# Sleepy mammals and bootstrapping

Group 9

Module 5 ICA3

## Introduction

Data frame `msleep` is a dataset available in `ggplot2`. It contains information on 83 mammals with regards to their sleep behavior. To get started, load packages `tidyverse` and `infer`.

Below is a basic custom theme. Feel free to try it out when you use `ggplot()`. Simply add it as a layer to your plot. Rather than using `theme_bw()` you can use `theme_custom()`. This custom theme increases the font point size on axes and their labels.

```r
theme_custom <- function() {
  theme_bw() +
  theme(axis.title = element_text(size = 16),
        title = element_text(size = 20),
        axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        plot.caption = element_text(size = 10))
}
```

Take a `glimpse()` at the data below

```r
attach(msleep)
glimpse(msleep)
```

```
Rows: 83
Columns: 11
$ name         <chr> "Cheetah", "Owl monkey", "Mountain beaver", "Greater shor~
$ genus        <chr> "Acinonyx", "Aotus", "Aplodontia", "Blarina", "Bos", "Bra~
$ vore         <chr> "carni", "omni", "herbi", "omni", "herbi", "herbi", "carn~
$ order        <chr> "Carnivora", "Primates", "Rodentia", "Soricomorpha", "Art~
$ conservation <chr> "lc", NA, "nt", "lc", "domesticated", NA, "vu", NA, "dome~
$ sleep_total  <dbl> 12.1, 17.0, 14.4, 14.9, 4.0, 14.4, 8.7, 7.0, 10.1, 3.0, 5~
$ sleep_rem    <dbl> NA, 1.8, 2.4, 2.3, 0.7, 2.2, 1.4, NA, 2.9, NA, 0.6, 0.8, ~
$ sleep_cycle  <dbl> NA, NA, NA, 0.1333333, 0.6666667, 0.7666667, 0.3833333, N~
$ awake        <dbl> 11.9, 7.0, 9.6, 9.1, 20.0, 9.6, 15.3, 17.0, 13.9, 21.0, 1~
$ brainwt      <dbl> NA, 0.01550, NA, 0.00029, 0.42300, NA, NA, NA, 0.07000, 0~
$ bodywt       <dbl> 50.000, 0.480, 1.350, 0.019, 600.000, 3.850, 20.490, 0.04~
```

For all the questions that follow, use a sequence of functions in package `infer`. For details on `msleep`, type `?msleep` in your console.

# Test of Hypothesis

Past research has shown that humans have a **median** sleep time of 7.5 hours per day. Researchers want to investigate if all other mammals have a higher median number of sleep hours per day. A random sample of 82 mammals revealed a median number of sleep hours per day to be 10.1 hours. Is this enough evidence to suggest mammals that are not human have a higher median number of sleep hours per day?

## Test for the median

State the null and alternative hypotheses given the problem above.

