Inference for numerical data

Kayleah Griffen

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)
```

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?

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"~
```

There are 13583 cases in this data set (sample). There are 13 observations taken on each high schooler who participated in this study, one row is one case.

Exploratory data analysis

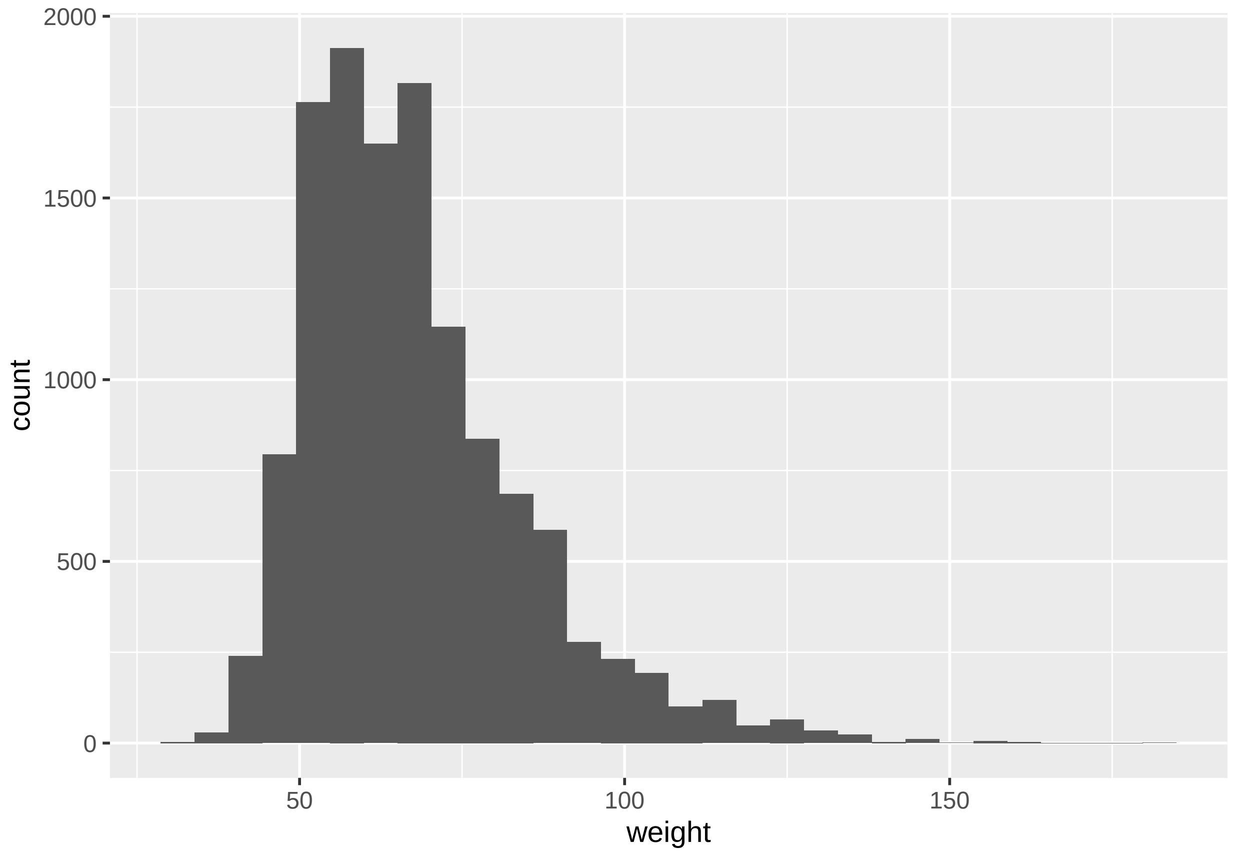
You will first start with analyzing the weight of the participants in kilograms: `weight`.

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
```

```
# To visualize the weights I will make a histogram
yrbss |>  ggplot(aes(x=weight)) + geom_histogram()
```



Looking at the histogram, the weight is normally distributed with a right skew to it.

2. How many observations are we missing weights from?

