# Data 606 Lab 5B

Joao De Oliveira

2025-10-11

### Overview

This report intends to explore the foundations of the statistical inference and confidence intervals.

#### Load Packages

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(openintro)
```

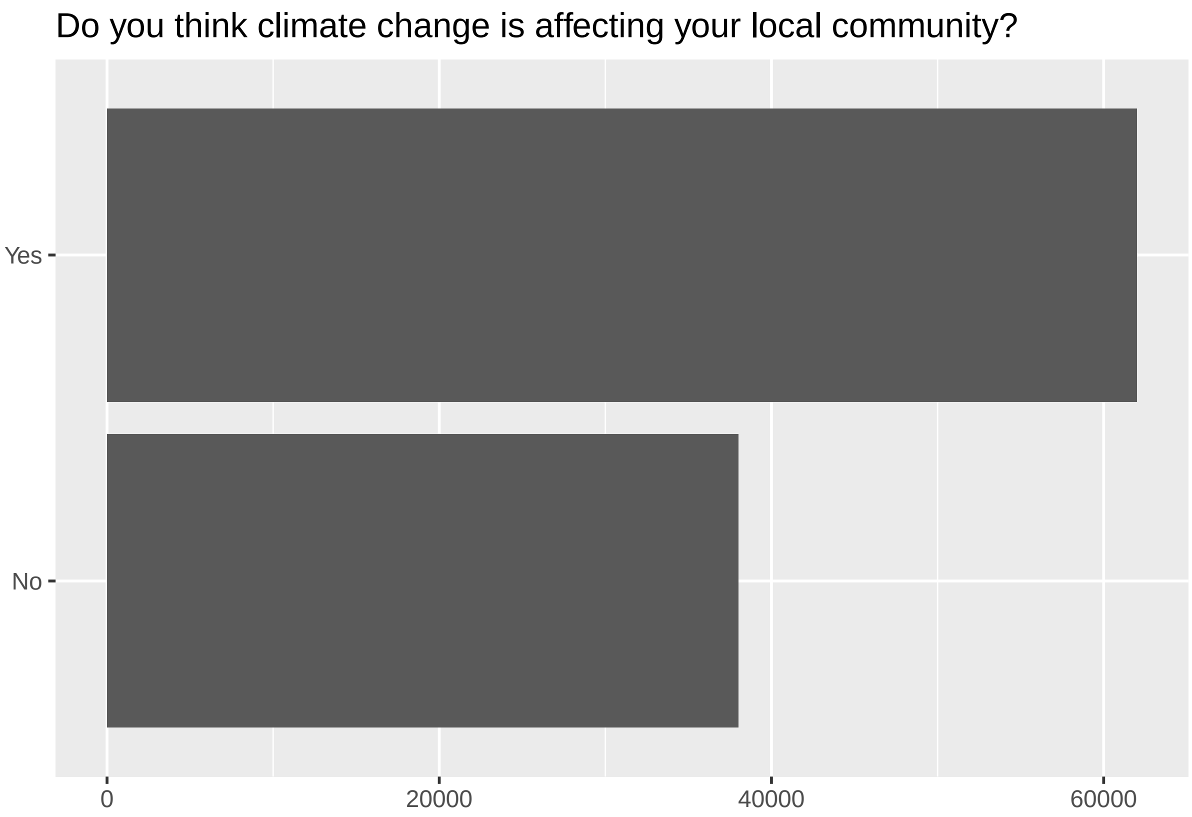
```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

```r
library(infer)
```

