Inference for numerical data

Alyssa Gurkas

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(lsr)
```

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 Response There are 13 variables (or columns) in this data set. For the raw unproccessed data, five variables are considered qualitative and are either in the double (dbl) or integer class (`age`,`height`,`weight`,`physically_active_7d`, `strength_training_7d`). Eight variables are considered qualitative and are in the character class (`gender`, `grade`,`hispanic`, `race`, `helmet_12m`, `text_while_driving_30d`, `hours_tv_per_school_day`, `school_night_hours_sleep`). In R, even if the data is mainly numeric, such as the `grade` column, if there are values that are characters. It will consider all the values within that variable to be characters.**

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

```r
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.

```r
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 Response Of the 13583 observations in the data set, 1004 observations are missing from the weight variable, or seven percent.**

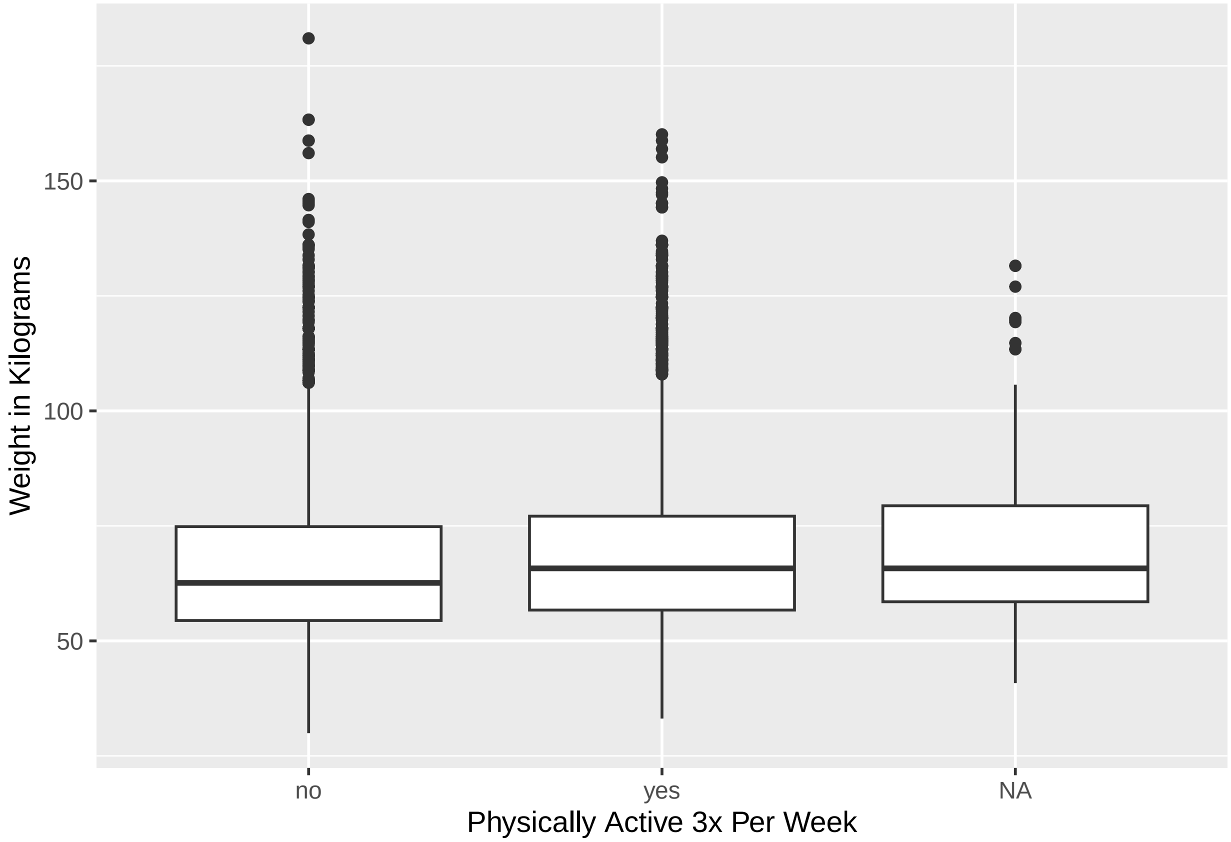
Next, consider the possible relationship between a high schooler's weight and their physical activity. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

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.

```r
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 Response**

```r
ggplot(yrbss, aes(x=physical_3plus, y=weight)) +
geom_boxplot() +
xlab("Physically Active 3x Per Week") +
ylab("Weight in Kilograms")
```



I expected that the people who are less active would weigh more on average. However, the box plot is showing that the people who are more active, weigh slightly more, on average.

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

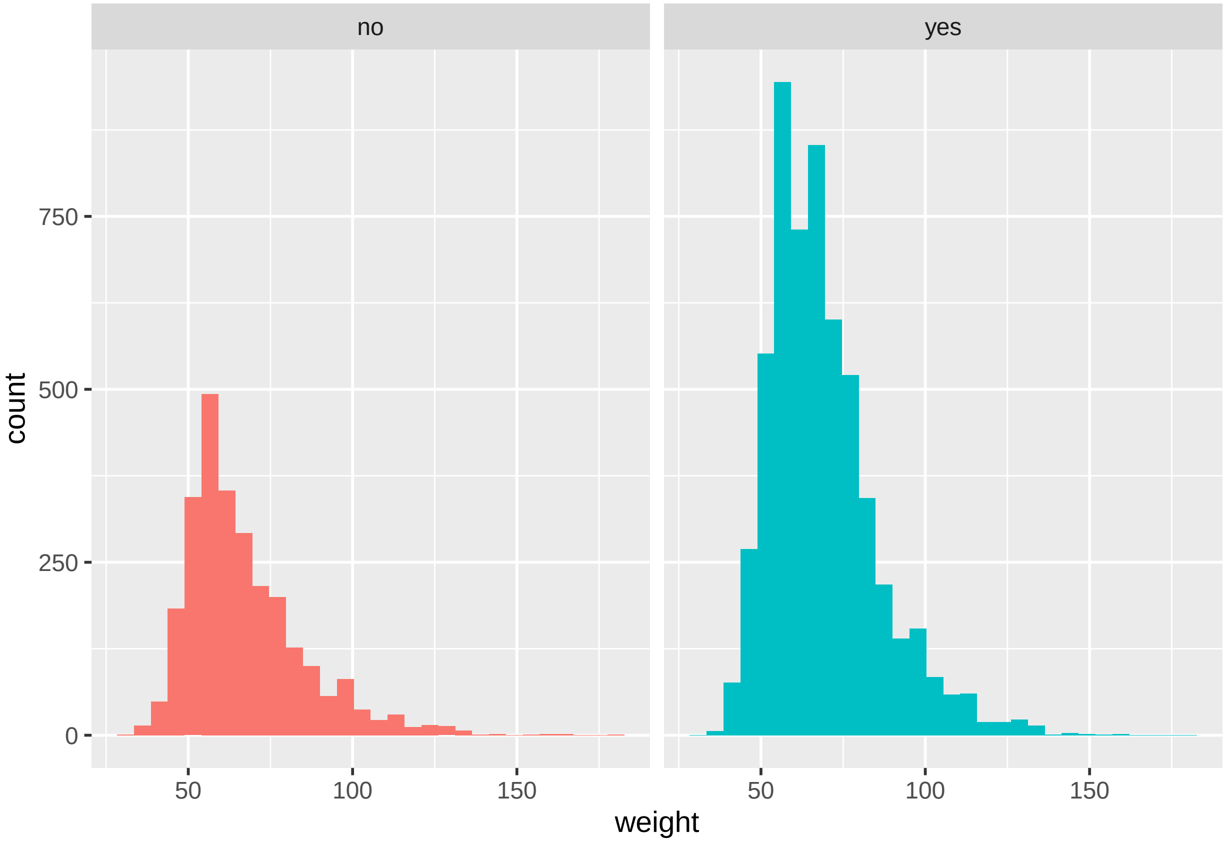
4. Are all conditions necessary for inference satisfied? Comment on each. You can compute the group sizes with the `summarize` command above by defining a new variable with the definition `n()`. **Exercise**

4 Response To apply the central limit theorem, there two conditions: (1) independence: this means that the observations must be independent, and the sample is randomized. (2) normality: this means that the sample size is sufficiently large such that np is greater than or equal to ten and n(1-p) is greater than ten. The CDC's data methodology documentation available at https://www.cdc.gov/mmwr/pdf/rr/rr6201.pdf certifies that the data collection method is representative of its respective population. This means we can consider the independence condition is met. However, we should ensure that the normality condition is met through some analysis. To do this, we will check if the sample size is less than 30, and if there are outliers in the data.

```r
# checking observations in the YRBSS dataset:
yrbss %>%
  group_by(physical_3plus) %>%
  drop_na() %>%
  summarise(n_obs = length(weight))
```

```
## # A tibble: 2 x 2
##   physical_3plus n_obs
##   <chr>          <int>
## 1 no              2656
## 2 yes             5695
```

```r
yrbss_v2 <- yrbss %>%
  drop_na()

ggplot(yrbss_v2, aes(x=weight, fill=physical_3plus)) +
    geom_histogram(position="dodge") +
    facet_wrap(vars(physical_3plus))+
    theme(
      legend.position = "none"
    )
```



Based on the analysis and plot, we can determine that the conditions necessary for inference are satisfied. The data seems to be normally distributed, and the sample size is sufficiently large.

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. **Exercise 5 Response H_0 Students who are physically active 3 or more days per week weigh the same as students who are not physically active. H_A Students who are physically active 3 or more days per week weigh less than those who don't exercise.**

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 %>%
  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`.

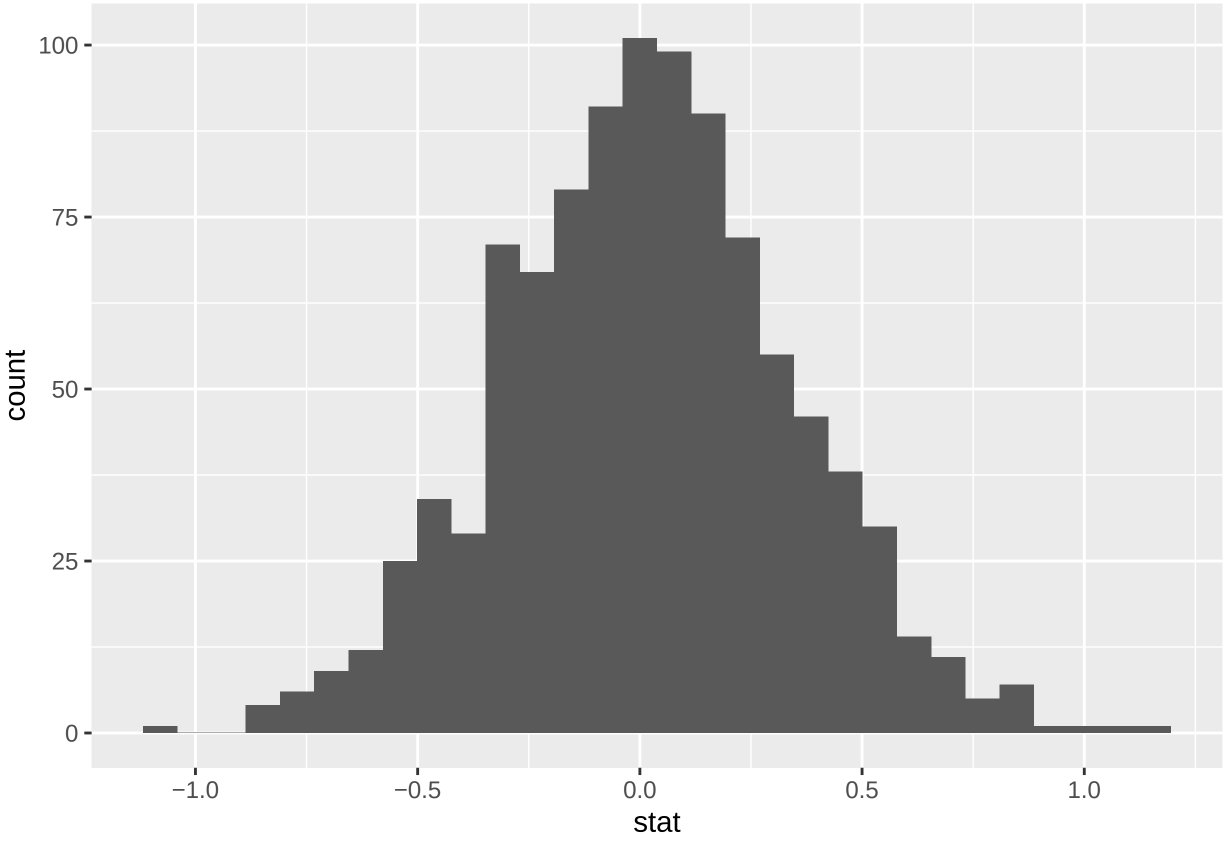
```r
null_dist <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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:

```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`? **Exercise 6 Response**

```r
null_dist |>
  filter(`stat` >= obs_diff) |>
  count()
```

```
## Response: weight (numeric)
## Explanatory: physical_3plus (factor)
## Null Hypothesis: independence
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

There are no permutations that have a difference of at least `obs_diff`, or 1.77.

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.

Question: this line of code generated the following warning: Warning message: Please be cautious in reporting a p-value of 0. This result is an approximation based on the number of `reps` chosen in the `generate()` step. See `get_p_value()` for more information.

What does this mean?

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 Response In this exercise, we are generating a confidence interval for a difference in means. Therefore, we will use the equation:**

Confidence interval = $(x1\breve{}x2) \pm \; t * \sqrt{((sp/n1) + (sp/n2))}$ Where: * x1~, x2~: sample 1 mean, sample 2 mean * t: the t-critical value based on the confidence level and (n1+n2-2) degrees of freedom * sp: pooled variance, calculated as $((n1-1)*s1^2 + (n2-1)*s2^2)/(n1+n2-2)$ * n1, n2: sample 1 size, sample 2 size

```
# creating sample for students who are physically active three days or more
active3xSamp <- yrbss |>
  filter(physical_3plus == "yes") |>
  drop_na()

# creating sample for students who are not physically active three days or more
nActiveSamp <- yrbss |>
  filter(physical_3plus == "no") |>
  drop_na()

# developing values for sample 1
n1 <- nrow(active3xSamp) # sample 1 size
x1 <- mean(active3xSamp$weight) # sample 1 mean
var1 <- var(active3xSamp$weight) # sample 1 variance

# developing values for sample 2
n2 <- nrow(nActiveSamp) # sample 2 size
```

```r
x2 <- mean(nActiveSamp$weight) # sample 2 mean
var2 <- var(nActiveSamp$weight) # sample 2 variance

# pooled variance for samples 1 and 2
sp <- ((n1 - 1) * var1 + (n2 - 1) * var2) / (n1 + n2 - 2)

# calculating the margin of error for a 95% level
margin <- qt(0.975,df=n1+n2-1)*sqrt(sp/n1 + sp/n2)

# lower confidence interval
lowerinterval <- (x1-x2) - margin

# upper confidence interval
upperinterval <- (x1-x2) + margin

lowerinterval
```

```
## [1] 0.747816
```

```r
upperinterval
```

```
## [1] 2.307636
```

```r
# taking t-test of weight and physical_3plus, using the values in the dataframe yrbss_v2.
# As a note, yrbss_v2 does not have any null values.
t.test(data = yrbss_v2, weight ~ physical_3plus)
```

```
##
##  Welch Two Sample t-test
##
## data:  weight by physical_3plus
## t = -3.7143, df = 4781.2, p-value = 0.0002061
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0
## 95 percent confidence interval:
##  -2.3340891 -0.7213629
## sample estimates:
##  mean in group no mean in group yes
##          67.14974          68.67747
```

Based on these results, we can conclude that the difference between the difference in weight in these two groups lies somewhere between .74 and 2.30. From the results of the t-test, we can conclude that that we are 95% certain that the mean of the two groups is somewhere between 67.14 and 68.67. Based on this, I would fail to reject the null hypothesis.

What would the confidence interval look like if the weights were not similar between the two groups?

More Practice

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

```
# Note:  yrbss_v2 is defined in line 138
# and is based off of the original yrbss dataset, with NAs dropped.

m_height <- mean(yrbss_v2$height) # mean height
sd_height <- sd(yrbss_v2$height) # standard deviation of height
n_height <- nrow(yrbss_v2) # sample size of height

lowerinterval_height <- m_height - 1.96*(sd_height/sqrt(n_height))
upperinterval_height <- m_height + 1.96*(sd_height/sqrt(n_height))

lowerinterval_height
```

```
## [1] 1.694811
```

```
upperinterval_height
```

```
## [1] 1.699298
```

From this exercise, we can conclude with 95% certainty that the mean value for students height is somewhere between 1.694811-2.307636 meters.

---

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 Response**

```
t <- qt(.95, n_height-1)

lowerinterval_height_90 <-
  m_height - t*(sd_height/sqrt(n_height))
upperinterval_height_90 <-
  m_height + t*(sd_height/sqrt(n_height))

lowerinterval_height_90
```

```
## [1] 1.695171
```

```
upperinterval_height_90
```

```
## [1] 1.698937
```

The range for the 95% confidence interval is larger than the range for the 90% confidence interval. This is because to be more confident the value lies in the dataset, you must have a larger range.

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 Response $\mathbf{H_0}$: The average height of the students that are physically active more than three days per week is the same as students who are not. $\mathbf{H_A}$: The average height of the students that are physically active more than three days per week is different than the students who are not.**

```
obs_diff_height <- yrbss_v2 %>%
  specify(height ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

null_dist_height <- yrbss_v2 %>%
  specify(height ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

null_dist_height |>
  filter(`stat` >= obs_diff_height) |>
  count()
```

```
## Response: height (numeric)
## Explanatory: physical_3plus (factor)
## Null Hypothesis: independence
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

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

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.942
```

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. **Exercise 11 Response**

```
unique(yrbss$hours_tv_per_school_day)
```

```
## [1] "5+"           "2"            "3"            "do not watch" "<1"
## [6] "4"            "1"            NA
```

```
length(unique(yrbss$hours_tv_per_school_day))
```

```
## [1] 8
```

There are eight different options in the dataset for the variable hours_tv_per_school_day. The options are: 5+, 2, 3, do not watch, <1, 4, 1, and NA (or null).

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. **Exercise 12 Response H_0: The average height of the students that sleep at least 8 hours a night is the same as students that do not. H_A: The average height of the students that sleep at least 8 hours a night is different than the students who do not.**

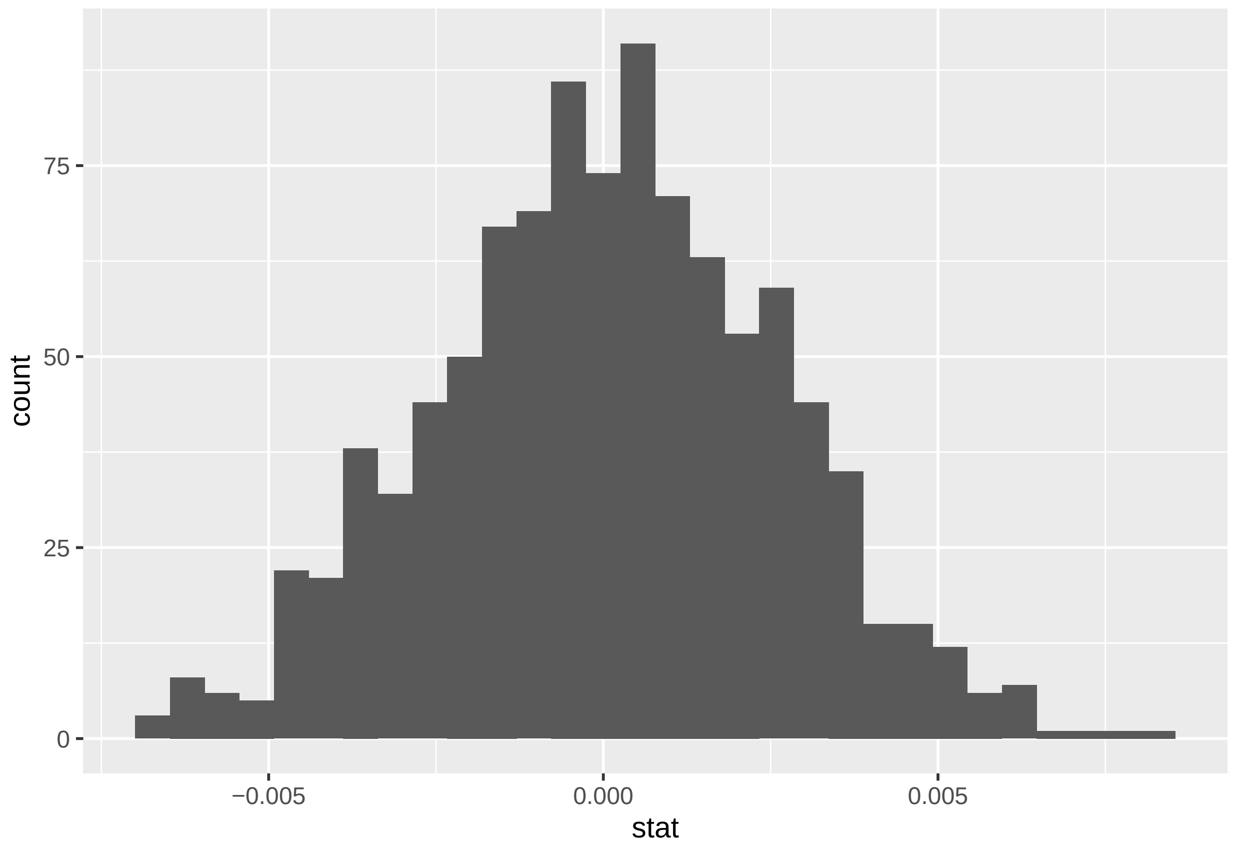
```
yrbss_v3 <- yrbss_v2 |>
  mutate(sleep_8plus = ifelse(yrbss_v2$school_night_hours_sleep >= 8, "yes", "no"))

obs_diff_sleep_height <- yrbss_v3 %>%
  specify(height ~ sleep_8plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

null_dist_sleep_height <- yrbss_v3 %>%
  specify(height ~ sleep_8plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

null_dist_height |>
  filter(`stat` >= obs_diff_sleep_height) |>
  count()
```

```
## Response: height (numeric)
## Explanatory: physical_3plus (factor)
## Null Hypothesis: independence
## # A tibble: 1 x 1
##       n
##   <int>
## 1  1000
```

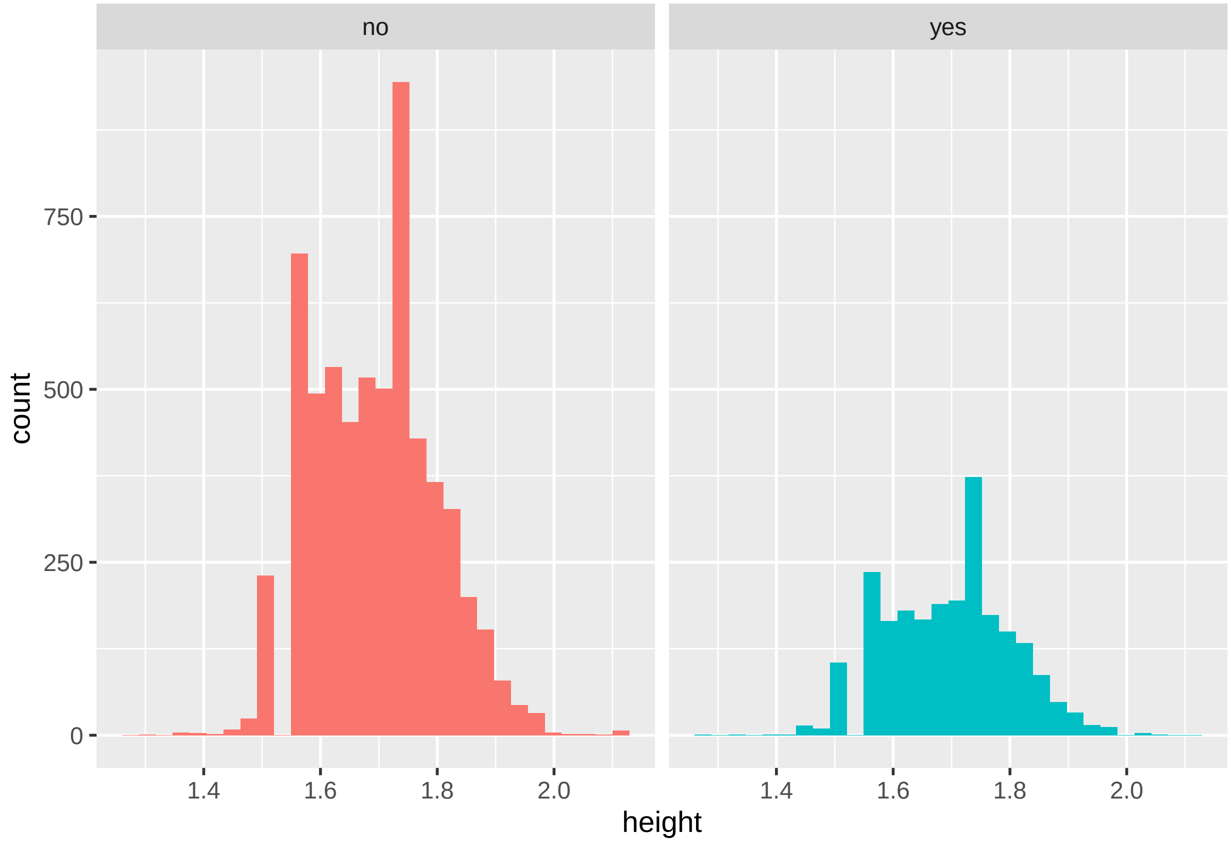
```
null_dist %>%
  get_p_value(obs_stat = obs_diff_sleep_height, direction = "two_sided")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.958
```

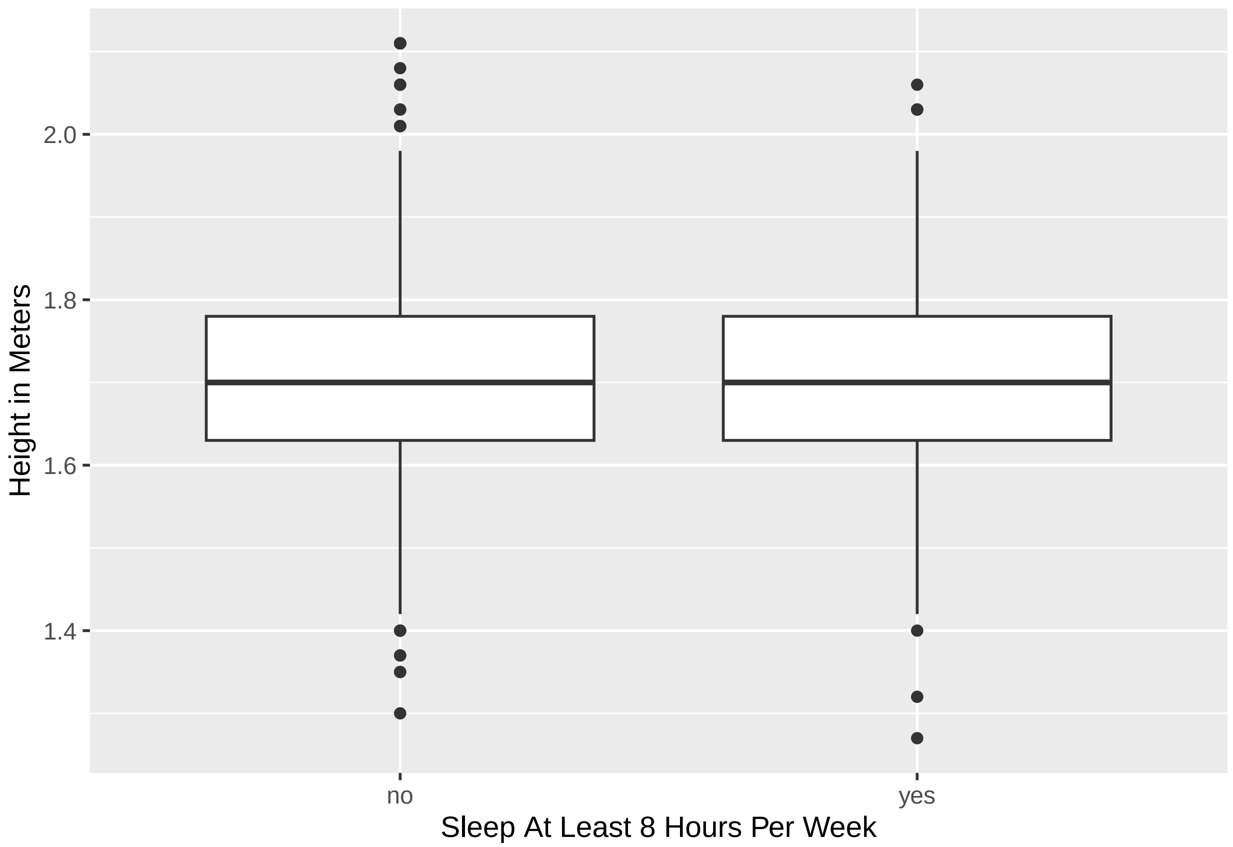
```
ggplot(data = null_dist_sleep_height, aes(x = stat)) +
  geom_histogram()
```



```r
ggplot(yrbss_v3, aes(x=height, fill=sleep_8plus)) +
    geom_histogram(position="dodge") +
    facet_wrap(vars(sleep_8plus))+
    theme(
      legend.position = "none"
    )
```



```
ggplot(yrbss_v3, aes(x=sleep_8plus, y=height)) +
geom_boxplot() +
xlab("Sleep At Least 8 Hours Per Week") +
ylab("Height in Meters")
```



```
t.test(data = yrbss_v3, height ~ sleep_8plus)
```

```
##
##  Welch Two Sample t-test
##
## data:  height by sleep_8plus
## t = -0.7719, df = 4097.2, p-value = 0.4402
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0
## 95 percent confidence interval:
##  -0.007041317  0.003063044
## sample estimates:
##  mean in group no mean in group yes
##          1.696508          1.698497
```

Based on this analysis, I would fail to reject the null hypothesis. It does not seem that students who sleep at least eight hours per week are taller than students who do not, on average. * * *