Lab5a - Sampling Distributions

Keith Colella

2023-03-12

Setup

Config

```
library(tidyverse)
library(openintro)
library(infer)
set.seed(789)
```

Data

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

Exercise 1

Question

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.

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

Response

```
data.frame(table(samp)) %>%
  mutate(Proportion = Freq / sum(Freq))
```

```
##   climate_change_affects Freq Proportion
## 1                     No   26  0.4333333
## 2                    Yes   34  0.5666667
```

In this sample, ~57% of adults believe that climate change affects their local community.

Exercise 2

Question

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

Response

As noted in part A, I would not expect another student's sample proportion to exactly match (unless our random seeds matched, but we'll discount that possibility for now). Random sampling should not result in the same observations being sampled across separate experiments. So we should expect to sample different subsets of individuals, driving different sample proportions. We should expect some degree of consistency, but not too much. A sample size of 50 to represent a population of 100,000 is likely to introduce significant variance.

Exercise 3

Question

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

Response

It means that, if we were to take 100 distinct sample groups, we would expect 95 of those to correctly encompass the true population mean. Similarly, if we took 20 sample groups, we would expect 1 to fail to captured the population mean.

Exercise 4

Question

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?

Response

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
```

```
## # A tibble: 1 × 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.433    0.683
```

Our 95% CI does indeed capture the true population proportion of US adults who think climate change affects their local community. I'm not conducting this lab in a classroom, but I would expect, in a class of 20, that 1 student would have a CI that does NOT cover the 62% population proportion.

Exercise 5

Question

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?

Response

I would expect 95% of those intervals to captured the true population proportion. So, in a class room of 30, I would expect 1 or 2 students' CIs to fail to captured 62%.

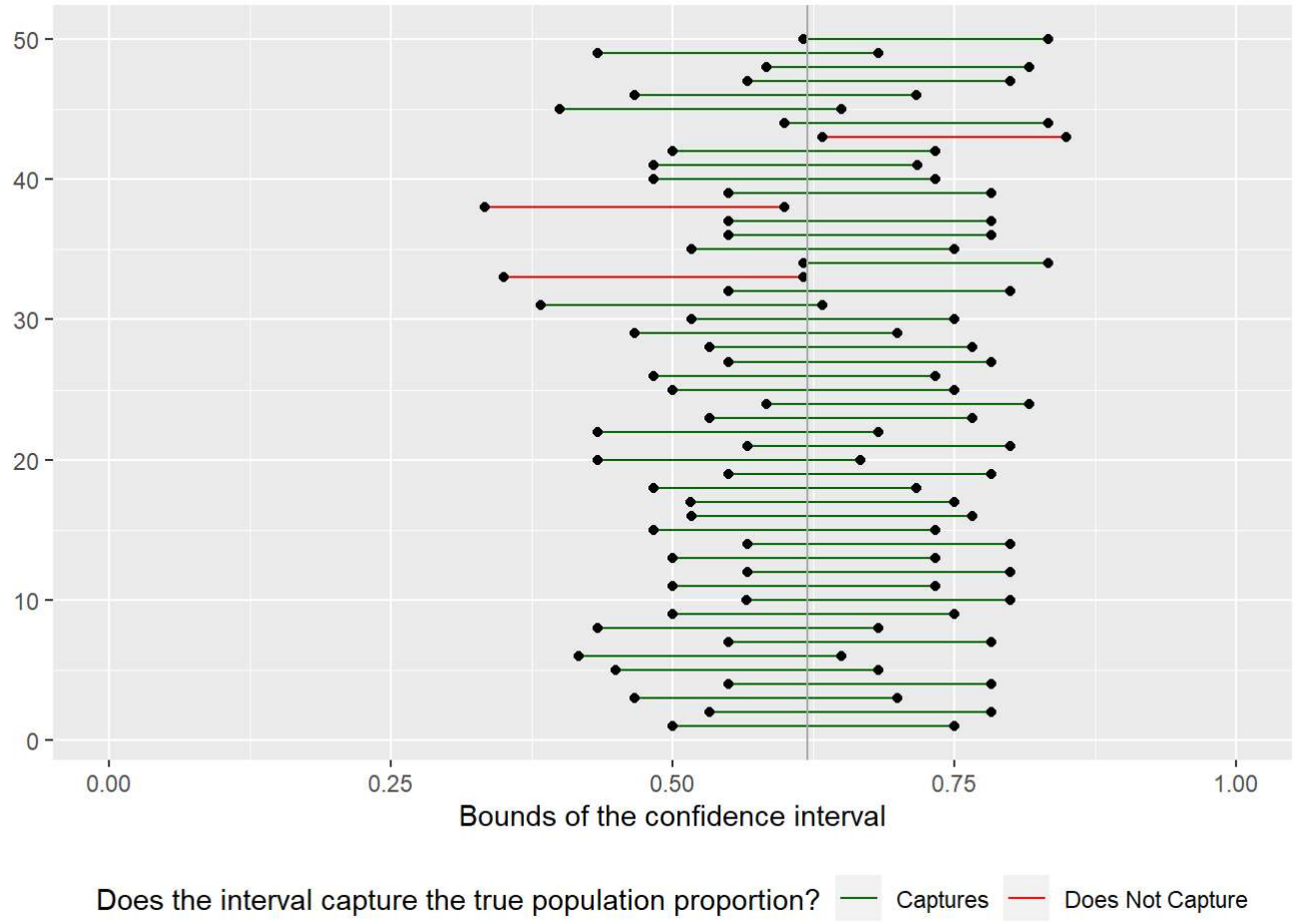
Exercise 6

Question

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.

Response

