DATA 606 Lab 6: Inference for categorical data

Kevin Kirby

Getting Started

Load packages

In this lab, we will explore and visualize the data using the **tidyverse** suite of packages, and perform statistical inference using **infer**. The data can be found in the companion package for OpenIntro resources, **openintro**.

Let's load the packages.

```
library(tidyverse)
library(openintro)
library(infer)
```

The data

You will be analyzing the same dataset as in the previous lab, where you delved into a sample from the Youth Risk Behavior Surveillance System (YRBSS) survey, which uses data from high schoolers to help discover health patterns. The dataset is called `yrbss`.

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

```
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 
```

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

55.99% of people who have texted while driving every day in the past 30 days have also never worn helmets.

```
text_30 <- sum(yrbss$text_while_driving_30d == "30", na.rm = TRUE)
text_30_nohelmet <- sum(yrbss$text_while_driving_30d == "30" & yrbss$helmet_12m == "never", na.rm = TRU
proportion_never_helmets <- round(text_30_nohelmet / text_30,4)

proportion_never_helmets
```

```
## [1] 0.5599
```

Remember that you can use `filter` to limit the dataset to just non-helmet wearers. Here, we will name the dataset `no_helmet`.

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

Also, it may be easier to calculate the proportion if you create a new variable that specifies whether the individual has texted every day while driving over the past 30 days or not. We will call this variable `text_ind`.

```
no_helmet <- no_helmet %>%
  mutate(text_ind = ifelse(text_while_driving_30d == "30", "yes", "no"))
```

Inference on proportions

When summarizing the YRBSS, the Centers for Disease Control and Prevention seeks insight into the population *parameters*. To do this, you can answer the question, "What proportion of people in your sample reported that they have texted while driving each day for the past 30 days?" with a statistic; while the question "What proportion of people on earth have texted while driving each day for the past 30 days?" is answered with an estimate of the parameter.

The inferential tools for estimating population proportion are analogous to those used for means in the last chapter: the confidence interval and the hypothesis test.

```
nohelmet_ci <- no_helmet %>%
  drop_na(text_ind) %>% # Drop missing values
  specify(response = text_ind, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
```

Note that since the goal is to construct an interval estimate for a proportion, it's necessary to both include the `success` argument within `specify`, which accounts for the proportion of non-helmet wearers than have consistently texted while driving the past 30 days, in this example, and that `stat` within `calculate` is here "prop", signaling that you are trying to do some sort of inference on a proportion.

3. 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?

Answer: The margin of error is going to be the difference between the upper_ci and lower_ci divided by 2.I turned your provided code above into a stored dataframe, allowing me to do this pretty easily.The margin of error is 0.006076

```
round((nohelmet_ci$upper_ci - nohelmet_ci$lower_ci)/2,6)
```

```
## [1] 0.006459
```

4. 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.

Answer: This first one establishes that 6 or less hours of sleep means they're not sleeping enough and "lacks sleep." I created a dataframe with a new indicator column and then calculated the confidence interval, which is 0.4305 for the lower_ci and 0.4482 for the upper_ci. The margin of error is 0.0088. I'm not terribly surprised 44% of people in this population aren't sleeping more than six hours a night. Middle school and high school kids are famous (notorious?) for being able to power through full days on minimal sleep. The hard life lessons about the importance of sleep don't come until later in life.

```
lacks_sleep <- yrbss %>%
  mutate(lacksleep_ind = ifelse(school_night_hours_sleep <= "6", "yes", "no"))

lacks_sleep_ci <- lacks_sleep %>%
  drop_na(lacksleep_ind) %>%
  specify(response = lacksleep_ind, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)

cat("The confidence interval is: \n")
```

```
## The confidence interval is:
```

```
lacks_sleep_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.431    0.449
```

```
cat("\nThe margin of error is: \n")
```

```
##
## The margin of error is:
```

```
round((lacks_sleep_ci$upper_ci - lacks_sleep_ci$lower_ci)/2,6)
```

```
## [1] 0.008877
```

The second one looks the proportion of males that don't wear helmets and watch 3 or more hours of TV a day. I've decided that this is an indicator of toxic masculinity. The lower_ci is 0.276 and the upper_ci is 0.298, with a margin of error of 0.0108. Considering I myself was once a 12-18 year male who did not wear a helmet and watched a lot of TV after school, I'm honestly surprised it's only 28% of this male subset.

```
masculinity <- yrbss %>%
  filter(gender == "male")

toxic_masculinty <- masculinity %>%
  mutate(toxic_ind = ifelse(helmet_12m == "never" & hours_tv_per_school_day >= 3, "yes", "no"))

toxic_masculinty_ci <- toxic_masculinty %>%
  drop_na(toxic_ind) %>%
  specify(response = toxic_ind, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)

cat("The confidence interval is: \n")
```

```
## The confidence interval is:
```

```
toxic_masculinty_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.276    0.299
```

```
cat("\nThe margin of error is: \n")
```

```
##
## The margin of error is:
```

```
round((toxic_masculinty_ci$upper_ci - toxic_masculinty_ci$lower_ci)/2,6)
```

```
## [1] 0.011146
```

How does the proportion affect the margin of error?

