Data 606 - Lab 7

Avery Davidowitz

2022-11-10

Inference for numerical data

Load Packages

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

Load Data

```
data('yrbss', package='openintro')
```

Exercise 1

What are the cases in this data set? How many cases are there in our sample?

```
glimpse(yrbss)
```

```
## Rows: 13,583
## Columns: 13
## $ age                      <int> 14, 14, 15, 15, 15, 15, 15, 14, 15, 15, 15, 1~
## $ gender                   <chr> "female", "female", "female", "female", "fema~
## $ grade                    <chr> "9", "9", "9", "9", "9", "9", "9", "9", "9", ~
## $ hispanic                 <chr> "not", "not", "hispanic", "not", "not", "not"~
## $ race                     <chr> "Black or African American", "Black or Africa~
## $ height                   <dbl> NA, NA, 1.73, 1.60, 1.50, 1.57, 1.65, 1.88, 1~
## $ weight                   <dbl> NA, NA, 84.37, 55.79, 46.72, 67.13, 131.54, 7~
## $ helmet_12m               <chr> "never", "never", "never", "never", "did not ~
## $ text_while_driving_30d   <chr> "0", NA, "30", "0", "did not drive", "did not~
## $ physically_active_7d     <int> 4, 2, 7, 0, 2, 1, 4, 4, 5, 0, 0, 0, 4, 7, 7, ~
## $ hours_tv_per_school_day  <chr> "5+", "5+", "5+", "2", "3", "5+", "5+", "5+",~
## $ strength_training_7d     <int> 0, 0, 0, 0, 1, 0, 2, 0, 3, 0, 3, 0, 0, 7, 7, ~
## $ school_night_hours_sleep <chr> "8", "6", "<5", "6", "9", "8", "9", "6", "<5"~
```

The rows in this data set represent 13 characteristics of a specific youth. There are 13,583 people in the data set.

Exercise 2

How many observations are we missing weights from?

```
summary(yrbss$weight)
```

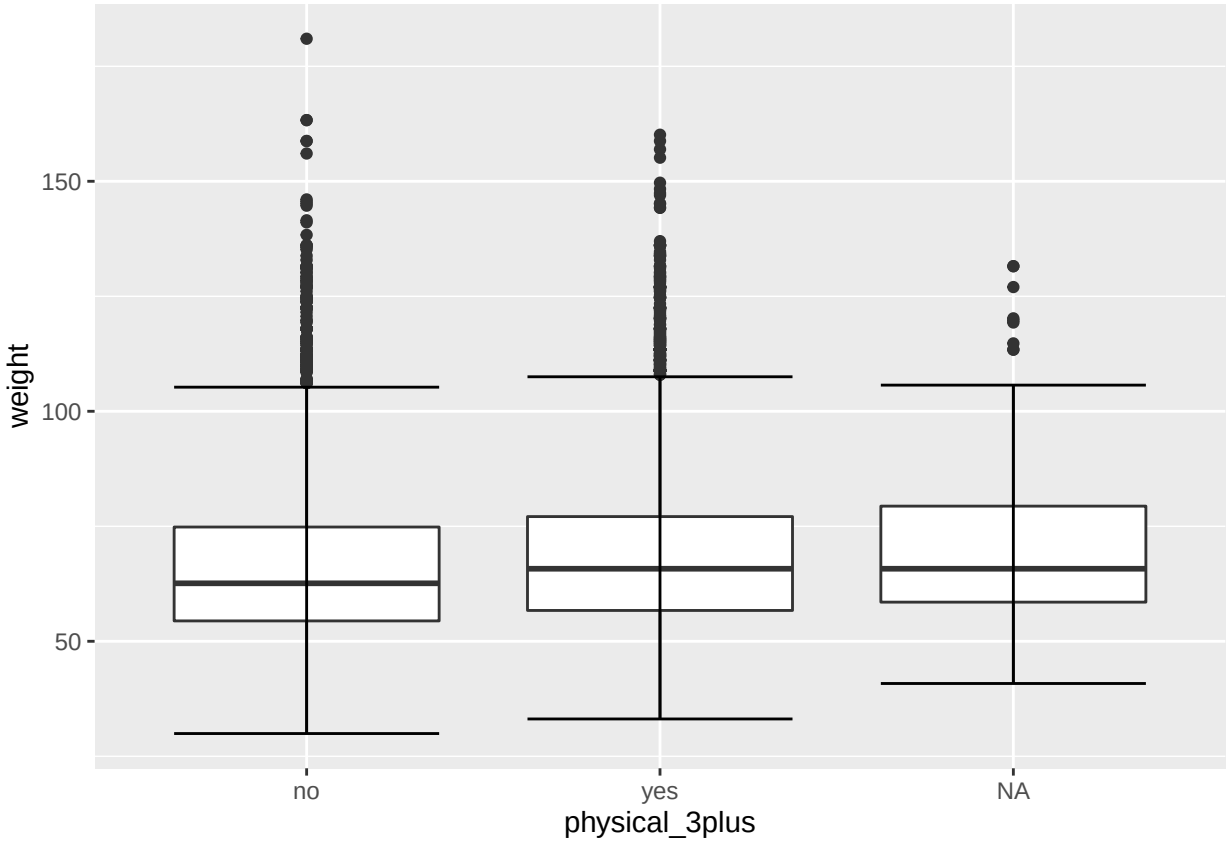
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   29.94   56.25   64.41   67.91   76.20  180.99    1004
```

We are missing 1004 weight observations.

Exercise 3

Make a side-by-side boxplot of physical_3plus and weight. Is there a relationship between these two variables? What did you expect and why?

