Inference for numerical data

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(DATA606)
```

```
##
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 4th Edition. You can read this by typing
## vignette('os4') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='DATA606') will list the demos that are available.
```

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:

There are 13,583 observations with 13 variables containing seen below.

```
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"~
```

```
# cases in the data set by column
colnames(yrbss)
```

```
##  [1] "age"                     "gender"
##  [3] "grade"                   "hispanic"
##  [5] "race"                    "height"
##  [7] "weight"                  "helmet_12m"
##  [9] "text_while_driving_30d"  "physically_active_7d"
## [11] "hours_tv_per_school_day" "strength_training_7d"
## [13] "school_night_hours_sleep"
```

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?

There are 1,004 observations missing from weights.

```
# missing observations from weights
sum(is.na(yrbss$weight))
```

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

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"))
```

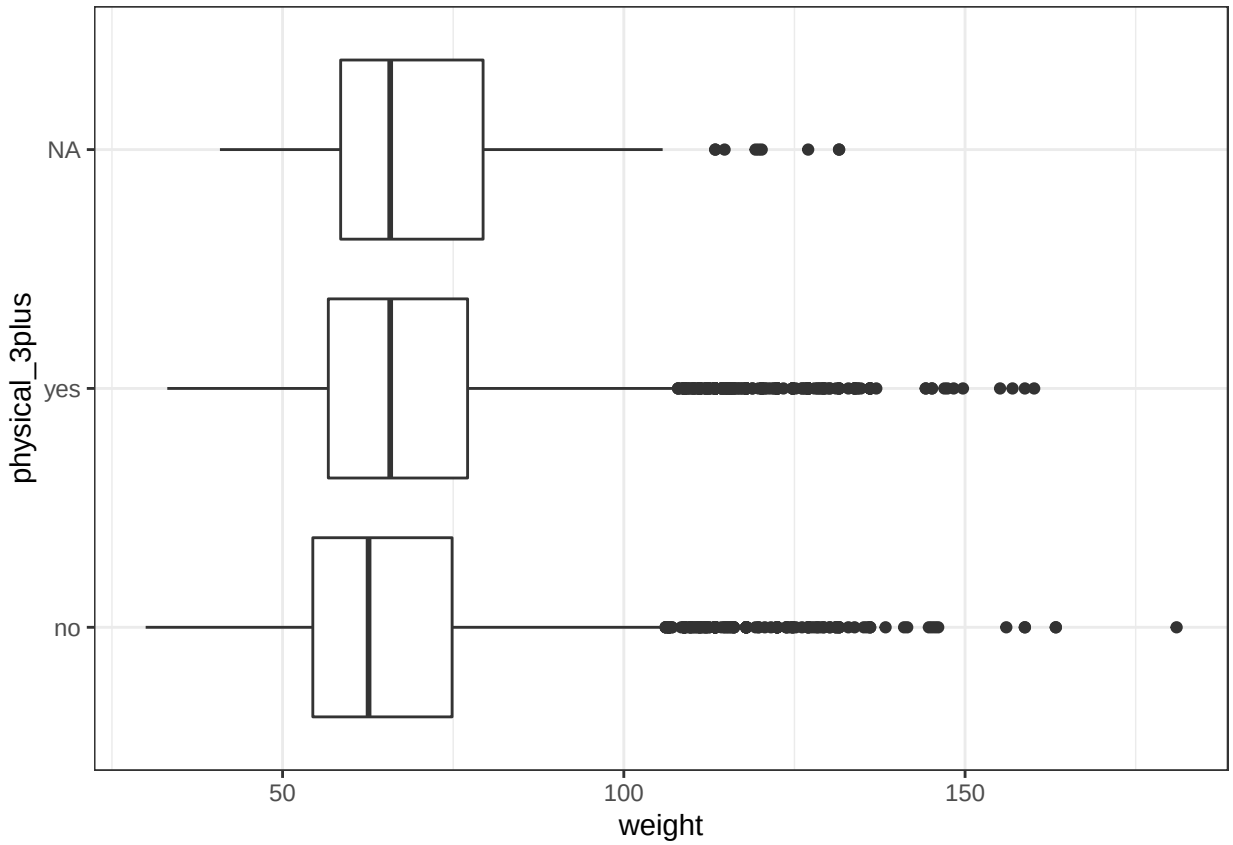
3. Make a side-by-side boxplot of `physical_3plus` and `weight`. Is there a relationship between these two variables? What did you expect and why?

