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)
set.seed(1234)
```

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?

Exercise 1

There are are 13,583 cases (observations) in this data set.

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?

Exercise 2

There are 1004 observations with missing weight.

```
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.

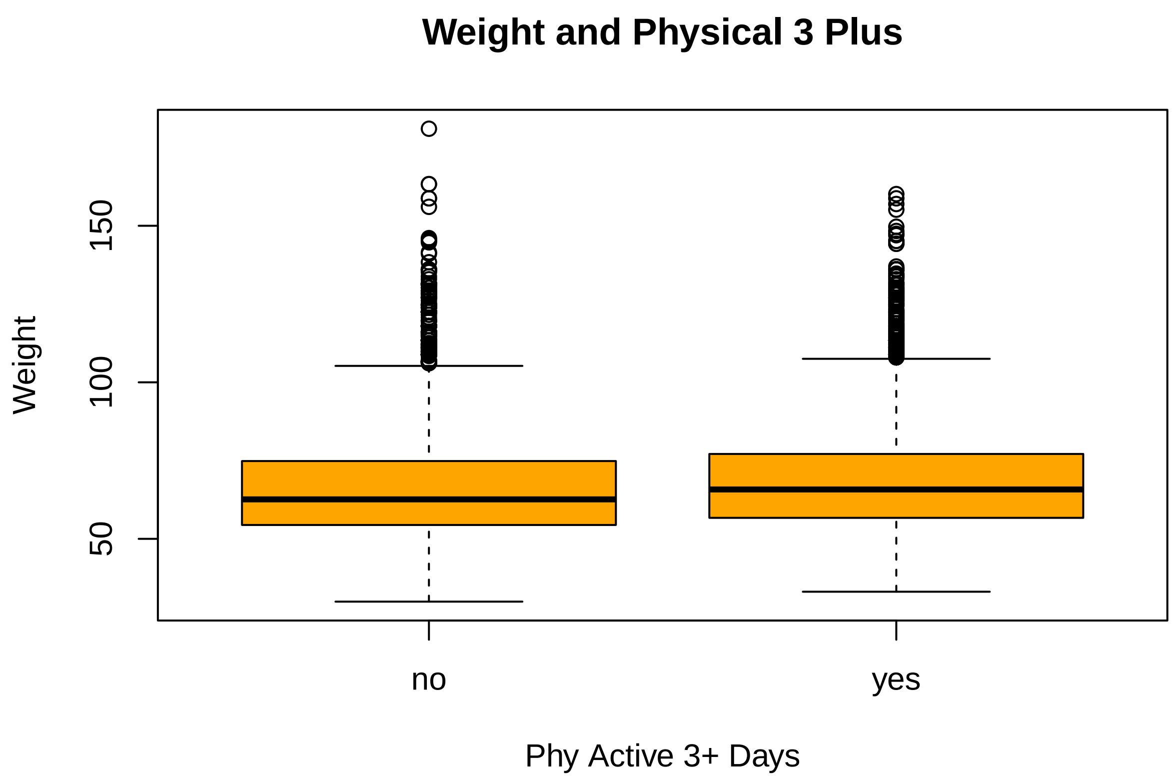
```
yrbss <- yrbss %>%
  mutate(physical_3plus = ifelse(yrbss$physically_active_7d > 2, "yes", "no"))
```

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?

Exercise 3

There seems to be no relationship between `physical_3plus` and `weight`. I was not expecting there to be much relationship except, perhaps, at the extreme weight outliers since it's possible that high schoolers who are extremely overweight might experience related health issues that might prevent them from exercising more often.

```
boxplot(yrbss$weight ~ yrbss$physical_3plus,
        col="orange",
        main="Weight and Physical 3 Plus",
        ylab="Weight",
        xlab="Phy Active 3+ Days",
        horizontal = FALSE)
```



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()`.

Exercise 4

1. **Random**: a random sample should be used to get the data. As stated above the samples are drawn at random from the population.

2. **Normal**: sampling distribution of the sample mean should be approximately normal. Since our sample size is > 30 this is true by CLT.

