Lab Exercises Week 06

Alison Lawyer

2024-10-16

```
# keep this chunk in all your RMarkdown scripts
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

```
# List required packages
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   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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
```

```
library(PresenceAbsence)
library(ggplot2)

# Load SPDATA from PresenceAbsence package
data("SPDATA")
```

LAB EXERCISES

In this week's lab you will be practicing random sampling to better understand the central limit theorem and its implications for estimating population parameters. You will also use and apply the concepts of standard error and confidence interval.

The dataset SPDATA, introduced in Session 05, contains presence/absence records of trees at 386 forest locations. You can load it, after installing the package PresenceAbsence, using the command data("SPDATA").

Exercise 1

In the first exercise, we want to take 5 random samples of 20 sites, and for each of the 5 random samples, measure the prevalence of Bigtooth maples (ACGR3) across sites. Then we want to calculate the proportion of sites in each sample where this species is present, so we will have 5 proportions, one for each of the samples, where each sample consists of 20 random sites.

Instructions: - Use set.seed(1234) to be sure everyone gets the same result. - Create two variables called sample_size and num_samples - Assign 20 to sample_size and 5 to num_samples - Filter your dataset and select columms SPECIES and OBSERVED, save this as a new dataframe - First start by conducting random sampling using the function sample() - Notice the result is a dataframe with just your random sample and the two columns you selected - Now, we need to repeat this operation 5 times. You can achieve this with the function replicate() or a for loop - By default, replicate() tries to simplify the result into a vector, which doesn't work very well in this case, so set the simplify argument to FALSE, which will result in a list, where each element of the list is one draw of 20 random rows - Combine the list into a dataframe. This can be achieved through bind_rows() function, feeding it the list object - To keep track of each of the 5 random draws, add an identifier to the bind_rows() function, like .id = "sample_id" - Check your output, which should be a dataframe with sample_id, SPECIES, and OBSERVED columns - Finally, calculate the proportion of sites with the species present for each of the five random samples. - Check that your final dataframe has 5 rows, consisting of sample_id and prevalence columns

```
set.seed(1234)
sample_size1 <- 20
num_samples1 <- 5

filtered_dataset <- SPDATA %>%
    select(c(SPECIES, OBSERVED)) %>%
    filter(SPECIES == "ACGR3")
sample_n(filtered_dataset, size = sample_size1, replace = FALSE)
```

```
##    SPECIES OBSERVED
## 1    ACGR3        0
## 2    ACGR3        0
## 3    ACGR3        0
## 4    ACGR3        0
## 5    ACGR3        0
## 6    ACGR3        0
## 7    ACGR3        0
## 8    ACGR3        1
## 9    ACGR3        0
## 10   ACGR3        0
## 11   ACGR3        0
## 12   ACGR3        0
## 13   ACGR3        0
## 14   ACGR3        0
## 15   ACGR3        0
## 16   ACGR3        0
## 17   ACGR3        0
## 18   ACGR3        0
## 19   ACGR3        0
## 20   ACGR3        0
```

```
set.seed(1234)
compiled_random_samples1 <- replicate(n = num_samples1, expr = sample_n(filtered_dataset,
    size = sample_size1, replace = FALSE), simplify = FALSE) %>%
    bind_rows(.id = "sample_id")

has_bigtooth_maples1 <- compiled_random_samples1 %>%
    group_by(sample_id) %>%
    summarize(prevalence = sum(ifelse(OBSERVED == 1, 1, 0))/n())
```

```
# add the total number of observations and divide by total
# number of rows
has_bigtooth_maples1
```