```
sum(is.na(yrbss$weight))
```

```
## [1] 1004
```

Summing all of the NA values for weight, I get that there are 1004 values for weight that are NA (or missing).

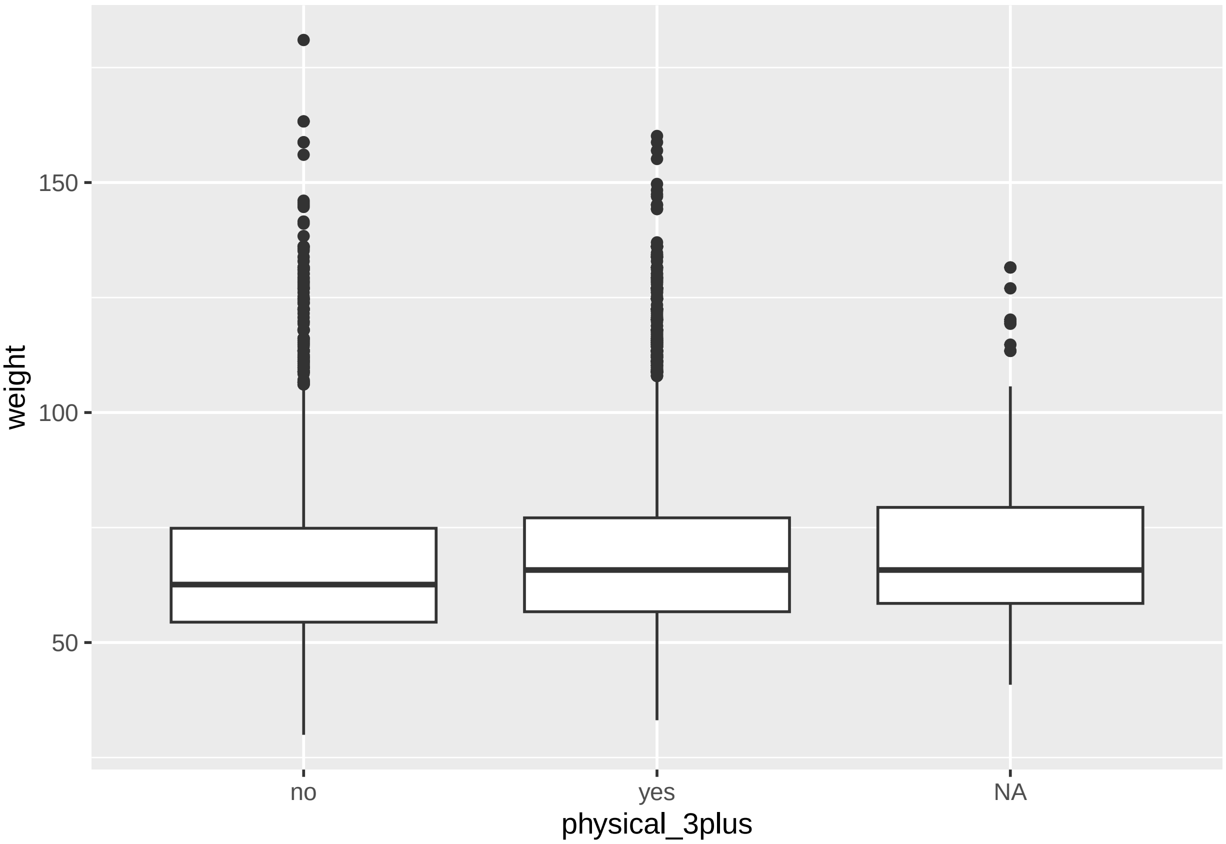
Next, consider the possible relationship between a high schooler's weight and their physical activity. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

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

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 |> ggplot(aes(physical_3plus, weight)) +
  geom_boxplot()
```



I was expecting that those who are physical at least 3 days a week would be on average lighter than those who are not - however looking at the box plot it appears that those who are physical at least 3 days a week are heavier than the less active high schoolers.

