Inference for categorical data

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)

data('yrbss', package='openintro')
seed <- 1234
```

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?

WJ Response:

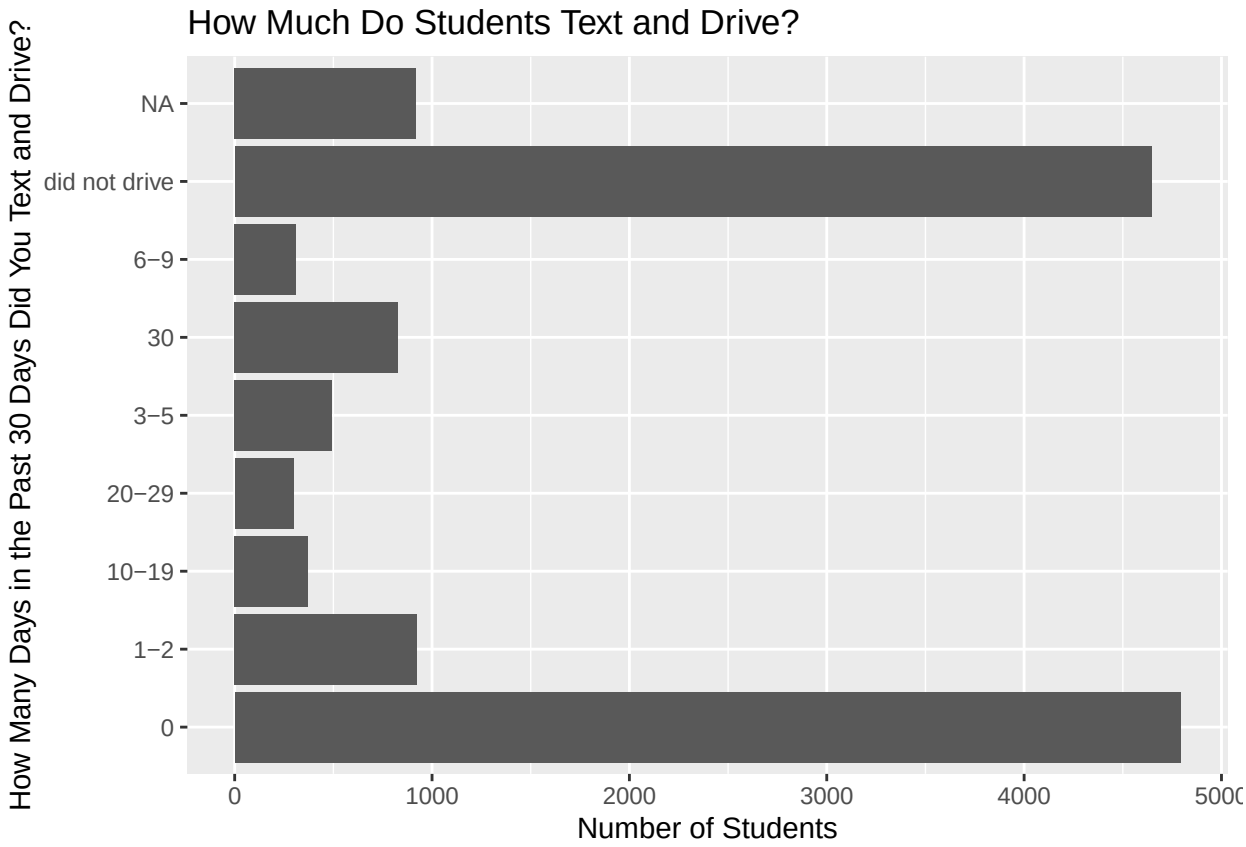
```
yrbss %>%
  count(text_while_driving_30d)
```

```
## # A tibble: 9 x 2
##   text_while_driving_30d     n
##   <chr>                  <int>
## 1 0                       4792
## 2 1-2                      925
## 3 10-19                    373
## 4 20-29                    298
## 5 3-5                      493
## 6 30                       827
## 7 6-9                      311
## 8 did not drive           4646
## 9 <NA>                     918
```

The output above above prints the number of students who selected each of the possible values in the `text_while_driving_30d` column. Most of the students selected 0, meaning that there were 0 days in the past 30 days in which they were texting and driving. The information is also displayed below in a bar graph:

```
plt_data <- yrbss %>%
  count(text_while_driving_30d)

ggplot(data=plt_data, aes(x=text_while_driving_30d, y=n)) +
  geom_bar(stat="identity") +
    labs(
      x = 'How Many Days in the Past 30 Days Did You Text and Drive?',
      y = 'Number of Students',
      title = 'How Much Do Students Text and Drive?'
    ) +
    coord_flip()
```



---

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

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