Imagine you've set out to survey 1000 people on two questions: are you at least 6-feet tall? and are you left-handed? Since both of these sample proportions were calculated from the same sample size, they should have the same margin of error, right? Wrong! While the margin of error does change with sample size, it is also affected by the proportion.

Think back to the formula for the standard error: $SE = \sqrt{p(1-p)/n}$. This is then used in the formula for the margin of error for a 95% confidence interval:

$$ME = 1.96 \times SE = 1.96 \times \sqrt{p(1-p)/n}\,.$$

Since the population proportion p is in this ME formula, it should make sense that the margin of error is in some way dependent on the population proportion. We can visualize this relationship by creating a plot of ME vs. p.

Since sample size is irrelevant to this discussion, let's just set it to some value ($n = 1000$) and use this value in the following calculations:

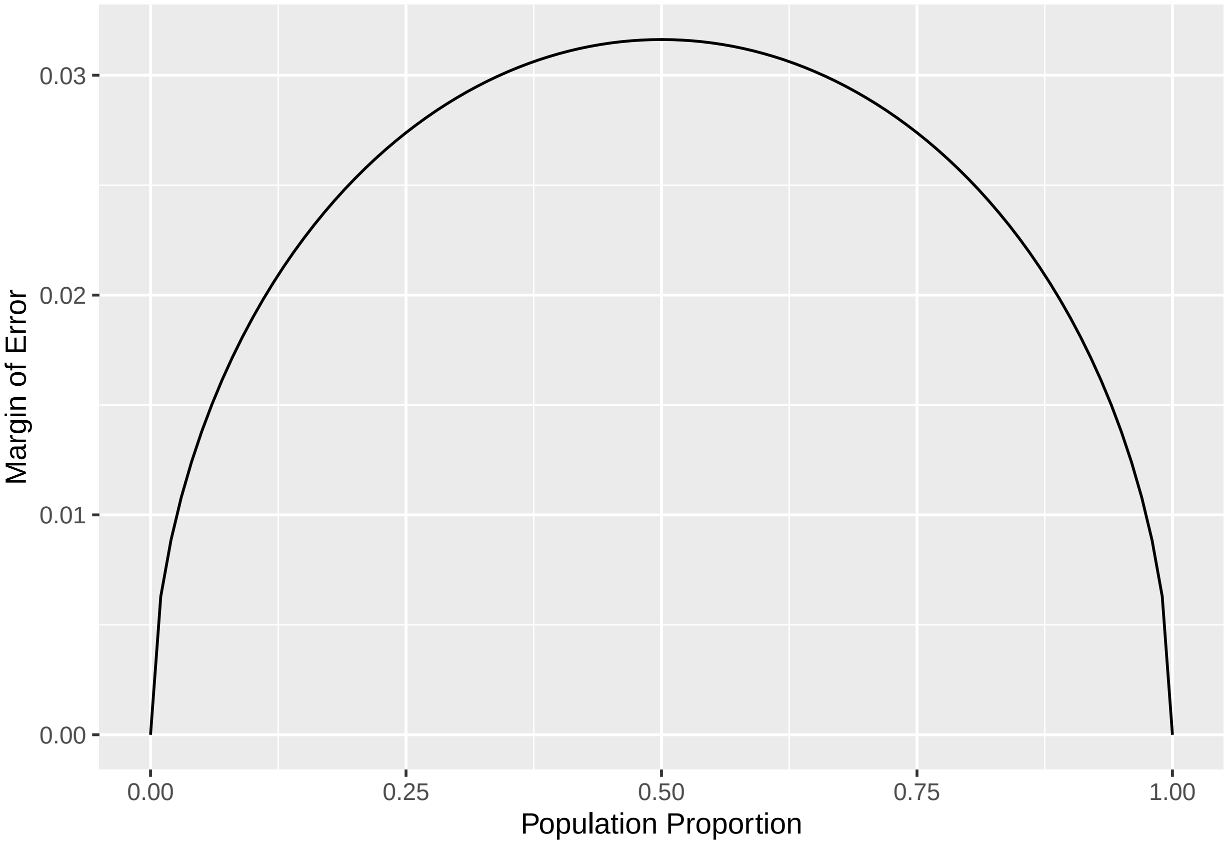
```
n <- 1000
```

The first step is to make a variable `p` that is a sequence from 0 to 1 with each number incremented by 0.01. You can then create a variable of the margin of error (`me`) associated with each of these values of `p` using the familiar approximate formula ($ME = 2 \times SE$).

```
p <- seq(from = 0, to = 1, by = 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
```