```
yrbss <- yrbss %>%
  mutate(physical_3plus = ifelse(yrbss$physically_active_7d > 2, "yes", "no"))

ggplot(yrbss, aes(x=physical_3plus, y=weight)) +
  geom_boxplot() + stat_boxplot(geom = 'errorbar')
```



It is unclear if there is any relationship between these variables from a plot alone.

```
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(mean_weight = mean(weight, na.rm = TRUE))
```

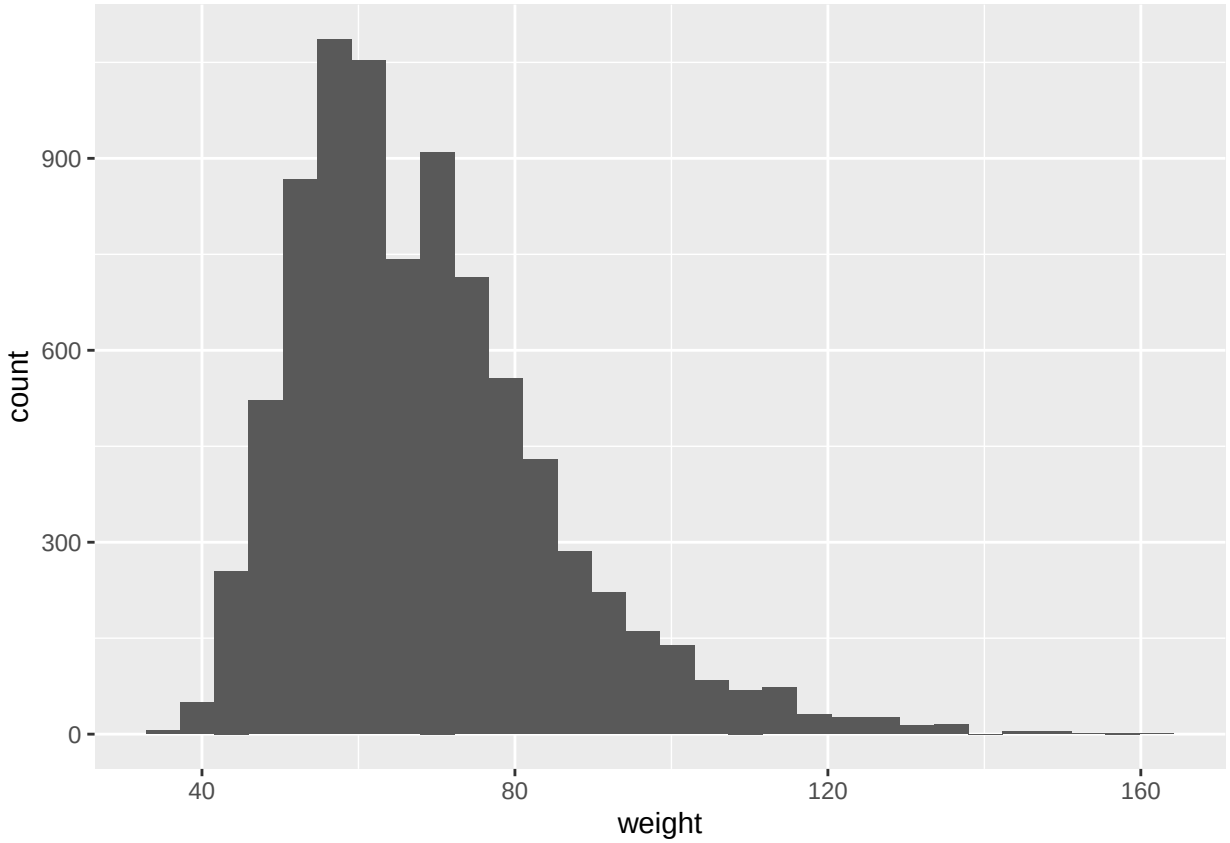
```
## # A tibble: 3 x 2
##   physical_3plus mean_weight
##   <chr>                <dbl>
## 1 no                    66.7
## 2 yes                   68.4
## 3 <NA>                  69.9
```

Exercise 4

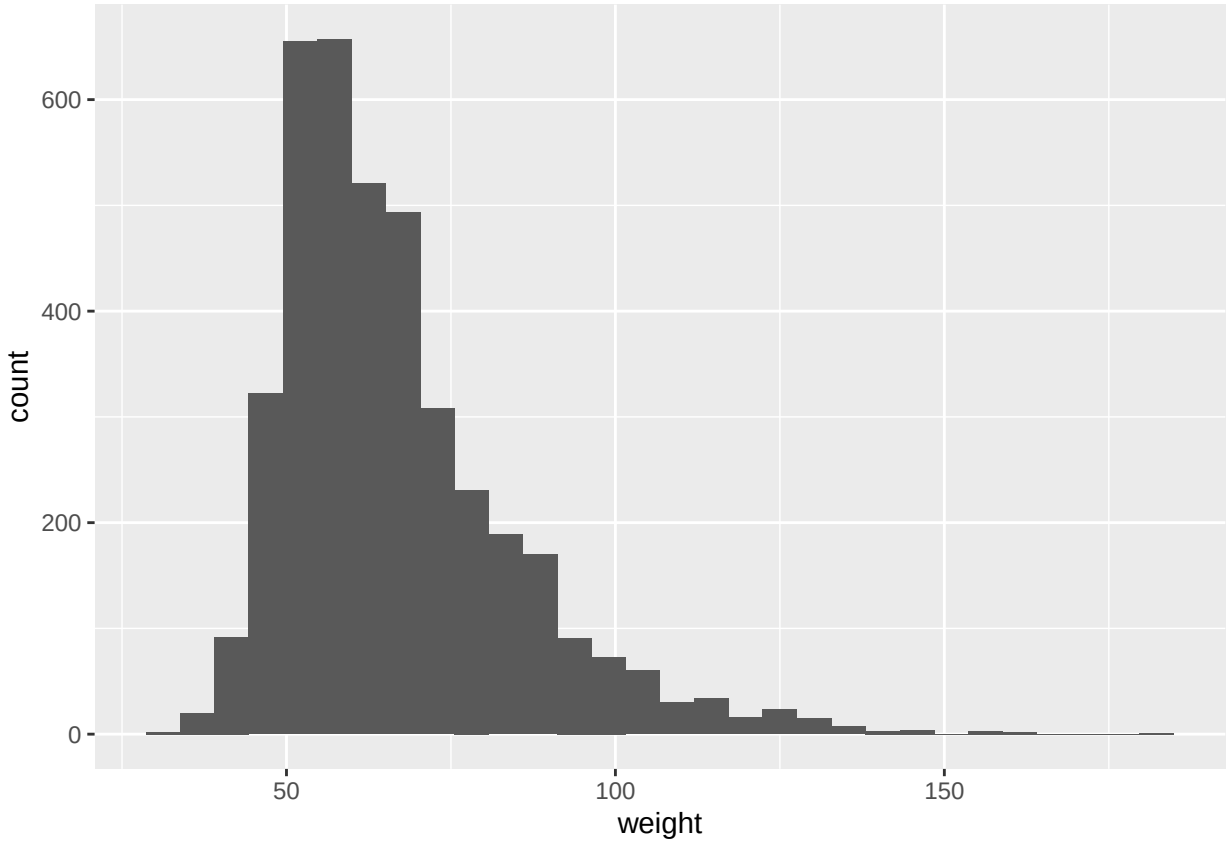
Are all conditions necessary for inference satisfied? Comment on each. You can compute the group sizes with the summarize command above by defining a new variable with the definition n().

Since the data is from a randomized experiment we can assume independence both between groups and within groups. However, there are many extreme outliers for both the yes and no categories of physical_3plus so it is unclear if this data would pass the normality test needed. Per the OpenIntro text using the t-distribution for the difference of two means requires that you check both groups for outliers. While the distributions for both appear to be nearly normal in shape, the whiskers on the box plot indicate many outliers. Both have skew. Since the outliers are not particularly extreme I assume that normality will hold.