```
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"))
```

WJ Response:

The code below determines the proportion of students who never wear helments and text while driving 30 days out of the month. It first does this by creating a column `no_helment_always_texting` in the `yrbss` dataframe, that is always `TRUE` for students who match those conditions. It then counts the number of these students and divides by the total number of students to generate a proportion:

```
yrbss <- yrbss %>%
  mutate(no_helmet_always_texting =
           ifelse(text_while_driving_30d == '30' &
                  helmet_12m == 'never', TRUE, FALSE))

yrbss %>%
  filter(!is.na(no_helmet_always_texting)) %>%
    filter(no_helmet_always_texting == TRUE) %>%
      nrow() / nrow(yrbss)
```

```
## [1] 0.03408673
```

As can be seen in the output above, about 3.4% percent of the students surveyed never wear helments and have texted while driving every of the last 30 days.

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.

```
set.seed(seed)

no_helmet <- yrbss %>%
  filter(helmet_12m == "never")

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

no_helmet <- no_helmet %>%
  filter(!is.na(text_ind))

no_helmet %>%
  specify(response = text_ind, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
```

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

```
## 1   0.0652   0.0777
```

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?

WJ Response:

Based on the above result, we can see that by using a 95% confidence interval we get a range between 0.0652 and 0.0777. These values can be used to determine the margin of error:

```
set.seed(seed)

ci <- no_helmet %>%
  specify(response = text_ind, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)

(ci$upper_ci - ci$lower_ci) / 2
```

```
## [1] 0.006229817
```

Using the upper and lower confidence intervals, we can see above that our margin of error is approximately 0.0062. This means that 95% of the time, our estimate of the proportion of non-helmet wearers that have texted while driving every day for the past 30 days to be within a range of +/- 0.0062.

---

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.

WJ Response:

What is the proportion of students who are physically active all 7 days of the week and sleep at least 8 hours a day?

```
set.seed(seed)

yrbss <- yrbss %>%
  mutate(active_sleep_well = ifelse(
    physically_active_7d == '7' &
    school_night_hours_sleep %in% c('8', '9', '10+'), TRUE, FALSE))

healthy <- yrbss %>%
  filter(!is.na(active_sleep_well))

ci <- healthy %>%
  specify(response = active_sleep_well, success = 'TRUE') %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
```

```
  get_ci(level = 0.95)

print(ci)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   0.0856   0.0952
```

The above output means that 95% of the time our estimate of the proportion of those who both sleep more than 8 hours a night and are physically active all 7 days of the week to falls within 0.0856 and 0.0952. Translating this into terms of margin of error:

```
(ci$upper_ci - ci$lower_ci) / 2
```

```
## [1] 0.004791605
```

Using the 95% confidence interval the margin of error is approximately 0.0048 (about 0.05%), meaning that 95% of the time we can expect the estimate of those who sleep at least 8 hours a night and are physically active 7 days a week to be within a range of +/- 0.0048.

Of those students who typically get 5 or less hours of sleep on a school night, what proportion of those students typically watch at least 3 hours of TV on those same school nights?

```
set.seed(seed)

bad_sleepers <- yrbss %>%
  filter(school_night_hours_sleep %in% c('5', '<5')) %>%
  mutate(watch_tv_alot = ifelse(
    hours_tv_per_school_day %in% c('3', '4', '5+'), TRUE, FALSE))

bad_sleepers <- bad_sleepers %>%
  filter(!is.na(watch_tv_alot))

ci <- bad_sleepers %>%
  specify(response = watch_tv_alot, success = 'TRUE') %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)

print(ci)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.358    0.397
```

The above output means that 95% of the time our estimate of the proportion of those students who sleep 5 hours or less a night that watch TV for at least three hours a day to falls within 0.358 and 0.0397. Translating this into terms of margin of error:

```
(ci$upper_ci - ci$lower_ci) / 2
```

```
## [1] 0.0192229
```

Using the 95% confidence interval the margin of error is approximately 0.019 (about 2%). This means that 95% of the time we can expect the estimate of the proportion of those students who sleep less than 5 hours a night who watch more than 3 hours to TV a day to be within a range of +/- 0.019.

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:

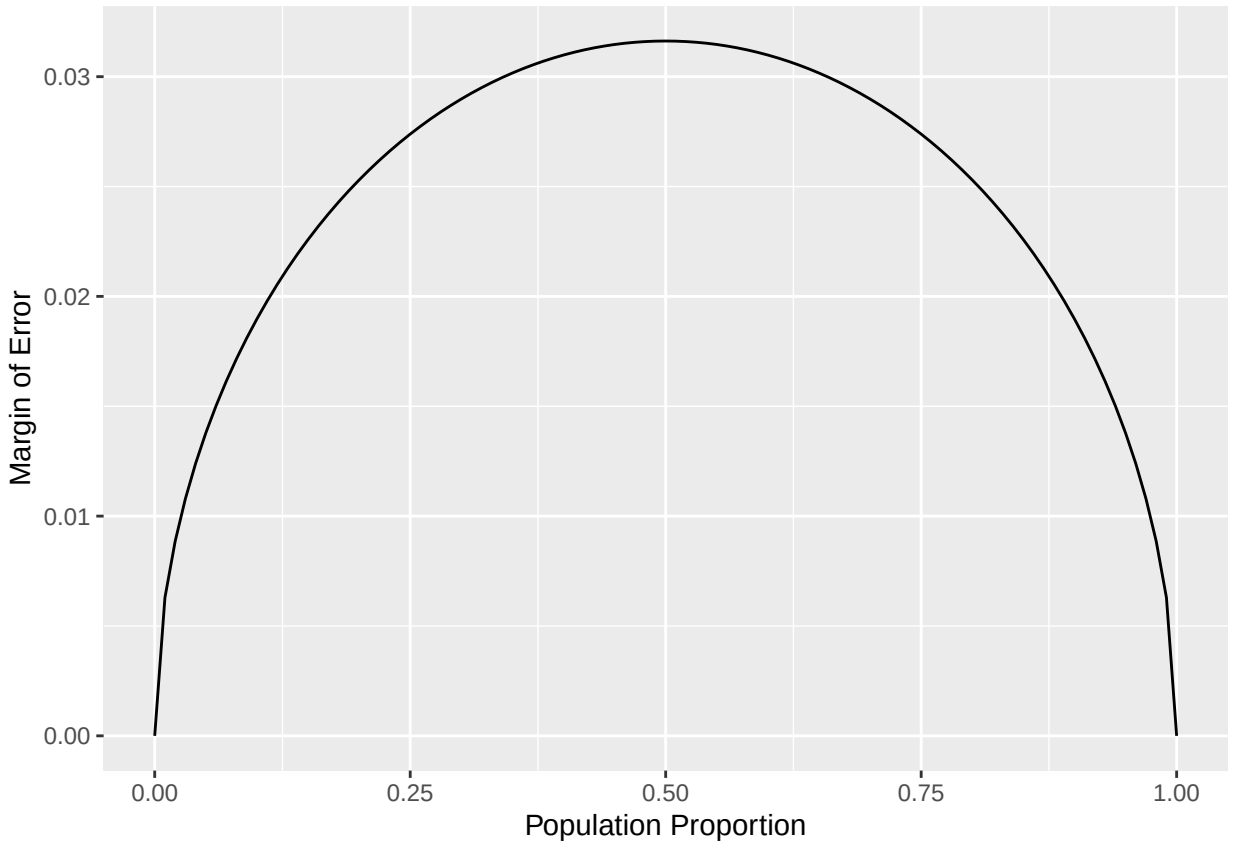
```
n <- 1000
```

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?

WJ Response:

Based on the above plot, the margin of error has a maximum when p has a value equal to 0.5 This is due to that fact that $p(1-p)$ reaches a maximum for that same value for values between 0 and 1. For values above and below 0.5, the margin of error decreases symmetrically.

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.

WJ Response:

Using the shiny app with those inputs described above, the sampling distribution exhibits the following characteristics:

- **Center:** The distribution is centered around the population proportion, which in this case is 0.1
- **Spread:** All values seem to fall within +/- 0.07 of the the population proportion.
- **Shape**: The distribution appears to be normally distributed, and centered around the population proportion value of 0.1.

---

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.

WJ Response:

The main change that occurs when shifting the value of p is the location of the center of the distribution, which should always be at p. The shape and spread of the distribution do not seem to change much except if the population proportion nears 0 or 1. This is due to the fact that the histogram is bound at these levels, and for such low and high values of p, np and $n(1-p)$ can decrease below 10, respectively. At these high and low population proportion values, we can start to see some of the normality of the distribution begin to wash out.

---

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

WJ Response:

The main effect of changing the value of n is on the spread of the sampling distribution. As n increases, the spread of the distribution decreases, and it begins to look more "perfectly" normal. Larger values of n make it possible for the normal appearance to be visible at the extreme values of p (close to 0 or 1), due to that fact that as n increases the values of p and $(1-p)$ required for np and $n(1-p)$ to be greater than or equal to 10 are smaller.

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.

WJ Response:

Let p_1 be the proportion of those students who sleep 10+ hours per day that strength train every day of the week. Let p_2 be the proportion of all other students who strength train every day of the week. The null and alternative hypotheses for this test is written below:

H_o: $p_1 - p_2 = 0$

H_A: $p_1 - p_2 \neq 0$

The hypothesis test is carried out below using 95% confidence intervals ($\alpha = 0.05$) on bootstrap samples of the test statistic ($p_1 - p_2$):

