DATA 606 Lab 7: Inference for numerical data

Kevin Kirby

Getting Started

Load packages

In this lab, we will explore and visualize the data using the **tidyverse** suite of packages, and perform statistical inference using **infer**. The data can be found in the companion package for OpenIntro resources, **openintro**.

Let's load the packages.

```
library(tidyverse)
library(openintro)
library(infer)
library(ggplot2)
```

The data

Every two years, the Centers for Disease Control and Prevention conduct the Youth Risk Behavior Surveillance System (YRBSS) survey, where it takes data from high schoolers (9th through 12th grade), to analyze health patterns. You will work with a selected group of variables from a random sample of observations during one of the years the YRBSS was conducted.

Load the `yrbss` data set into your workspace.

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

There are observations on 13 different variables, some categorical and some numerical. The meaning of each variable can be found by bringing up the help file:

```
?yrbss
```

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

Answer:

The data set is select variables from the CDC's Youth Risk Behavior Surveillance System. There are 13,583 cases, which I determined both from the help page and by counting the number of rows.

```
cat("I used help to view the author-provided explanation: \n")
```

```
## I used help to view the author-provided explanation:
```

```
help("yrbss")

cat("\nI reviewed the structure of the data with this:")
```

```
##
## I reviewed the structure of the data with this:
```

```
str(yrbss)
```

```
## tibble [13,583 x 13] (S3: tbl_df/tbl/data.frame)
##  $ age                     : int [1:13583] 14 14 15 15 15 15 15 14 15 15 ...
##  $ gender                  : chr [1:13583] "female" "female" "female" "female" ...
##  $ grade                   : chr [1:13583] "9" "9" "9" "9" ...
##  $ hispanic                : chr [1:13583] "not" "not" "hispanic" "not" ...
##  $ race                    : chr [1:13583] "Black or African American" "Black or African American" "
##  $ height                  : num [1:13583] NA NA 1.73 1.6 1.5 1.57 1.65 1.88 1.75 1.37 ...
##  $ weight                  : num [1:13583] NA NA 84.4 55.8 46.7 ...
##  $ helmet_12m              : chr [1:13583] "never" "never" "never" "never" ...
##  $ text_while_driving_30d  : chr [1:13583] "0" NA "30" "0" ...
##  $ physically_active_7d    : int [1:13583] 4 2 7 0 2 1 4 4 5 0 ...
##  $ hours_tv_per_school_day : chr [1:13583] "5+" "5+" "5+" "2" ...
##  $ strength_training_7d    : int [1:13583] 0 0 0 0 1 0 2 0 3 0 ...
##  $ school_night_hours_sleep: chr [1:13583] "8" "6" "<5" "6" ...
```

```
cat("\nI counted cases by counting the number of rows in the dataset: \n")
```

```
##
## I counted cases by counting the number of rows in the dataset:
```

```
cat("Number of cases: ", nrow(yrbss))
```

```
## Number of cases:  13583
```

Remember that you can answer this question by viewing the data in the data viewer or by using the following command:

```
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"~
```

Exploratory data analysis

You will first start with analyzing the weight of the participants in kilograms: `weight`.

Using visualization and summary statistics, describe the distribution of weights. The `summary` function can be useful.

```
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
```

2. How many observations are we missing weights from?

Answer: 1,004

Next, consider the possible relationship between a high schooler's weight and their physical activity. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