```
plot_ci <- function(sample_size, conf_level, iterations, bootstraps = 1000) {
  df_ci <- data.frame(matrix(nrow = 0, ncol = 4))
  colnames(df_ci) <- c('lower', 'upper', 'index', 'capture')

  for (i in 1:iterations) {
    samp <- us_adults %>%
      sample_n(size = sample_size)

    ci <- samp %>%
      specify(response = climate_change_affects, success = "Yes") %>%
      generate(reps = bootstraps, type = "bootstrap") %>%
      calculate(stat = "prop") %>%
      get_ci(level = conf_level) %>%
      mutate(index = i,
             capture = if_else(lower_ci < 0.62 & upper_ci > 0.62,
                               "Captures", "Does Not Capture"))

    df_ci[i,] <- ci
  }

  p <- ggplot(df_ci) +
    geom_segment(aes(x = lower, y = index, xend = upper, yend = index, color = capture)) +
    geom_point(aes(x = lower, y = index)) +
    geom_point(aes(x = upper, y = index)) +
    geom_vline(xintercept = 0.62, color = "darkgray") +
    labs(y = "", x = "Bounds of the confidence interval",
         color = "Does the interval capture the true population proportion?") +
    scale_color_manual(values = c('darkgreen','red')) +
    xlim(0,1) +
    theme(legend.position = "bottom")