## Exercise 1

What percent of the adults in your sample think climate change affects their local community? Hint: Just like we did with the population, we can calculate the proportion of those in this sample who think climate change affects their local community.

```
set.seed(1234)

us_adults <- tibble(
  climate_change_affects = c(rep("Yes", 62000), rep("No", 38000))
)

ggplot(us_adults, aes(x = climate_change_affects)) +
  geom_bar() +
  labs(
    x = "", y = "",
    title = "Do you think climate change is affecting your local community?"
  ) +
  coord_flip()
```



```
# summary stats population
us_adults %>%
  count(climate_change_affects) %>%
  mutate(p = n /sum(n))
```

```
## # A tibble: 2 x 3
##   climate_change_affects     n     p
##   <chr>                  <int> <dbl>
## 1 No                     38000  0.38
## 2 Yes                    62000  0.62
```

```r
n <- 60
samp <- us_adults %>%
  sample_n(size = n)

# summary stats sample with size 60
samp %>%
  count(climate_change_affects) %>%
  mutate(p = n / sum(n))
```

```
## # A tibble: 2 x 3
##   climate_change_affects     n     p
##   <chr>                  <int> <dbl>
## 1 No                        23 0.383
## 2 Yes                       37 0.617
```

### Answer 1

In the sample with size 60, the percent of adults who think climate change affects their local community is 61.7%, which is a very close value to the population proportion of 62%.

### Exercise 2

Would you expect another student's sample proportion to be identical to yours? Would you expect it to be similar? Why or why not?

### Answer 2

I would expect another student's sample proportion to be similar to mine since we are using the same population (with proportion 62%) but not the same since the sampling method is random, so my sample of 60 adults is different than another student's sample.

```r
samp %>%
  specify(response = climate_change_affects, 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    0.483     0.75
```

### Exercise 3

In the interpretation above, we used the phrase "95% confident". What does "95% confidence" mean?

### Answer 3

A 95% confidence level means that the method used is able to generate intervals that capture the true proportion population value 95% of the time.

### Exercise 4

Does your confidence interval capture the true population proportion of US adults who think climate change affects their local community? If you are working on this lab in a classroom, does your neighbor's interval capture this value?

### Answer 4

My 95% confidence interval for the proportion of US adults who think climate change affects their local community was from 0.483 up to 0.75. Since we know that the true population proportion is 0.62, this value falls within my confidence interval, which means my interval successfully captured the true population parameter. I am not in a classroom but I assume that 95% of my colleagues would have a confidence interval that included the true population proportion, while 5% wouldn't due to the randomness of sample selection.

### Exercise 5

Each student should have gotten a slightly different confidence interval. What proportion of those intervals would you expect to capture the true population mean? Why?

### Answer 5

I would expect that 95% of students' intervals would have the true proportion value within their intervals since the 95% confidence interval means that the method used would have the true parameter in around 95% of those intervals.

### Exercise 6

Given a sample size of 60, 1000 bootstrap samples for each interval, and 50 confidence intervals constructed (the default values for the above app), what proportion of your confidence intervals include the true population proportion? Is this proportion exactly equal to the confidence level? If not, explain why. Make sure to include your plot in your answer.

```r
p_true <- 0.62

n <- 60
reps_boot <- 1000
n_intervals <- 50

# sample and bootstrap a 95% confidence interval
one_sample_ci <- function() {
  samp <- us_adults %>% sample_n(n)
  p_hat <- mean(samp$climate_change_affects == "Yes")
  ci <- samp %>%
    specify(response = climate_change_affects, success = "Yes") %>%
    generate(reps = reps_boot, type = "bootstrap") %>%
    calculate(stat = "prop") %>%
    get_ci(level = 0.95, type = "percentile")
  tibble(p_hat = p_hat, lower = ci$lower_ci, upper = ci$upper_ci)
}

# 50 confidence intervals
```

