DS606-Lab7

Code

George Cruz

2020-10-18

#Inference for Numerical Data

Hide

```
library(tidyverse)
library(openintro)
library(infer)
data(yrbss)
```

Exercise 1

What are the cases in this data set? How many cases are there in our sample? Individual Youth behavior. There are 13,583 cases in our sample. Each case represents an individual student's behavior and demographic info.

Hide

```
glimpse(yrbss)
```

```
## Rows: 13,583
## Columns: 13
## $ age                       <int> 14, 14, 15, 15, 15, 15, 15, 14, 15, 15, 15...
## $ gender                    <chr> "female", "female", "female", "female", "f...
## $ grade                     <chr> "9", "9", "9", "9", "9", "9", "9", "9", "9...
## $ hispanic                  <chr> "not", "not", "hispanic", "not", "not", "n...
## $ race                      <chr> "Black or African American", "Black or Afr...
## $ height                    <dbl> NA, NA, 1.73, 1.60, 1.50, 1.57, 1.65, 1.88...
## $ weight                    <dbl> NA, NA, 84.37, 55.79, 46.72, 67.13, 131.54...
## $ helmet_12m                <chr> "never", "never", "never", "never", "did n...
## $ text_while_driving_30d    <chr> "0", NA, "30", "0", "did not drive", "did ...
## $ physically_active_7d      <int> 4, 2, 7, 0, 2, 1, 4, 4, 5, 0, 0, 0, 4, 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, ...
## $ school_night_hours_sleep <chr> "8", "6", "<5", "6", "9", "8", "9", "6", "...
```

Hide

```
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
```

Exercise 2

How many observations are we missing weights from? We're missing 1,004.

Hide

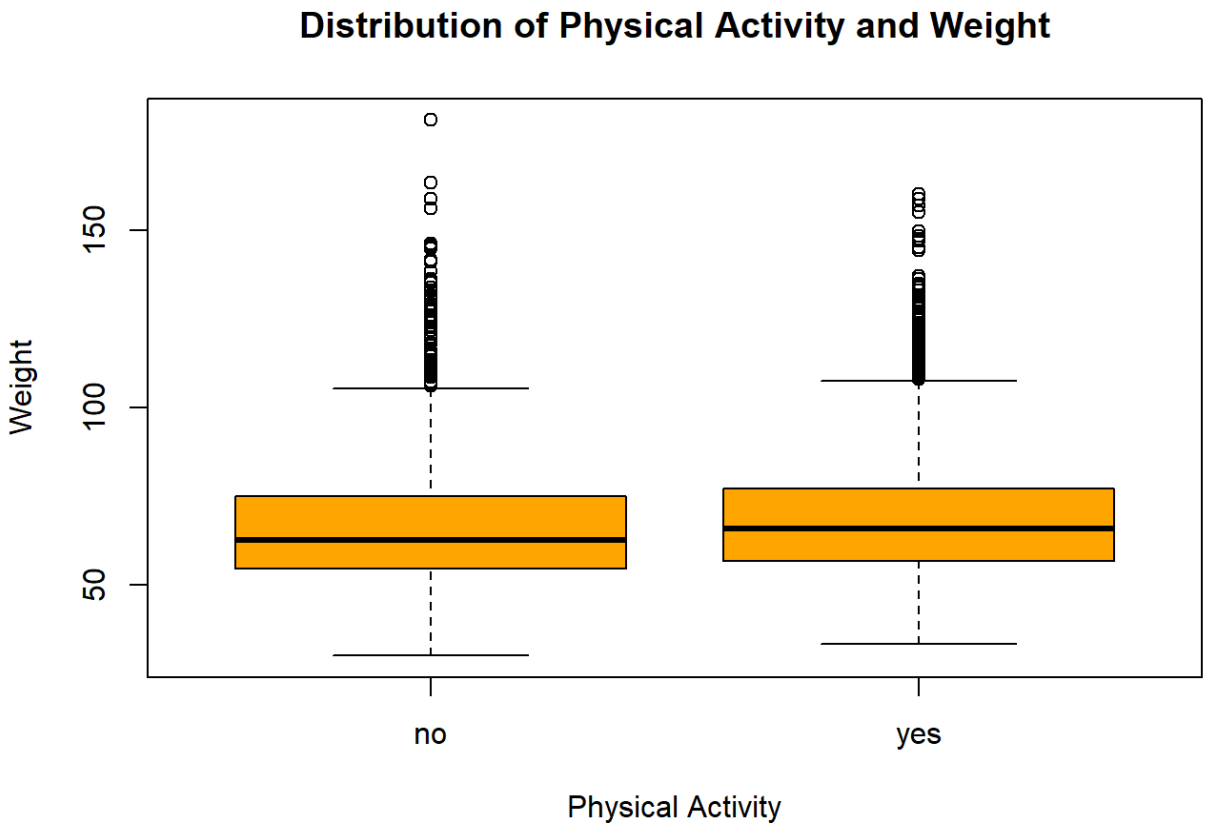
```
yrbss <- yrbss %>%
  mutate(physical_3plus = ifelse(yrbss$physically_active_7d > 2, "yes", "no"))
```

Excercise 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?

