss %>%
  filter(text_while_driving_30d != 'did not drive') %>%
  select(text_while_driving_30d) %>%
  table() %>% as.data.frame() %>%
  mutate(Prop = Freq / sum(Freq)) %>%
  arrange(factor(text_while_driving_30d,
                 levels = c('0','1-2','3-5','6-9','10-19','20-29','30')))
```

```
##   text_while_driving_30d Freq       Prop
## 1                      0 4792 0.59758075
## 2                    1-2  925 0.11535104
## 3                    3-5  493 0.06147899
## 4                    6-9  311 0.03878289
## 5                  10-19  373 0.04651453
## 6                  20-29  298 0.03716174
## 7                     30  827 0.10313007
```

The table above summarizes the frequency of responses under each category. Based on these responses, the majority of respondents who drove within the past 30 days did so without texting. Good to hear!

Exercise 2

Question

What is the proportion of people who have texted while driving every day in the past 30 days and never wear helmets?

Response

```
yrbss %>%
  select(text_while_driving_30d, helmet_12m) %>%
  table() %>% as.data.frame() %>%
  mutate(Prop = Freq / sum(Freq)) %>%
  filter(text_while_driving_30d == '30',
         helmet_12m == 'never')
```

```
##   text_while_driving_30d helmet_12m Freq       Prop
## 1                     30      never  463 0.03738695
```

Of the respondents, 463 answered that they texted while driving on 30 out of the last 30 days and never wore a helmet over the last 12 months, representing ~3.74% of total respondents.

Exercise 3

Question

What is the margin of error for the estimate of the proportion of non-helmet wearers that have texted while driving each day for the past 30 days based on this survey?

Response

```
yrbss <- yrbss %>%
  mutate(text_binary = if_else(text_while_driving_30d == '30',
                               'yes','no'))

yrbss %>%
  filter(helmet_12m == 'never',
         !is.na(text_binary)) %>%
  specify(response = text_binary, success = 'yes') %>%
  generate(reps = 1000, type = 'bootstrap') %>%
  calculate(stat = 'prop') %>%
  get_ci(level = 0.95) %>%
  mutate(margin_error = (upper_ci - lower_ci) / 2)
```

```
## # A tibble: 1 × 3
##   lower_ci upper_ci margin_error
##      <dbl>    <dbl>        <dbl>
## 1   0.0649   0.0770      0.00608
```

The margin of error for the proportion of non-helmet wearers who text while driving everyday is ~0.65%. This formulation removed all observations for which `text_while_driving_30d` was NA. Different treatments of NAs may drive slightly different results.

Exercise 4

Question

Using the `infer` package, calculate confidence intervals for two other categorical variables (you'll need to decide which level to call "success", and report the associated margins of error. Interpet the interval in context of the data. It may be helpful to create new data sets for each of the two countries first, and then use these data sets to construct the confidence intervals.

Response

I calculate confidence intervals for the proportion of high schoolers who (i) got 0 hours of physical activity per week and (ii) watched more than 5 hours of TV per school day. On the first point, I estimate between 15.#% and 16.#% of high schoolers get 0 hours of physical activity a week. On the second, I estimate between 11.#% and 12.#% of high schoolers watch more than 5 hours of TV a day.

```
yrbss %>%
  filter(!is.na(physically_active_7d)) %>%
  mutate(phys_act_binary  = if_else(
    physically_active_7d == 0,'yes','no')) %>%
  specify(response = phys_act_binary, success = 'yes') %>%
  generate(reps = 1000, type = 'bootstrap') %>%
  calculate(stat = 'prop') %>%
  get_ci(level = 0.95) %>%
  mutate(margin_error = (upper_ci - lower_ci) / 2)
```

```
## # A tibble: 1 × 3
##   lower_ci upper_ci margin_error
##      <dbl>    <dbl>        <dbl>
## 1    0.157    0.170      0.00676
```

```
yrbss %>%
  filter(!is.na(hours_tv_per_school_day )) %>%
  mutate(hour_tv_binary  = if_else(
    hours_tv_per_school_day == '5+','yes','no')) %>%
  specify(response = hour_tv_binary, success = 'yes') %>%
  generate(reps = 1000, type = 'bootstrap') %>%
  calculate(stat = 'prop') %>%
  get_ci(level = 0.95) %>%
  mutate(margin_error = (upper_ci - lower_ci) / 2)
```