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

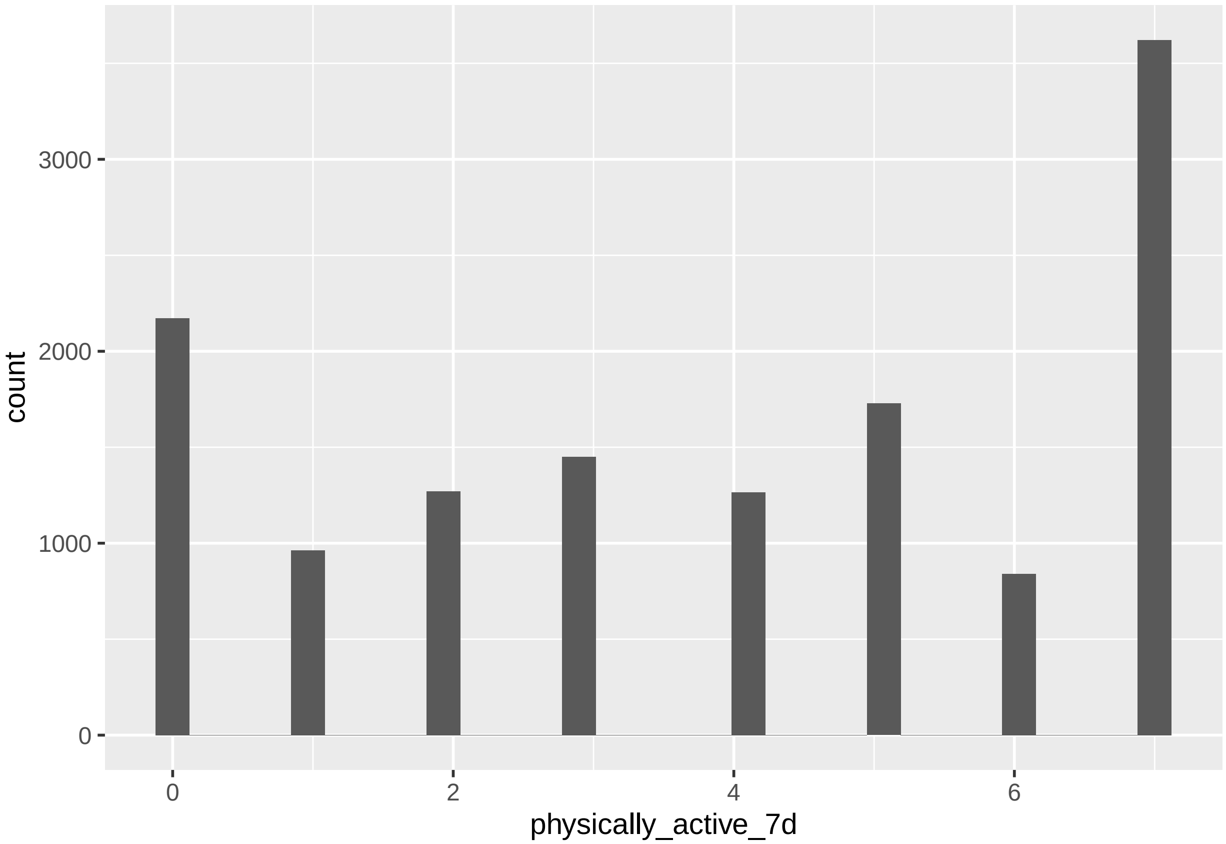
There is an observed difference, but is this difference statistically significant? In order to answer this question we will conduct a hypothesis test.

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()`.

According to the OS textbook, two conditions are required to apply the central limit theorem for a sample mean. The samples must be independent. In this case the sample is a random sample and its not greater than 10% of the population so it is independent. Next it needs to be normal. Taking a look at the weights the weight appeared normally distributed (with the right skew). I can take a look at physical activity to see if it is normally distributed.

```
yrbss |>  ggplot(aes(x=physically_active_7d)) + geom_histogram()
```



The physical activity doesn't look normally distributed. However according to the textbook you can perform a normality check if n >= 30 and there are no particulary extreme outliers it can be considered normal.

```
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(count = n())
```

```
## # A tibble: 3 x 2
##   physical_3plus count
##   <chr>          <int>
## 1 no              4404
## 2 yes             8906
## 3 <NA>             273
```

Counting the physical activity - there are 4404 nos, 8906 yess and 273 NAs. Therefor, the conditions for inference are satisfied for this case.

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.

For the hypothesis test, the null hypothesis is that nothing is going on - therefor in this case the null hypothesis is that the average weights are NOT different for those who exercise at least 3 times a week and those who don't. The alternative hypothesis is that there IS a difference in the average weight for those who exercise at least 3 times a week and those who don't.

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(weight,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`.