Based on the box plots below, those students who are more active for at least 3 days a week weight more than those students who do not. These results actually don't surprise me because those students who are physically active can have more muscle mass thus resulting in weighing more than those students who do not.

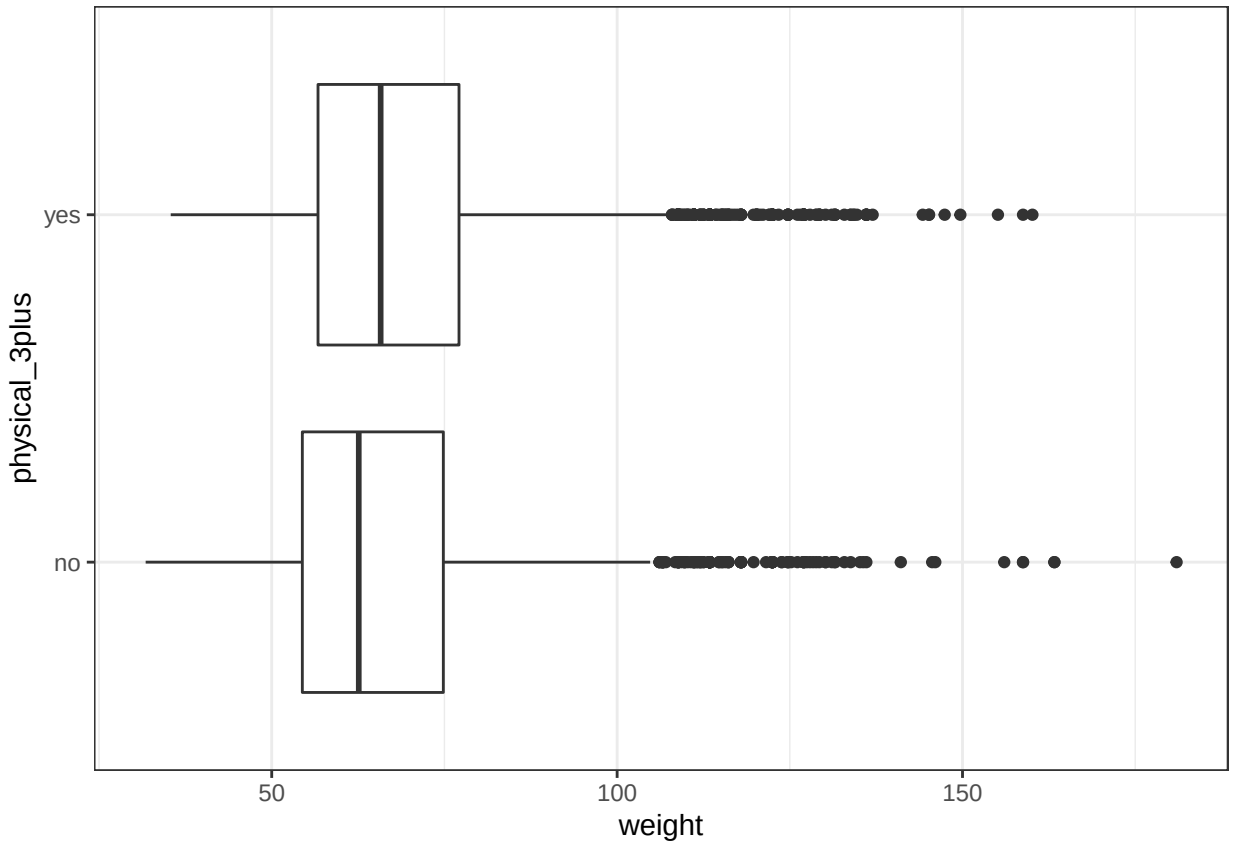
```
# missing values in physical_3plus
sum(is.na(yrbss$physical_3plus))
```

```
## [1] 273
```

```
# side-by-side box plot with missing values
ggplot(yrbss, aes(x=weight, y=physical_3plus)) + geom_boxplot() + theme_bw()
```




```
# side-by-side box plot without missing values
yrbss2 <- yrbss %>%
  mutate(physical_3plus = ifelse(yrbss$physically_active_7d > 2, "yes", "no")) %>%
  na.exclude()
ggplot(yrbss2, aes(x=weight, y=physical_3plus)) + geom_boxplot() + theme_bw()
```



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()`.