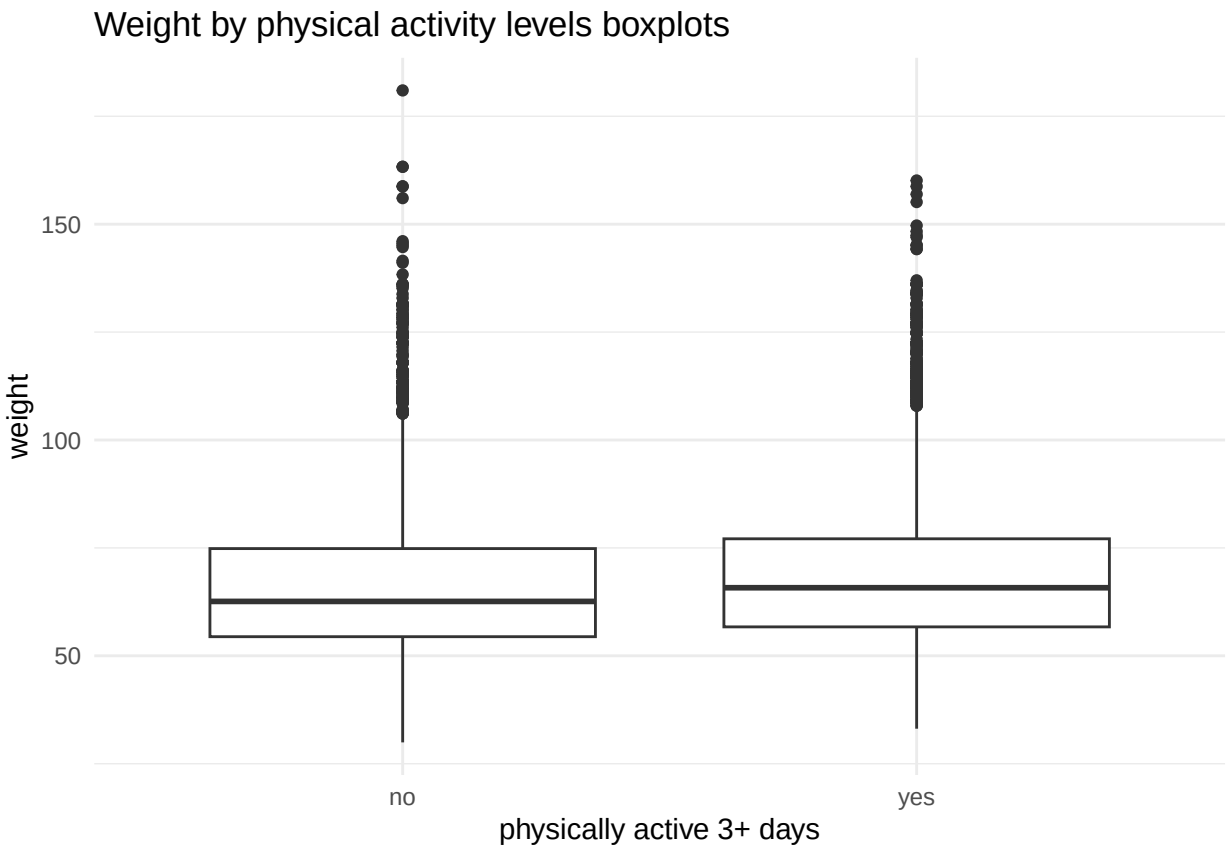
First, let's create a new variable `physical_3plus`, which will be coded as either "yes" if they are physically active for at least 3 days a week, and "no" if not.

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

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?

Answer: This isn't a meaningful relationship between these two variables since the box plots, save small differences in the median and the ends of the tails, are almost the same. I wasn't expecting these two variables to show a meaningful relationship because it's premised that the only positive outcome from being physically active is weight loss. If you lift weights five days a week, your weight is going to increase.

```
ggplot(yrbss, aes(x = physical_3plus, y = weight)) +
  geom_boxplot() +
  labs(x = "physically active 3+ days", y = "weight", title = "Weight by physical activity levels boxpl
  theme_minimal()
```



The box plots show how the medians of the two distributions compare, but we can also compare the means of the distributions using the following to first group the data by the `physical_3plus` variable, and then calculate the mean `weight` in these groups using the `mean` function while ignoring missing values by setting

the `na.rm` argument to `TRUE`.

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

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

There is an observed difference, but is this difference statistically significant? In order to answer this question we will conduct a hypothesis test.

Inference

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()`.

Answer: To aide reviewing the conditions, I can calculate group sizes, average weights, and standard deviation for the two groups.

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

```
## # A tibble: 2 x 4
##   physical_3plus mean_weight group_size sd_weight
##   <chr>                <dbl>      <int>     <dbl>
## 1 no                    66.7       4022      17.6
## 2 yes                   68.4       8342      16.5
```

Condition: the data should come from a random sample or a randomized experiment. Determination: the Youth Risk Behavior Surveillance System is a CDC dataset and is a proper scientific survey with random sampling.

- Condition: the observations should be independent within each group, and each group should be independent from the other.
  - Determination: independence between those those who engage in physical activity three or more times a week and those who don't can be assumed if the groupings are based on distinct individuals. There's nothing to suggest the dataset has overlapping individuals.
- Condition: The distribution of the response variable, weight, should be approximately normal within each group, or the sample sizes should be large enough (usually above 30) to apply the Central Limit Theorem. Determination: each group is well over 30, with 4,022 nos and 8,342 yeses.
- Condition: The variances (or standard deviations) of the two groups should be approximately equal.
  - Determination: the SD for the nos is 17.6 and the yeses is 16.5. These are relatively close and the assumption of equal variances is satisfied.

