DS606-Lab5B

Code

George Cruz

2020-10-04

Hide

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

Exercise 1

Hide

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

Hide

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

Hide

```
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                        22 0.367
## 2 Yes                       38 0.633
```

58.3% of people in my sample think climate change affects their local community.

Exercise 2

I would expect another student's proportion to be similar but not identical. The reason for it is we cannot be certain the sample will take the same proportion of answers every time.

Exercise 3

Hide

```
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    0.517     0.75
```

95% Confidence means we are 95% confident that the Yes answer is between the lower and upper confidence interval.

Exercise 4

Yes, it does capture the real proportion.

Exercise 5

I would expect 95% of the intervals to capture the true population.

Exercise 6

I could not find the app referred to. Trying to answer the question by programmatically running the described steps:

1. Obtain a random sample.
2. Calculate the sample proportion, and use these to calculate and store the lower and upper bounds of the confidence intervals.
3. Repeat these steps 50 times.

Hide

```
df <- data.frame(NA, NA)

for (i in 1:50) { #repeat 50 times
  temp_vector <- samp %>%
                      specify(response = climate_change_affects, success = "Yes") %>%
                      generate(reps = 1000, type = "bootstrap") %>%
                      calculate(stat = "prop") %>%
                      get_ci(level = 0.95)
  if (i == 1) {
    df <- temp_vector
  } else {
    df <- rbind(df, temp_vector)
  }

}
```

Hide

```
df %>%
  mutate(within_range = case_when(lower_ci <= 0.62 & upper_ci >= 0.62 ~ "Yes",
                                  lower_ci > 0.62 | upper_ci < .62 ~ "No"))%>%
  count(within_range) %>%
  mutate(p = n /sum(n))
```

```
## # A tibble: 1 x 3
##   within_range     n     p
##   <chr>        <int> <dbl>
## 1 Yes             50     1
```

The proportion is higher than the confidence level. Not sure of the reason.

Exercise 7

Hide

```
samp %>%
  specify(response = climate_change_affects, success = "Yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.5)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.583    0.667
```

I would expect the interval to be narrower because the confidence level is lower.

Exercise 8

I chose 50% confidence level and the boundaries were very narrow. I expect this to be due to the fact that we raised our uncertainty.

Exercise 9

Hide

```
df <- data.frame(NA, NA)

for (i in 1:50) { #repeat 50 times
  temp_vector <- samp %>%
                      specify(response = climate_change_affects, success = "Yes") %>%
                      generate(reps = 1000, type = "bootstrap") %>%
                      calculate(stat = "prop") %>%
                      get_ci(level = 0.50)
  if (i == 1) {
    df <- temp_vector
  } else {
    df <- rbind(df, temp_vector)
  }

}


df %>%
  mutate(within_range = case_when(lower_ci <= 0.62 & upper_ci >= 0.62 ~ "Yes",
                                  lower_ci > 0.62 | upper_ci < .62 ~ "No"))%>%
  count(within_range) %>%
  mutate(p = n /sum(n))
```

```
## # A tibble: 1 x 3
##   within_range     n     p
##   <chr>        <int> <dbl>
## 1 Yes             50     1
```

By lowering our confidence level, we got a proportion of samples withing range of 68%, opposite to the 100% we got before.

Exercise 10

Hide

```
df <- data.frame(NA, NA)

for (i in 1:50) { #repeat 50 times
  temp_vector <- samp %>%
                      specify(response = climate_change_affects, success = "Yes") %>%
                      generate(reps = 1000, type = "bootstrap") %>%
                      calculate(stat = "prop") %>%
                      get_ci(level = 0.20)
  if (i == 1) {
    df <- temp_vector
  } else {
    df <- rbind(df, temp_vector)
  }

}


df %>%
  mutate(within_range = case_when(lower_ci <= 0.62 & upper_ci >= 0.62 ~ "Yes",
                                  lower_ci > 0.62 | upper_ci < .62 ~ "No"))%>%
  count(within_range) %>%
  mutate(p = n /sum(n))
```

```
## # A tibble: 2 x 3
##   within_range     n     p
##   <chr>        <int> <dbl>
## 1 No               1  0.02
## 2 Yes             49  0.98
```

When I try 20% confidence level, I don't get a single observation to agree with the real proportion.

…