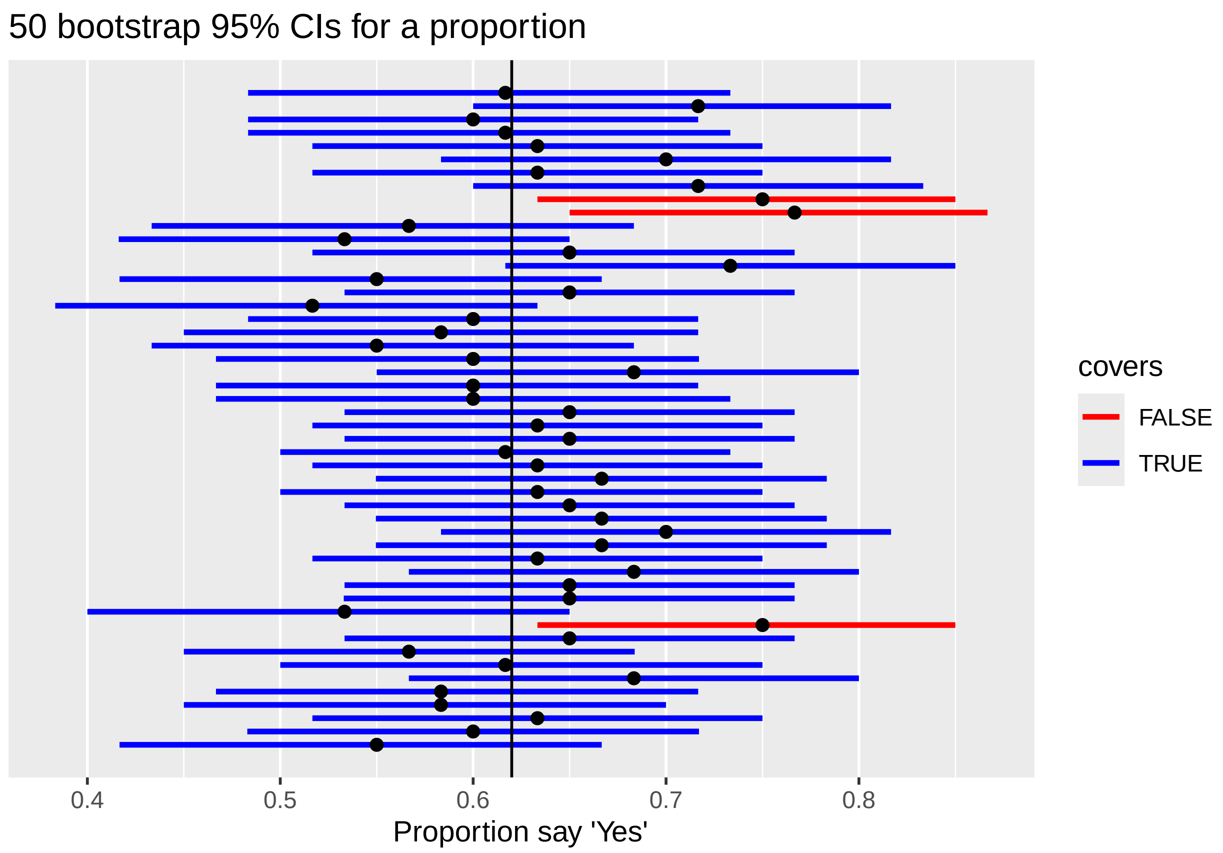
```r
confidence_intervals <- map_dfr(1:n_intervals, ~ one_sample_ci()) %>%
  mutate(id = row_number(),
         covers = (lower <= p_true & upper >= p_true))

coverage <- mean(confidence_intervals$covers)
coverage
```

```
## [1] 0.94
```

```r
# table counts
table(covers = confidence_intervals$covers)
```

```
## covers
## FALSE  TRUE
##     3    47
```

```r
# plot
ggplot(confidence_intervals, aes(y = id)) +
  geom_segment(aes(x = lower, xend = upper, yend = id, color = covers), linewidth = 1) +
  geom_point(aes(x = p_hat), size = 1.8) +
  geom_vline(xintercept = p_true) +
  scale_y_continuous(NULL, breaks = NULL) +
  scale_color_manual(values = c(`TRUE` = "blue", `FALSE` = "red")) +
  labs(
    x = "Proportion say 'Yes'",
    title = "50 bootstrap 95% CIs for a proportion"
  )
```



**Answer 6**

I got 47 of the 50 intervals that included the true population proportion value of 0.62 (62%). This gives us a 94% rate, which is very close to the expected 95% confidence level. This small difference is explained by the random sample variablity but, at the same time, by the small number of intervals (50). I assume that more intervals would make this number (94) even closer to 95.