  return(p)
}

plot_ci(sample_size = 60, conf_level = 0.95, iterations = 50)
```



The proportion of CIs that capture the population proportion is 47 out of 50 or 94%, roughly conforming to the confidence level, as expected. The match, however, is not exact. First, we don't have 100 confidence intervals. More importantly, there is a degree of randomness to these simulations, so we can't expect that results exactly conform to expected values.

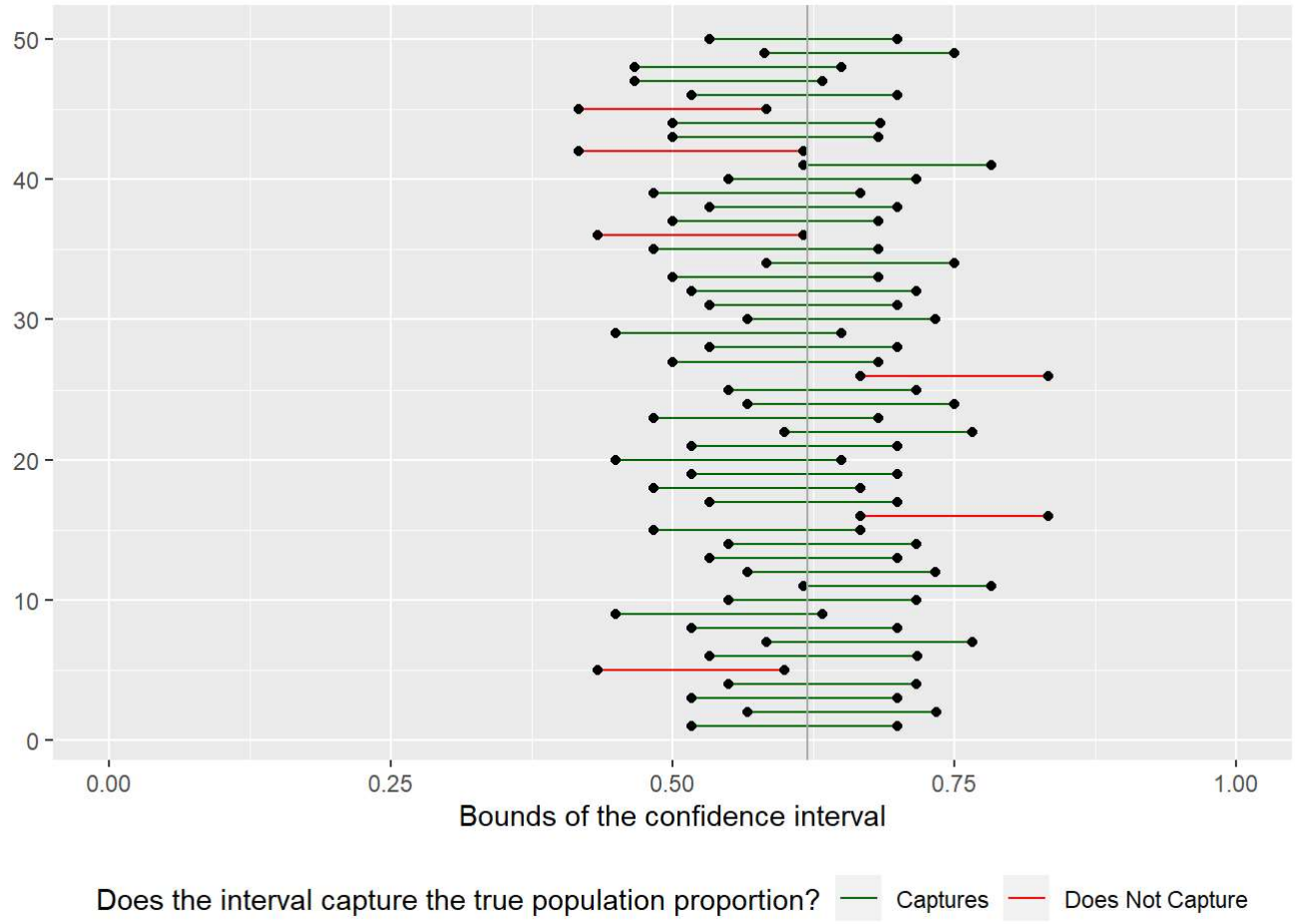
Exercise 7

Question

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.

Response

```
plot_ci(sample_size = 60, conf_level = 0.85, iterations = 50)
```



We run the same exercise to generate 50 confidence intervals with a confidence level of 85%. With a lower confidence level, we can produce tighter intervals. However, we must also expect more “misses” (or CIs that do not capture the true mean). This plot confirms that expectation, as we now have only 44 intervals (88%) that cover the 62% population proportion.

Exercise 8

Question

Using code from the infer package and data from 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.

Response

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

samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.99)
```

```
## # A tibble: 1 × 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1      0.5    0.817
```

Based on our sample of 60, we expect with 99% confidence that the population proportion of adults who believe climate change is affecting their local community is between 50% and 81.7%. By “99% confidence”, we mean that if we performed this survey 100 times, we would expect only one of the results confidence intervals to not correctly captured to the true population proportion.