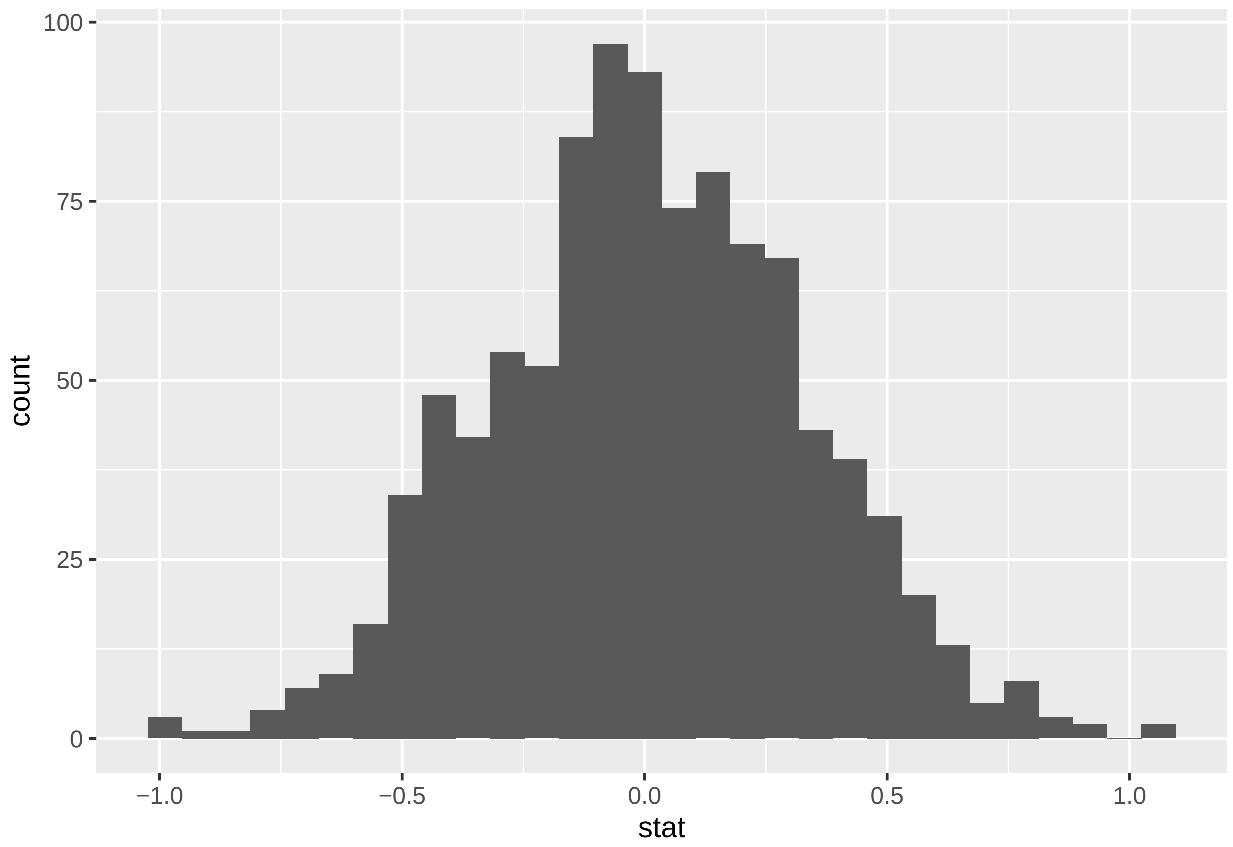
```r
null_dist <- yrbss %>%
  drop_na(weight,physical_3plus) |>
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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`, which is the argument when generating a null distribution for a hypothesis test.

We can visualize this null distribution with the following code:

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



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

If the Obs_stat is the `obs_diff` stat component that would be 1.77. The amount of `null` permutations that have a difference of at least `obs_stat` can be calculated.

```
null_dist |> filter(stat > obs_diff$stat) |>
  summarise(count = n())
```

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

There are 0 null permutations that have a difference of at least obs_stat.

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`.

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

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

The reported p value is 0 - but there is a warning message saying to be cautious in reporting that as this is an approximation based on the number of reps chosen. The p value likely is small and is rounding to 0. To reject the null hypothesis a p value of less than p = 0.05 is needed. Therefor in this case the null hypothesis will be rejected meaning that there IS evidence there IS a difference in the average weight for those who exercise at least 3 times a week and those who don't.

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.

In theory, the hypothesis test and the confidence interval should point to the same conclusion. So in this case I expect that the confidence interval will indicate a difference in average weight between those who exercise and those who do not.