```
## # A tibble: 1 × 3
##   lower_ci upper_ci margin_error
##      <dbl>    <dbl>        <dbl>
## 1    0.115    0.126      0.00555
```

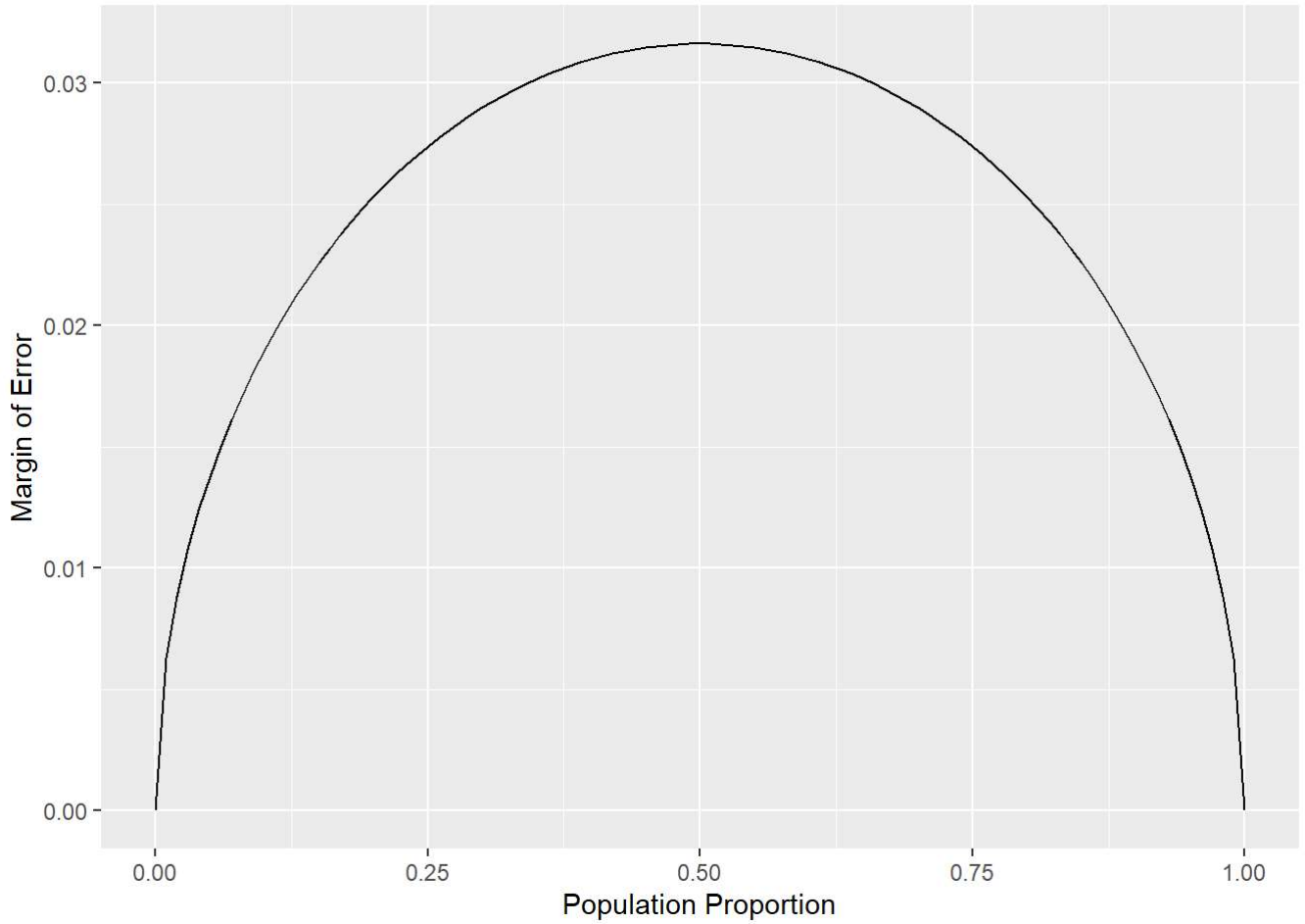
Exercise 5

Question

Describe the relationship between `p` and `me`. Include the margin of error vs. population proportion plot you constructed in your answer. For a given sample size, for which value of p is margin of error maximized?

Response