**Exercise 7**

Choose a different confidence level than 95%. Would you expect a confidence interval at this level to me wider or narrower than the confidence interval you calculated at the 95% confidence level? Explain your reasoning.

```r
samp %>%
    specify(response = climate_change_affects, success = "Yes") %>%
    generate(reps = 1000, type = "bootstrap") %>%
    calculate(stat = "prop") %>%
    get_ci(level = 0.90, type = "percentile")
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.517    0.717
```

### Answer 7

I chose 90%. I was expecting a 90% confidence level interval to be narrower since I am asking for something that is less certain to be truth. In other words, the 90% confidence level leaves 10% of possible sample proportions outside of the equation, while a 95% confidence only leaves 5%.

### Exercise 8

Using code from the infer package and data fromt the one sample you have (samp), find a confidence interval for the proportion of US Adults who think climate change is affecting their local community with a confidence level of your choosing (other than 95%) and interpret it.

```
p_hat_yes <- mean(samp$climate_change_affects == "Yes")

confidence_intervals_90 <- samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 2000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.9, type = "percentile")

p_hat_yes
```

```
## [1] 0.6166667
```

```
confidence_intervals_90
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.517    0.717
```

```
table(samp$climate_change_affects)
```

```
##
##  No Yes
##  23  37
```

### Answer 8

The sample proportion who answered that climate change affects their community is 0.6167. A 90% bootstrap percentile confidence interval is between 0.517 and 0.717, which means that if we continue to take samples of the same size and with the same method, around 90% of those intervals would contain the true population proportion. The interval is narrower than for 95% confidence level since we have less confidence in that interval than in a 95% confidence level interval.

### Exercise 9

Using the app, calculate 50 confidence intervals at the confidence level you chose in the previous question, and plot all intervals on one plot, and calculate the proportion of intervals that include the true population proportion. How does this percentage compare to the confidence level selected for the intervals?