The conditions for inference are independence and normality. Based on the data below, we see that it is a representative sample of students across national, state, tribal and local school

systems. The students are independent and the sample size and distributions appear to be normal. With a large enough sample size we can assume that all conditions for inference are satisfied.

```
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(freq = table(weight)) %>%
  summarise(n = sum(freq))
```

```
## # A tibble: 3 x 2
##   physical_3plus     n
##   <chr>          <int>
## 1 no              4022
## 2 yes             8342
## 3 <NA>             215
```

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.

Ho: Students who are physically active 3 or more days per week have the same average weight as those who are not physically active.

Ha: Students who are physically active 3 or more days per week have different average weight compared to those who are not physically active.

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`.

```
obs_diff <- yrbss %>%
  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`.

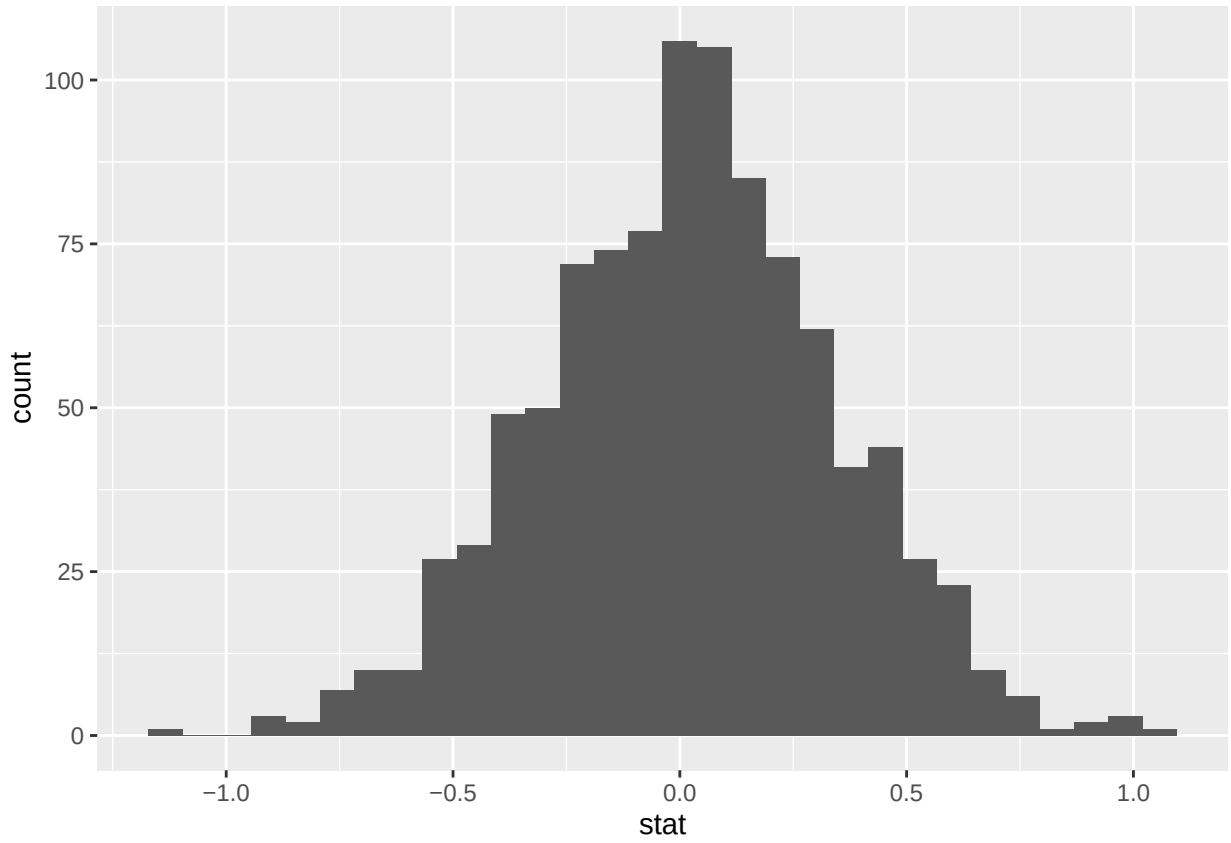
```
null_dist <- yrbss %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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:

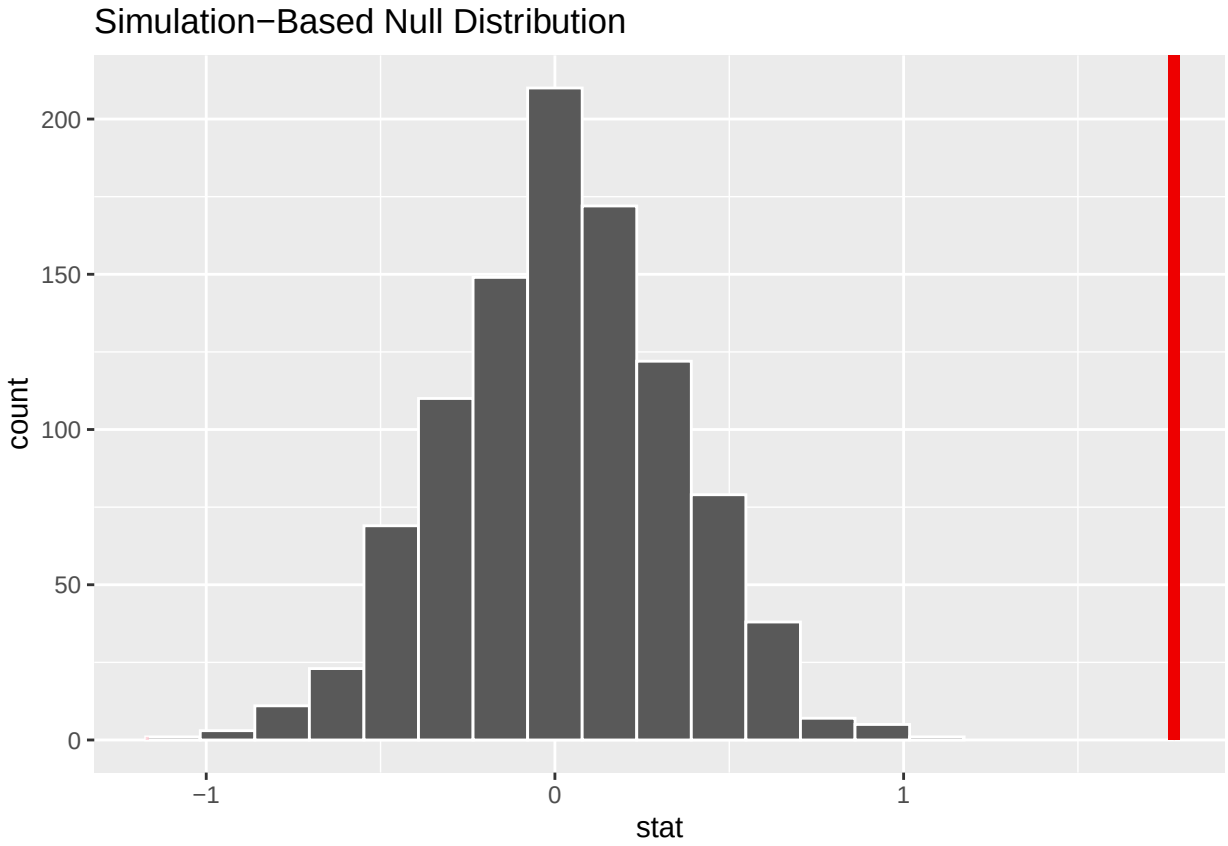
```
ggplot(data = null_dist, aes(x = stat)) +
  geom_histogram()
```



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

With the red line being our indicator of the obs_stat it does appear to be far from the data.

```
visualize(null_dist) +
  shade_p_value(obs_stat = obs_diff, direction = "two_sided")
```



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
```

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.

at a 95% confidence interval, those students who are active at least three times a week have an average weight between 68.05 kg and 68.75 kg. Those students who are not active at least three times a week have an average weight between 66.16 kg and 67.24 kg.