```
yrbss |>
  drop_na(weight,physical_3plus) |>
  specify(response = weight, explanatory = physical_3plus) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "diff in means", order = c("yes", "no")) |>
  get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     1.10     2.46
```

In my hypothesis test my lower CI is 1.19 and upper CI is 2.43 - because 0 is not included in the CI - there is evidence to reject my null hypothesis. I would say that there IS evidence that the average weight is different for those who work out at least 3 days a week and those who don't.

More Practice

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

```
ycis <- yrbss |>
  drop_na(height) |>
  specify(response = height) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "mean") |>
  get_ci(level = 0.95)
ycis
```

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

```
ycis$upper_ci- ycis$lower_ci
```

```
## [1] 0.003408001
```

The rounded lower CI and upper CI are both 1.69 meters. These reported values are the same but there actually is a difference, 0.003624016, between the upper and lower CI.

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.

```
ycis <- yrbss |>
  drop_na(height) |>
  specify(response = height) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "mean") |>
  get_ci(level = 0.90)
ycis
```

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

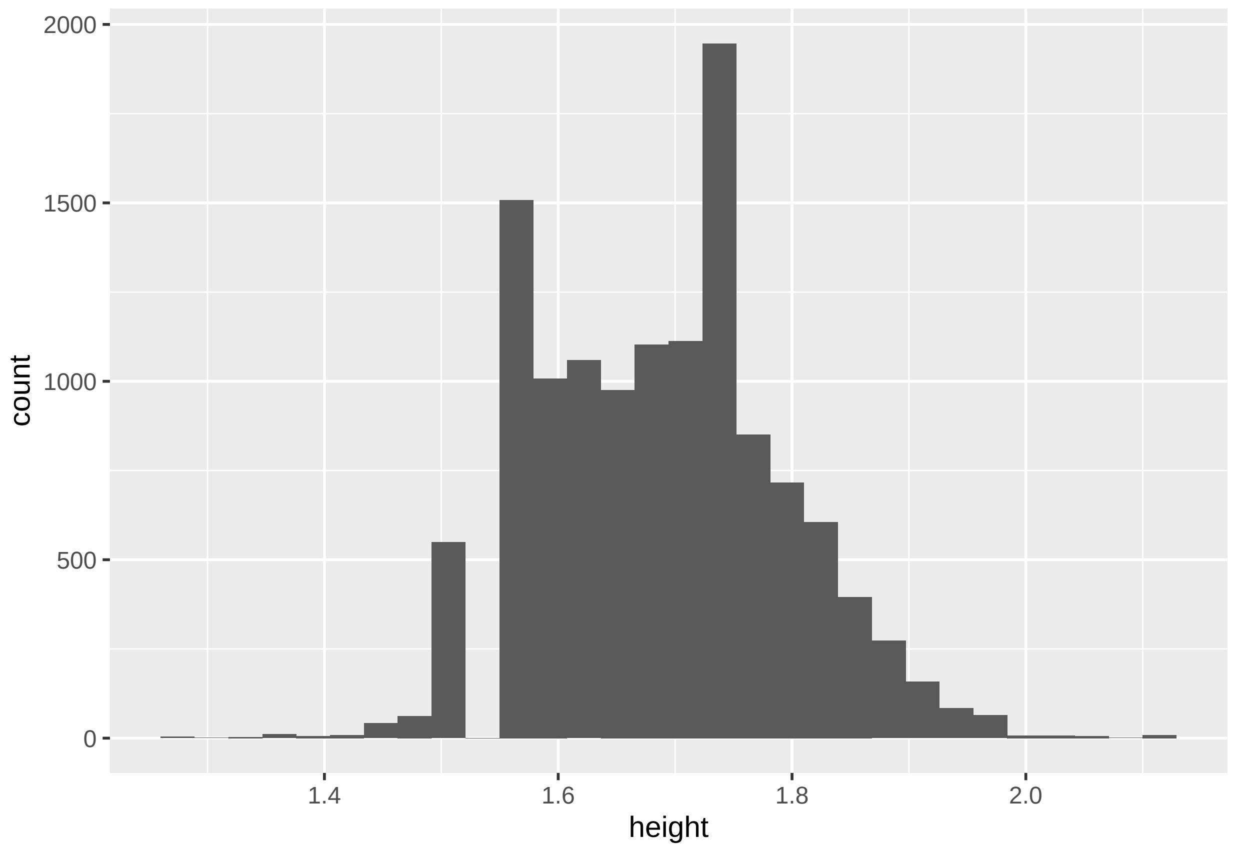
```
ycis$upper_ci- ycis$lower_ci
```

```
## [1] 0.003038238
```

For the 90% confidence interval I got the same reported values as the 95%, the lower CI of 1.69 and upper CI of 1.69. Even though the reported values are the same as before - there is a difference like before, but this time 0.003122824. In theory the 90% confidence interval should be narrower than the 95% confidence interval and this holds true (the difference in CIs for the 90% confidence interval is less than the 95%).

I'd like to take a look at the height histogram to see if it is normally distributed.