Lastly, you can plot the two variables against each other to reveal their relationship. To do so, we need to first put these variables in a data frame that you can call in the `ggplot` function.

```
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")
```

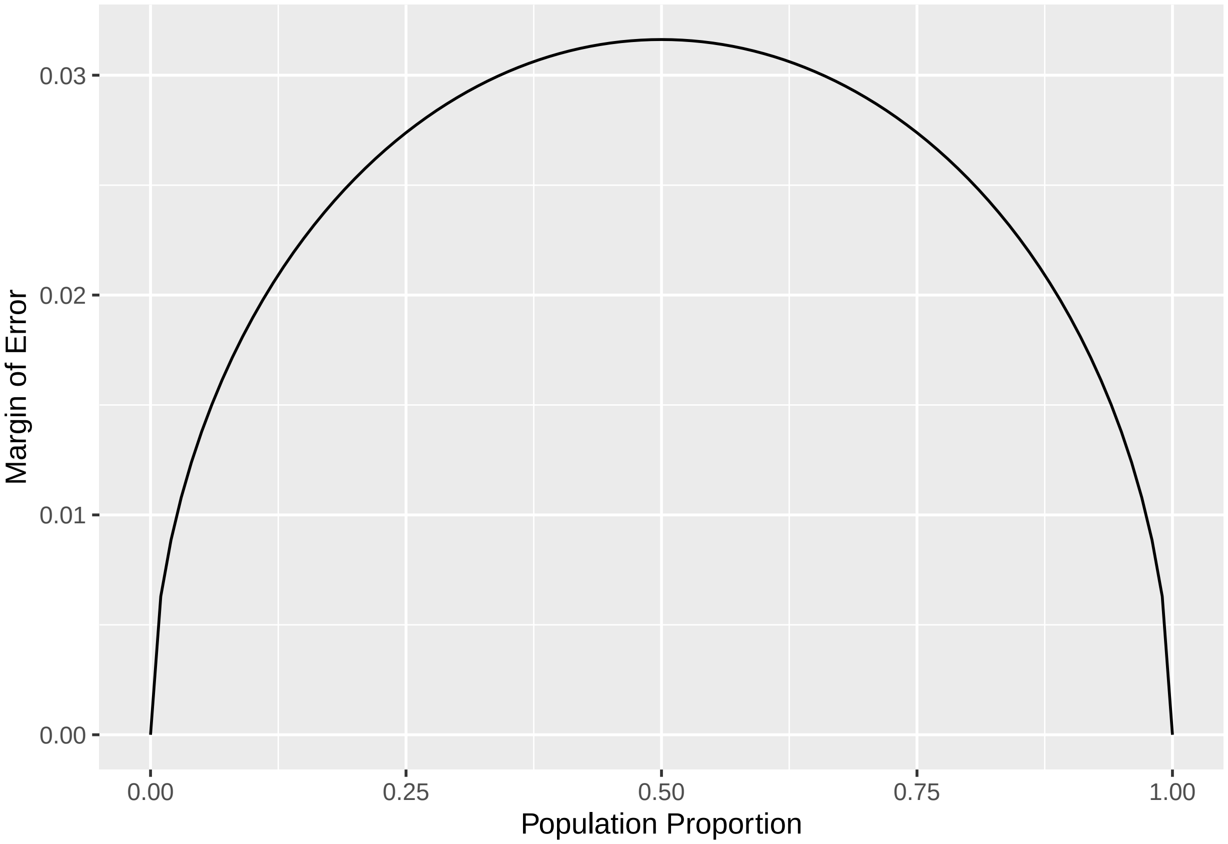


5. Describe the relationship between `p` and `me`. Include the margin of error vs. population proportion plot you constructed in your answer. For a given sample size, for which value of `p` is margin of error maximized?

Answer: Here's the plot showing the relationship betwen the porportion of the population and the margin of error. The plot follows the shape of a Laffer Curve, although entirely unrelated to the Laffer Curve.