```
p3plus_yes <- dplyr::filter(yrbss, physical_3plus=="yes")
p3plus_no <- dplyr::filter(yrbss, physical_3plus=="no")
ggplot(p3plus_yes, aes(x=weight)) +
  geom_histogram()
```



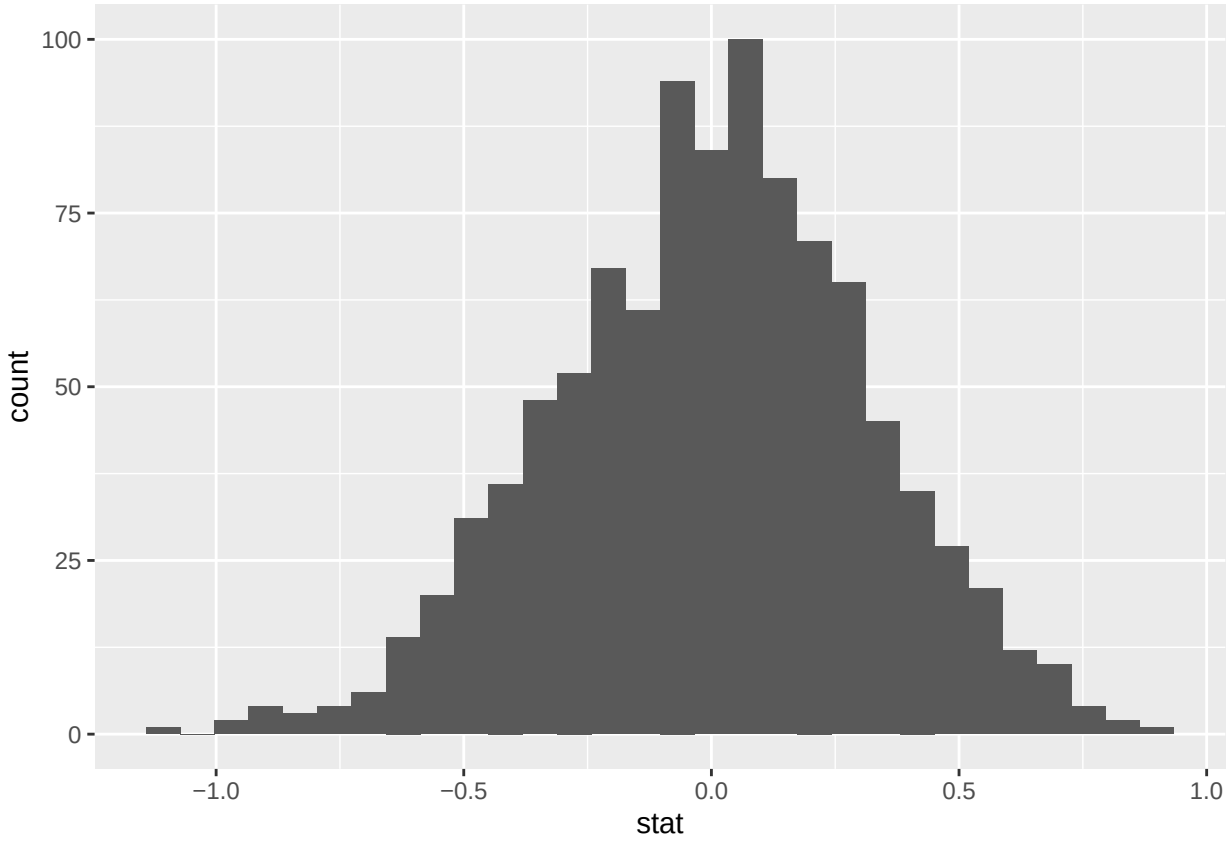