```
n <- 1000
p <- seq(from = 0, to = 1, by = 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
dd <- data.frame(p = p, me = me)
ggplot(data = dd, aes(x = p, y = me)) +
  geom_line() +
  labs(x = "Population Proportion", y = "Margin of Error")
```



For a given sample size, the margin of error is smallest as the population proportion nears 0 and 1. Conversely, the margin of error greater as the proportion nears 0.5. It appears the margin of error is greatest exactly at the proportion of 0.5.

Exercise 6

Question

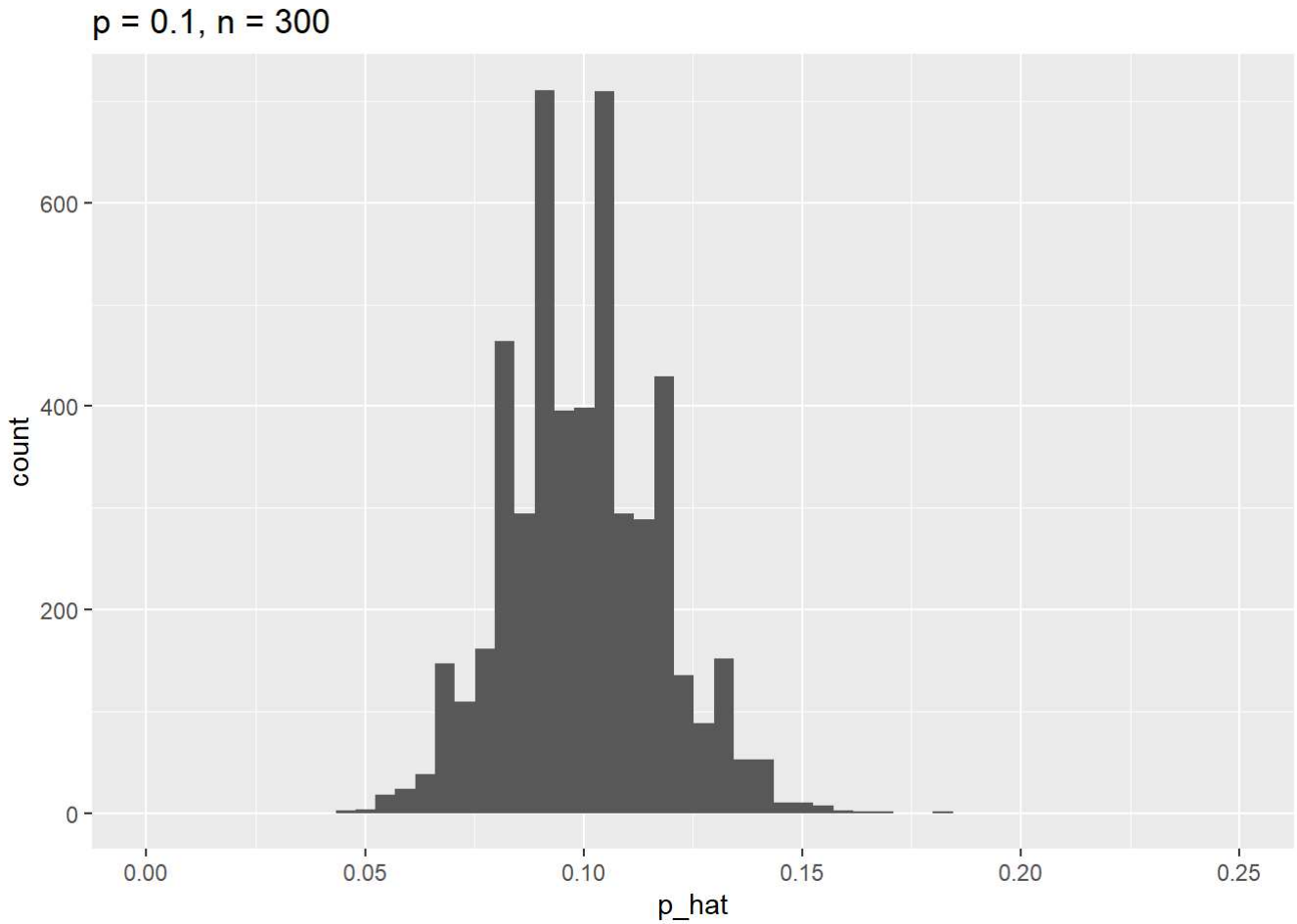
Describe the sampling distribution of sample proportions at $n = 300$ and $p = 0.1$. Be sure to note the center, spread, and shape.

Response