```
#Standard deviation
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(sd_weight = sd(weight, na.rm = TRUE))
```

```
## # A tibble: 3 x 2
##   physical_3plus sd_weight
##   <chr>              <dbl>
## 1 no                  17.6
## 2 yes                 16.5
## 3 <NA>                17.6
```

```
#Mean
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
```

```
#Sample size N
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(freq = table(weight)) %>%
  summarise(n = sum(freq))
```

```
## # A tibble: 3 x 2
##   physical_3plus     n
##   <chr>          <int>
## 1 no              4022
## 2 yes             8342
## 3 <NA>             215
```

```
# not Active
not_active_mean <- 66.7
not_active_sd <- 17.6
not_active_n <- 4022

# active
active_mean <- 68.4
active_sd <- 16.5
active_n <- 8342

z <- 1.96

# confidence interval for not active
upper_not_active <- not_active_mean + z * (not_active_sd / sqrt(not_active_n))
upper_not_active
```

```
## [1] 67.24394
```

```r
lower_not_active <- not_active_mean - z * (not_active_sd / sqrt(not_active_n))
lower_not_active
```

```
## [1] 66.15606
```

```r
# confidence interval for active
upper_active <- active_mean + z * (active_sd / sqrt(active_n))
upper_active
```

```
## [1] 68.75408
```

```r
lower_active <- active_mean - z * (active_sd / sqrt(active_n))
lower_active
```

```
## [1] 68.04592
```

More Practice

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

At a 95% confidence interval, the average height in meters for the students is between 1.689411 m and 1.693071 m.

```r
height_table <- as.data.frame(table(yrbss$height))
height_freq <- sum(height_table$Freq)

# mean, standard deviation and sample size
height_mean <- mean(yrbss$height, na.rm = TRUE)
height_mean
```

```
## [1] 1.691241
```

```r
height_sd <- sd(yrbss$height, na.rm = TRUE)
height_sd
```

```
## [1] 0.1046973
```

```r
height_n <- yrbss %>%
  summarise(freq = table(height)) %>%
  summarise(n = sum(freq, na.rm = TRUE))
height_n
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 12579
```

```r
z_height <- 1.96

# confidence interval for height
upper_height<- height_mean + z_height * (height_sd / sqrt(height_n))
upper_height
```

```
##          n
## 1 1.693071
```

```r
lower_height <- height_mean - z_height * (height_sd / sqrt(height_n))
lower_height
```

```
##          n
## 1 1.689411
```

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.

At a 90% confidence interval, the average height in meters for the students is between 1.689701 m and 1.692781. Comparing both intervals at a 90% and 95% there is a slight difference where the range of the 95% confidence interval is slightly larger.

```r
# set z value to 1.65 for 90% confidence interval
z_90 <- 1.65

#confidence interval for height
upper_height_90 <- height_mean + z_90 * (height_sd / sqrt(height_n))
upper_height_90
```

```
##          n
## 1 1.692781
```

```r
lower_height_90 <- height_mean - z_90 * (height_sd / sqrt(height_n))
lower_height_90
```

```
##          n
## 1 1.689701
```

```r
# difference between both confidence intervals
range_95 <- (upper_height - lower_height)
range_95
```

```
##             n
## 1 0.003659302
```

```r
range_90 <- (upper_height_90 - lower_height_90)
range_90
```

```
##             n
## 1 0.003080535
```

```r
# difference between the two ranges
diff_range <- range_95 - range_90
diff_range
```