```
## # A tibble: 5 x 2
##   sample_id prevalence
##   <chr>          <dbl>
## 1 1               0.05
## 2 2               0.05
## 3 3               0.15
## 4 4               0.05
## 5 5               0.15
```

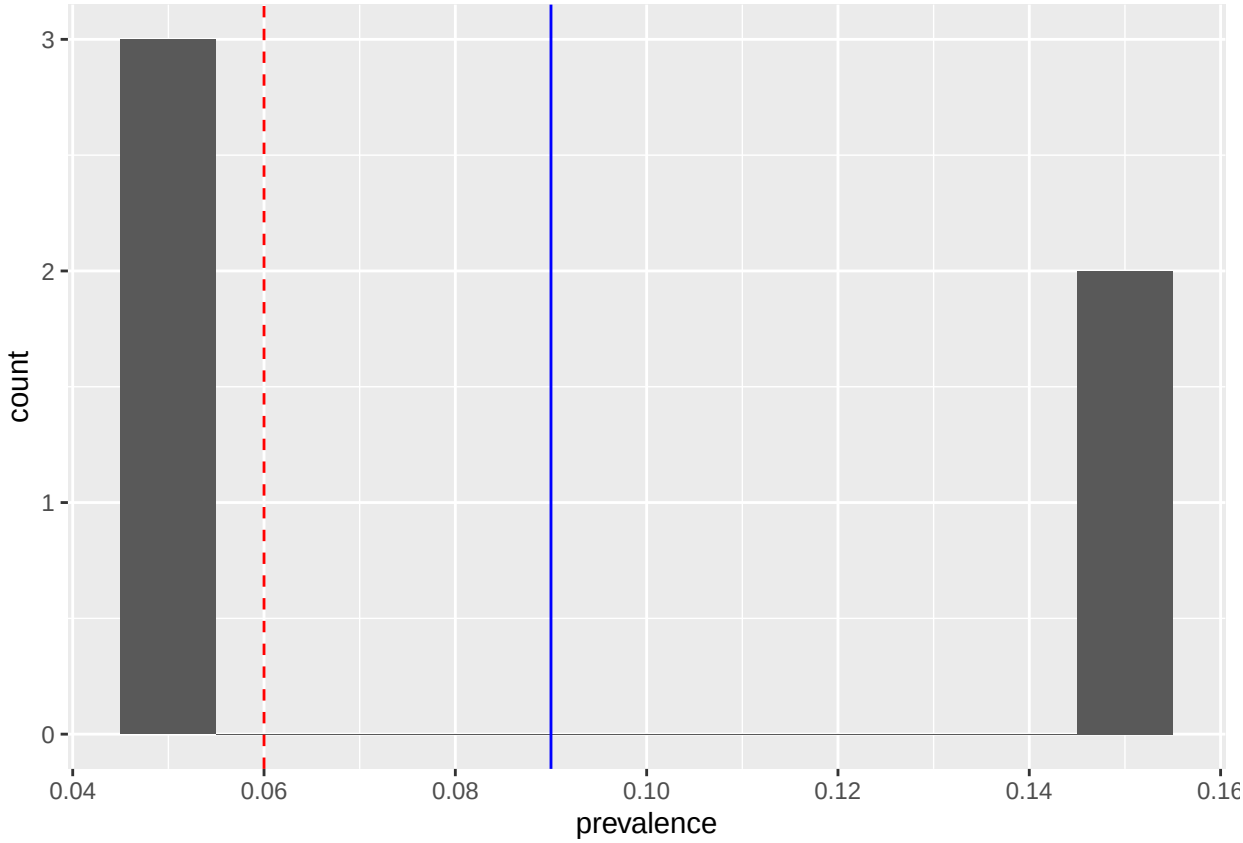
```
# mutate(prevalence = ifelse(OBSERVED > 0, sum(OBSERVED) /
# sample_size, 0))
```

Exercise 2

Plot the distribution of estimated prevalence from above using a histogram. Adjust the binwidth parameter to be 0.01 instead of the default, which is 30 bins over the range of values. Add a vertical line for the overall prevalence of the species (0.06), and a second line for the actual mean of prevalence observed across your replicated random samples. Make sure both lines are uniquely identified. You can but do not need to label the lines.

Describe what you see.

```
ggplot(has_bigtooth_maples1, aes(x = prevalence)) + geom_histogram(binwidth = 0.01) +
    geom_vline(aes(xintercept = 0.06), color = "red", linetype = "dashed") +
    geom_vline(aes(xintercept = mean(prevalence)), color = "blue",
        linetype = "solid")
```



Answer:

Exercise 3

Repeat exercises 1 & 2, but this time increase the number of replicates to 50 instead of 5. Be sure to again include the same set.seed(1234) from above. Save your dataframe holding the prevalence estimates in a new object. You will be needing it later.

Describe the distribution. What implications does this have for your ability to study the distribution of Bigtooth maples?