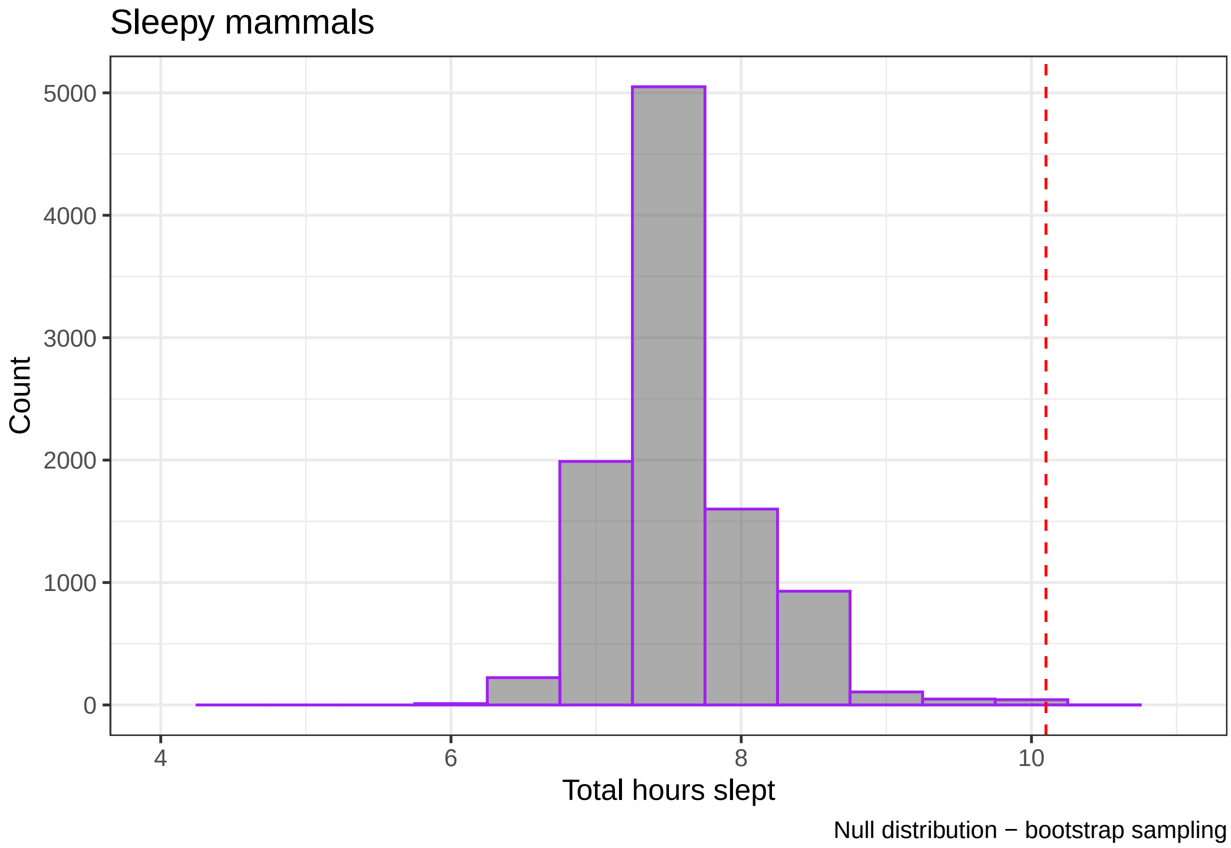
The null hypothesis states that the population median is 7.5 hours per day.

The alternative hypothesis states that the population median is not 7.5 hours per day.

### Simulated null distribution

Plot a histogram of the simulated null distribution and place a vertical line at the value of the observed sample median of 10.1.

```
Warning: Removed 2 rows containing missing values (geom_bar).
```



### Understanding the histogram

In the histogram, what does each bar represent? For example, the bar at 8 having a little more than 1500 counts, what does that represent?

This represents that around 1500 animals that are not human sleep about 8 hours a day.

Where is the center of the histogram? What do you think will change in the shape of the histogram if we had a list of 150 mammals (excluding human)? The center of the histogram is the median. I think that the histogram would shift more towards normal distibution with 150 mammals.

### Compute the p-value

Use the simulated null distribution to compute the p-value. Recall that the p-value is the probability of observing data at least as favorable to the alternative hypothesis as the current data set, given that the null hypothesis is true.

```r
null.dist %>%
  filter(stat >= 10.1) %>%
  summarise(p_value = 2 * n() / nrow(null.dist))
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1  0.0004
```

### Conclusion

State your conclusion in the context of the problem.

We have rejected the null hypothesis. This allows to have sufficient to confirm that the median of our sample is greater than 7.5 which is therefore greater then the amount of the median for humans. Animals sleep a greater amount of time than humans.

## How about the mean?

We have discussed about the average sleep hours in our class lecture in context of a two sided hypothesis test. Let us examine how it changes (if at all) when the test is one sided.

On average, it is believed that humans need 8 hours of sleep per day. Using the same sample as above, it was found that the average number of sleep hours per day to be 10.4 hours for all other mammals except humans. Is this enough evidence to suggest mammals that are not human have a higher mean number of sleep hours per day?

### Hypotheses

State the null and alternative hypotheses.

The null hypothesis states that the population mean is 8 hours per day.

The alternative hypothesis states that the population median is not 8 hours per day.