The relation is symmetric, with the largest margin of error peaking around 0.05 for population proportion of 0.5. It seems that when the proportion is near 0, the margin of error is small because there is little uncertainty when only a small number of people exhibit the characteristic. On the other hand, the MOE peaks at 50% proportion because there's a there is a near-equal chance of individuals in the sample falling into either category.

```
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")
```



Success-failure condition

We have emphasized that you must always check conditions before making inference. For inference on proportions, the sample proportion can be assumed to be nearly normal if it is based upon a random sample of independent observations and if both $np \geq 10$ and $n(1-p) \geq 10$. This rule of thumb is easy enough to follow, but it makes you wonder: what's so special about the number 10?

The short answer is: nothing. You could argue that you would be fine with 9 or that you really should be using 11. What is the "best" value for such a rule of thumb is, at least to some degree, arbitrary. However, when np and $n(1-p)$ reaches 10 the sampling distribution is sufficiently normal to use confidence intervals and hypothesis tests that are based on that approximation.

You can investigate the interplay between n and p and the shape of the sampling distribution by using simulations. Play around with the following app to investigate how the shape, center, and spread of the distribution of $\hat{p}$ changes as n and p changes.

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

Answer: * Center: aroud 0.1, with a count around 750 * Spread: Most values are between 0.05 and 0.15, indicating low variability * Shape: A symmetric bell-shaped curve that looks like a normal distribution

This is a sample with an unremarkable normal curve.

7. 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.

Answer: p: 0.25 * Center: aroud 0.23, with a count around 580 * Spread: Most values are between 0.12 and 0.38, which is increased variability when compared to p 0.1 * Shape: A symmetri bell-shaped curve that looks like a normal distribution

The shape remained relatively stabble but the center became higher on the x-axis and lower on the y-axis, which is also supported by the increased spread. Based on the previous chart plotting the relationship between proportion and MOE, I would expect these measures of uncertainty to keep increasing until p 0.5 and then start to come down again.

8. Now also change n. How does n appear to affect the distribution of $\hat{p}$?

Answer: p: 0.25 n: 1,000 * Center: 0.25, with a count around 480 * Spread:All values are between 0.22 and 0.28, which is decreased variability when compared to p 0.25 and n 300 * Shape: A very compact normal curve that looks like the Burj Khalifa in Dubai. There's no tail on either side.

The increase of n from 300 to 1000 had the effect of dramatically reducing the spread and margin of error. The increased sample size offset the increased uncertainty that came from increasing p from 0.1 to 0.25 earlier.

More Practice

For some of the exercises below, you will conduct inference comparing two proportions. In such cases, you have a response variable that is categorical, and an explanatory variable that is also categorical, and you are comparing the proportions of success of the response variable across the levels of the explanatory variable. This means that when using `infer`, you need to include both variables within `specify`.

9. 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.