Hide

```
boxplot(yrbss$weight ~ yrbss$physical_3plus, col="orange", main="Distribution of Physi
        cal Activity and Weight", ylab="Weight", xlab="Physical Activity")
```



There appears to be minimal change based on the physical activity. The mean weight of the “yes” is slightly higher than the “no” and it appears to be a more normalized distribution. I expected the mean weight of the “no” to be higher based on the physical activity.

Hide

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

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

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

###Exercise 4 The conditions we need for inference on one proportion are: Random: The data comes from a random or randomized experiment. Normal: The sampling distribution of p^ needs to be approximately normal — needs at least 10 expected successes and 10 expected failures. Independent: Individual observations need to be independent.

###Exercise 5 H0: There’s no statistical difference in weight due to exercise. H1: The average weights are different between those who exercise at least 3 times a week.

Hide

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

```
## Warning: Removed 1219 rows containing missing values.
```

Hide

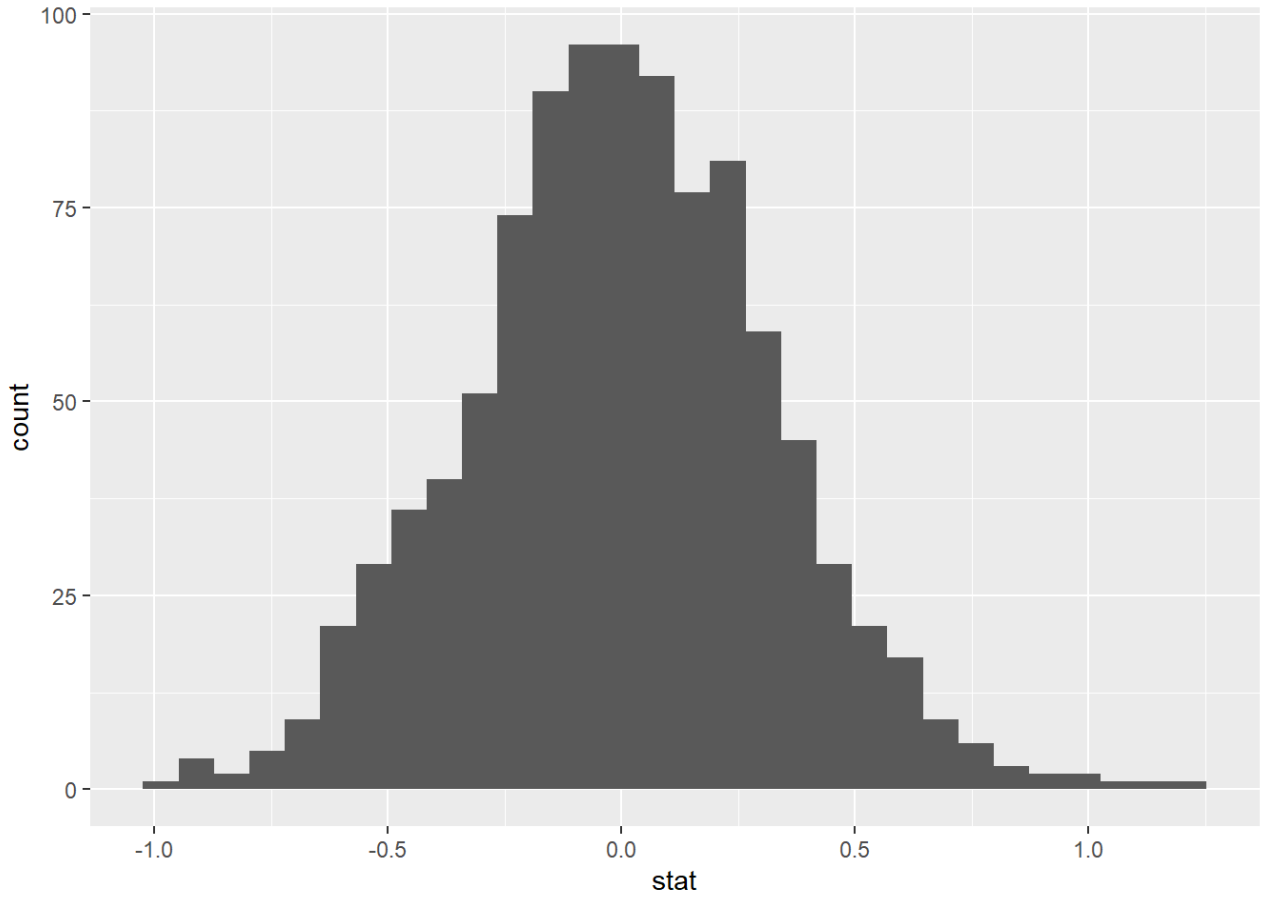
```
null_dist <- yrbss %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
## Warning: Removed 1219 rows containing missing values.
```