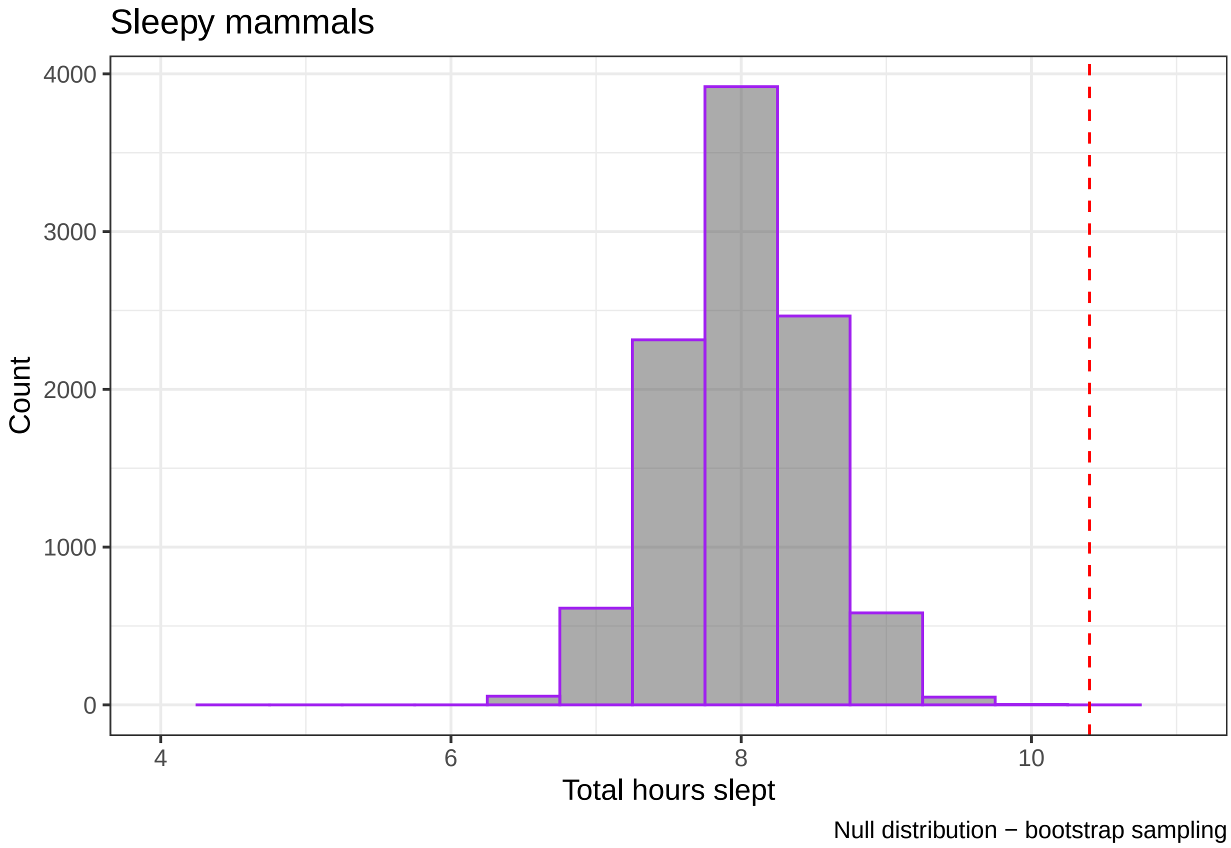
### Simulated null distribution

Plot a histogram of the simulated null distribution and place a vertical line at the value of the observed sample mean of 10.4.

```
  null.dist <- msleep %>%
  filter(name != "Human") %>%
  specify(response = sleep_total) %>%
  hypothesize(null = "point" , mu = 8) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

# Plot for the null distribution
null.dist %>%
  ggplot(mapping = aes(x = stat)) +
  geom_histogram(binwidth = .5, color = "purple", alpha = .5) +
  labs(x = "Total hours slept", y = "Count",
       title = "Sleepy mammals",
       caption = "Null distribution - bootstrap sampling") +
  xlim(4,11)+
  geom_vline(xintercept = 10.4, color="red", linetype= "dashed")+
  theme_bw()
```

```
Warning: Removed 2 rows containing missing values (geom_bar).
```



### Compute the p-value

Use the simulated null distribution to compute the p-value. Recall that the p-value is the probability of observing data at least as favorable to the alternative hypothesis as the current data set, given that the null hypothesis is true.

```
null.dist %>%
  filter(stat >= 10.4) %>%
  summarise(p_value = 2 * n() / nrow(null.dist))
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1       0
```

### Conclusion

Did the conclusion change when the parameter was changed from median to mean? Justify your answer.