```
ggplot(p3plus_no, aes(x=weight)) +
  geom_histogram()
```



Exercise 5

Write the hypotheses for testing if the average weights are different for those who exercise at least times a week and those who don't.

```
p3plus <- yrbss %>% dplyr::filter(!is.na(physical_3plus))
obs_diff <- p3plus |>
  specify(weight ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

null_dist <- p3plus %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

ggplot(data = null_dist, aes(x = stat)) +
  geom_histogram()
```



The null hypothesis is that the difference of means of yes(works out 3 times a week) and no(does not work out 3 times a week) are equal to 0 and the alternate hypothesis is that they are different.

Exercise 6

How many of these null permutations have a difference of at least obs_stat?

```
null_dist %>%
  get_p_value(obs_stat = obs_diff, direction = "two_sided")
```

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

None of the null distribution have the observed difference. Therefore, we reject the null hypothesis and claim that the difference in means is statistically significant.

Exercise 7

Construct and record a confidence interval for the difference between the weights of those who exercise at least three times a week and those who don't, and interpret this interval in context of the data.

```r
get_ci(null_dist, level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.630    0.605
```

We would be 95% confident that the difference between the means should fall in the confidence interval if the null hypothesis was true. Since our observed difference doesn't fall into the CI we have to reject the null.

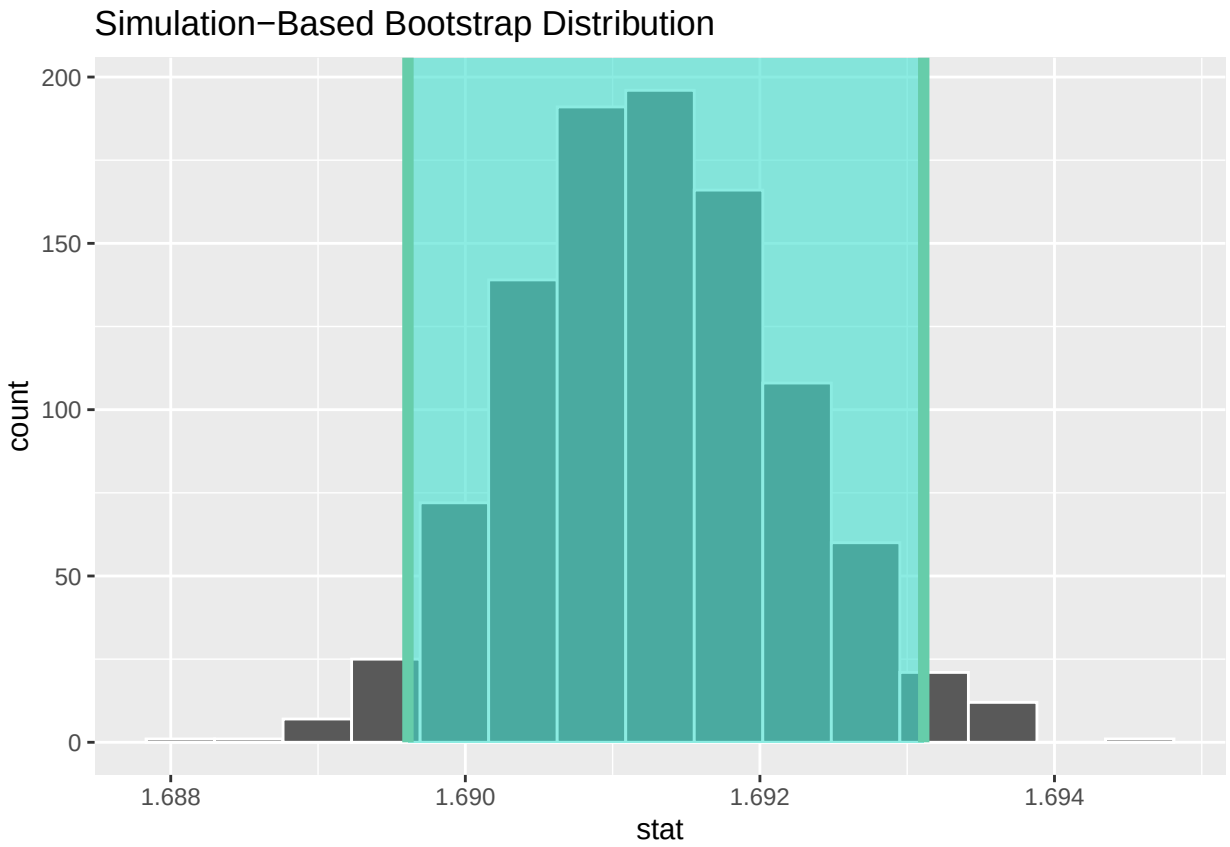
Exercise 8

Calculate a 95% confidence interval for the average height in meters (height) and interpret it in context.