```
set.seed(seed)

yrbss %>%
  mutate(
    very_sleepy = ifelse(school_night_hours_sleep == '10+', TRUE, FALSE),
    bodybuilder = ifelse(strength_training_7d == '7', TRUE, FALSE)
  ) %>%
  filter(!is.na(very_sleepy) & !is.na(bodybuilder)) %>%
    specify(bodybuilder ~ very_sleepy, success = 'TRUE') %>%
    generate(reps = 1000, type = "bootstrap") %>%
    calculate(stat = "diff in props", order = c('FALSE', 'TRUE')) %>%
    get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.154  -0.0579
```

Thus given an α of 0.05, we can conclude that there is a statistical difference between these two proportions given that the confidence interval shown above does not bound 0 (we reject the null hypothesis). The conclusion to be drawn from this test is that the proportion of students who sleep more than 10+ hours a day are less likely to work out 7 days a week compared to the rest of the student body.

While a bootstrapping methodology was implemented above, these conclusions are confirmed when using the actual formula for the confidence interval of the difference between two proportions:

$$CI = (p_1 - p_2) \pm z^* \cdot \sqrt{\frac{p_1(1-p1)}{n1} + \frac{p_2(1-p2)}{n2}}$$

This formula is carried out for our example in R in the code chunk below:

```
yrbss_tmp <- yrbss %>%
  filter(!is.na(school_night_hours_sleep) & !is.na(strength_training_7d))

bad_sleepers <- yrbss_tmp %>%
  filter(school_night_hours_sleep == '10+')

bad_sleepers_strong <- bad_sleepers %>%
  filter(strength_training_7d == '7')

others <- yrbss_tmp %>%
  filter(school_night_hours_sleep != '10+')

others_strong <- others %>%
  filter(strength_training_7d == '7')
```

```r
p1 <- nrow(bad_sleepers_strong) / nrow(yrbss_tmp)
n1 <- nrow(bad_sleepers_strong)

p2 <- nrow(others_strong) / nrow(yrbss_tmp)
n2 <- nrow(others_strong)

confidence_level <- 0.95
z <- qnorm(1-(1-confidence_level)/2)

ci_lower <- (p1-p2) - z * sqrt((p1*(1-p1)/n1) + (p2*(1-p2)/n2))
ci_upper <- (p1-p2) + z * sqrt((p1*(1-p1)/n1) + (p2*(1-p2)/n2))

cat(ci_lower, ci_upper)
```

```
## -0.1773725 -0.1293629
```

Once again, we see that this confidence interval does not bound 0 and is close to the interval obtained from the bootstrapping methodology.

---

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 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.

WJ Response:

The question asked here can be simply boiled down to: what is the chance of getting a false positive result (type 1 error)? In this case, a false positive means that we detect a statistical difference between the proportion of students who strength train 7 days a week when comparing the populations of those who sleep 10+ hours a day and those who don't, even when there isn't a difference in actuality. Fortunately, the significance level used is exactly the incidence that we might obtain a false positive result. For example, using a significance level of 0.05 means there is a 5% chance we will obtain a false positive result.

---

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.

WJ Response:

This value can be obtained by using the formula to determine the margin of error for proportions (ME):

$ME = z^* \cdot \sqrt{\frac{p(1-p)}{n}}$

With the help of a little algebra, we can solve this equation in terms of n:

$n = \frac{z^{*2} p(1-p)}{ME^2}$

This equation can now be used to determine the sample size given to reach any margin of error, given that the probability is known. In this case, since the probability is not known, we can use $p = 0.5$ as it is the value at which the margin of error is at a maximum (the worse case scenario). Thus, using $p = 0.5$ ensures

that the formula will generate an n that will be less than or equal to the desired margin of error. The n required for a 1% margin of error is calculated below:

```
moe <- 0.01
z <- 1.96
p <- 0.5

z^2 * p * (1-p) / moe^2
```

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

Thus, to obtain a margin of error of 1% using a 95% confidence interval we would need to have a sample size >= 9,604.