No, becuase the p-value is less thenthe significance level, the data leads us to reject the null hypothesis.

## One sided vs. two sided

If our research question was: is this enough evidence to suggest that mean number of sleep hours per day for mammals that are not human is different from humans, how would our hypothesis test change?

We would have to perform a two sided test proving that the sleep for mammals is different than for with humans.We would go to the extreme high and extreme low. ### Compute the p-value

Using the same the simulated null distribution for the mean, calculate the p-value.

```
null.dist %>%
  filter(stat >= 10.1) %>%
  summarise(p_value = 2 * n() / nrow(null.dist))
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1       0
```

### Which test needs stronger evidence?

Based on the calculated p-values for the one sided and two-sided tests, which test needs stronger evidence against the null hypothesis in order to reject it?

The two sided test would need stronger evidence against the null hypothesis.

## Plausible value of the population parameter?

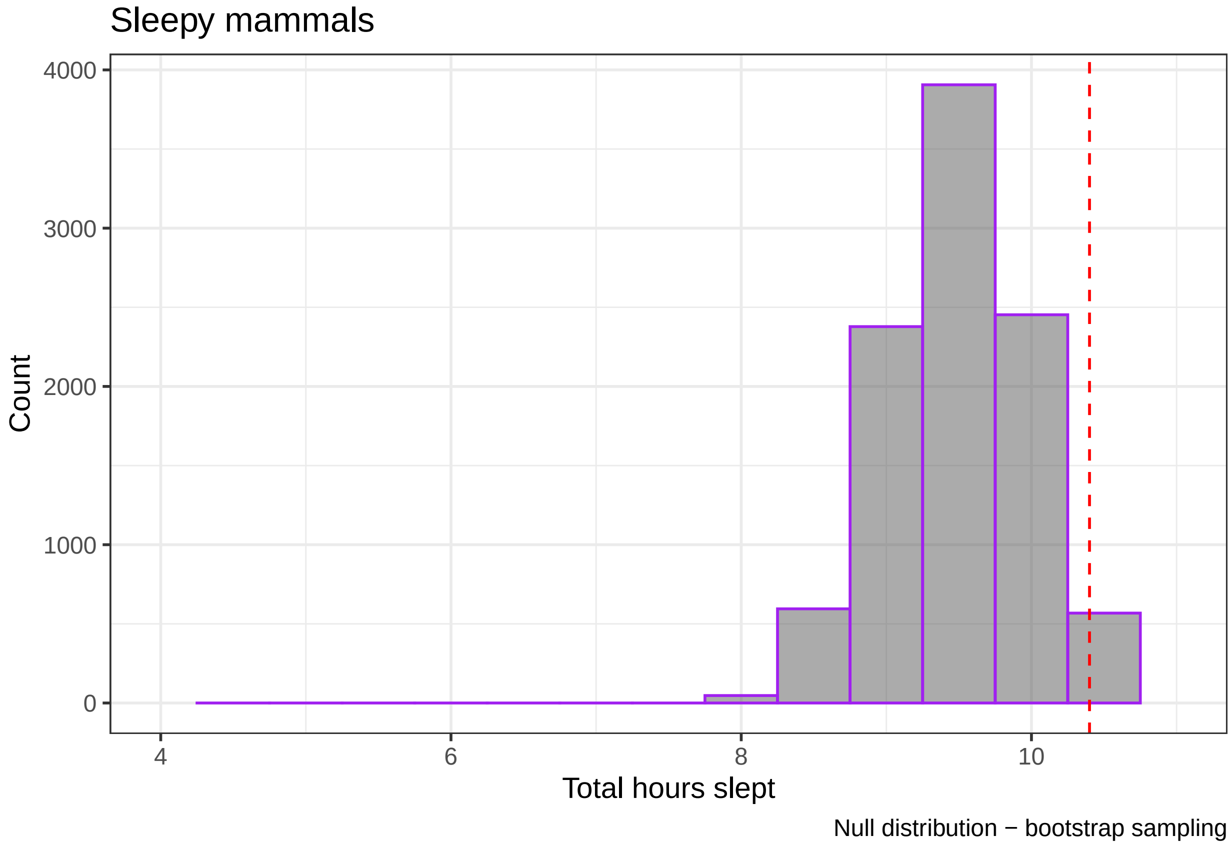
Based on the given data, had our hypothesized mean been 9.5, would our conclusion have changed? Perform both the one-sided and two-sided test and state the conclusions.

```
  null.dist <- msleep %>%
  filter(name != "Human") %>%
  specify(response = sleep_total) %>%
  hypothesize(null = "point" , mu = 9.5) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

null.dist %>%
  ggplot(mapping = aes(x = stat)) +
  geom_histogram(binwidth = .5, color = "purple", alpha = .5) +
  labs(x = "Total hours slept", y = "Count",
       title = "Sleepy mammals",
       caption = "Null distribution - bootstrap sampling") +
  xlim(4,11)+
  geom_vline(xintercept = 10.4, color="red", linetype= "dashed")+
  theme_bw()
```

```
Warning: Removed 3 rows containing non-finite values (stat_bin).

Warning: Removed 2 rows containing missing values (geom_bar).
```



```
null.dist %>%
  filter(stat >= 10.1) %>%
  summarise(p_value = 2 * n() / nrow(null.dist))
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.226
```