3. **Independent**: Individual observations should be independent. If drawing samples without replacement, sample size should not be $> 10\%$ of population. We can safely assume that less than 10% of all high schoolers were sampled. Therefore, both conditions have been met.

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

```
## # A tibble: 3 x 3
##   physical_3plus mean_weight count
##   <chr>                <dbl> <int>
## 1 no                    66.7  4404
## 2 yes                   68.4  8906
## 3 <NA>                  69.9   273
```

5. Write the hypotheses for testing if the average weights are different for those who exercise at least 3 times a week and those who don't.

Exercise 5

Ho: the average weight of those who exercise at least 3 times a week = average weight of those who don't exercise at least 3 times a week

H1: the average weight of those who exercise at least 3 times a week $\neq$ average weight of those who don't exercise at lesst 3 times a week

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 %>%
  filter(! is.na(weight)) %>%
  filter(! is.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`.

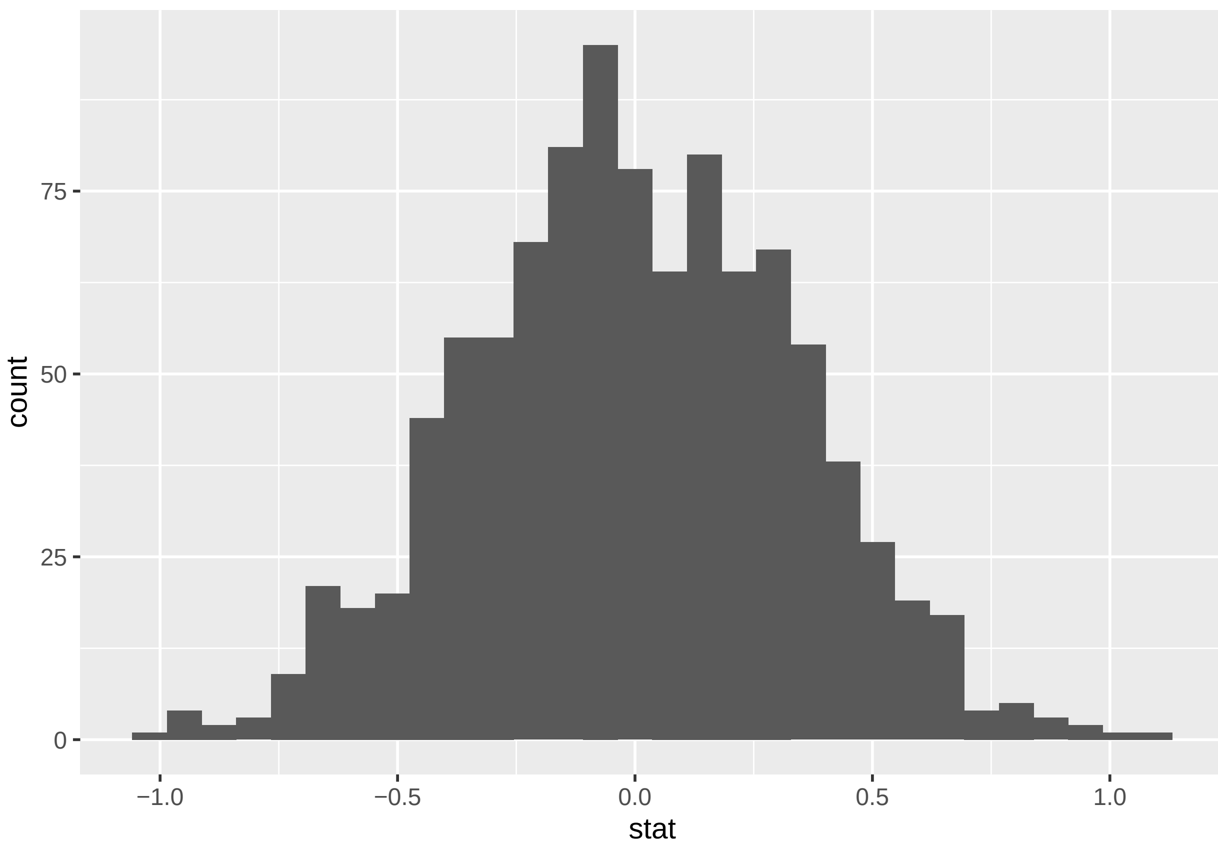
```
null_dist <- yrbss %>%
  filter(! is.na(weight)) %>%
  filter(! is.na(physical_3plus)) %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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`, 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`?

Exercise 6

There are no `null` permutations that have a difference greater than or equal to `obs_stat`.