```r
height_df <- yrbss %>% dplyr::filter(!is.na(height))
x_bar <- height_df %>%
  specify(response = height) %>%
  calculate(stat = "mean")

boot_dist <- height_df %>%
   specify(response = height) %>%
   generate(reps = 1000, type = "bootstrap") %>%
   calculate(stat = "mean")

percentile_ci <- get_ci(boot_dist, level = .95)

percentile_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     1.69     1.69
```

```r
visualize(boot_dist) +
  shade_confidence_interval(endpoints = percentile_ci)
```



We are 95% confident that the population mean is between 1.689345 and 1.693021.

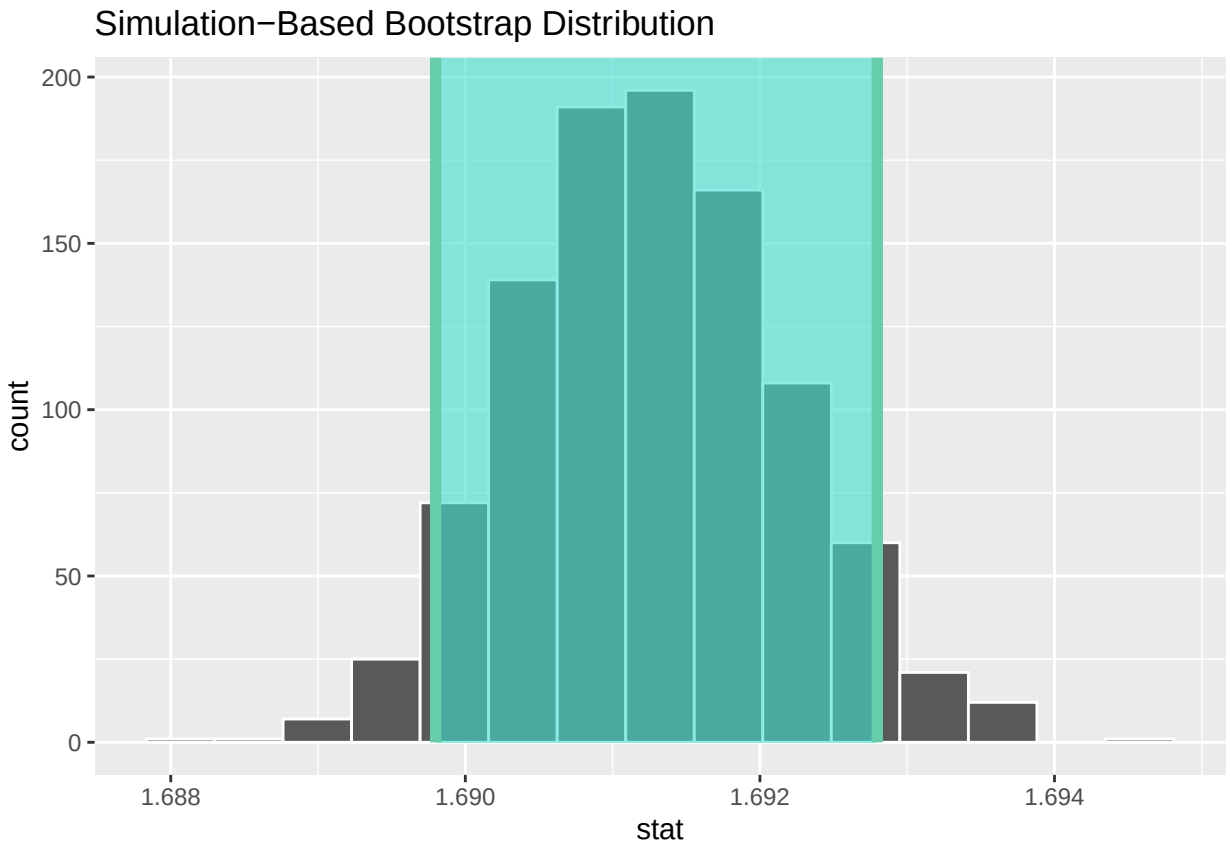
Exercise 9

Calculate a new confidence interval for the same parameter at the 90% confidence level. Comment on the width of this interval versus the one obtained in the previous exercise.

```
percentile_ci_90 <- get_ci(boot_dist, level = .90)

percentile_ci_90
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     1.69     1.69
```