We do not have enough evidence to disapprove the null hypothesis which therefore we cannot confidently say that the animals have a different mean than humans.

# Estimation

## Population mean

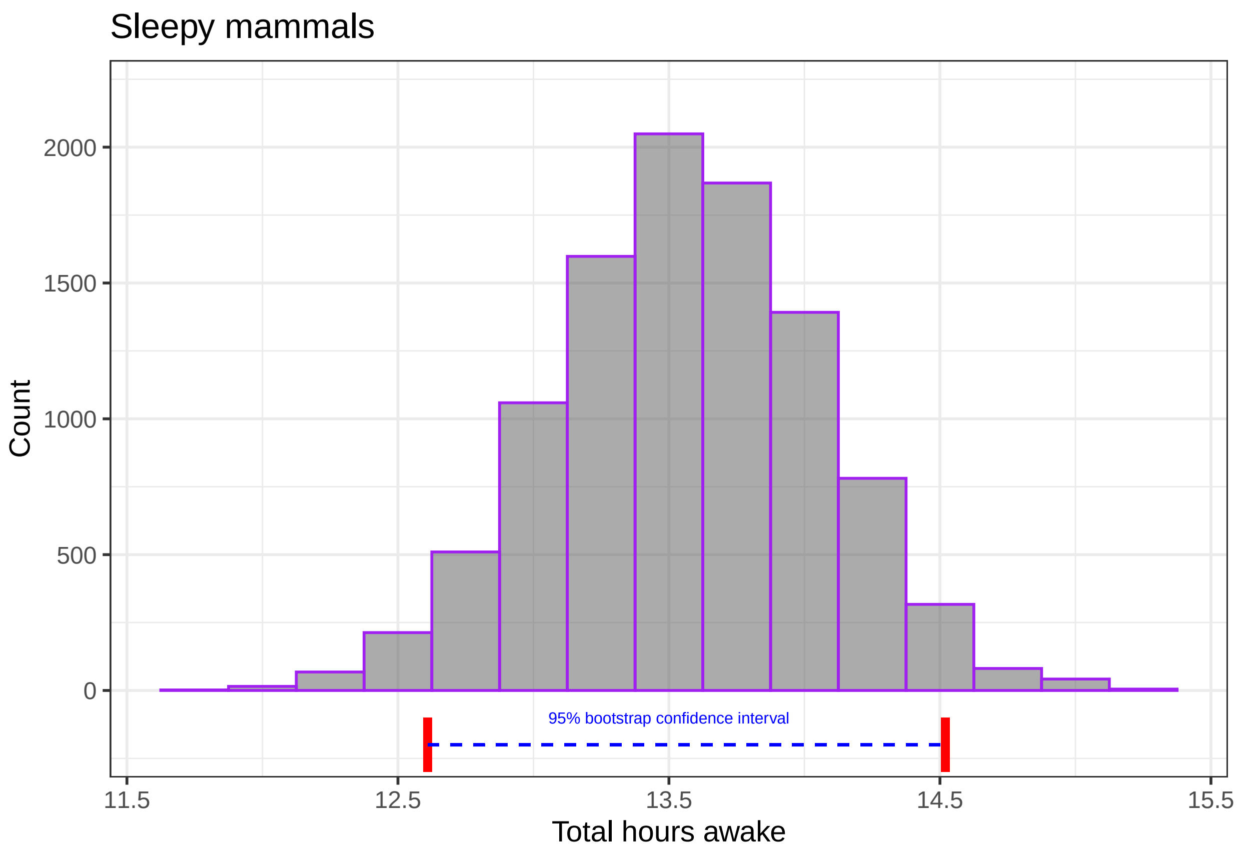
1. Create a 95% confidence interval for the mean amount of hours all mammals are awake per day. Also, plot the simulated bootstrap distribution.Interpret the 95% confidence interval.

```
boot.means <- msleep %>%
  specify(response = awake) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")
boot.means %>%
  summarise(lower95 = quantile(stat, probs = .025),
            upper95 = quantile(stat, probs = .975),
            lower99 = quantile(stat, probs = 0.005),
            upper99 = quantile(stat, probs = .995))
```

```
# A tibble: 1 x 4
  lower95 upper95 lower99 upper99
    <dbl>   <dbl>   <dbl>   <dbl>
1    12.6    14.5    12.3    14.8
```