```
set.seed(1234)
sample_size3 <- 20
num_samples3 <- 50

filtered_dataset <- SPDATA %>%
    select(c(SPECIES, OBSERVED)) %>%
    filter(SPECIES == "ACGR3")
sample_n(filtered_dataset, size = sample_size3, replace = FALSE)
```

```
##    SPECIES OBSERVED
## 1    ACGR3        0
## 2    ACGR3        0
## 3    ACGR3        0
## 4    ACGR3        0
```

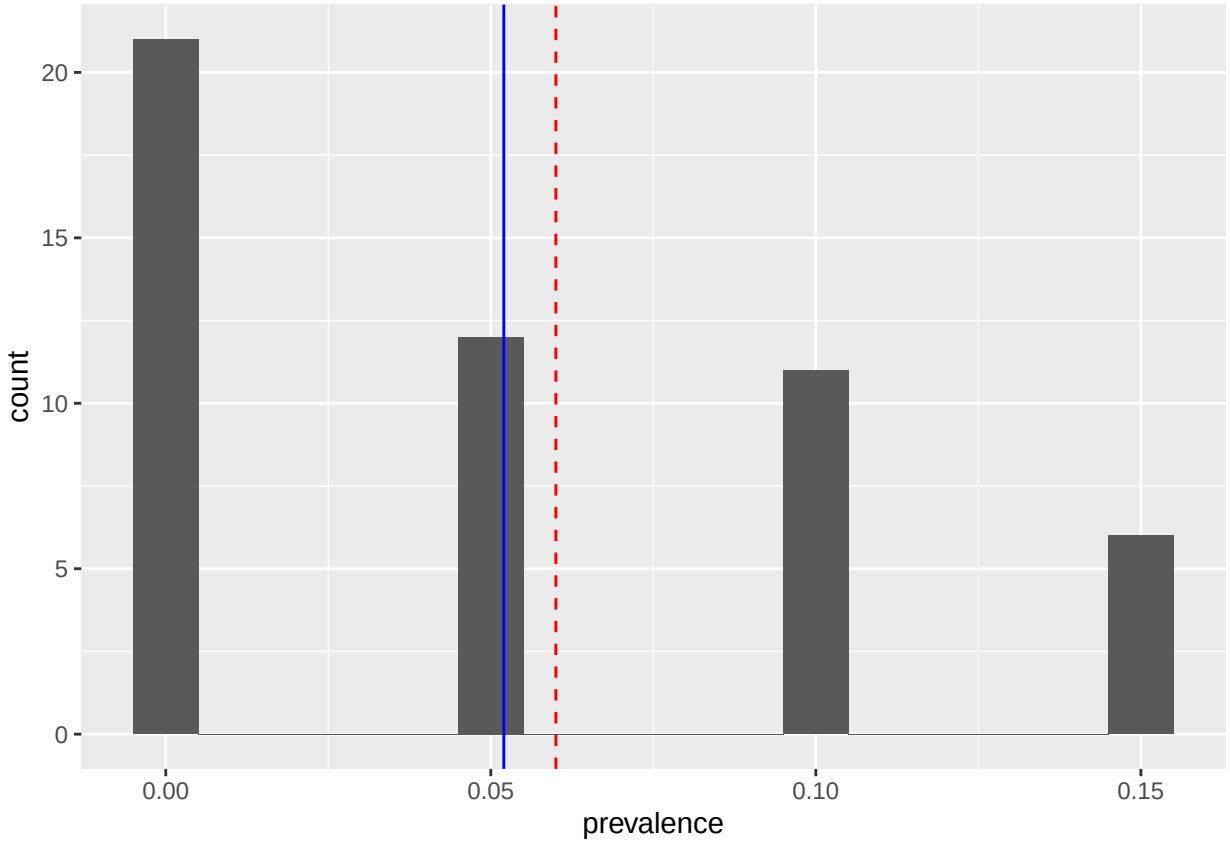
```
## 5    ACGR3        0
## 6    ACGR3        0
## 7    ACGR3        0
## 8    ACGR3        1
## 9    ACGR3        0
## 10   ACGR3        0
## 11   ACGR3        0
## 12   ACGR3        0
## 13   ACGR3        0
## 14   ACGR3        0
## 15   ACGR3        0
## 16   ACGR3        0
## 17   ACGR3        0
## 18   ACGR3        0
## 19   ACGR3        0
## 20   ACGR3        0
```