Hide

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Exercise 6

Hide

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

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

We found no permutation with a difference with at least obs_stat.

Hide

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

```
## Warning: 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.
```

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

What this warning means is that, although the result in calculations is 0, p-value is an *approximation* and not the real zero. As the number of repetitions grow, the p-value becomes increasingly small or approaches 0.

Exercise 7

Hide

```
p1 <- yrbss %>%
    filter(physical_3plus == "yes") %>%
    summarise(mean_weight = mean(weight, na.rm = TRUE),
              total = n())
p2 <- yrbss %>%
  filter(physical_3plus == "no") %>%
    summarise(mean_weight = mean(weight, na.rm = TRUE),
              total = n())

n <- 13583 - 1004

prop1 <- p1$total/n
prop2 <- p2$total/n

mean_diff <- p1$mean_weight - p2$mean_weight

se <- ( ((prop1 * (1 - prop1)) / n) +  ((prop2 * (1 - prop2)) / n)) ** 0.5
z <- 1.96

me <-z * se

ci_95 <- c(mean_diff - me, mean_diff + me)

ci_95
```

```
## [1] 1.763068 1.786101
```

The 95% confidence interval of the difference between those who exercise regularly and those who do not is between 1.76 and 1.79kg. This means that there is a difference between those who exercise and those who do not and we should reject the null hypothesis.

Exercise 8

Hide

```
z = 1.96
mean_height <- mean(yrbss$height, na.rm = TRUE)
sd_height <- sd(yrbss$height, na.rm = TRUE)
total_height <- yrbss %>%
  filter(!is.na(height)) %>%
  summarise(n = n())

total_height <- total_height$n

ci_low <- mean_height - (z * (sd_height/sqrt(total_height)))
ci_high <- mean_height + (z * (sd_height/sqrt(total_height)))

ci_95 <- c(ci_low, ci_high)
ci_95
```

```
## [1] 1.689411 1.693071
```

The 95% confidence interval is between 1.6894 and 1.6930

Exercise 9

Hide

```
z = 1.64
#mean_height <- mean(yrbss$height, na.rm = TRUE)
#sd_height <- sd(yrbss$height, na.rm = TRUE)
#total_height <- yrbss %>%
#  filter(!is.na(height)) %>%
#  summarise(n = n())

#total_height <- total_height$n

ci_low <- mean_height - (z * (sd_height/sqrt(total_height)))
ci_high <- mean_height + (z * (sd_height/sqrt(total_height)))

ci_90 <- c(ci_low, ci_high)
ci_90
```

```
## [1] 1.689710 1.692772
```

