Inference for numerical data

Donald Butler

10/24/2021

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?

The cases are the high school students that took the survey and there are 13,583 observations in the dataset.

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?

```
yrbss %>%
  filter(is.na(weight)) %>%
  count(weight)
```

```
## # A tibble: 1 x 2
##   weight     n
##    <dbl> <int>
## 1     NA  1004
```

There are 1004 observations with weight missing.

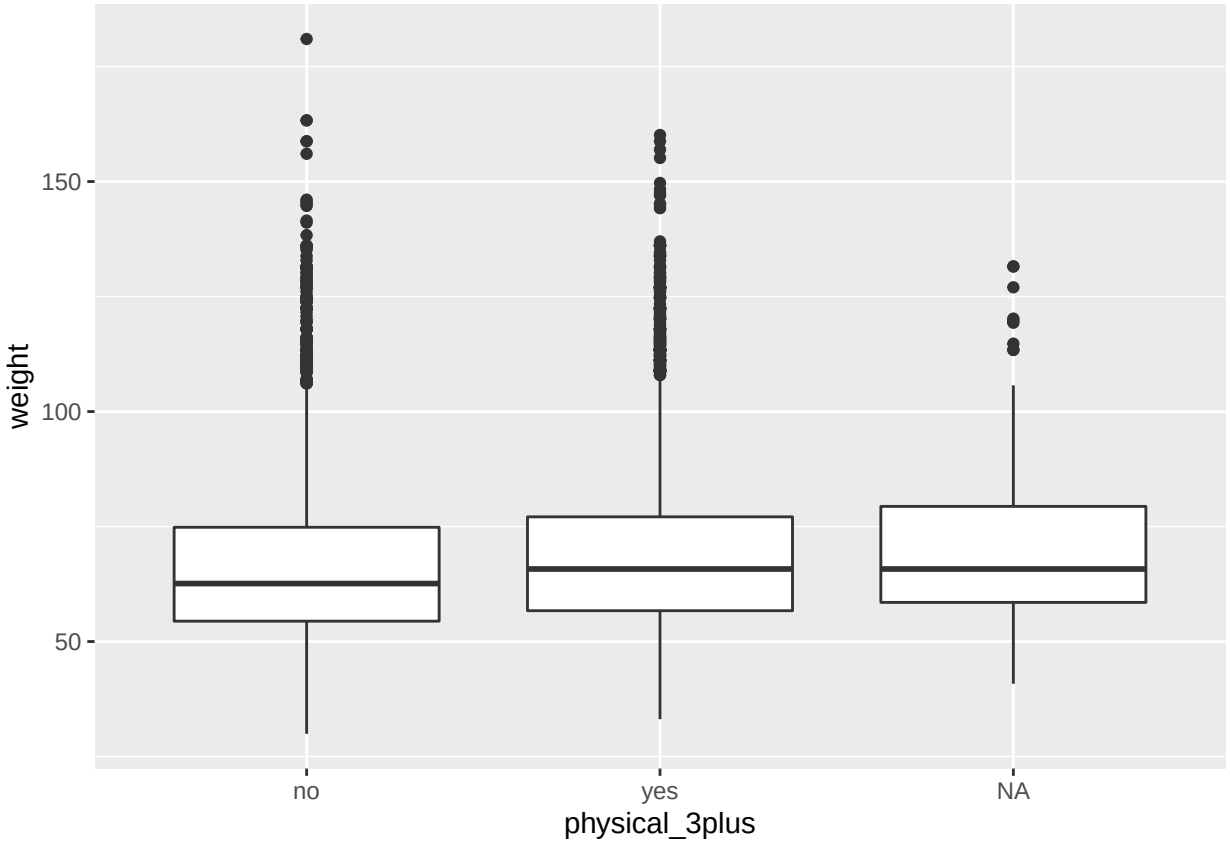
Next, consider the possible relationship between a high schooler's weight and their physical activity. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

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(x = physical_3plus, y = weight)) + geom_boxplot()
```



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),n = n())
```

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

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()`.

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

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

The groups of students with physical activity at least 3 times a week and those that don't are independent. The groups are drawn from a random sample and sufficiently large to be about normal.

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.

H_0 : The average weight for students that exercise at least 3 times per week is the same as students that don't. H_A : The average weight for students that exercise at least 3 times per week is not the same as students that 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`.

```
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`.

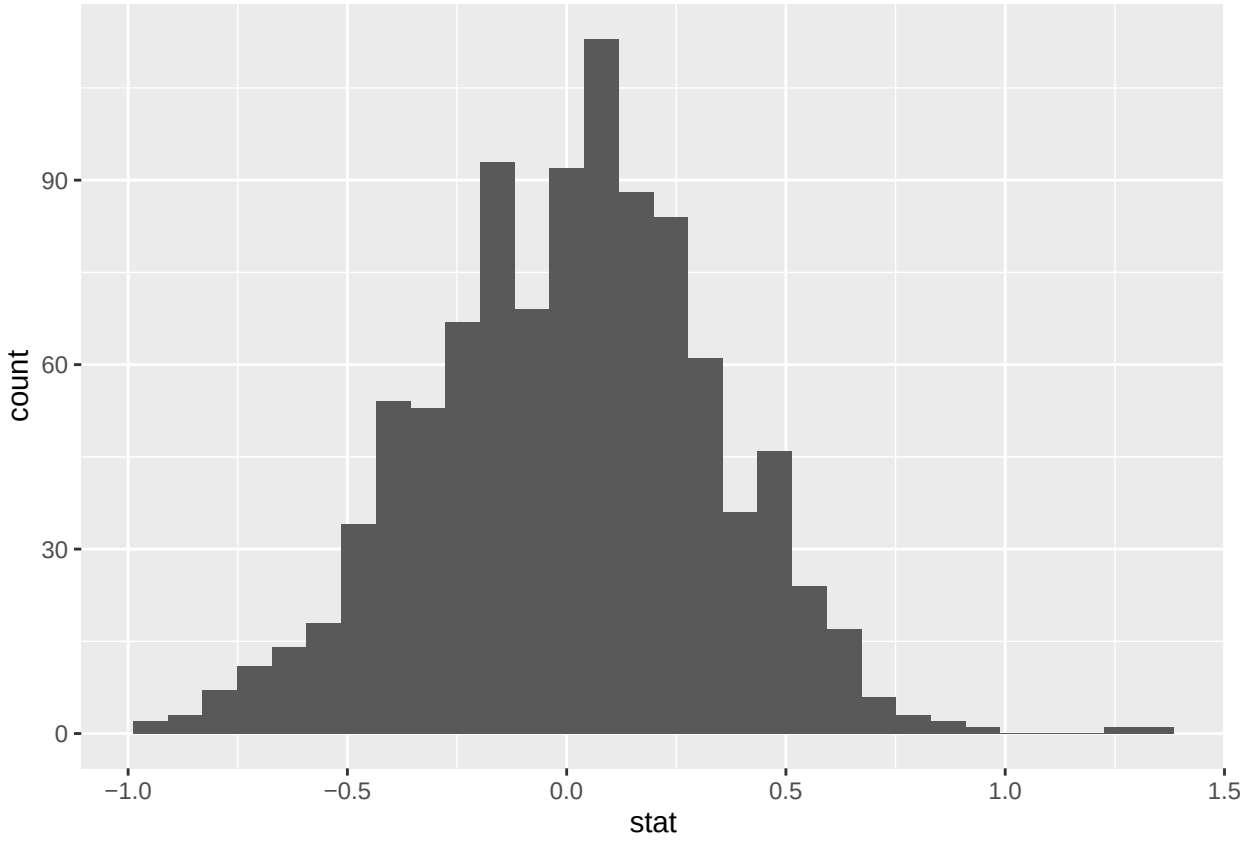
```
null_dist <- yrbss %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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`?

What is `obs_stat`, do you mean `obs_diff`?

```
null_dist %>%
  mutate(atleasttest = ifelse(stat >= 1.77,1,0)) %>%
  count(atleasttest)
```

```
## Response: weight (numeric)
## Explanatory: physical_3plus (factor)
## Null Hypothesis: independence
## # A tibble: 1 x 2
##   atleasttest     n
##         <dbl> <int>
## 1           0  1000
```

None of the null permutations have a difference of at least 1.77458426448825

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

```
## # A tibble: 3 x 4
##   physical_3plus mean_weight     n sd_weight
##   <chr>                <dbl> <int>     <dbl>
## 1 no                    66.7  4404      17.6
## 2 yes                   68.4  8906      16.5
## 3 <NA>                  69.9   273      17.6
```

```
ME <- 1.96 * sqrt(17.6^2/4404 + 16.5^2/8906)

c(obs_diff - ME,obs_diff + ME)
```

```
## $stat
## [1] 1.151978
##
## $stat
## [1] 2.39719
```

We are 95% confident that the difference in mean weight between students that exercise at least 3 times per week, and those that don't is between 1.2 and 2.4 kgs.

More Practice

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

```
yrbss %>%
  summarise(mean_height = mean(height, na.rm = TRUE),n = n(),sd_height = sd(height,na.rm = TRUE))
```

```
## # A tibble: 1 x 3
##   mean_height     n sd_height
##         <dbl> <int>     <dbl>
## 1        1.69 13583     0.105
```

```
ME <- 1.96 * .105 / sqrt(13583)
c(1.69 - ME,1.69 + ME)
```

```
## [1] 1.688234 1.691766
```

We are 95% confident that the mean height for high school students is between 1.688 and 1.692 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.

```
ME90 <- 1.64 * .105 / sqrt(13583)
c(1.69 - ME90,1.69 + ME90)
```

```
## [1] 1.688522 1.691478
```

```
2 * ME90 - 2 * ME
```

```
## [1] -0.0005765957
```

The width of the 90% confidence interval reduced by .00057 meters.

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.

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

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

null_dist_height %>%
  mutate(atleasttest = ifelse(stat >= .0376,1,0)) %>%
  count(atleasttest)
```

```
## Response: height (numeric)
## Explanatory: physical_3plus (factor)
## Null Hypothesis: independence
## # A tibble: 1 x 2
##   atleasttest     n
##         <dbl> <int>
## 1           0  1000
```

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

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

```
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(mean_height = mean(height, na.rm = TRUE),n = n(),sd_height = sd(height,na.rm = TRUE))
```

```
## # A tibble: 3 x 4
##   physical_3plus mean_height     n sd_height
##   <chr>                <dbl> <int>     <dbl>
## 1 no                    1.67  4404     0.103
## 2 yes                   1.70  8906     0.103
## 3 <NA>                  1.71   273     0.107
```

```
ME_height <- 1.96 * sqrt(.103^2/4404 + .103^2/8906)

c(obs_diff_height - ME_height,obs_diff_height + ME_height)
```

```
## $stat
## [1] 0.03390696
##
## $stat
## [1] 0.04134481
```

H_0 : The average height for students that exercise at least 3 times per week is the same as students that don't. H_A : The average height for students that exercise at least 3 times per week is not the same as students that don't.

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
```

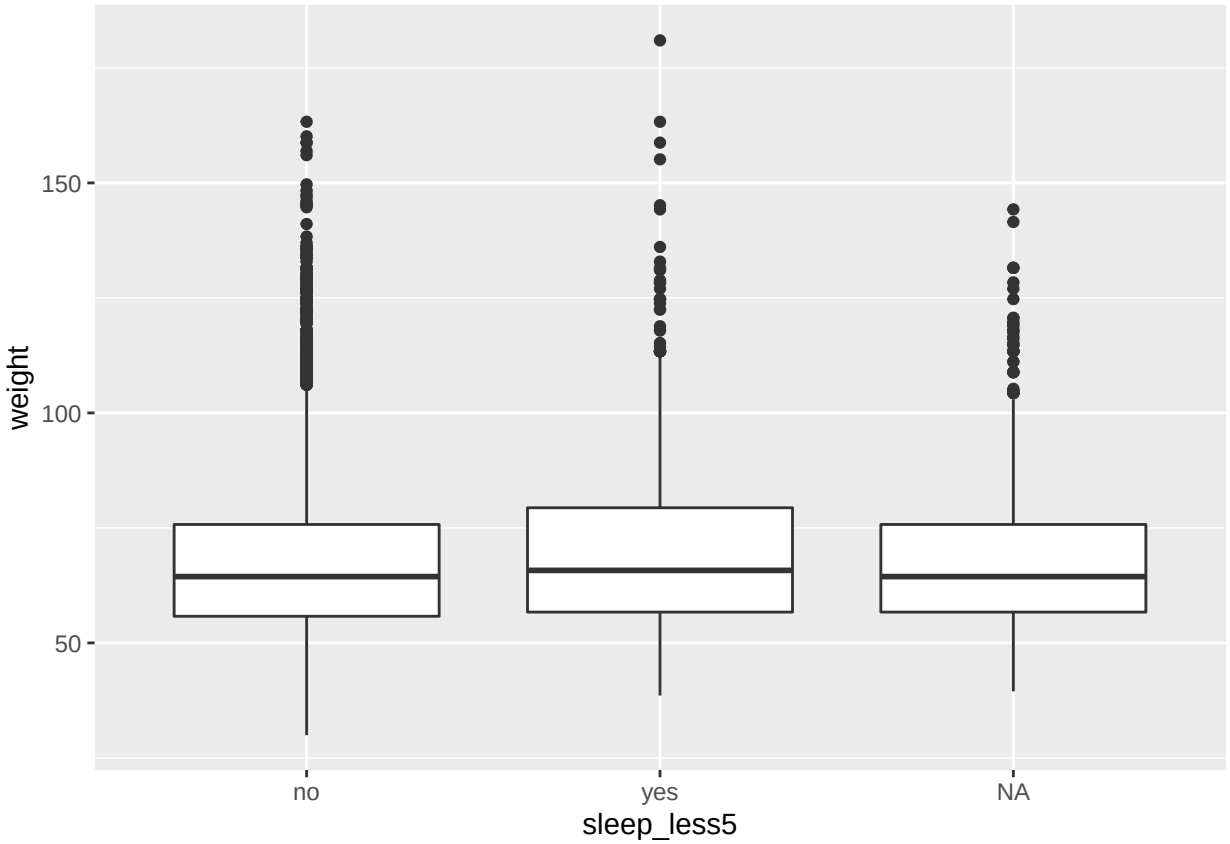
There are 7 different options in the dataset for hours_tv_per_school_day not including the missing observations.

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.

Is the average weight of students that sleep less than 5 hours per night different than those that sleep longer? With a 95% confidence interval, $\alpha = .05$.

```
yrbss <- yrbss %>%
  mutate(sleep_less5 = ifelse(school_night_hours_sleep == '<5', "yes", "no"))

yrbss %>%
  ggplot(aes(x = sleep_less5, y = weight)) + geom_boxplot()
```




```
obs_diff_sleep <- yrbss %>%
  specify(weight ~ sleep_less5) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

null_dist_sleep <- yrbss %>%
  specify(weight ~ sleep_less5) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))

null_dist_sleep %>%
```

```
  mutate(atleasttest = ifelse(stat >= 2.6,1,0)) %>%
  count(atleasttest)
```

```
## Response: weight (numeric)
## Explanatory: sleep_less5 (factor)
## Null Hypothesis: independence
## # A tibble: 1 x 2
##   atleasttest     n
##         <dbl> <int>
## 1           0  1000
```

```
null_dist_sleep %>%
  get_p_value(obs_stat = obs_diff_sleep, direction = "two_sided")
```

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

```
yrbss %>%
  group_by(sleep_less5) %>%
  summarise(mean_weight = mean(weight, na.rm = TRUE),n = n(),sd_weight = sd(weight,na.rm = TRUE),max_wei
```

```
## # A tibble: 3 x 5
##   sleep_less5 mean_weight     n sd_weight max_weight
##   <chr>             <dbl> <int>     <dbl>      <dbl>
## 1 no                 67.7 11370      16.7       163.
## 2 yes                70.3   965      19.5       181.
## 3 <NA>               68.0  1248      16.2       144.
```

```
ME_sleep <- qt(.975,964) * sqrt(16.7^2/11370 + 19.5^2/965)

c(obs_diff_sleep - ME_sleep,obs_diff_sleep + ME_sleep)
```

```
## $stat
## [1] 1.328535
##
## $stat
## [1] 3.867798
```

H_0 : The average weight for students that sleep less than 5 hours per night is the same as students that sleep longer. H_A : The average weight for students that sleep less than 5 hours per night is not the same as students that sleep longer.

The groups are independent and were taken from a random sample. There are some significant outliers that occur 5+ standard deviations from the mean.

The 95% confidence interval (1.33,3.87) does not contain 0 which would indicate that we should reject H_0 and conclude that there is a correlation between weight and sleeping less than 5 hours per night in high school students.