```
set.seed(1234)
compiled_random_samples3 <- replicate(n = num_samples3, expr = sample_n(filtered_dataset,
    size = sample_size3, replace = FALSE), simplify = FALSE) %>%
    bind_rows(.id = "sample_id")

has_bigtooth_maples3 <- compiled_random_samples3 %>%
    group_by(sample_id) %>%
    summarize(prevalence = sum(ifelse(OBSERVED == 1, 1, 0))/n())
# add the total number of observations and divide by total
# number of rows
has_bigtooth_maples3
```

```
## # A tibble: 50 x 2
##    sample_id prevalence
##    <chr>          <dbl>
##  1 1               0.05
##  2 10              0
##  3 11              0.05
##  4 12              0.05
##  5 13              0.1
##  6 14              0.1
##  7 15              0.1
##  8 16              0.1
##  9 17              0.05
## 10 18              0.15
## # i 40 more rows
```

```
ggplot(has_bigtooth_maples3, aes(x = prevalence)) + geom_histogram(binwidth = 0.01) +
    geom_vline(aes(xintercept = 0.06), color = "red", linetype = "dashed") +
    geom_vline(aes(xintercept = mean(prevalence)), color = "blue",
        linetype = "solid")
```



Answer:

Exercise 4

Let's repeat this one more time. Now keep the number of replicates at 50 but increase the sample size each time to 100 sites. Be sure to again include the same set.seed(1234) from above.

Describe the distribution. What implications does this have for your ability to study the distribution of Bigtooth maples?

```
set.seed(1234)
sample_size4 <- 100
num_samples4 <- 50

filtered_dataset <- SPDATA %>%
    select(c(SPECIES, OBSERVED)) %>%
    filter(SPECIES == "ACGR3")
sample_n(filtered_dataset, size = sample_size4, replace = FALSE)
```

```
##     SPECIES OBSERVED
## 1     ACGR3        0
## 2     ACGR3        0
## 3     ACGR3        0
## 4     ACGR3        0
## 5     ACGR3        0
```