```
plot_samp_dist <- function(n,p,x_min,x_max) {
  pp <- data.frame(p_hat = rep(0, 5000))
      for(i in 1:5000){
        samp <- sample(c(TRUE, FALSE), n, replace = TRUE,
                       prob = c(p, 1 - p))
        pp$p_hat[i] <- sum(samp == TRUE) / n
      }
      bw <- diff(range(pp$p_hat)) / 30
      ggplot(data = pp, aes(x = p_hat)) +
        geom_histogram(binwidth = bw) +
        xlim(x_min, x_max) +
        ggtitle(paste0("p = ", p, ", n = ", n))
}

plot_samp_dist(300,0.1,0,0.25)
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



The distribution appears centered on 0.1, aligning to p. Most values are between 0.5 and 0.15. In terms of shape, the distribution appears approximately normal, but we do see some quasi-modes appear different from the mean.

Exercise 7

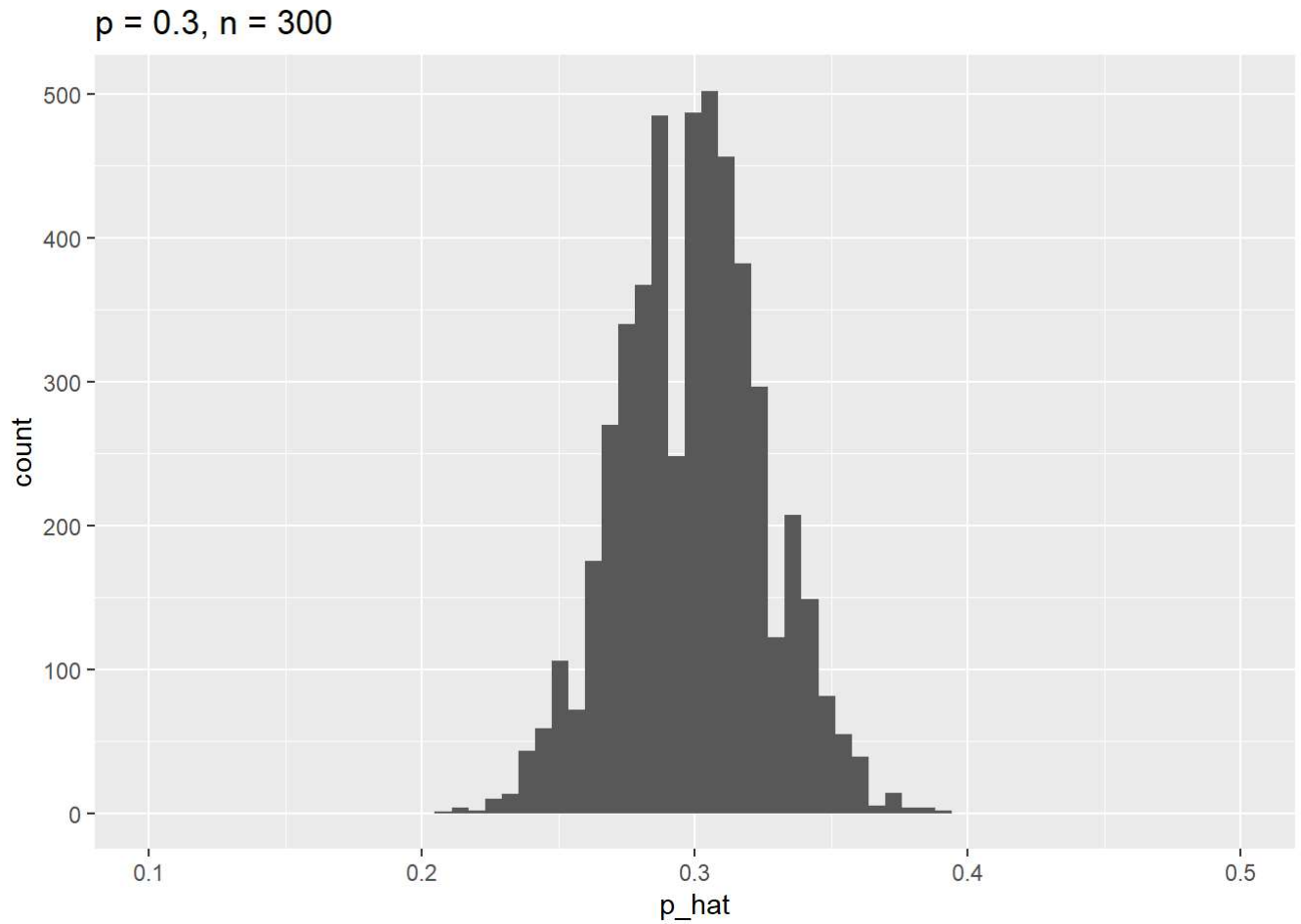
Question

Keep n constant and change p. How does the shape, center, and spread of the sampling distribution vary as p changes. You might want to adjust min and max for the x-axis for a better view of the distribution.