2. Create 90% and 99% confidence intervals for the mean amount of hours all mammals are awake per day. What do you notice about the widths of the three intervals?

```
boot.means %>%
  ggplot(mapping = aes(x = stat)) +
  geom_histogram(binwidth = .25, color = "purple", alpha = .5) +
  geom_segment(x = 12.61, xend =  12.61,
               y = -300, yend = -100, color = "red", size = 1.5) +
  geom_segment(x = 14.52, xend =  14.52,
               y = -300, yend = -100, color = "red", size = 1.5) +
  geom_segment(x = 12.61, xend = 14.52, y = -200, yend = -200,
               color = "blue", linetype = "dashed") +
  scale_y_continuous(limits = c(-200, 2200)) +
  annotate("text", x = 13.5, y = -100,
           label = "95% bootstrap confidence interval",
           color = "blue", size = 2) +
  labs(x = "Total hours awake", y = "Count",
       title = "Sleepy mammals") +
  theme_bw()
```



## Beyond mean: Population standard deviation

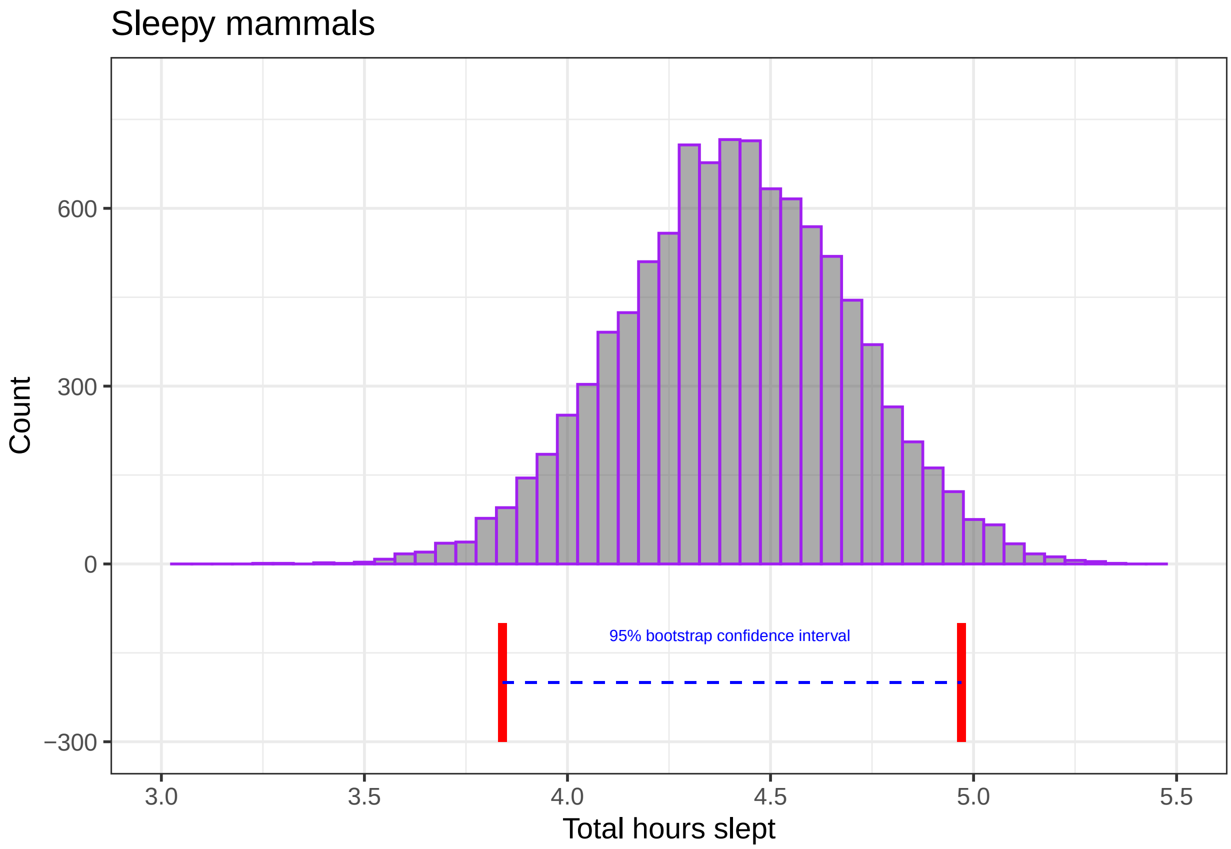
1. Create a 95% confidence interval for the standard deviation in terms of the hours mammals sleep per day.

```
# bootstrap samples change the stat to sd
boot.sd <- msleep %>%
  specify(response = sleep_total) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "sd")

boot.sd %>%
  summarise(lower95 = quantile(stat, probs = .025),
            upper95 = quantile(stat, probs = .975),