```
## 6     ACGR3        0
## 7     ACGR3        0
## 8     ACGR3        1
## 9     ACGR3        0
## 10    ACGR3        0
## 11    ACGR3        0
## 12    ACGR3        0
## 13    ACGR3        0
## 14    ACGR3        0
## 15    ACGR3        0
## 16    ACGR3        0
## 17    ACGR3        0
## 18    ACGR3        0
## 19    ACGR3        0
## 20    ACGR3        0
## 21    ACGR3        0
## 22    ACGR3        0
## 23    ACGR3        0
## 24    ACGR3        0
## 25    ACGR3        0
## 26    ACGR3        0
## 27    ACGR3        0
## 28    ACGR3        0
## 29    ACGR3        0
## 30    ACGR3        0
## 31    ACGR3        0
## 32    ACGR3        0
## 33    ACGR3        0
## 34    ACGR3        0
## 35    ACGR3        1
## 36    ACGR3        0
## 37    ACGR3        0
## 38    ACGR3        0
## 39    ACGR3        0
## 40    ACGR3        0
## 41    ACGR3        0
## 42    ACGR3        0
## 43    ACGR3        0
## 44    ACGR3        0
## 45    ACGR3        0
## 46    ACGR3        0
## 47    ACGR3        0
## 48    ACGR3        0
## 49    ACGR3        1
## 50    ACGR3        0
## 51    ACGR3        1
## 52    ACGR3        1
## 53    ACGR3        0
## 54    ACGR3        0
## 55    ACGR3        0
## 56    ACGR3        0
## 57    ACGR3        0
## 58    ACGR3        0
## 59    ACGR3        0
```

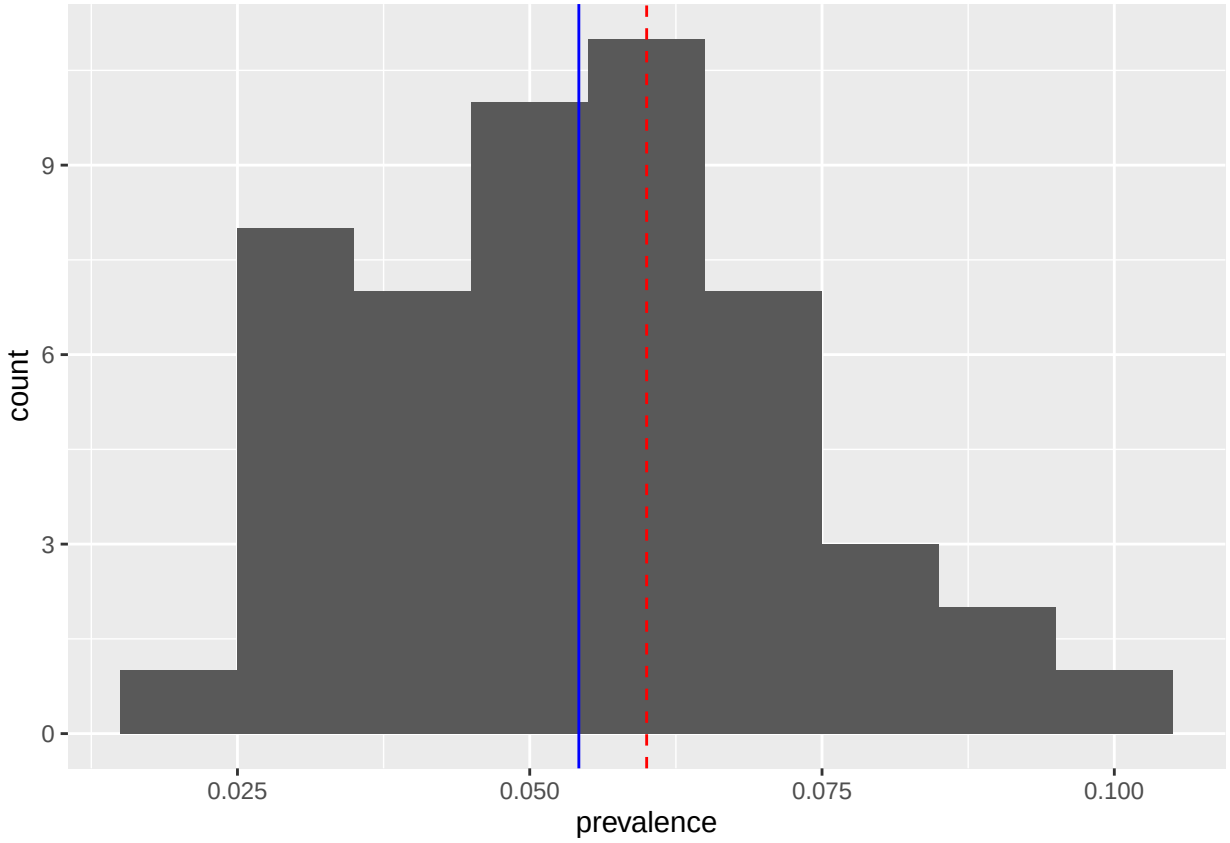
```
## 60    ACGR3        0
## 61    ACGR3        0
## 62    ACGR3        0
## 63    ACGR3        0
## 64    ACGR3        0
## 65    ACGR3        1
## 66    ACGR3        0
## 67    ACGR3        0
## 68    ACGR3        0
## 69    ACGR3        0
## 70    ACGR3        0
## 71    ACGR3        0
## 72    ACGR3        0
## 73    ACGR3        0
## 74    ACGR3        0
## 75    ACGR3        0
## 76    ACGR3        0
## 77    ACGR3        0
## 78    ACGR3        0
## 79    ACGR3        0
## 80    ACGR3        0
## 81    ACGR3        0
## 82    ACGR3        0
## 83    ACGR3        0
## 84    ACGR3        0
## 85    ACGR3        0
## 86    ACGR3        0
## 87    ACGR3        0
## 88    ACGR3        0
## 89    ACGR3        0
## 90    ACGR3        0
## 91    ACGR3        0
## 92    ACGR3        0
## 93    ACGR3        0
## 94    ACGR3        0
## 95    ACGR3        0
## 96    ACGR3        0
## 97    ACGR3        0
## 98    ACGR3        0
## 99    ACGR3        0
## 100   ACGR3        1
```