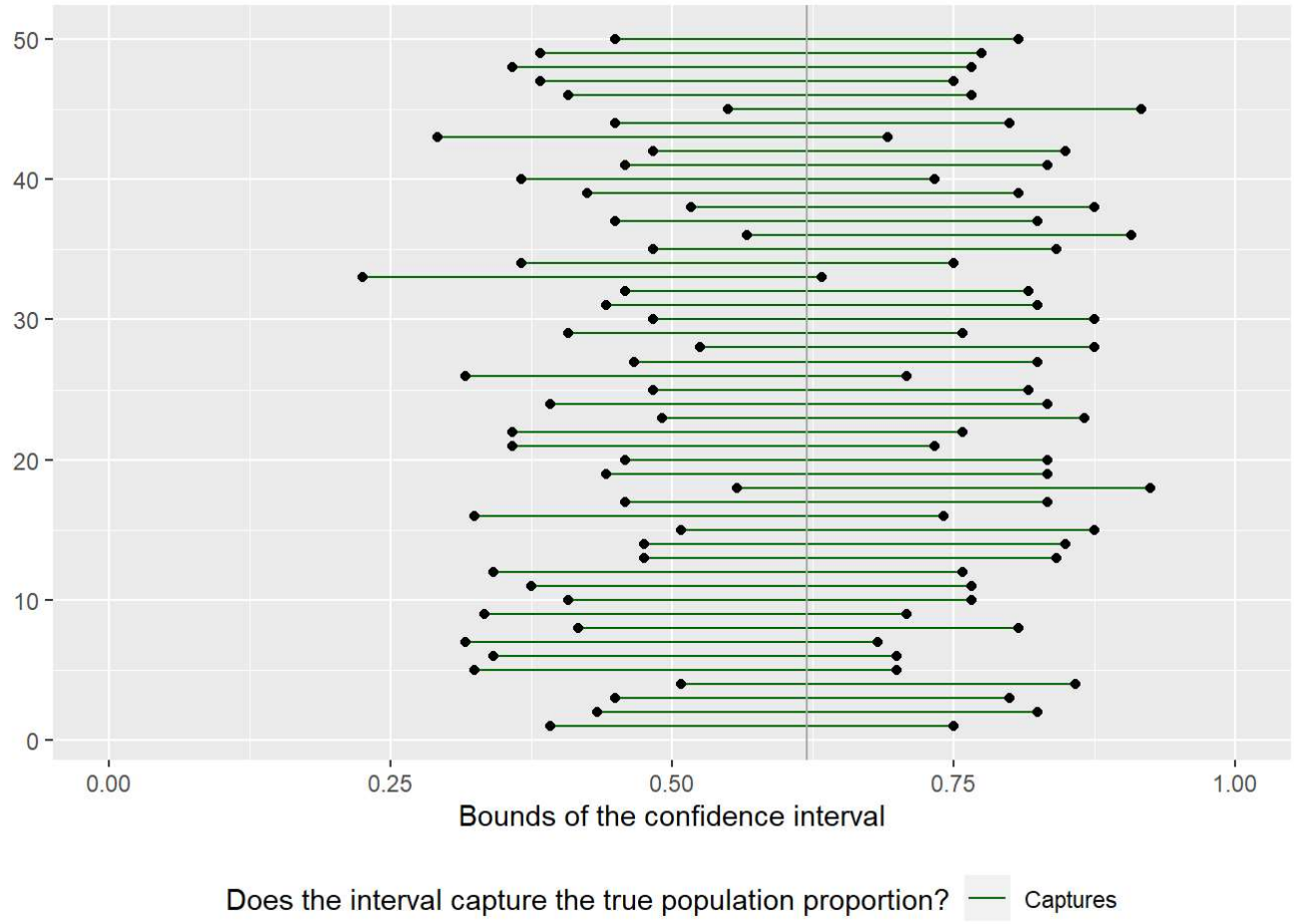
Exercise 9

Question

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?

Response

```
plot_ci(sample_size = 60, conf_level = 0.999, iterations = 50)
```



We produce another plot at the 99.9% confidence level. This time, all the intervals encompass the 62% population mean. However, our intervals are much larger than the intervals produced at lower confidence levels, diminishing the ability to infer much.