```
yrbss |>  ggplot(aes(x=height)) + geom_histogram()
```



The histogram does not appear normally distributed - this would affect the accuracy of the confidence intervals as the confidence intervals apply to normally distributed data.

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.

Before conducting my test - I need to confirm independence (check) and normality. I already checked before that there is more than 30 in the physical_3plus group. Looking at the height histogram above - as mentioned previously - it does not appear normally distributed. I will go ahead and conduct the hypothesis test anyways - but I'd like to note the caveat that for high schoolers the height doesn't look normal.

My null hypothesis is that there IS NOT a difference in average height for those who exercise at least 3 times a week and those who don't. I can conduct this test using the method we learned earlier.

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

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

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

```
## [1] 0
```

I got 0 for my p-value. This p-value would lead me to REJECT my null hypothesis. However, because I noted before that the height doesn't appear normally distrubuted - I would say that overall this test is invalid and inconclusive.

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.

```
yrbss |> group_by(hours_tv_per_school_day) |>
  summarise(count = n())
```

```
## # A tibble: 8 x 2
##   hours_tv_per_school_day count
##   <chr>                   <int>
## 1 1                        1750
## 2 2                        2705
## 3 3                        2139
## 4 4                        1048
## 5 5+                       1595
## 6 <1                       2168
## 7 do not watch             1840
## 8 <NA>                      338
```

There are 7 different options for hours_tv_per_school_day (8 if you count NA). The options are <1,1,2,3,4,5+, and do not watch.

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.

Check out the survey options for sleep:

```
yrbss |> group_by(school_night_hours_sleep) |> summarise(count = n())
```

```
## # A tibble: 8 x 2
##   school_night_hours_sleep count
##   <chr>                    <int>
## 1 10+                        316
## 2 5                         1480
## 3 6                         2658
## 4 7                         3461
## 5 8                         2692
## 6 9                          763
## 7 <5                         965
## 8 <NA>                      1248
```

My alpha level is 0.05. According to the American Academy of Sleep Medicine teenagers aged 13-18 should sleep 8-10 hours per 24 hours. My research question is: is there a difference in the average weight for those who sleep at least 8 hours a night and those who do not.

To carry this out - I need to create a new variable.

```
yrbss <- yrbss %>%
  mutate(sleep_8plus = ifelse((yrbss$school_night_hours_sleep == "8" |yrbss$school_night_hours_sleep ==
```

Check how many are in each group.

```
yrbss |> group_by(sleep_8plus) |> summarise(count = n())
```

```
## # A tibble: 3 x 2
##   sleep_8plus count
##   <chr>       <int>
## 1 no           8564
## 2 yes          3771
## 3 <NA>         1248
```

I previously checked that weight was normally distributed, and I can note that there are at least 30 in each group. Also the observations are independent. Now that the conditions for inference are met - I can conduct my analysis.

First I will do my analysis as a hypothesis test and then I will do it as a confidence interval - both should agree.

```
obs_diff <- yrbss %>%
  drop_na(weight, sleep_8plus) |>
  specify(weight ~ sleep_8plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

null_dist <- yrbss %>%
  drop_na(weight, sleep_8plus) |>
  specify(weight ~ sleep_8plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

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

```
## [1] 0.006
```

The p value reported is 0.004. My alpha is 0.05. Because the p values is less than 0.05 I will reject my null hypothesis. Therefor there is evidence that there IS a difference in average weight for those who sleep 8 or more hours a night and those who don't.

Now I'd like to find my confidence intervals.

```r
yrbss |>
  drop_na(weight,sleep_8plus) |>
  specify(response = weight, explanatory = sleep_8plus) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "diff in means", order = c("yes", "no")) |>
  get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    -1.64   -0.367
```

The lower CI is -1.64 and upper CI is -0.35. Because 0 is not contained in the CI I can say that there IS evidence that there IS a difference in average weight for those who sleep 8 or more hours a night and those who don't. This is in agreement with my hypothesis test.