```
visualize(boot_dist) +
  shade_confidence_interval(endpoints = percentile_ci_90)
```

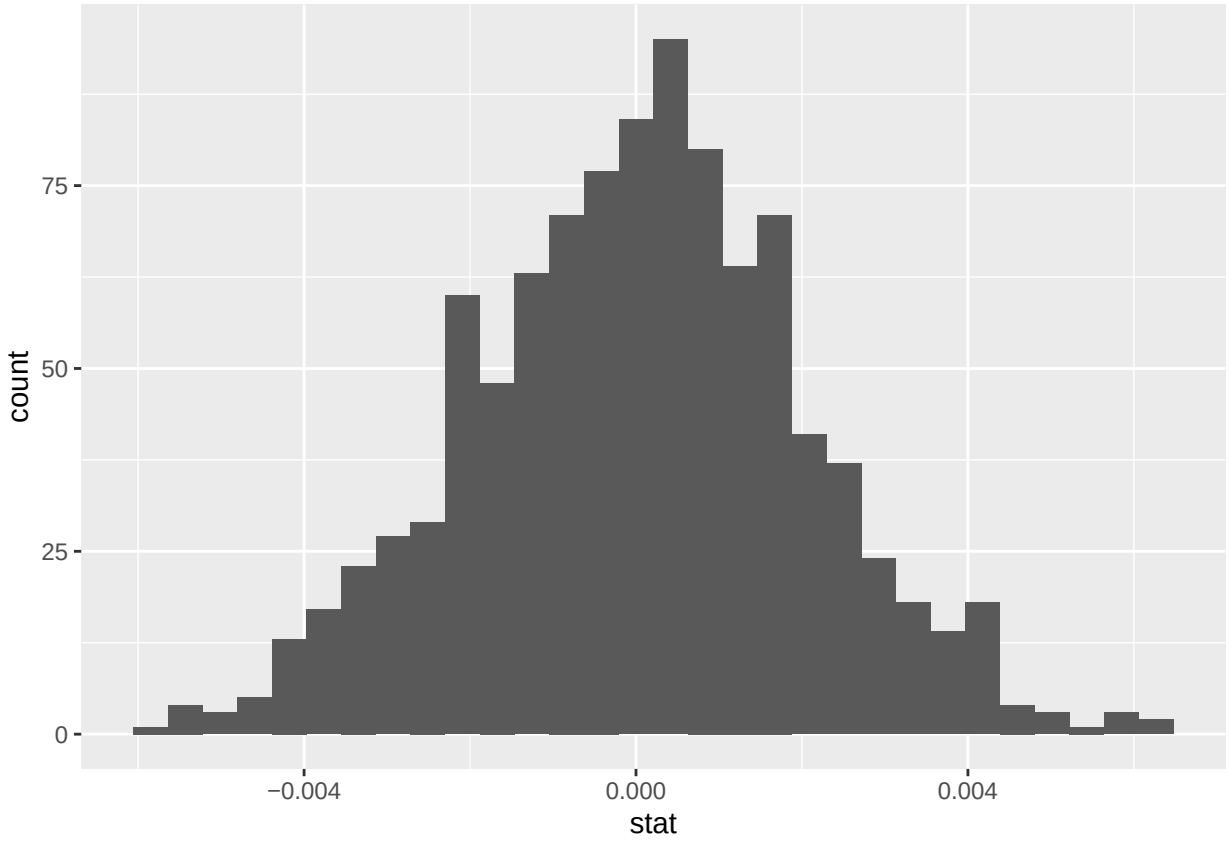


The range of the confidence interval got tighter because we used a lower 90% threshold for confidence.

Exercise 10

Conduct a hypothesis test evaluating whether the average height is different for those who exercise at least three times a week and those who don't.

```
obs_diff2 <- p3plus |>
  specify(height ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

null_dist2 <- p3plus %>%
  specify(height ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

ggplot(data = null_dist2, aes(x = stat)) +
  geom_histogram()
```




```
null_dist %>%
  get_p_value(obs_stat = obs_diff2, direction = "two_sided")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1    0.94
```

```
get_ci(null_dist2, level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1 -0.00401  0.00407
```

Exercise 11

Now, a non-inference task: Determine the number of different options there are in the dataset for the hours_tv_per_school_day there are.

```
tv_values <- dplyr::distinct(yrbss,hours_tv_per_school_day)
tv_values
```

```
## # A tibble: 8 x 1
##   hours_tv_per_school_day
##   <chr>
## 1 5+
## 2 2
## 3 3
## 4 do not watch
## 5 <1
## 6 4
## 7 1
## 8 <NA>
```

Exercise 12

Come up with a research question evaluating the relationship between height or weight and sleep. Formulate the question in a way that it can be answered using a hypothesis test and/or a confidence interval. Report the statistical results, and also provide an explanation in plain language. Be sure to check all assumptions, state your alpha level, and conclude in context.

```
sleep_values <- dplyr::distinct(yrbss,school_night_hours_sleep)
sleep_values
```