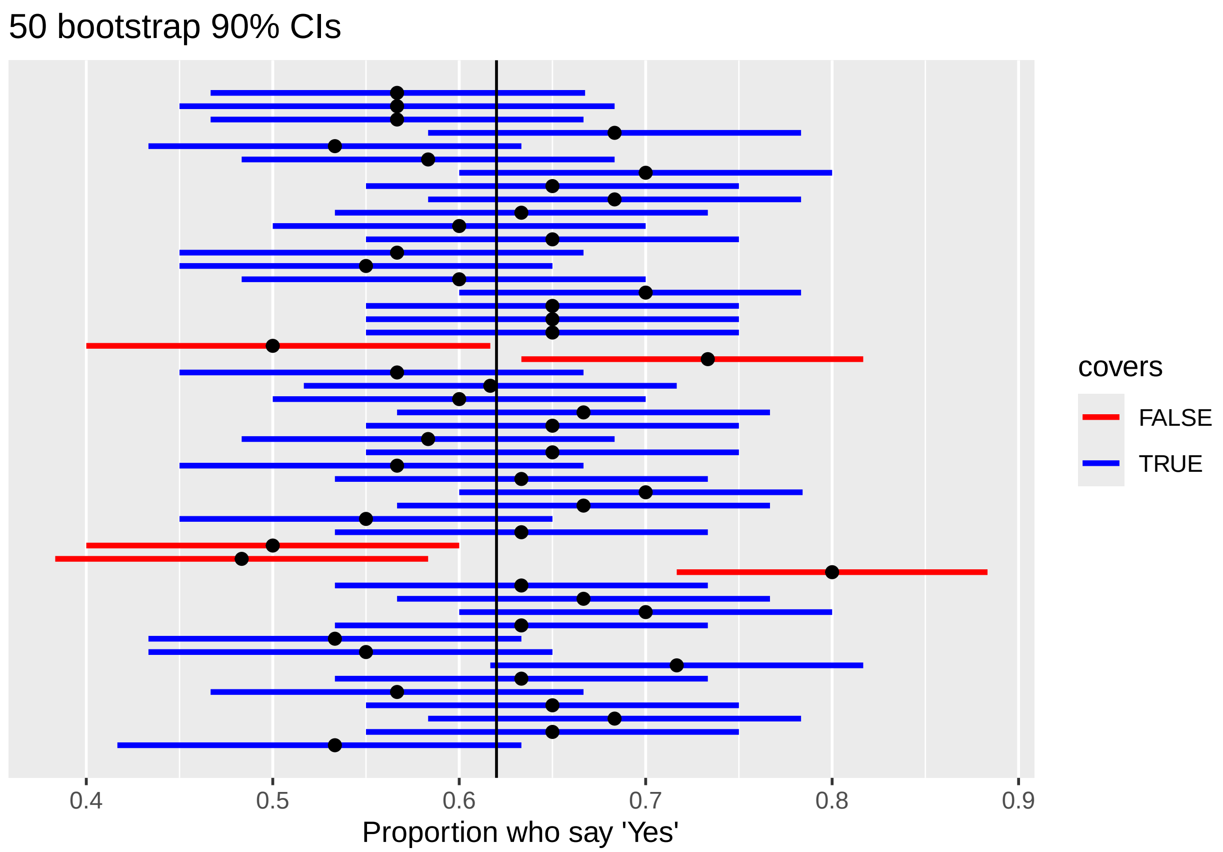
```r
n <- 60
reps_boot <- 2000
n_intervals <- 50
conf_level <- 0.90

one_ci_90 <- function() {
  samp <- us_adults %>% sample_n(n)
  p_hat <- mean(samp$climate_change_affects == "Yes")
  ci <- samp %>%
    specify(response = climate_change_affects, success = "Yes") %>%
    generate(reps = reps_boot, type = "bootstrap") %>%
    calculate(stat = "prop") %>%
    get_ci(level = conf_level, type = "percentile")
  tibble(p_hat = p_hat, lower = ci$lower_ci, upper = ci$upper_ci)
}

cis_90 <- map_dfr(1:n_intervals, ~ one_ci_90()) %>%
  mutate(id = row_number(),
         covers = (lower <= p_true & upper >= p_true))

coverage90 <- mean(cis_90$covers)
coverage90
```

```
## [1] 0.9
```

```r
table(covers = cis_90$covers)
```

```
## covers
## FALSE  TRUE
##     5    45
```

```r
# plot of all 50 intervals
ggplot(cis_90, aes(y = id)) +
  geom_segment(aes(x = lower, xend = upper, yend = id, color = covers), linewidth = 1) +
  geom_point(aes(x = p_hat), size = 1.8) +
  geom_vline(xintercept = p_true) +
  scale_y_continuous(NULL, breaks = NULL) +
  scale_color_manual(values = c('TRUE' = "blue", 'FALSE' = "red")) +
  labs(
    x = "Proportion who say 'Yes'",
    title = "50 bootstrap 90% CIs")
```



```
round(100*coverage90, 1)
```

```
## [1] 90
```

**Answer 9**

Using 50 bootstrap 90% CIs (n = 60 and 2000 bootstrap resamples), 90% of the intervals contained the true population proportion of 0.62, which exactly the expected percentage (I tried with 1000 bootstrap resamples and I got 94%).

**Exercise 10**

Lastly, try one more (different) confidence level. First, state how you expect the width of this interval to compare to previous ones you calculated. Then, calculate the bounds of the interval using the infer package and data from samp and interpret it. Finally, use the app to generate many intervals and calculate the proportion of intervals that are capture the true population proportion.