Response

```
plot_samp_dist(300,0.3,0.1,0.5)
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



This distributions appears to adhere more closely to a normal distribution. The mean and mode are much closer, and the overall shape is more symmetrical. There is, however, more spread, as most values now appear between a wider range of 0.2 and 0.4.

Exercise 8

Question

Now also change n. How does n appear to affect the distribution of p^?

Response

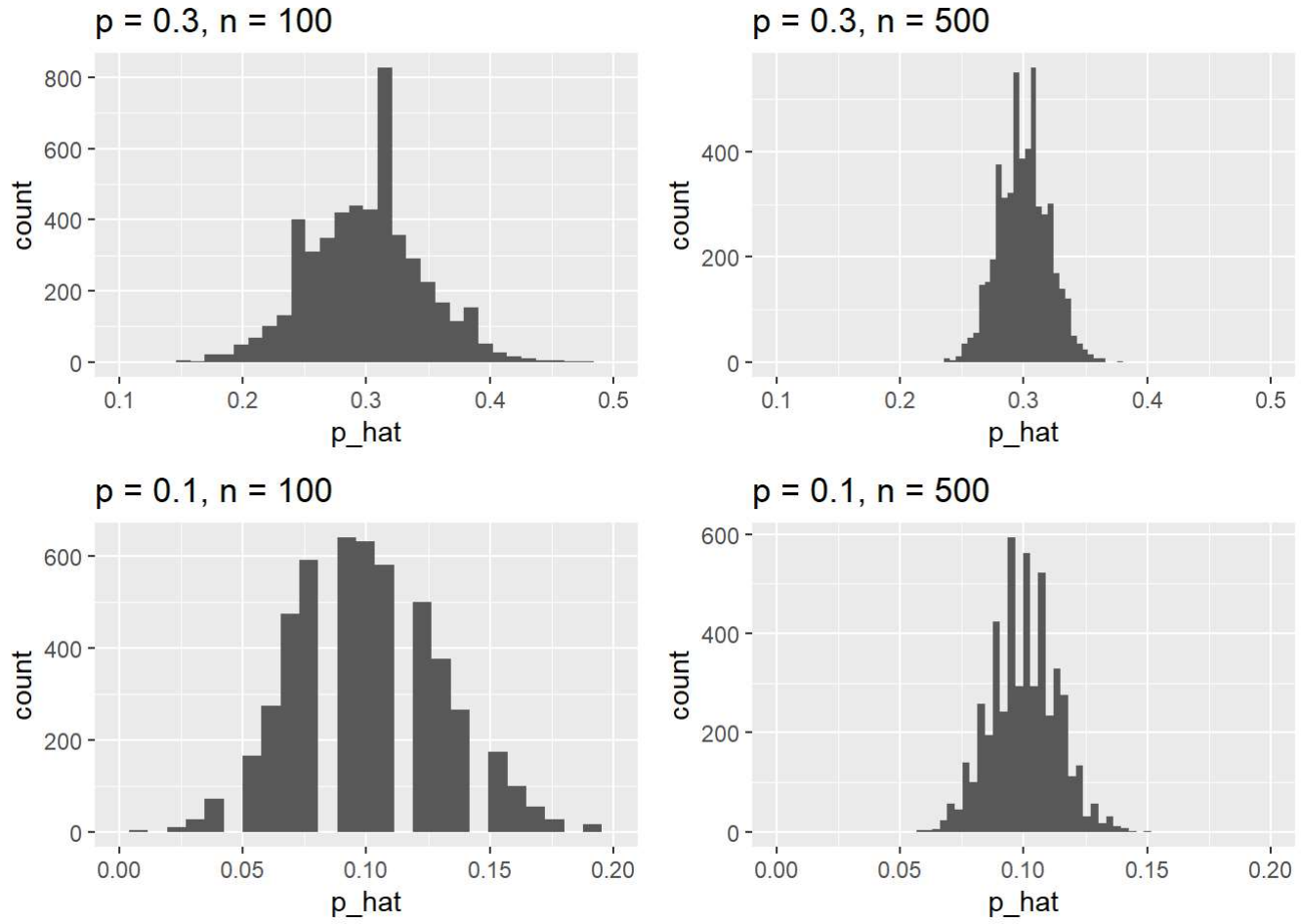
```
p1 <- plot_samp_dist(100,0.3,0.1,0.5)
p2 <- plot_samp_dist(500,0.3,0.1,0.5)
p3 <- plot_samp_dist(100,0.1,0.0,0.2)
p4 <- plot_samp_dist(500,0.1,0.0,0.2)