```
## # A tibble: 8 x 1
##   school_night_hours_sleep
##   <chr>
## 1 8
## 2 6
## 3 <5
## 4 9
## 5 10+
## 6 7
## 7 5
## 8 <NA>
```

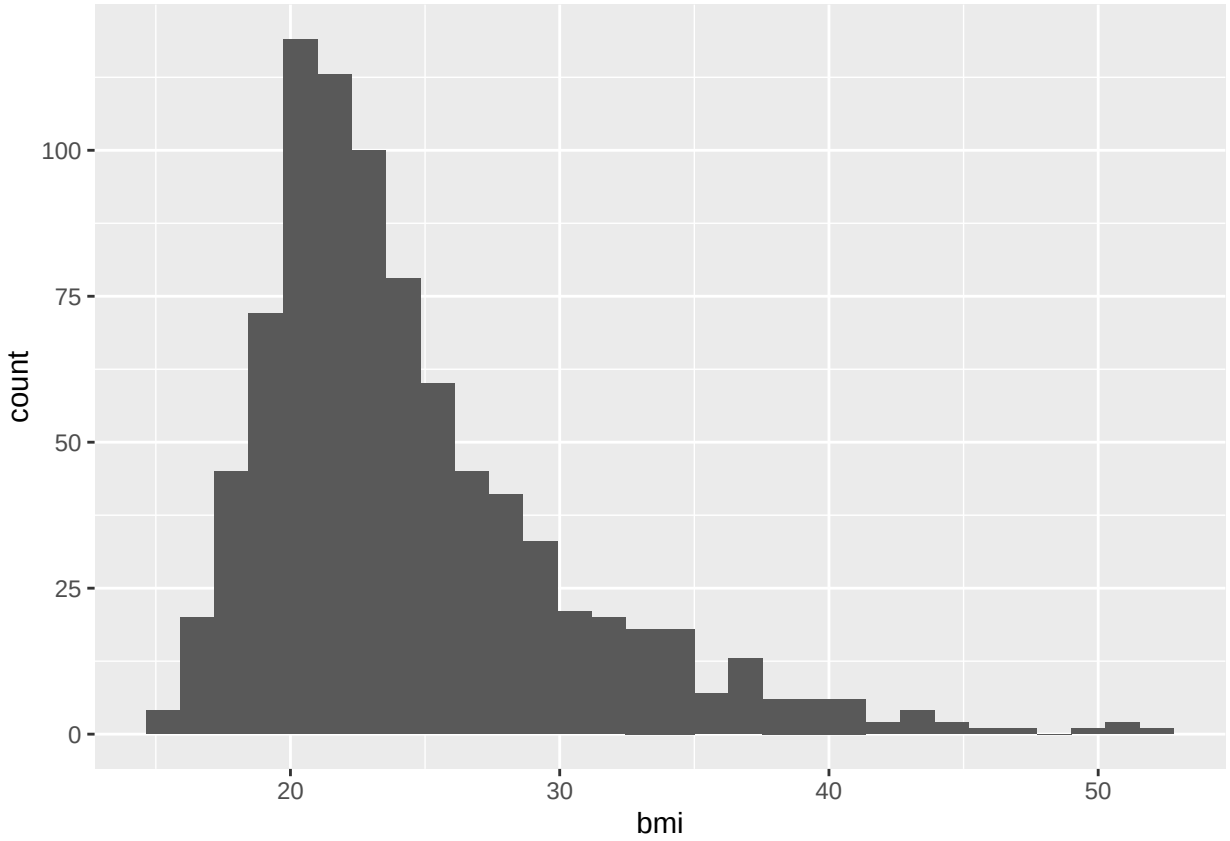
I will evaluate if youths who get the least sleep of less than 5 hours on school nights have a statistically significant difference in mean BMI(kilograms divided by height in meters squared) compared to the rest of their peers.

Setting up the dataframe for analysis

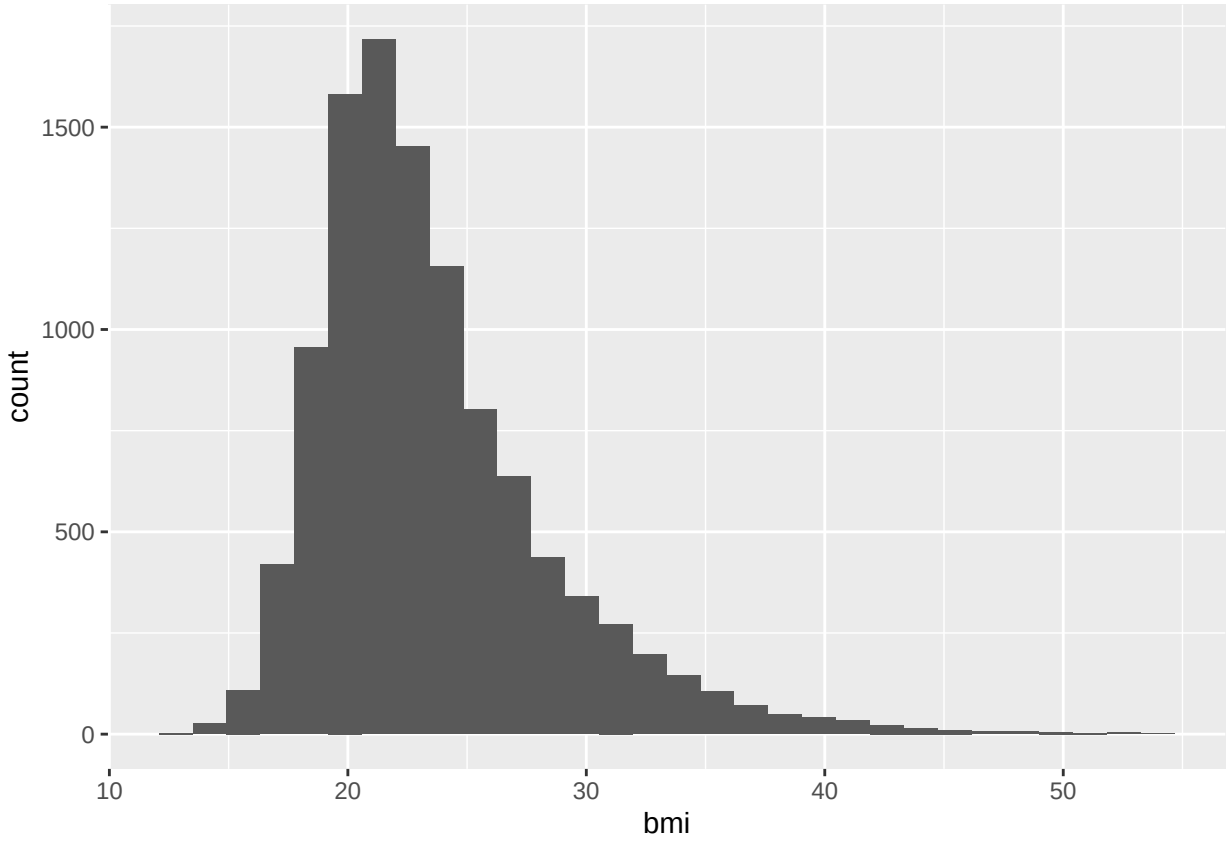
```
bmi_sleep_df <- yrbss |> dplyr::filter(!is.na(height) & !is.na(weight) & !is.na(school_night_hours_slee
bmi_sleep_df <- bmi_sleep_df %>%
  dplyr::mutate(bmi = weight/height^2 ) |>
  dplyr::mutate(sleep_less5 = ifelse(bmi_sleep_df$school_night_hours_sleep=="<5", "yes", "no"))
```

Since the data is from a random sample independence is given both between and in groups. The distributions for both groups appear nearly normal with a slight skew.