```
##              n
## 1 0.0005787672
```

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.

Ho: There is no difference in the average height of those who are physically active 3 or more days per week.

Ha: There is a difference in the average height of those who are physically active 3 or more days per week.

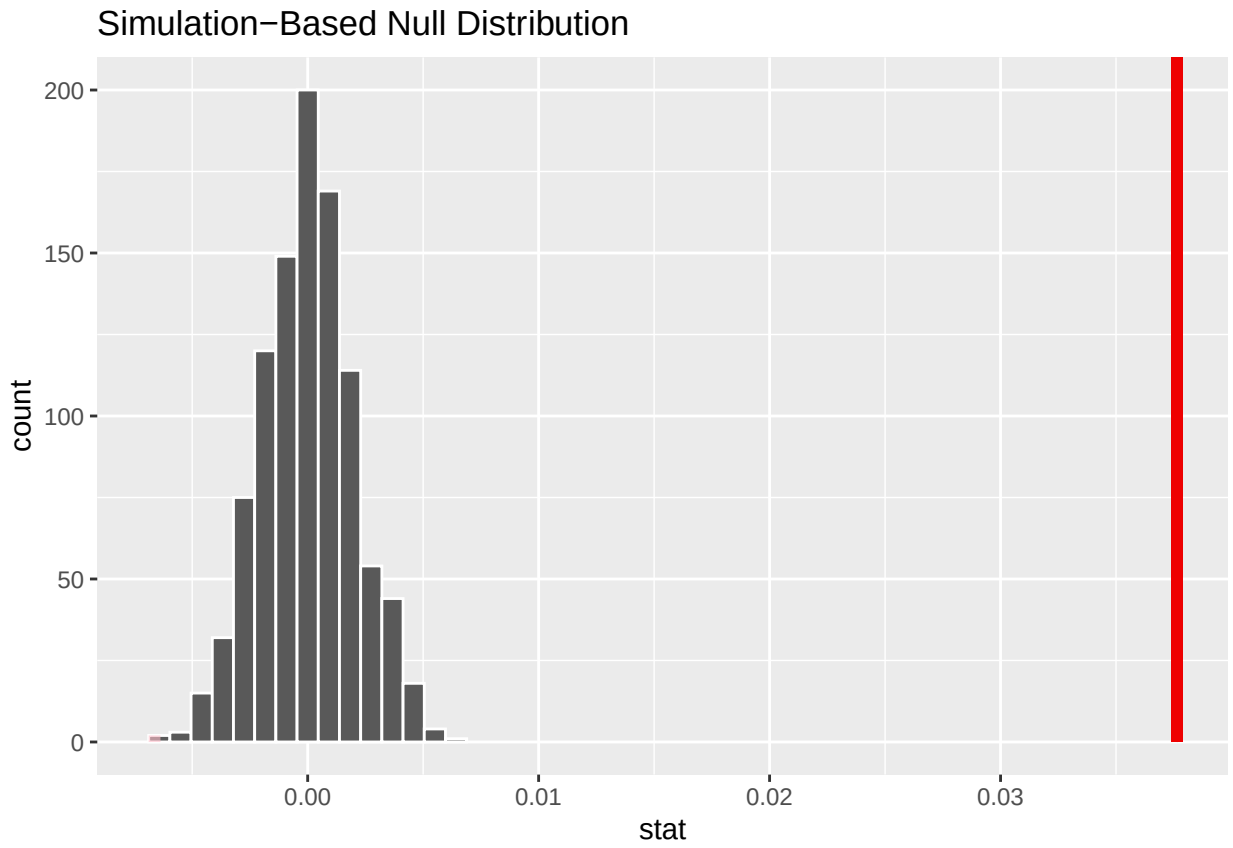
With a 95% confidence interval, the average heights of those students who are not physically active 3 or more days per week is between 1.66 m and 1.67 m. While for those students who are physically active is between 1.701 m and 1.705 m.

Since the p-values is below 0.05, we reject the null hypothesis. There is an a difference in the average height of the students who are physically active and those who are not.

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

set.seed(87)
null_dist_hgt <- yrbss %>%
  specify(height ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```r
visualize(null_dist_hgt) +
  shade_p_value(obs_stat = obs_diff_hgt, direction = "two_sided")
```




```
null_dist_hgt %>%
  get_p_value(obs_stat = obs_diff_hgt, direction = "two_sided")
```

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

```
# not Active
height_not_active_mean <- 1.665
height_not_active_sd <- 0.1029
height_not_active_n <- 4022

# active
height_active_mean <- 1.7032
height_active_sd <- 0.1033
height_active_n <- 8342

z_height <- 1.96

# confidence interval for not active
height_upper_not_active <- height_not_active_mean + z * (height_not_active_sd / sqrt(height_not_active_n
height_upper_not_active
```

```
## [1] 1.66818
```

```
height_lower_not_active <- height_not_active_mean - z * (height_not_active_sd / sqrt(height_not_active_n))
height_lower_not_active
```

```
## [1] 1.66182
```

```
# confidence interval for active
height_upper_active <- height_active_mean + z_height * (height_active_sd / sqrt(height_active_n))
height_upper_active
```

```
## [1] 1.705417
```

```
height_lower_active <- height_active_mean - z_height * (height_active_sd / sqrt(height_active_n))
height_lower_active
```

```
## [1] 1.700983
```

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

There are 7 different options for the data set `hours_tv_per_school_day`and 1 option for NA.

```
yrbss %>%
  group_by(hours_tv_per_school_day)%>%
  summarise(n())
```

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

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.

Question: Do student's who are shorter than the mean height sleep less than those students who are taller?

Ho: There is no relationship between the mean height and sleep of students.

Ha: There is a relationship between the mean height and sleep of students.