The 90% confidence interval is more narrow: 1.6897 and 1.6927

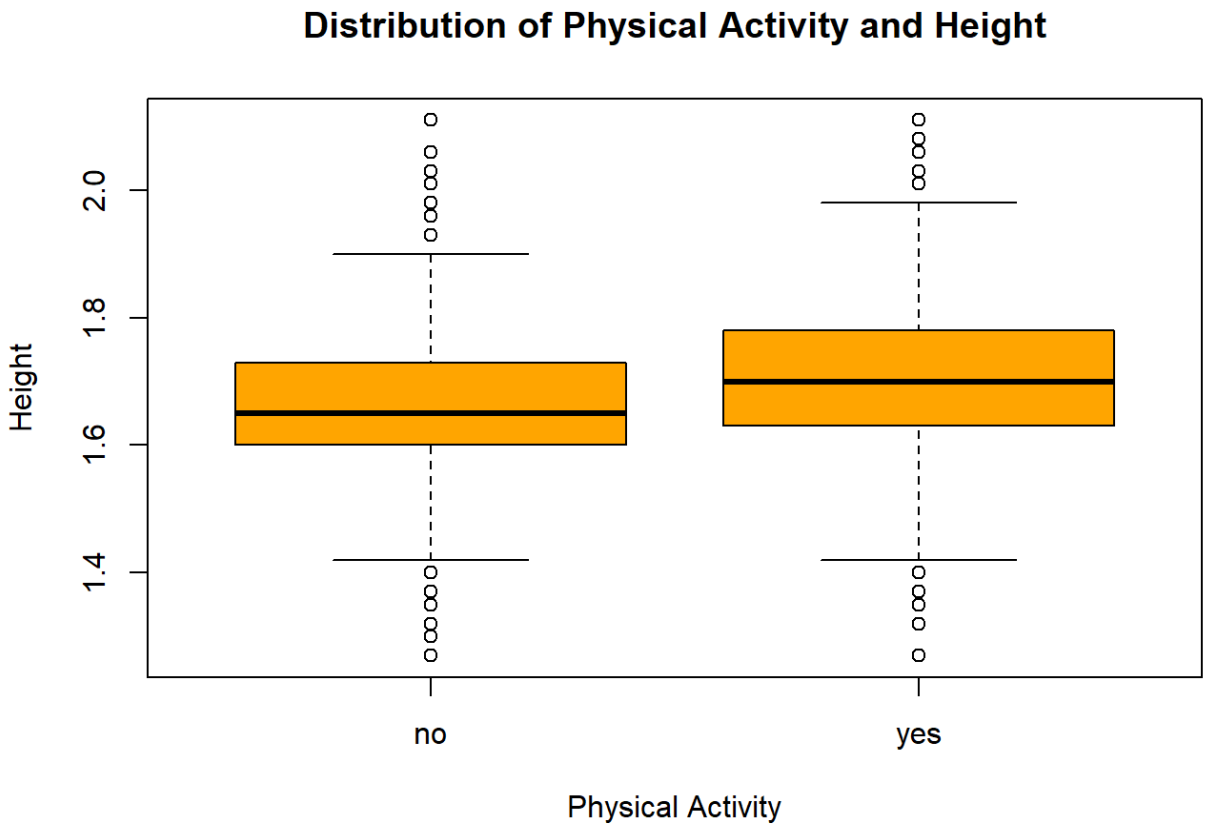
Exercise 10

H0 = There's no statistical difference in height between those who exercise regularly and those who don't.

H1 = There is a difference in height between the two groups.

Hide

```
boxplot(yrbss$height ~ yrbss$physical_3plus, col="orange", main="Distribution of Physi
        cal Activity and Height", ylab="Height", xlab="Physical Activity")
```



There appears to be a difference in height between the two groups.

Hide

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

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

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

Hide

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

```
## Warning: Removed 1219 rows containing missing values.
```

Hide

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

```
## Warning: Removed 1219 rows containing missing values.
```

Hide

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

```
## Warning: 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.
```

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

What this warning means is that, although the result in calculations is 0, p-value is an *approximation* and not the real zero. As the number of repetitions grow, the p-value becomes increasingly small or approaches 0.

Hide

```
p1 <- yrbss %>%
    filter(physical_3plus == "yes") %>%
    summarise(mean_height = mean(height, na.rm = TRUE),
              total = n())
p2 <- yrbss %>%
  filter(physical_3plus == "no") %>%
    summarise(mean_height = mean(height, na.rm = TRUE),
              total = n())

n <- 12364

prop1 <- p1$total/n
prop2 <- p2$total/n

mean_diff <- p1$mean_height - p2$mean_height

se <- ( ((prop1 * (1 - prop1)) / n) +  ((prop2 * (1 - prop2)) / n)) ** 0.5
z <- 1.96

me <-z * se

ci_95 <- c(mean_diff - me, mean_diff + me)

ci_95
```

```
## [1] 0.02605665 0.04919512
```

The 95% confidence interval of the difference between those who exercise regularly and those who do not is between 0.02 and 0.05. This means that there is a difference between those who exercise and those who do not and we should reject the null hypothesis.

Exercise 11

Hide

```
yrbss %>%
  group_by(hours_tv_per_school_day) %>%
  summarise(n = n(), .groups="drop_last")
```

```
## # 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
```

7 options, unless we count NA, then 8.

Exercise 12

Research question: Is there evidence that students weight is affected by the sleep time?