```
null_dist%>%
    filter(stat>=obs_diff) %>%
    tally()
```

```
## # A tibble: 1 x 1
##       n
##   <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`.

```
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.

Exercise 7

Mean

```
yrbss %>%
  filter(! is.na(physical_3plus)) %>%
  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
```

Standard Deviation

```
yrbss %>%
  filter(! is.na(physical_3plus)) %>%
  group_by(physical_3plus) %>%
  summarise(sd_weight = sd(weight, na.rm = TRUE))
```

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

Sample size

```
yrbss %>%
  filter(! is.na(physical_3plus)) %>%
  group_by(physical_3plus) %>%
  summarise(freq = table(weight)) %>%
  summarise(n = sum(freq))
```

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

```r
mean_not_active <- 66.7
sd_not_active <- 17.6
n_not_active <- 4022

mean_active <- 68.4
sd_active <- 16.5
n_active <- 8342

z = 1.96

#CI for those not active
sqrt_n_not_active <- sqrt(n_not_active)
sd_n_not_active_over_sqrt_n_not_active <- sd_not_active / sqrt_n_not_active

upper_ci_not_act <- mean_not_active + z*(sd_n_not_active_over_sqrt_n_not_active)

lower_ci_not_act <- mean_not_active - z*(sd_n_not_active_over_sqrt_n_not_active)

#CI for those active
sqrt_n_active <- sqrt(n_active)
sd_n_active_over_sqrt_n_active <- sd_active / sqrt_n_active

upper_ci_act <- mean_active + z*(sd_n_active_over_sqrt_n_active)

lower_ci_act <- mean_active - z*(sd_n_active_over_sqrt_n_active)
```

```r
sprintf("CI(not active): %f %f", lower_ci_not_act, upper_ci_not_act)
```

```
## [1] "CI(not active): 66.156064 67.243936"
```

```r
sprintf("CI(active): %f %f", lower_ci_act, upper_ci_act)
```

```
## [1] "CI(active): 68.045917 68.754083"
```

More Practice

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

Exercise 8

```r
z <- 1.96

mean_height <- mean(yrbss$height, na.rm = TRUE)

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

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

height_upper_95 <- mean_height + z*(sd_height/sqrt(sample_height))

height_lower_95 <- mean_height - z*(sd_height/sqrt(sample_height))
```

```r
sprintf("95%% CI(Average Height): %f %f, Diff: %f", height_lower_95, height_upper_95, height_upper_95 -
```

```
## [1] "95% CI(Average Height): 1.689411 1.693071, Diff: 0.003659"
```

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.

Exercise 9

The width of the interval for 95% CI is wider than that for 90%.

```r
z <- 1.645

mean_height <- mean(yrbss$height, na.rm = TRUE)

sd_height <- sd(yrbss$height, na.rm = TRUE)

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

height_upper_90 <- mean_height + z*(sd_height/sqrt(sample_height))

height_lower_90 <- mean_height - z*(sd_height/sqrt(sample_height))
```

```r
sprintf("90%% CI(Average Height): %f %f, Diff: %f", height_lower_90, height_upper_90, height_upper_90 -
```