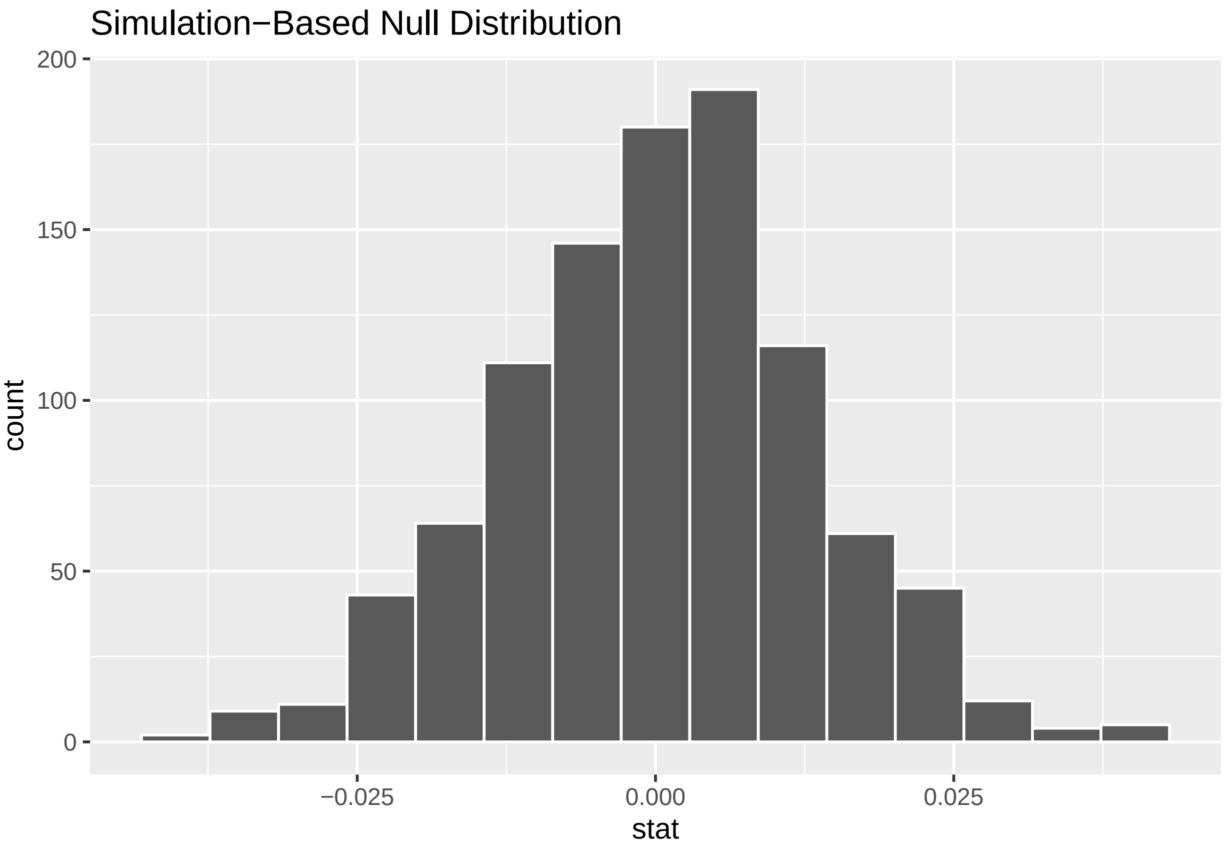
Answer: Here's the code to determine this. Required analysis beneath it.

```r
sleep_ten_bi <- yrbss %>%
  mutate(sleep_ten_ind = ifelse(school_night_hours_sleep >= 10, "yes", "no")) %>%
  mutate(strength_seven_ind = ifelse(strength_training_7d == 7, "yes", "no"))

ss_inf <- sleep_ten_bi %>%
  drop_na(sleep_ten_ind, strength_seven_ind) %>%
  specify(response = strength_seven_ind, explanatory = sleep_ten_ind, success = "yes")

ssnull_dist <- ss_inf %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in props", order = c("yes", "no"))

visualize(ssnull_dist)
```



```r
ss_obs <- ss_inf %>%
  calculate(stat = "diff in props", order = c("yes", "no"))

ss_pv <- ssnull_dist %>%
  get_p_value(obs_stat = ss_obs, direction = "two-sided")

ss_ci <- ss_inf %>%
  drop_na(sleep_ten_ind, strength_seven_ind) %>%
  specify(response = strength_seven_ind, explanatory = sleep_ten_ind, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in props", order = c("yes", "no")) %>%
  get_ci(level = 0.95, type = "percentile")

cat("The P value is:\n")
```

```
## The P value is:
```

```r
ss_pv
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.622
```

```r
cat("\nThe confidence interval is:\n")
```

```
##
## The confidence interval is:
```

```
ss_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1  -0.0177   0.0312
```

```
cat("\nThe margin of error is:\n")
```

```
##
## The margin of error is:
```

```
round((ss_ci$upper_ci - ss_ci$lower_ci)/2,6)
```

```
## [1] 0.024443
```

Analysis: The data shows no convincing evidence that sleeping 10+ hours a night increases the likelihood of strength training every day. The confidence interval above spans from -0.0194 to positive 0.0308 and includes zero, which means there's no statistically significant difference. The MOE of 0.0251 is narrow but inconclusive when placed in the context of the CI. Given this, the null hypothesis is not rejected and I don't conclude that sleep duration influences daily strength training.

10. 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 probablity 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.

This would be a Type 1 error if the null hypothesis is rejected but it'sd actually true. The probability of this happening would be equal to the significant level of the test, in this case 0.05. This means the probability of detecting a change simply by chance is 5%, making the probability of a Type 1 error also 5%.

11. 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.

Answer: For a 95% confidence level, I calculated the z-score using qnorm with 0.975 as the input. 0.975 is the input for a 95% confidence because I want a z-score that leaves 2.5% in each tail of the normal distribution. The confidence interval is symmetric so the leftover 5% is split equally between the two tails.

```
z_95 <- qnorm(0.975)

cat("The z-score is:\n")
```

```
## The z-score is:
```

```
z_95
```

```
## [1] 1.959964
```

```
moe <- 0.01
p_vic <- 0.5

n_vic <- round((z_95^2 * p_vic * (1 - p_vic)) / (moe^2),0)

cat("The required sample size given 95% confidence and 1% margin of error is:\n")
```

```
## The required sample size given 95% confidence and 1% margin of error is:
```

```
n_vic
```

```
## [1] 9604
```