plot_grid(p1, p2, p3, p4, nrow = 2, ncol = 2)
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
## Removed 2 rows containing missing values (`geom_bar()`).
```

```
## Warning: Removed 5 rows containing non-finite values (`stat_bin()`).
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
## Removed 2 rows containing missing values (`geom_bar()`).
```



We can see that increasing the sample size reinforces the normality of the sampling distribution, for a given p. So, as p nears 0.5, the normality appears to increase, but so does the spread (as note previously in viewing the interaction of p and the margin of error). Similarly, increasing the sample sizes increases the normality and *decreases* spread.

Exercise 9

Question

Is there convincing evidence that those who sleep 10+ hours per day are more likely to strength train every day of the week? As always, write out the hypotheses for any tests you conduct and outline the status of the conditions for inference. If you find a significant difference, also quantify this difference with a confidence interval.

Response

We are testing whether the proportion of high schoolers who sleep 10+ hours a day and workout 7 days a week is different from the proportion of high schoolers who do *not* sleep 10+ hours a day and workout 7 days a week. Our null hypothesis is that the difference between these proportions is zero, and out alternative hypothesis is that the difference is *not* zero.

$H_0 : p_1 - p_2 = 0$
$H_\alpha : p_1 - p_2 \neq 0$

```
yrbss %>%
  filter(!is.na(school_night_hours_sleep),
         !is.na(strength_training_7d)) %>%
  mutate(sleep_10 = if_else(school_night_hours_sleep == '10+',
                            'yes','no'),
         strength_7 = if_else(strength_training_7d == '7',
                            'yes','no')) %>%
  specify(response = strength_7, explanatory = sleep_10, success = 'yes') %>%
  generate(reps = 1000, type = 'bootstrap') %>%
  calculate(stat = 'diff in props', order = c('no','yes')) %>%
  get_ci(level = 0.95)
```

```
## # A tibble: 1 × 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.156  -0.0530
```

Our resulting confidence interval does *not* include zero, so we can conclude with 95% confidence that the difference between the two populations is not zero (i.e. we reject H_0). The negative values indicate that the proportion of heavy trainers who also sleep 10+ hours a day is *lower* than the proportion of heavy trainers who sleep <10 hours a day (i.e. $p_1 < p_2$). In other words, we conclude that high schoolers who sleep 10+ hours per day are *less* likely to strength train every day of the week. Our confidence interval indicates that heavy sleepers are ~6% to ~16% less likely to train sever days a week.

Exercise 10

Question

Let's say there has been no difference in likeliness to strength train every day of the week for those who sleep 10+ hours. What is the probability that you could detect a change (at a significance level of 0.05) simply by chance? *Hint:* Review the definition of the Type 1 error.

Response

The probability of wrongly rejecting a null hypothesis that is actually true (i.e. a Type 1 error) is the same as the significance level. So, in our example, we would expect that, there is a 5% likelihood that we observed a difference in the two proportions simply by chance.

Exercise 11

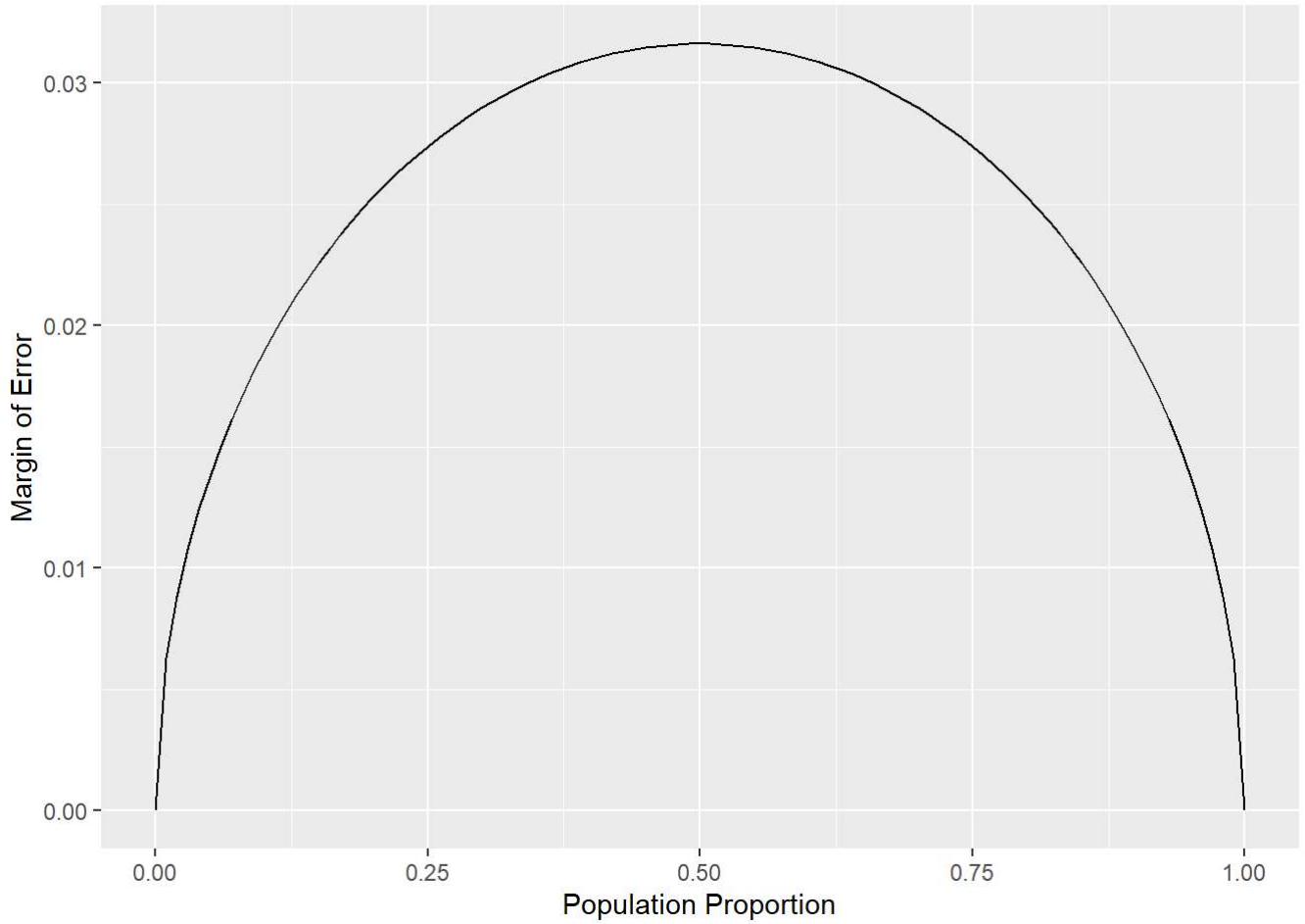
Question

Suppose you're hired by the local government to estimate the proportion of residents that attend a religious service on a weekly basis. According to the guidelines, the estimate must have a margin of error no greater than 1% with 95% confidence. You have no idea what to expect for p. How many people would you have to sample to ensure that you are within the guidelines? *Hint:* Refer to your plot of the relationship between p and margin of error. This question does not require using a dataset.

Response

If we reconstruct the earlier plot of p against margin of error, we see that the maximum margin of error occurs at $p = 0.5$.

```
n <- 1000
p <- seq(from = 0, to = 1, by = 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
dd <- data.frame(p = p, me = me)
ggplot(data = dd, aes(x = p, y = me)) +
  geom_line() +
  labs(x = "Population Proportion", y = "Margin of Error")
```



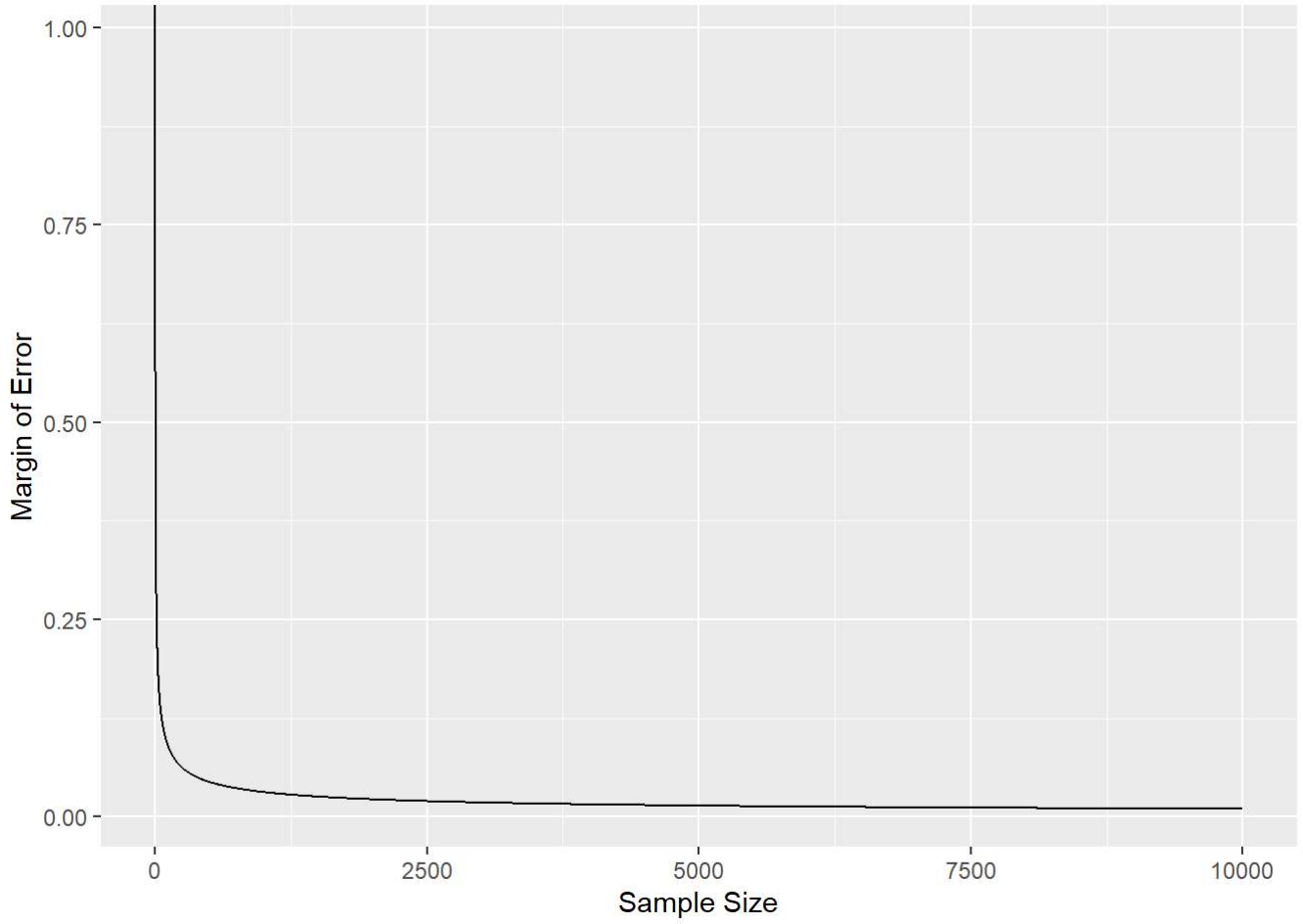
So, we can get a sense of the largest margin of error by assuming that $p = 50$. Moreover, we know that we'll be using a confidence level of 95%, so we know that our z-value will be 1.96. We can then use the formula for margin of hour to back out the minimum sample size needed to ensure our margin of error does not exceed 1%.

$$z * \sqrt{\frac{p(1-p)}{n}}$$

We adapt the approach above exploring the interaction between p and the margin of error to plot n and the margin of error. We then isolate the sample size that gives us a ~1% margin of error.

```
n <- seq(from = 0, to = 10000, by = 1)
p <- 0.5
me <- 1.96 * sqrt(p * (1 - p) / n)
dd <- data.frame(n = n, me = me)

ggplot(data = dd, aes(x = n, y = me)) +
  geom_line() +
  labs(x = "Sample Size", y = "Margin of Error")
```




```
dd %>%
  filter(round(me,4) == 0.0100) %>%
  head()
```

```
##      n         me
## 1 9509 0.01004983
## 2 9510 0.01004930
## 3 9511 0.01004877
## 4 9512 0.01004824
## 5 9513 0.01004772
## 6 9514 0.01004719
```

We conclude that we need a sample size of >9,509 to obtain a margin of error no larger than ~1.00% with a 95% confidence level.