```
sleep_less5_yes <- dplyr::filter(bmi_sleep_df, sleep_less5=="yes")
sleep_less5_no <- dplyr::filter(bmi_sleep_df, sleep_less5=="no")
ggplot(sleep_less5_yes, aes(x=bmi)) +
  geom_histogram()
```




```
ggplot(sleep_less5_no, aes(x=bmi)) +
  geom_histogram()
```



I will test with a alpha of .05 if there is a difference in mean bmi of those who get less than 5 hours sleep compared to the mean bmi of everyone else.

```
obs_diff3 <- bmi_sleep_df |>
  specify(bmi ~ sleep_less5) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

null_dist3 <- bmi_sleep_df %>%
  specify(bmi ~ sleep_less5) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

null_dist3 %>%
  get_p_value(obs_stat = obs_diff3, direction = "two_sided")
```

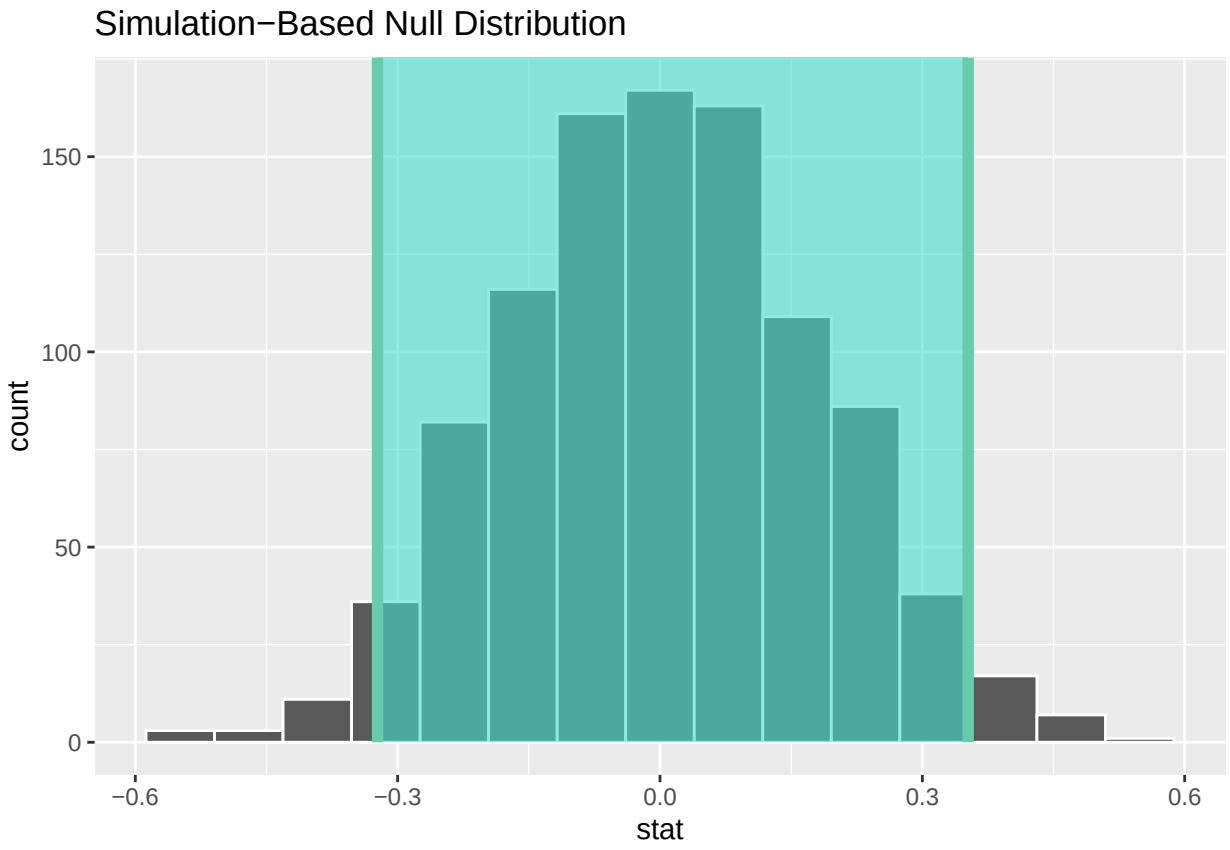
```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

```
ci_95 <- get_ci(null_dist3, level = 0.95)
ci_95
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
```

```
##       <dbl>    <dbl>
## 1   -0.323    0.352
```

```
visualize(null_dist3) +
  shade_confidence_interval(endpoints = ci_95)
```



The observed difference of .913 between the mean BMIs is statistically significant with a 95% confidence interval. The expected range of the CI is between -0.3371378 and 0.3629425. The fact that our observed mean BMI is so much higher that the range indicates that those that get less sleep are more likely to have a higher BMI.