```r
set.seed(1234)
compiled_random_samples4 <- replicate(n = num_samples4, expr = sample_n(filtered_dataset,
    size = sample_size4, replace = FALSE), simplify = FALSE) %>%
    bind_rows(.id = "sample_id")

has_bigtooth_maples4 <- compiled_random_samples4 %>%
    group_by(sample_id) %>%
    summarize(prevalence = sum(ifelse(OBSERVED == 1, 1, 0))/n())
# add the total number of observations and divide by total
# number of rows
has_bigtooth_maples4
```

```
## # A tibble: 50 x 2
##    sample_id prevalence
##    <chr>          <dbl>
##  1 1               0.07
##  2 10              0.06
##  3 11              0.03
##  4 12              0.08
##  5 13              0.05
##  6 14              0.07
##  7 15              0.04
##  8 16              0.05
##  9 17              0.05
## 10 18              0.09
## # i 40 more rows
```

```
ggplot(has_bigtooth_maples4, aes(x = prevalence)) + geom_histogram(binwidth = 0.01) +
    geom_vline(aes(xintercept = 0.06), color = "red", linetype = "dashed") +
    geom_vline(aes(xintercept = mean(prevalence)), color = "blue",
        linetype = "solid")
```



Answer:

Exercise 5

Calculate the standard error and 95% confidence interval of sample prevalence from Exercises 1, 3 and 4. Which of the confidence intervals includes the "true" prevalence, if any? What do you make of this finding?

Discuss the implications in terms of accuracy and precision of your prevalence estimates.

```
first_prevalence <- has_bigtooth_maples1$prevalence
third_prevalence <- has_bigtooth_maples3$prevalence
fourth_prevalence <- has_bigtooth_maples4$prevalence
se1 <- sd(first_prevalence)/sqrt(length((first_prevalence)))
se3 <- sd(third_prevalence)/sqrt(length((third_prevalence)))
se4 <- sd(fourth_prevalence)/sqrt(length((fourth_prevalence)))

(mean(first_prevalence) + 1.96) * se1
```

```
## [1] 0.05021454
```

```
(mean(third_prevalence) + 1.96) * se3
```

```
## [1] 0.01519842
```

```
(mean(fourth_prevalence) + 1.96) * se4
```

```
## [1] 0.005182248
```

```
(mean(first_prevalence) - 1.96) * se1
```

```
## [1] -0.04580546
```

```
(mean(third_prevalence) - 1.96) * se3
```

```
## [1] -0.01441282
```

```
(mean(fourth_prevalence) - 1.96) * se4
```

```
## [1] -0.00490335
```

```
has_bigtooth_maples1 %>%
    summarize(mean = mean(prevalence), se = sd(prevalence)/sqrt(n()),
        lower_ci = mean - 1.96 * se, upper_ci = mean + 1.96 *
            se)
```

```
## # A tibble: 1 x 4
##    mean     se lower_ci upper_ci
##   <dbl>  <dbl>    <dbl>    <dbl>
## 1  0.09 0.0245   0.0420    0.138
```

```
has_bigtooth_maples3 %>%
    summarize(mean = mean(prevalence), se = sd(prevalence)/sqrt(n()),
        lower_ci = mean - 1.96 * se, upper_ci = mean + 1.96 *
            se)
```

```
## # A tibble: 1 x 4
##    mean      se lower_ci upper_ci
##   <dbl>   <dbl>    <dbl>    <dbl>
## 1 0.052 0.00755   0.0372   0.0668
```

```
has_bigtooth_maples4 %>%
    summarize(mean = mean(prevalence), se = sd(prevalence)/sqrt(n()),
        lower_ci = mean - 1.96 * se, upper_ci = mean + 1.96 *
            se)
```

```
## # A tibble: 1 x 4
##     mean      se lower_ci upper_ci
##    <dbl>   <dbl>    <dbl>    <dbl>
## 1 0.0542 0.00257   0.0492   0.0592
```

Answer: