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)
```

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"~
```

The cases can be explored above. We have a total number of 13,583 observations.

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?

```
colSums(is.na(yrbss))
```

```
##                       age                    gender                     grade
##                        77                        12                        79
##                  hispanic                      race                    height
##                       231                      2805                      1004
##                    weight                helmet_12m   text_while_driving_30d
##                      1004                       311                       918
##     physically_active_7d  hours_tv_per_school_day     strength_training_7d
##                       273                       338                      1176
## school_night_hours_sleep
##                      1248
```

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.

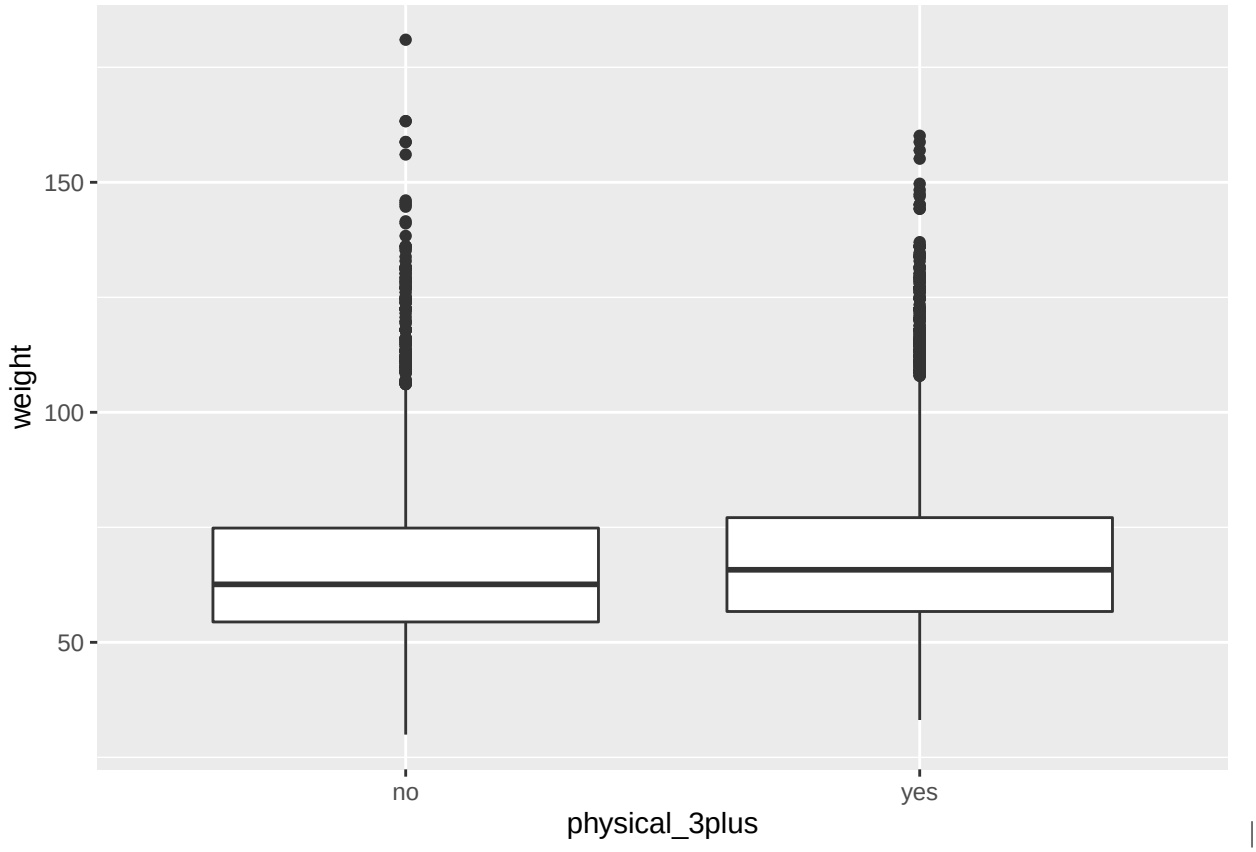
First, let's create a new variable `physical_3plus`, which will be coded as either "yes" if they are physically active for at least 3 days a week, and "no" if not.

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

ggplot(aes(physical_3plus,weight)) + geom_boxplot(na.rm = T)
```



The boxplot shows us that the median weight of though that are physically active for at least 3 days a week, is higher than those that do not. You might expect that those who were active would also weigh less, however according to this that is not the case. Is this a significant difference though?

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 data is taken from a random sample of a larger population, so we can assume that our observations are independent. The boxplots shown above show no sign of a skewed distribution either, so know that our groups are normally distributed.

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: The weight of the students who exercise at least three days a week is equal to those who do not

Ha: THe weight of the students who exercise at least three days a week is different than those who do not

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`.

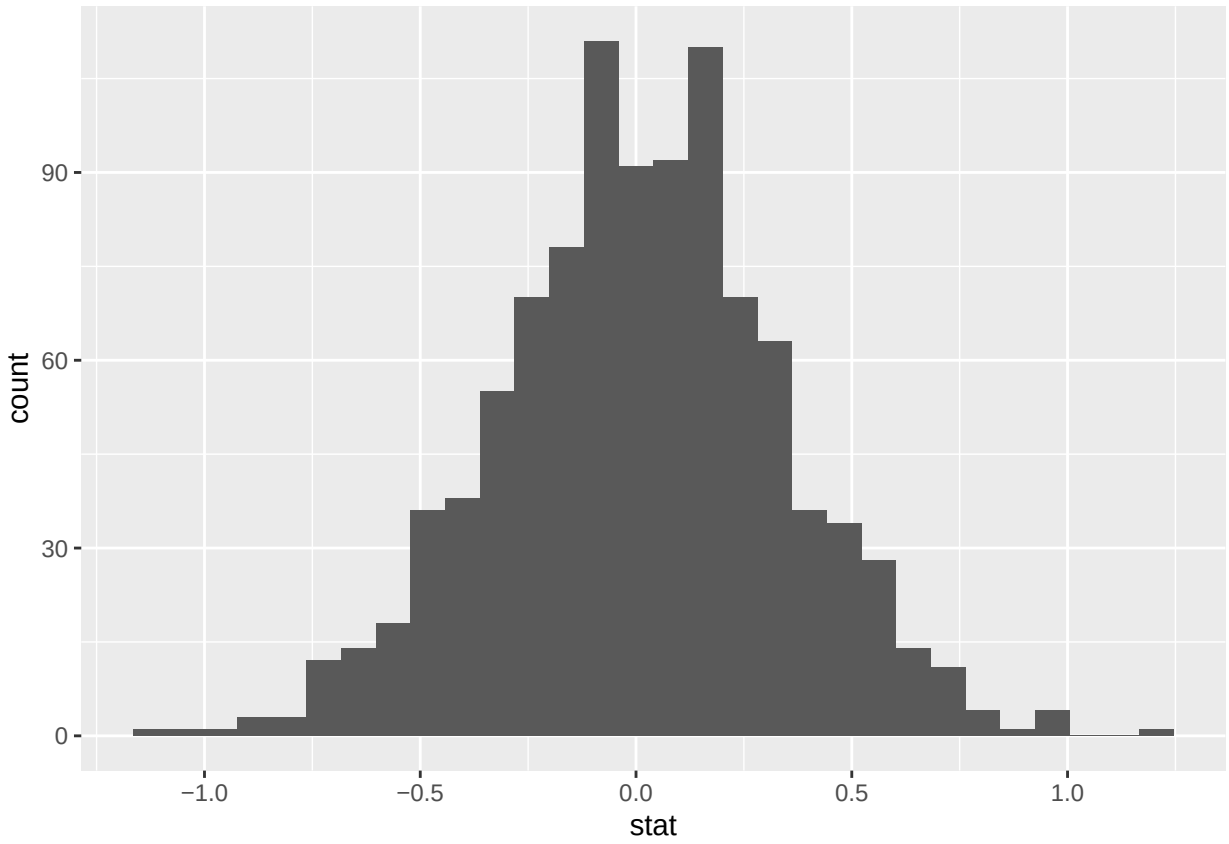
```
null_dist <- yrbss %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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`, 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`?

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

```
## # A tibble: 0 x 2
## # ... with 2 variables: replicate <int>, stat <dbl>
```

As shown above, none of the null differences are at least equal to the obs_stat value.

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.

```
yrbss %>%
 filter(physical_3plus=="yes") %>%
 specify(response =weight ) %>%
 generate(reps = 1000, type = "bootstrap") %>%
 calculate(stat = "mean") %>%
 get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     68.1     68.8
```

```
yrbss %>%
 filter(physical_3plus=="no") %>%
 specify(response =weight ) %>%
 generate(reps = 1000, type = "bootstrap") %>%
 calculate(stat = "mean") %>%
 get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     66.1     67.2
```

```
yrbss %>%
 specify(weight ~ physical_3plus) %>%
 hypothesize(null = "independence") %>%
 generate(reps = 1000, type = "permute") %>%
 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   -0.656    0.639
```

As shown above, we generated both the individual confidence intervals for each groups mean value as well as the difference in mean confidence interval. Through looking at our results, we can conclude we are fairly confident that there is a significant difference in weight as both the individual values are significantly different, and the combined difference in means interval overlaps 0, indicating that there is a significant difference occurring.

More Practice

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

```
yrbss %>%
  specify(response =height ) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
  get_ci(level = 0.95)
```

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

We are 95% confident that the average height of the population is between 1.689 and 1.693

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.

```
yrbss %>%
  specify(response =height ) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
  get_ci(level = 0.90)
```

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

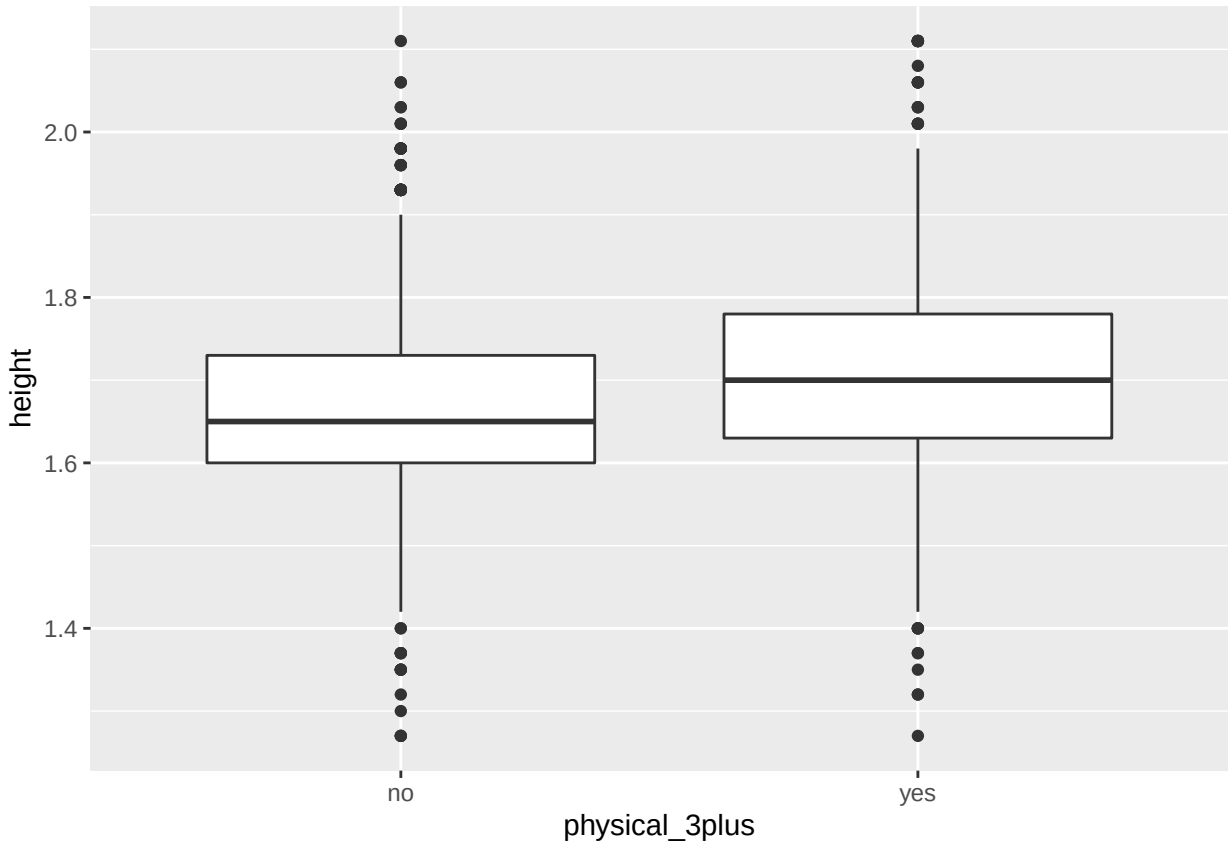
10. Conduct a hypothesis test evaluating whether the average height is different for those who exercise at least three times a week and those who don't.

H0: The average heights of those who exercise at least three times a week and those who do not, are the same.

HA: The average heights of those who exercise at least three times a week and those who do not, are different.

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

ggplot(aes(physical_3plus,height)) + geom_boxplot(na.rm = T)
```




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

```
## # A tibble: 3 x 2
##   physical_3plus mean_weight
##   <chr>                <dbl>
## 1 no                    1.67
## 2 yes                   1.70
## 3 <NA>                  1.71
```

```
 yrbss %>%
  filter(physical_3plus=="yes") %>%
  specify(response =height ) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
  get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     1.70     1.71
```

```
yrbss %>%
 filter(physical_3plus=="no") %>%
 specify(response =height ) %>%
 generate(reps = 1000, type = "bootstrap") %>%
 calculate(stat = "mean") %>%
 get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     1.66     1.67
```

```
yrbss %>%
 specify(height~ physical_3plus ) %>%
 hypothesize(null = "independence") %>%
 generate(reps = 1000, type = "permute") %>%
 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 -0.00374  0.00404
```

As shown above, our interval contains 0 within it, so we fail to reject the null hypothesis. It's clear that the averages for both groups are slightly different, and the interval hovering so close to 0 represents this. It seems however, that the values for both groups are similar enough to conclude that is no significant difference.

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 %>%
count(hours_tv_per_school_day)
```

```
## # 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.

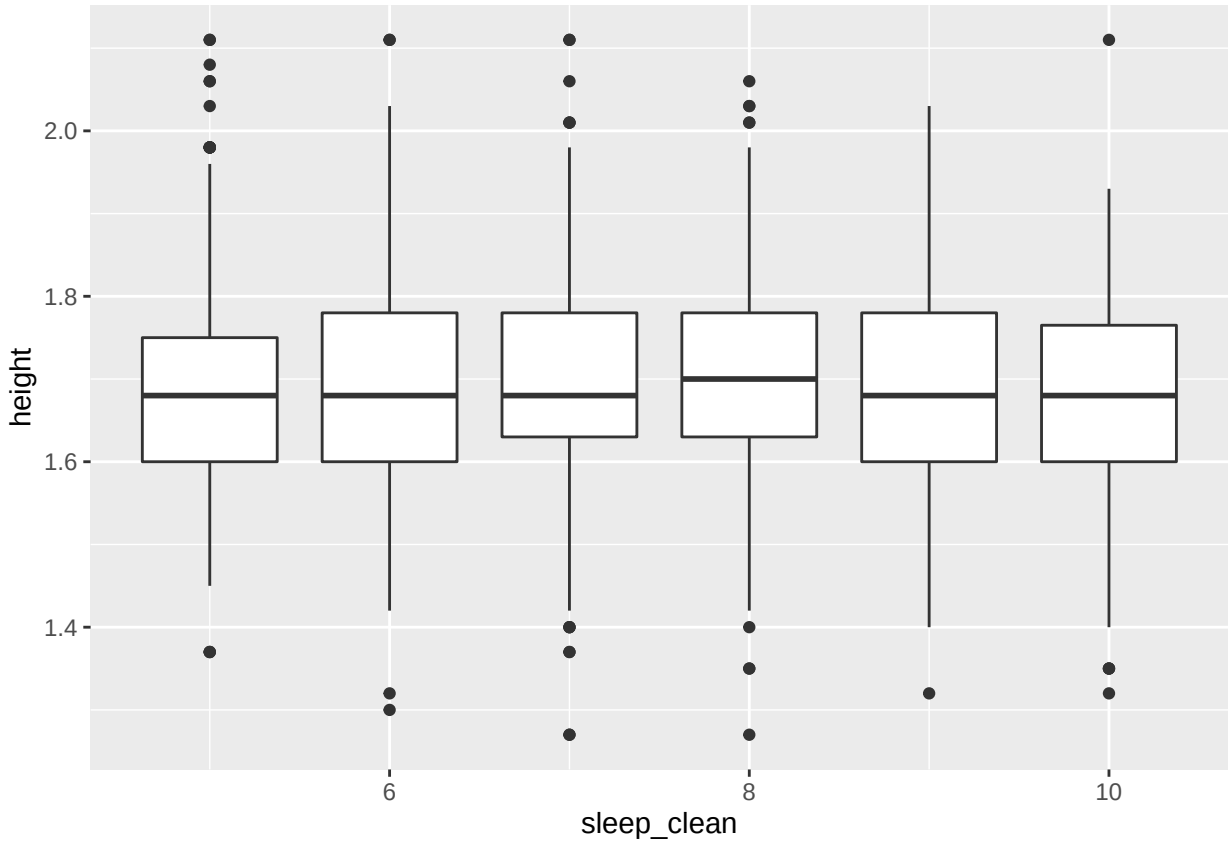
H0: There is no association between weight and sleep.

HA: There is a relationship between weight and sleep

```
yrbss <- yrbss %>%
  mutate(sleep_clean =as.numeric(gsub("[^0-9.-]", "", school_night_hours_sleep))  )
```

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

ggplot(aes(sleep_clean,height,group=sleep_clean)) + geom_boxplot(na.rm = T)
```




```
yrbss %>%

  specify(height~ sleep_clean ) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "correlation")  %>%
  get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1  -0.0186   0.0183
```

As our confidence interval includes 0 within it, we are fairly certain that we fail to reject the null hypothesis and we can conclude that there is no significant association between sleep and weight.