Exercise 10

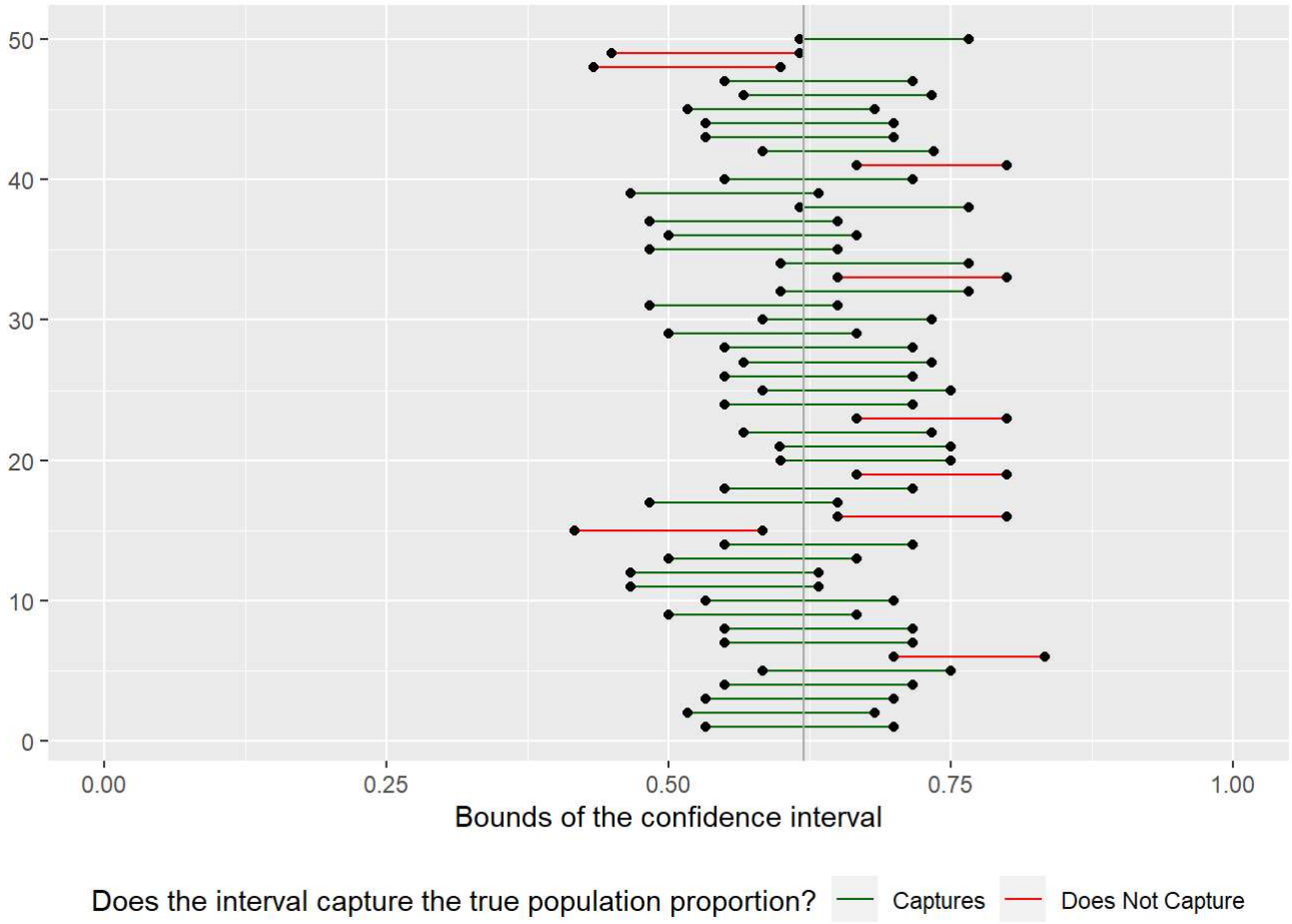
Question

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.

Response

I'll use an 80% confidence interval. With a lower confidence level, I would expect much tigher intervals, but a much higher miss rate.

```
plot_ci(sample_size = 60, conf_level = 0.80, iterations = 50)
```



The plot confirms my expectations. The intervals are much tighter than those observed above with a 99.9% confidence level, but only 41 of the intervals (82%) capture the true mean.