Conclusion: All necessary conditions for inference are satisfied. The sample sizes are large, standard deviations are similar, and the data comes from a well-designed survey.

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.

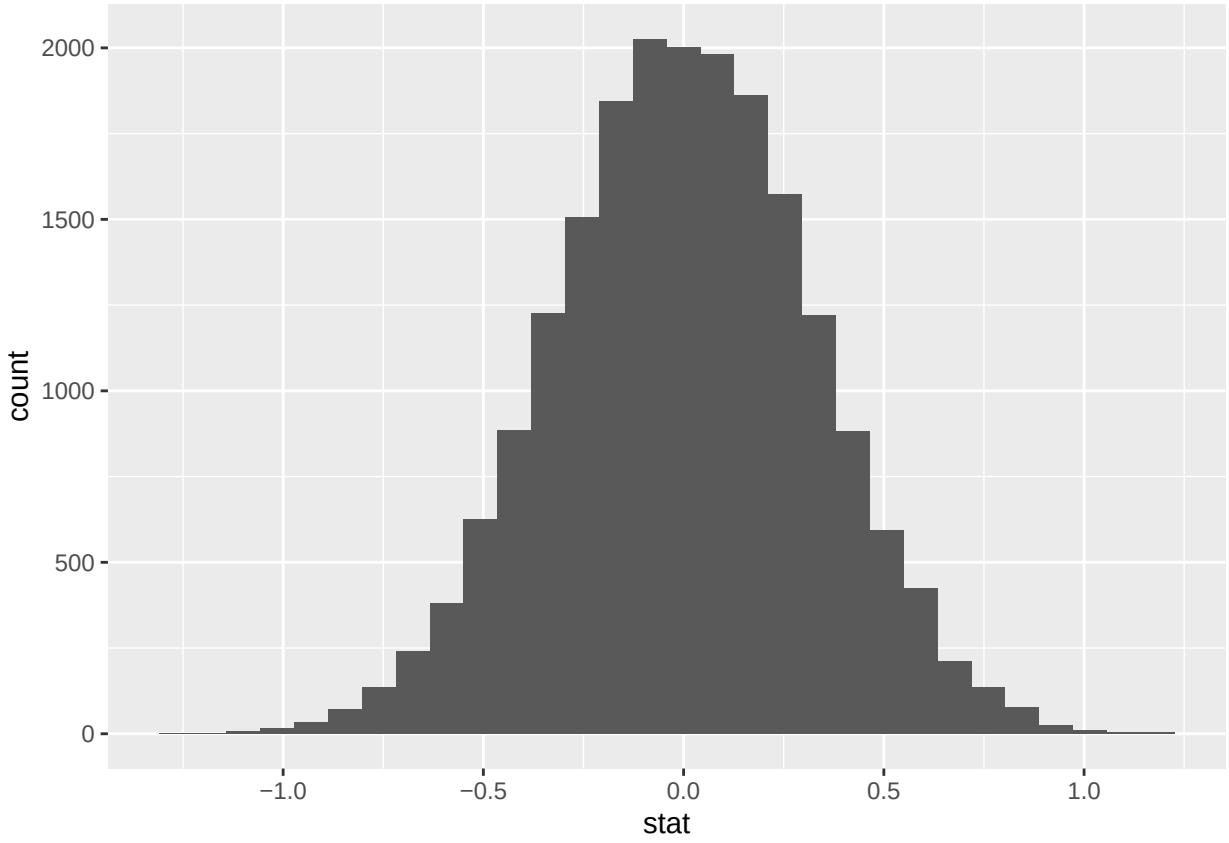
Answer: The p-value is reported as zero. I increased reps from 1,000 to 10,000 to 20,000 and still got the result so I'm accepting it as the true read out. This means that the true p-value is, per the infer documentation, "less than 3/reps(based on a poisson approximation)."

```
act_e3d <- yrbss %>%
  drop_na(physical_3plus, weight) %>%
  specify(response = weight, explanatory = physical_3plus)

act_obs <- yrbss %>%
  specify(weight ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

act_null_dist <- act_e3d %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 20000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

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