```
## [1] "90% CI(Average Height): 1.689705 1.692777, Diff: 0.003071"
```

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.

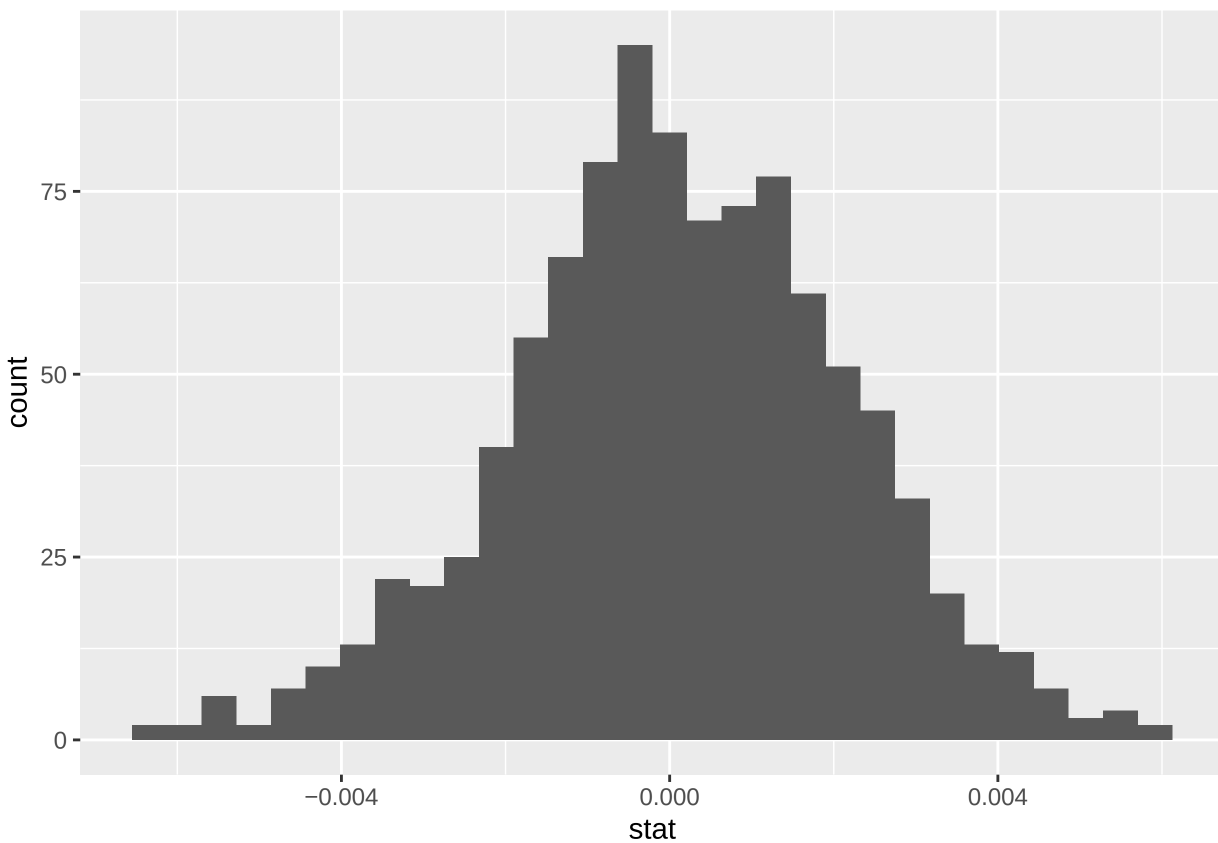
Exercise 10

HO: There is no difference in average height of those who are physically active at least 3 days per week, and those who aren't.
HA: There is a difference in average height of those who are physically active at least 3 days per week, and those who aren't.

```
obs_diff_ht <- yrbss %>%
  filter(! is.na(height)) %>%
  filter(! is.na(physical_3plus)) %>%
  specify(height ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
null_dist_ht <- yrbss %>%
  filter(! is.na(height)) %>%
  filter(! is.na(physical_3plus)) %>%
  specify(height ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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




```
yrbss %>%
  filter(! is.na(height)) %>%
  filter(! is.na(physical_3plus)) %>%
  group_by(physical_3plus) %>%
  summarise(mean_height = mean(height, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
```

```
##   physical_3plus mean_height
##   <chr>                <dbl>
## 1 no                    1.67
## 2 yes                   1.70
```

```
yrbss %>%
  filter(! is.na(height)) %>%
  filter(! is.na(physical_3plus)) %>%
  group_by(physical_3plus) %>%
  summarise(sd_height = sd(height, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##   physical_3plus sd_height
##   <chr>              <dbl>
## 1 no                 0.103
## 2 yes                0.103
```

```
yrbss %>%
  filter(! is.na(physical_3plus)) %>%
  group_by(physical_3plus) %>%
  summarise(freq = table(height)) %>%
  summarise(n = sum(freq))
```

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

```
z <- 1.96

# Active
mean_height_a <- 1.70
samples_a <- 8342
sd_height_a <- 0.103
height_upper_a <- mean_height_a + z*(sd_height_a/sqrt(samples_a))
height_lower_a <- mean_height_a - z*(sd_height_a/sqrt(samples_a))

mean_height_na <- 1.67
samples_na <- 4022
sd_height_na<- 0.103
height_upper_na <- mean_height_na + z*(sd_height_na/sqrt(samples_na))
height_lower_na <- mean_height_na - z*(sd_height_na/sqrt(samples_na))
```

```
sprintf("95%% CI(Average Height of Active): %f %f, Diff: %f", height_lower_a, height_upper_a, height_upp
```

```
## [1] "95% CI(Average Height of Active): 1.697790 1.702210, Diff: 0.004421"
```

```
sprintf("95%% CI(Average Height of Not Active): %f %f, Diff: %f", height_lower_na, height_upper_na, heig
```

```
## [1] "95% CI(Average Height of Not Active): 1.666817 1.673183, Diff: 0.006367"
```

With a confidence level of 95%, the average height of students who are physically active at least 3 days/week, is between ~1.697m and 1.702m. The average height of students who are not physically active is between ~1.666m and ~1.673m.

Since the p-value is less than 0.05 we reject the null hypothesis and conclude that there is a difference in average height of those who are physically active at least 3 days per week, and those who aren't.

```
null_dist_ht %>%
  get_p_value(obs_stat = obs_diff_ht, direction = "two_sided")
```

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

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.

There are 7 options is the dataset for the `hours_tv_per_school_day` variable.

```
yrbss %>%
    filter(! is.na(hours_tv_per_school_day)) %>%
    group_by(hours_tv_per_school_day)%>%
    summarise(n())
```

```
## # A tibble: 7 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
```

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: Is there evidence to support the hypothesis that students who are heavier than the mean weight sleep more than students who weight less than the mean weight?

HO: Students who are heavier than the mean weight do not sleep more than students whose weight is less than the mean weight HA: Students who are heavier than the mean weight sleep more than students whose weight is less than the mean weight

Assumptions: 1. Random sampling: check. 1. Normality: sample size is large and CLT applies.

α: 0.05

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

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

```
mean_wt <- yrbss %>%
            filter(! is.na(weight)) %>%
            filter(! is.na(school_night_hours_sleep)) %>%
            summarise(mean = mean(weight)) %>%
            pull(mean)
mean_wt
```

```
## [1] 67.89322
```

We will make define sleeping less implies as less than 6 hours of sleep per night. Add sleeping less indicator to data frame.

```
yrbss <- yrbss %>%
    filter(! is.na(school_night_hours_sleep)) %>%
    mutate(sleep_less_ind = ifelse(school_night_hours_sleep < 6, 'yes', 'no'))
```