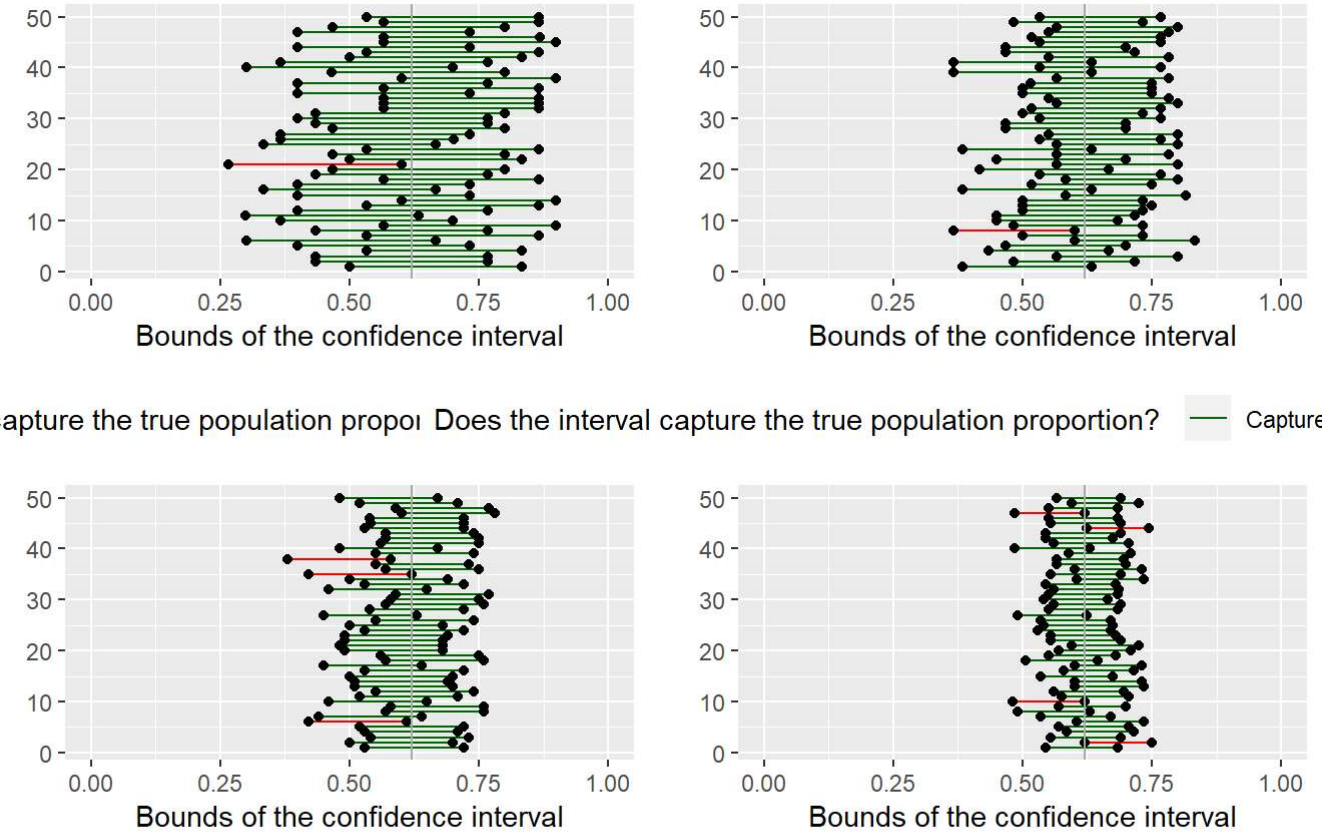
Exercise 11

Question

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

Response

```
p1 <- plot_ci(sample_size = 30, conf_level = 0.95, iterations = 50)
p2 <- plot_ci(sample_size = 60, conf_level = 0.95, iterations = 50)
p3 <- plot_ci(sample_size = 100, conf_level = 0.95, iterations = 50)
p4 <- plot_ci(sample_size = 200, conf_level = 0.95, iterations = 50)

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



Holding the confidence levels constant, we can see that increasing sample sizes drive tighter intervals. However, the “miss rate” appears relatively consistent, with 1 to 5 intervals missing the true population mean. So, we can infer that greater sample sizes drive tighter intervals, but only the confidence level determines the expected “miss rate”.

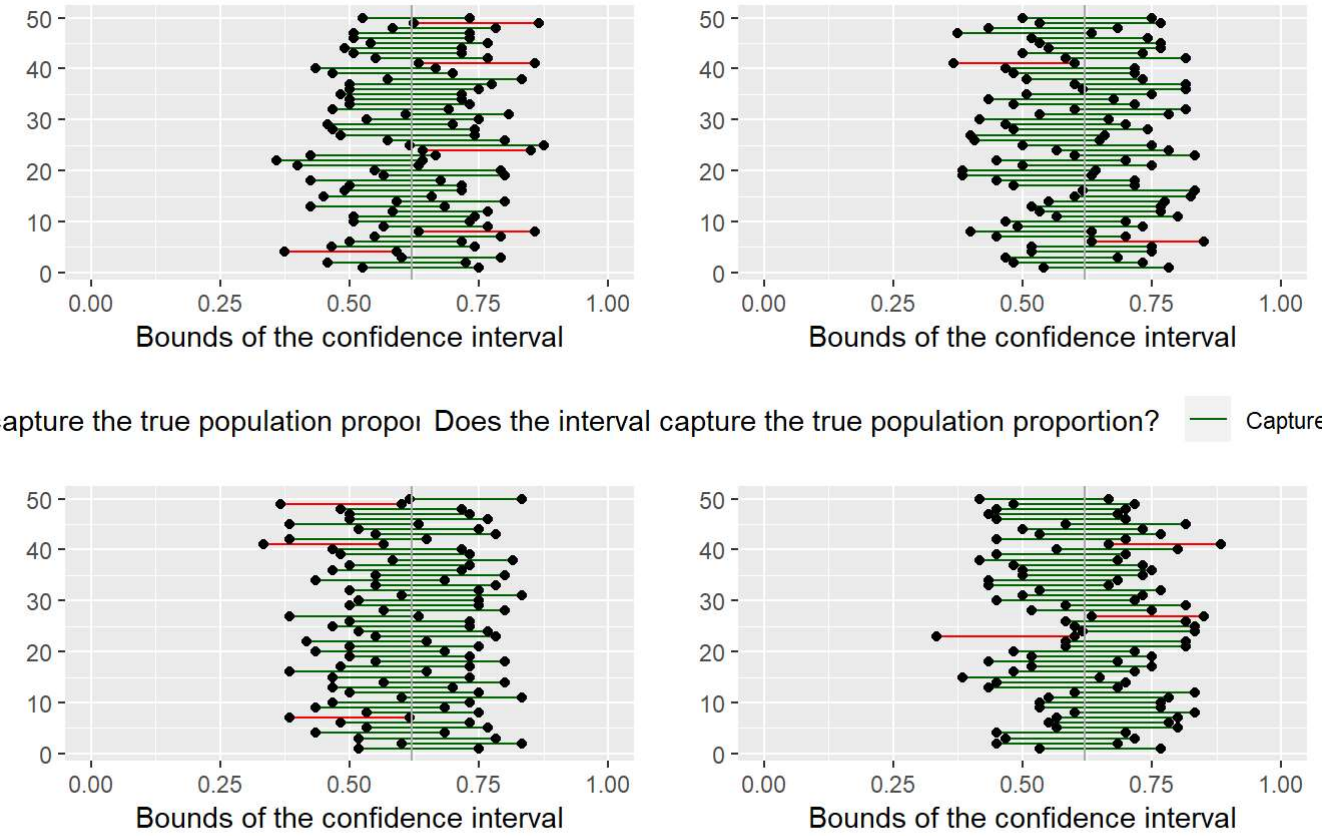
Exercise 12

Question

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 bootstrap samples affect the standard error?

Response

```
p1 <- plot_ci(sample_size = 60, conf_level = 0.95, iterations = 50, bootstraps = 100)
p2 <- plot_ci(sample_size = 60, conf_level = 0.95, iterations = 50, bootstraps = 500)
p3 <- plot_ci(sample_size = 60, conf_level = 0.95, iterations = 50, bootstraps = 1000)
p4 <- plot_ci(sample_size = 60, conf_level = 0.95, iterations = 50, bootstraps = 5000)

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



The bootstrap size does not seem to have a discernible impact on the width of confidence intervals or the “miss rate”. If anything, it appears to drive more consistent intervals. Whereas the intervals produces with bootstraps >=1000 move around quite a bit, the intervals produced with 5000 bootstraps appear more consistently centered on the population mean.