```
act_pv <- get_p_value(act_null_dist, act_obs, direction = "two-sided")

act_ci <- act_e3d %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("yes", "no")) %>%
  get_ci(level = 0.95, type = "percentile")
```

```r
cat("The P value is:\n")
```

```
## The P value is:
```

```r
act_pv
```

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

```r
cat("\nThe confidence interval is:\n")
```

```
##
## The confidence interval is:
```

```r
act_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     1.09     2.40
```

```r
cat("\nThe margin of error is:\n")
```

```
##
## The margin of error is:
```

```r
round((act_ci$upper_ci - act_ci$lower_ci)/2,6)
```

```
## [1] 0.652312
```

Next, we will introduce a new function, `hypothesize`, that falls into the `infer` workflow. You will use this method for conducting hypothesis tests.

But first, we need to initialize the test, which we will save as `obs_diff`.

```r
obs_diff <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

Notice how you can use the functions `specify` and `calculate` again like you did for calculating confidence intervals. Here, though, the statistic you are searching for is the difference in means, with the order being `yes - no != 0`.

After you have initialized the test, you need to simulate the test on the null distribution, which we will save as `null`.

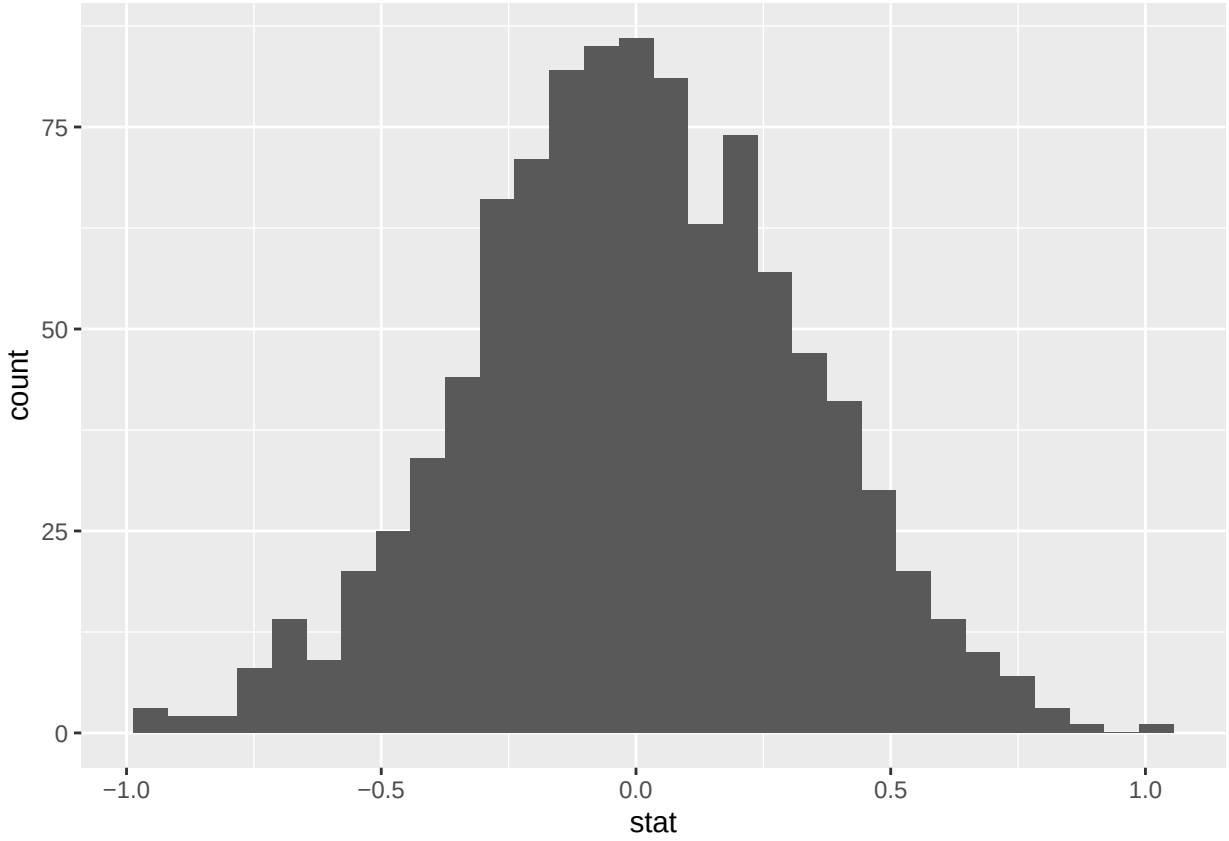
```r
null_dist <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

Here, `hypothesize` is used to set the null hypothesis as a test for independence. In one sample cases, the `null` argument can be set to "point" to test a hypothesis relative to a point estimate.

Also, note that the `type` argument within `generate` is set to `permute`, whichis the argument when generating a null distribution for a hypothesis test.

We can visualize this null distribution with the following code:

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



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

Answer:

```
obs_count <- null_dist %>%
  filter(abs(stat) >= abs(obs_diff)) %>%
  summarise(count = n())

cat("\nThe number of null permutations with a difference of at least the observed stat is:\n")
```

```
##
## The number of null permutations with a difference of at least the observed stat is:
```

```
obs_count
```

```
## # A tibble: 1 x 1
##   count
##   <int>
## 1     0
```

Now that the test is initialized and the null distribution formed, you can calculate the p-value for your hypothesis test using the function `get_p_value`.

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

```r
cat("The P value is:\n")
```

```
## The P value is:
```

```r
act_pv_rev
```

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

This the standard workflow for performing hypothesis tests.

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.

Answer:

The p-value is zero. This means that, in the 1,000 permutations, none had a difference in mean as large as the observed difference. This is strong evidence against the null hypothesis because the observed difference is large enough that, under the null hypothesis assumption of no difference is very unlikely. Additionally, the confidence interval of 1.15 to 2.43 with a margin of error of 0.6368 is relatively small in the context of the data set, where weight values can stretch up towards 180. This adds additional strength to the case that there's statistically significant difference in the average weights between the groups.

```r
act_ci_rev <- act_e3d %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("yes", "no")) %>%
  get_ci(level = 0.95, type = "percentile")

cat("\nThe confidence interval is:\n")
```

```
##
## The confidence interval is:
```

```r
act_ci_rev
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     1.11     2.41
```

```r
cat("\nThe margin of error is:\n")
```

```
##
## The margin of error is:
```

```r
round((act_ci_rev$upper_ci - act_ci_rev$lower_ci)/2, 6)
```

```
## [1] 0.650199
```

More Practice

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

Answer: When I did it the first time and just saw 1.69 for upper and lower, I thought something had gone wrong. I ran the test again, increasing the reps from 1,000 to 5,000 but still got the same 1.69 for upper and lower. I looked at the results in a new tab instead of just printed out in the console and saw that the lower CI was being rounded up from 1.689 and the upper CI was being rounded down from 1.692. I also reviewed

the summay stats afterwards and saw that the mean and median were both right near 1.69, validating the results I got.

Overall, this data set has an intense enough clustered right around 1.69 that the other values aren't moving the needle much. Ad you can see, the minimum value is 1.27 and the maximum value is 2.11, showing that there are unique values covering a wide spread.

```
height_ci <- yrbss %>%
  drop_na(height) %>%
  specify(response = height) %>%
  generate(reps = 5000, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
  get_ci(level = 0.95, type = "percentile")

cat("The lower confidence interval is:\n")
```

```
## The lower confidence interval is:
```

```
round(height_ci$lower_ci, 6)
```

```
## [1] 1.689151
```

```
cat("\nThe upper confidence interval is\n")
```

```
##
## The upper confidence interval is
```

```
round(height_ci$upper_ci, 6)
```

```
## [1] 1.692862
```

```
cat("The summary stats for height data in YRBSS are:\n")
```

```
## The summary stats for height data in YRBSS are:
```

```
summary(yrbss$height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.270   1.600   1.680   1.691   1.780   2.110
```

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.

Answer:

```
height_ci_90 <- yrbss %>%
  drop_na(height) %>%
  specify(response = height) %>%
  generate(reps = 5000, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
  get_ci(level = 0.90, type = "percentile")

cat("The lower confidence interval is:\n")
```

```
## The lower confidence interval is:
```

```
round(height_ci_90$lower_ci, 6)
```

```
## [1] 1.689425
```

```
cat("The difference between 95% and 90% confidence lower CI is:\n")
```

```
## The difference between 95% and 90% confidence lower CI is:
```

```
 height_ci$lower_ci - height_ci_90$lower_ci
```

```
## [1] -0.0002742438
```

```
cat("The difference between 95% and 90% confidence upper CI is:\n")
```

```
## The difference between 95% and 90% confidence upper CI is:
```

```
 height_ci$upper_ci - height_ci_90$upper_ci
```

```
## [1] 0.0003324167
```

```
 cat("The spread changed from ", height_ci$upper_ci - height_ci$lower_ci," for 95% confidence to ",
    height_ci_90$upper_ci - height_ci_90$lower_ci, " for 90% confidence. ", "a difference of around 6 t
```