**Answer 10(part 1)**

I picked a 99% confidence level interval, which I expect to be wider than the 90 or 95% intervals. Since the interval has a 99% confidence/reliability, it is natural that the range of values increases.

```
p_hat_yes <- mean(samp$climate_change_affects == "Yes")

conf_level = 0.99

ci_99 <- samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 2000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = conf_level, type = "percentile")

p_hat_yes
```

```
## [1] 0.6166667
```

```
ci_99
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.467    0.767
```

The 99% bootstrap percentile CI for the population proportion is between 0.467 and 0.767, which is wider than the previous intervals for 90% and 95%. As explained above, if we continue to take samples using the same method, around 99% of those intervals would include the true population proportion.

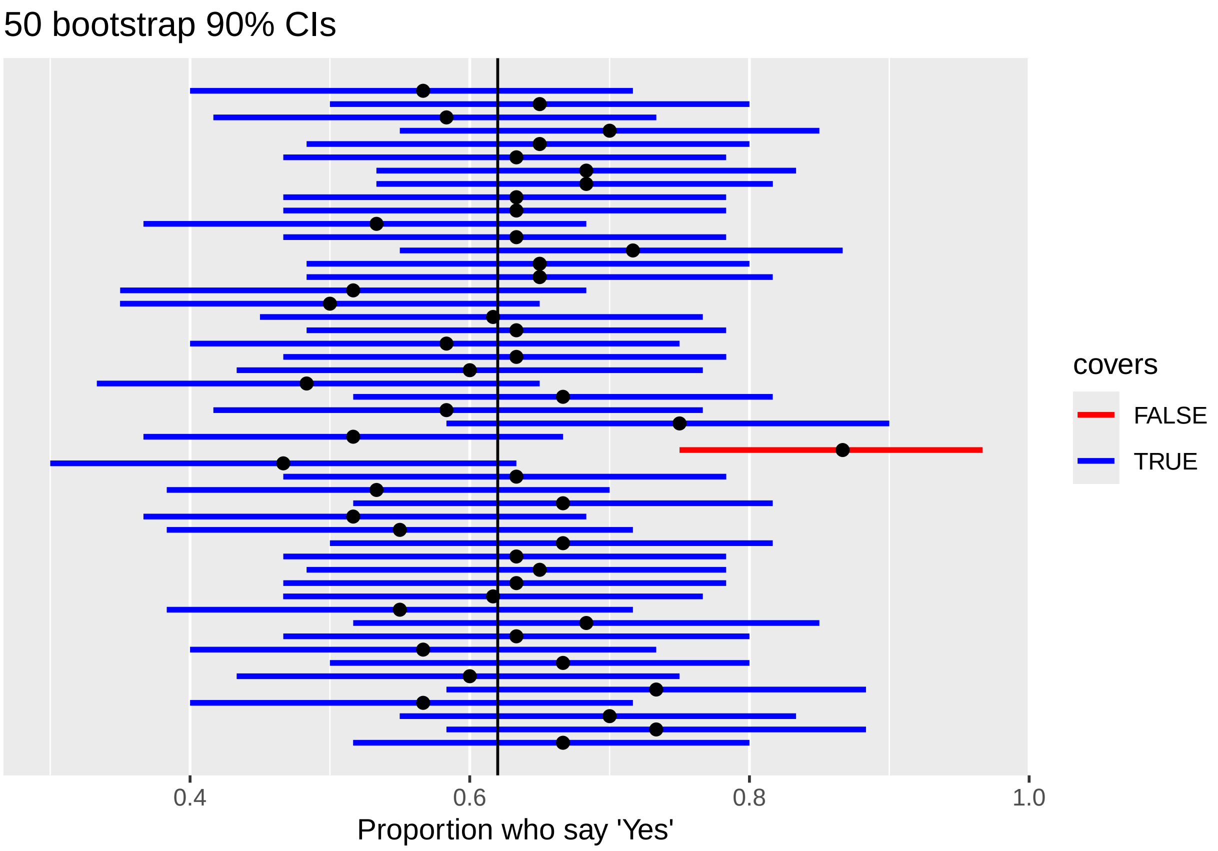
```
p_true <- 0.62
n <- 60
reps_boot <- 2000
n_intervals <- 50
conf_level <- 0.99

one_ci_99 <- function() {
  samp <- us_adults %>% sample_n(n)
  p_hat <- mean(samp$climate_change_affects == "Yes")
  ci <- samp %>%
    specify(response = climate_change_affects, success = "Yes") %>%
    generate(reps = reps_boot, type = "bootstrap") %>%
    calculate(stat = "prop") %>%
    get_ci(level = conf_level, type = "percentile")
  tibble(p_hat = p_hat, lower = ci$lower_ci, upper = ci$upper_ci)
}

cis_99 <- map_dfr(1:n_intervals, ~ one_ci_99()) %>%
  mutate(id = row_number(),
         covers = (lower <= p_true & upper >= p_true))

coverage99 <- mean(cis_99$covers)
coverage99
```

```
## [1] 0.98
```

```
table(covers = cis_99$covers)
```

```
## covers
## FALSE  TRUE
##     1    49
```

```
# plot of all 50 intervals
ggplot(cis_99, aes(y = id)) +
  geom_segment(aes(x = lower, xend = upper, yend = id, color = covers), linewidth = 1) +
  geom_point(aes(x = p_hat), size = 1.8) +
  geom_vline(xintercept = p_true) +
  scale_y_continuous(NULL, breaks = NULL) +
  scale_color_manual(values = c('TRUE' = "blue", 'FALSE' = "red")) +
  labs(
    x = "Proportion who say 'Yes'",
    title = "50 bootstrap 90% CIs")
```



```
round(100*coverage99, 1)
```

```
## [1] 98
```