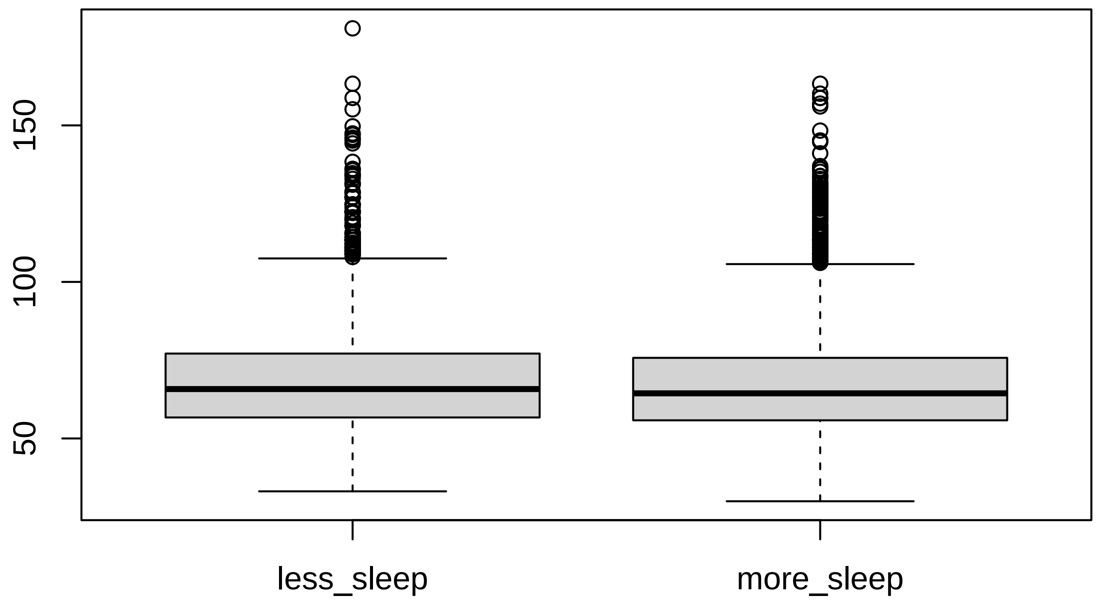
Split up population weights into those who sleep less and those who don't.

```
weight_sleep_less <- yrbss %>%
  filter(! is.na(weight)) %>%
  filter(! is.na(sleep_less_ind)) %>%
  filter(sleep_less_ind == "yes") %>%
  select(weight)

weight_sleep_more <- yrbss %>%
  filter(! is.na(weight)) %>%
  filter(! is.na(sleep_less_ind)) %>%
  filter(sleep_less_ind == "no") %>%
  select(weight)
```

Boxplot population weights.

```
boxplot(weight_sleep_less$weight, weight_sleep_more$weight,
        names = c("less_sleep", "more_sleep"))
```



Compute summary stats for weight of those who sleep less.

```
summary_weight_sleep_less <- summary(weight_sleep_less$weight)
summary_weight_sleep_less
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   33.11   56.70   65.77   69.16   77.11  180.99
```

Compute summary stats for weight of those who sleep more.

```
summary_weight_sleep_more <- summary(weight_sleep_more$weight)
summary_weight_sleep_more
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.94   55.79   64.41   67.54   75.75  163.30
```

Compute sd, and diff in mean of those who sleep more and those who sleep less. Compute standard deviation

```
mean_diff <- summary_weight_sleep_more["Mean"] - summary_weight_sleep_less["Mean"]
sd <-
  sqrt(
  ((summary_weight_sleep_more["Mean"]^2) / nrow(weight_sleep_more)) +
  ((summary_weight_sleep_less["Mean"]^2) / nrow(weight_sleep_less))
  )
```

Compute t-dist and confidence intervals.

```
degrees_of_freedom <- NROW(weight_sleep_less) - 1

t <- qt(0.05/2, df = degrees_of_freedom, lower.tail = FALSE)

lower_ci <- mean_diff - t * sd
upper_ci <- mean_diff + t * sd
sprintf("CI: %f %f, Diff: %f", lower_ci, upper_ci, upper_ci - lower_ci)
```

```
## [1] "CI: -4.666506 1.442799, Diff: 6.109305"
```

Compute `p_val`. Since `p_val` is equal to significance level (0.05), the null hypothesis can be rejected.

```
p_val <- 2 * pt(t, degrees_of_freedom, lower.tail = FALSE)
p_val
```

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