```
## The spread changed from  0.003711724  for 95% confidence to  0.003105063  for 90% confidence.  a dif
```

Depending on your approach to rounding numbers, there's practically no difference between those numbers. The lower CI would still round up to 1.69 and the upper would still round down to 1.69. In the context of data, this is pretty unremarkable. It leads to a different question though: why is this happening? I would like to understand the approach to determining who to sample, other than targeting a specific age group. Additionally, it would be worth comparing this against the average heights for this group in the whole population to see if this is in line or a departure.

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.

Answer:

```
dh3_obs <- yrbss %>%
  drop_na(physical_3plus, height) %>%
  specify(height ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

dh3null_dist <- yrbss %>%
  drop_na(physical_3plus, height) %>%
  specify(height ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

dh3_pv <- dh3null_dist %>%
  get_p_value(obs_stat = dh3_obs, direction = "two_sided")

dh3_ci <- yrbss %>%
  drop_na(physical_3plus, height) %>%
  specify(height ~ physical_3plus) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("yes", "no")) %>%
  get_ci(level = 0.95, type = "percentile")

cat("The P value is:\n")
```

```
## The P value is:
```

```
dh3_pv
```

```
## # A tibble: 1 x 1
```

```
##   p_value
##     <dbl>
## 1       0
```

```
cat("\nThe confidence interval is:\n")
```

```
##
## The confidence interval is:
```

```
dh3_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   0.0338   0.0414
```

```
cat("\nThe margin of error is:\n")
```

```
##
## The margin of error is:
```

```
round((dh3_ci$upper_ci - dh3_ci$lower_ci)/2,6)
```

```
## [1] 0.003797
```

The p-value of 0 points to a statistically significant difference in average height between those who exercise at least three times a week and those who don't. The confidence interval of 0.0337 to 0.0412 with a margin of error of 0.003758 is strong evidence that the actual difference is both positive and quite small, with minimal uncertainty involved. These results suggest a small but statistically significant difference in average height between the two groups.

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.

Answer: There are 7 unique non-NA values. 8 if you include NA.

```
n_distinct(yrbss$hours_tv_per_school_day, na.rm = TRUE)
```

```
## [1] 7
```

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 α level, and conclude in context.

Answer: Question: does the average weight of individuals differ based on how many hours they sleep on school nights?

```
ws_obs <- yrbss %>%
  drop_na(weight, school_night_hours_sleep) %>%
  specify(weight ~ school_night_hours_sleep) %>%
  calculate(stat = "F")

ws_null <- yrbss %>%
  drop_na(weight, school_night_hours_sleep) %>%
  specify(weight ~ school_night_hours_sleep) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "F")

ws_pv <- ws_null %>%
```

```
  get_p_value(obs_stat = ws_obs, direction = "greater")

ws_ci <- yrbss %>%
  drop_na(weight, school_night_hours_sleep) %>%
  specify(weight ~ school_night_hours_sleep) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "F") %>%
  get_ci(level = 0.95, type = "percentile")

cat("The P value is:\n")
```

```
## The P value is:
```

```
ws_pv
```

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

```
cat("\nThe confidence interval is:\n")
```

```
##
## The confidence interval is:
```

```
ws_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     3.83     12.6
```

```
cat("\nThe margin of error is:\n")
```

```
##
## The margin of error is:
```

```
round((ws_ci$upper_ci - ws_ci$lower_ci)/2,6)
```

```
## [1] 4.389319
```

The confidence interval for the F-statistic of 3.92 to 12.5 is not insignificant, even in the context of a weight value that can go up to 180+ KG. This uncertainty is offset a bit by the 0 p-value, which supports statistically significant differences in weight based on hours of sleep. The two together mean that there's a lot of variability but that the data is still statistically significant. Overall, I would say there's a decent relationship but that I would want to investigate more and focus on population subsets, such as male versus female, to bring more nuance.