Confidence interval: 95%

Conditions: Independent sample: Yes, Normality: Yes

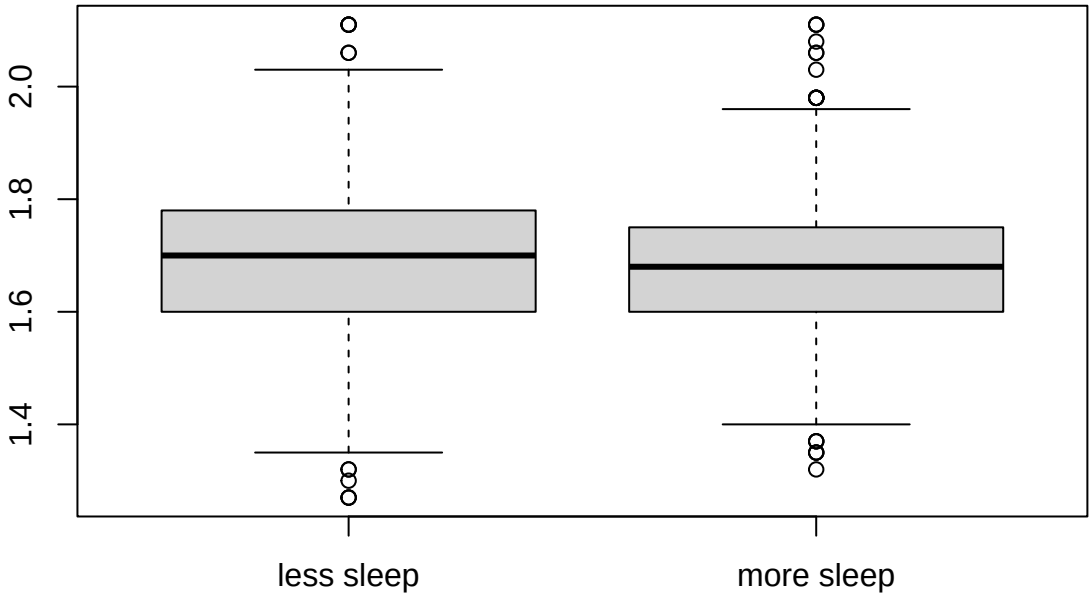
Based on the results, the p-value is 0.05 so we can reject the null hypothesis. There is a relationship between the mean height and sleep of students.

```
yrbss <- yrbss %>%
  mutate(sleep_less = ifelse(yrbss$school_night_hours_sleep < 6, "yes", "no"))

height_less <- yrbss %>%
  select(height, sleep_less) %>%
  filter(sleep_less == "no") %>%
  na.omit()

height_more <- yrbss %>%
  select(height, sleep_less) %>%
  filter(sleep_less == "yes") %>%
  na.omit()
```

```
boxplot(height_less$height, height_more$height,
        names = c("less sleep", "more sleep"))
```




```
# less sleep
less_sleep_mean <- mean(height_less$height)
less_sleep_mean
```

```
## [1] 1.692256
```

```
less_sleep_sd <- sd(height_less$height)
less_sleep_sd
```

```
## [1] 0.1042161
```

```
max <- max(height_less$height)
max
```

```
## [1] 2.11
```

```r
# more sleep
more_sleep_mean <- mean(height_more$height)
more_sleep_mean
```

```
## [1] 1.685185
```

```r
more_sleep_sd <- sd(height_more$height)
more_sleep_sd
```

```
## [1] 0.1059036
```

```r
max_2 <- max(height_more$height)
max_2
```

```
## [1] 2.11
```

```r
# difference
diff_mean <- more_sleep_mean - less_sleep_mean
diff_mean
```

```
## [1] -0.0070715
```

```r
diff_sd <- sqrt(((more_sleep_mean^2) / nrow(height_more)) + ((less_sleep_mean^2) / nrow(height_less)))
diff_sd
```

```
## [1] 0.03818596
```

```r
sleep_df <- 2492-1
t_sleep <- qt(.05/2, sleep_df, lower.tail = FALSE)

# confidence interval
upper_sleep<- diff_mean + t_sleep * diff_sd
upper_sleep
```

```
## [1] 0.06780798
```

```r
lower_sleep<- diff_mean - t_sleep * diff_sd
lower_sleep
```

```
## [1] -0.08195098
```

```r
# p-value
p_value_sleep <- 2 * pt(t_sleep, sleep_df, lower.tail = FALSE)
p_value_sleep
```

```
## [1] 0.05
```