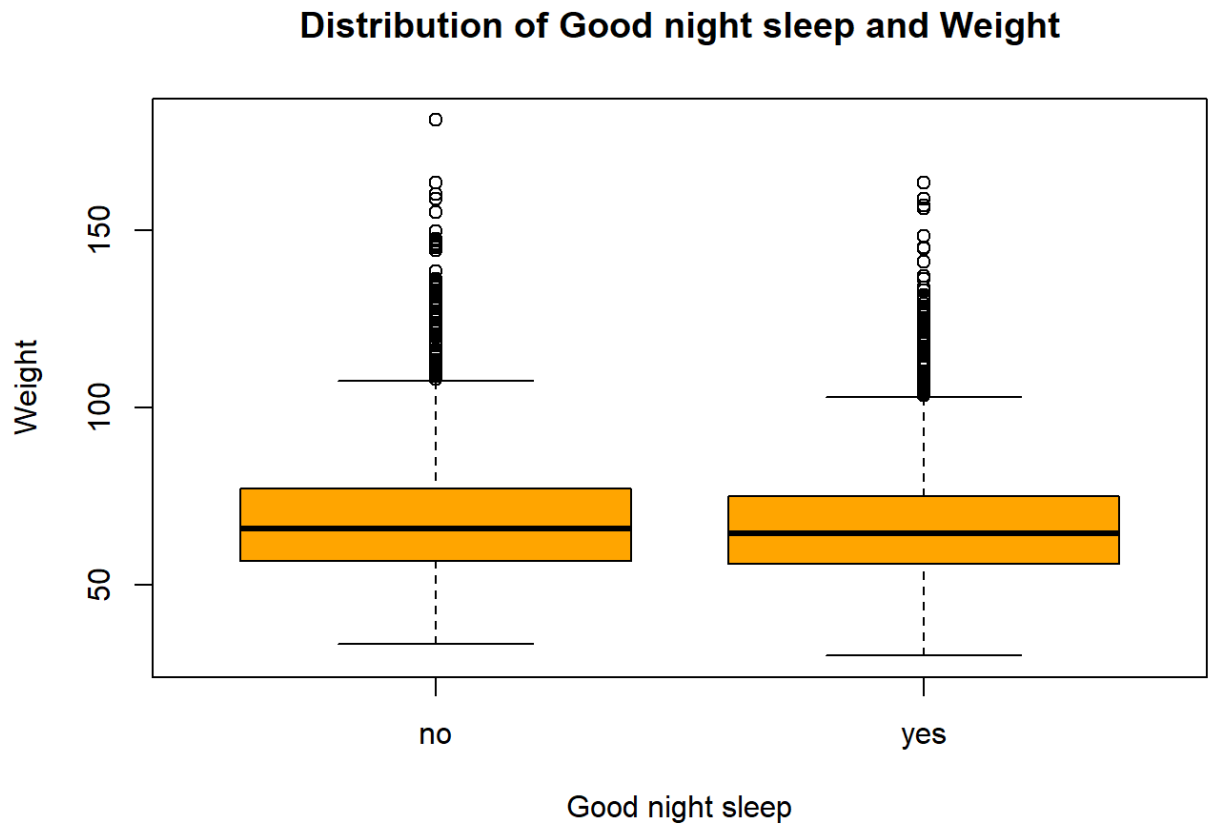
H0: There is no relationship between weight and sleep for students. H1: There is a relationship between weight and sleep.

α Level: .05 (95% confident)

Hide

```
yrbss <- yrbss %>%
  mutate(sleep_7plus = ifelse(yrbss$school_night_hours_sleep > 6, "yes", "no"))

boxplot(yrbss$weight ~ yrbss$sleep_7plus, col="orange", main="Distribution of Good nig
        ht sleep and Weight", ylab="Weight", xlab="Good night sleep")
```



The boxplot points to a slight difference in weight.

Hide

```
weight_no <- yrbss %>%
    filter(sleep_7plus == "no") %>%
    summarise(w_avg = mean(weight, na.rm = TRUE),
              w_sd = sd(weight, na.rm=TRUE),
              total = n())

weight_yes <- yrbss %>%
    filter(sleep_7plus == "yes") %>%
    summarise(w_avg = mean(weight, na.rm = TRUE),
              w_sd = sd(weight, na.rm=TRUE),
              total = n())

mean_diff <- weight_no$w_avg - weight_yes$w_avg

se <-
  sqrt(
  ((weight_no$w_avg**2) / weight_no$total) +
  ((weight_yes$w_avg**2) / weight_yes$total))

ci_low <- mean_diff - 1.96 * se
ci_high <- mean_diff + 1.96 * se

ci_95 <- c(ci_low, ci_high)

ci_95
```

```
## [1] -0.9169308  3.9251705
```

Because the CI is between -0.92 and 3.92 we will accept the null hypothesis that there is no relationship between weight and sleep. These results are somewhat surprising. …