**Answer 10 (part 2)**

Using 50 bootstrap 99% CIs (n = 60 and 2000 bootstrap resamples), 98% of the intervals contained the true population proportion of 0.62, which is very similar to the expected percentage of 99%.

### Exercise 11

Using the app, experiment with different sample sizes and comment on how the widths of intervals change as sample size changes (increases and decreases).

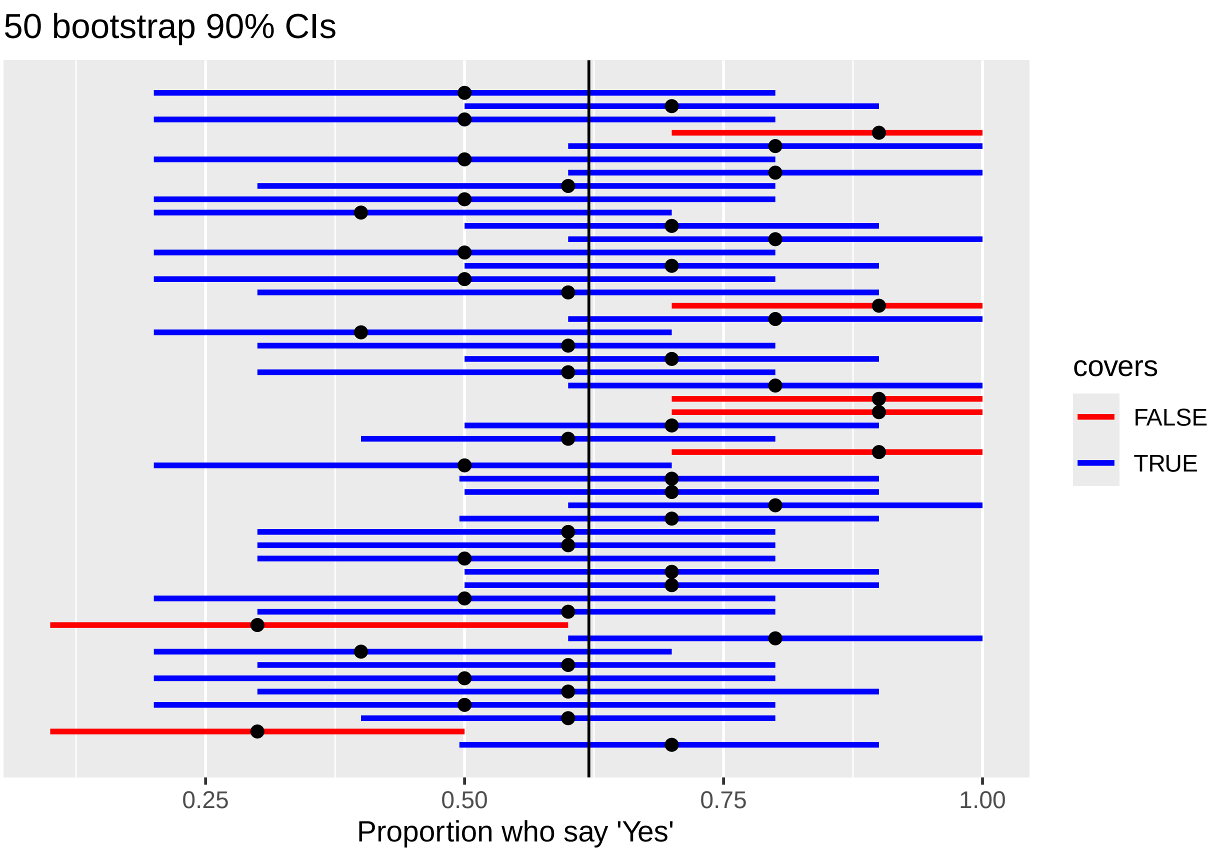
```
n <- 10
reps_boot <- 2000
n_intervals <- 50
conf_level <- 0.90

one_ci_90 <- function() {
  samp <- us_adults %>% sample_n(n)
  p_hat <- mean(samp$climate_change_affects == "Yes")
  ci <- samp %>%
    specify(response = climate_change_affects, success = "Yes") %>%
    generate(reps = reps_boot, type = "bootstrap") %>%
    calculate(stat = "prop") %>%
    get_ci(level = conf_level, type = "percentile")
  tibble(p_hat = p_hat, lower = ci$lower_ci, upper = ci$upper_ci)
}

cis_90 <- map_dfr(1:n_intervals, ~ one_ci_90()) %>%
  mutate(id = row_number(),
         covers = (lower <= p_true & upper >= p_true))

coverage90 <- mean(cis_90$covers)
coverage90
```

```
## [1] 0.86
```

```
table(covers = cis_90$covers)
```

```
## covers
## FALSE  TRUE
##     7    43
```

```
# plot of all 50 intervals
ggplot(cis_90, aes(y = id)) +
  geom_segment(aes(x = lower, xend = upper, yend = id, color = covers), linewidth = 1) +
  geom_point(aes(x = p_hat), size = 1.8) +
  geom_vline(xintercept = p_true) +
  scale_y_continuous(NULL, breaks = NULL) +
  scale_color_manual(values = c('TRUE' = "blue", 'FALSE' = "red")) +
  labs(
    x = "Proportion who say 'Yes'",
    title = "50 bootstrap 90% CIs")
```



```
round(100*coverage90, 1)
```

```
## [1] 86
```

**Answer 11**

I changed the sample size using both the 90, 95, and 99% confidence intervals. The behavior was similar: higher sample sizes gave me narrower intervals, while smaller sample sizes gave me wider intervals. This behavior is not surprising since the higher the sample size, less chance for sampling errors. It is important to remember that the standard error is inversely proportional to the sample size.

**Exercise 12**

Finally, given a sample size (say, 60), how does the width of the interval change as you increase the number of bootstrap samples. Hint: Does changing the number of bootstap samples affect the standard error?

**Answer 12**

The width of the interval doesn't change with the number of bootstrap samples . The standard error only affected by sample size and true proportion value, not the number of bootstrap samples. Increasing the bootstrap samples changes the percentiles of the bootstrap distribution.