            lower99 = quantile(stat, probs = 0.005),
            upper99 = quantile(stat, probs = .995))

boot.sd %>%
  ggplot(mapping = aes(x = stat)) +
  geom_histogram(binwidth = .05, color = "purple", alpha = .5) +
  geom_segment(x = 3.84, xend =  3.84,
               y = -300, yend = -100, color = "red", size = 1.5) +
  geom_segment(x = 4.97, xend =  4.97,
               y = -300, yend = -100, color = "red", size = 1.5) +
  geom_segment(x = 3.84, xend = 4.97, y = -200, yend = -200,
                color = "blue", linetype = "dashed") +
  xlim(3.0, 5.5)+
  scale_y_continuous(limits = c(-300, 800)) +
  annotate("text", x = 4.4, y = -120,
           label = "95% bootstrap confidence interval",
           color = "blue", size = 2) +
  labs(x = "Total hours slept", y = "Count",
       title = "Sleepy mammals") +
  theme_bw()
```

```
Warning: Removed 2 rows containing missing values (geom_bar).
```



```
# save as vector

# visualize bootstrap sample means
```

```
# A tibble: 1 x 4
  lower95 upper95 lower99 upper99
    <dbl>   <dbl>   <dbl>   <dbl>
1    3.85    4.96    3.66    5.11
```

# References

1. V. M. Savage and G. B. West. A quantitative, theoretical framework for understanding mammalian sleep. Proceedings of the National Academy of Sciences, 104 (3):1051-1056, 2007.

2. https://cran.r-project.org/web/packages/infer/vignettes/flights_examples.html