Code ▾

Lab 7: Inference for Numerical Data

Julian Adames-Ng

2022-04-02

Hide

```
library(tidyverse)
library(openintro)
library(infer)

data('yrbss', package='openintro')

yrbss
```

```
## # A tibble: 13,583 × 13
##      age gender grade hispanic race    height weight helmet_12m text_while_driv…
##    <int> <chr>  <chr> <chr>    <chr>    <dbl>  <dbl> <chr>      <chr>
##  1    14 female 9     not      Black …  NA       NA   never      0
##  2    14 female 9     not      Black …  NA       NA   never      <NA>
##  3    15 female 9     hispanic Native…   1.73   84.4 never      30
##  4    15 female 9     not      Black …   1.6    55.8 never      0
##  5    15 female 9     not      Black …   1.5    46.7 did not r… did not drive
##  6    15 female 9     not      Black …   1.57   67.1 did not r… did not drive
##  7    15 female 9     not      Black …   1.65  132.  did not r… <NA>
##  8    14 male   9     not      Black …   1.88   71.2 never      <NA>
##  9    15 male   9     not      Black …   1.75   63.5 never      <NA>
## 10    15 male   10    not      Black …   1.37   97.1 did not r… <NA>
## # … with 13,573 more rows, and 4 more variables: physically_active_7d <int>,
## #   hours_tv_per_school_day <chr>, strength_training_7d <int>,
## #   school_night_hours_sleep <chr>
```

Hide

```
nrow(yrbss)
```

```
## [1] 13583
```

Exercise 1

What are the cases in this data set? How many cases are there in our sample?

-Each row corresponds to a case. There are 13583 cases in this data set.

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?

-There are 1004 NA values which correspond to missing weights.

…

Hide

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

yrbss
```

```
## # A tibble: 13,583 × 14
##      age gender grade hispanic race    height weight helmet_12m text_while_driv…
##    <int> <chr>  <chr> <chr>    <chr>    <dbl>  <dbl> <chr>      <chr>
##  1    14 female 9     not      Black …  NA      NA   never      0
##  2    14 female 9     not      Black …  NA      NA   never      <NA>
##  3    15 female 9     hispanic Native…   1.73   84.4 never      30
##  4    15 female 9     not      Black …   1.6    55.8 never      0
##  5    15 female 9     not      Black …   1.5    46.7 did not r… did not drive
##  6    15 female 9     not      Black …   1.57   67.1 did not r… did not drive
##  7    15 female 9     not      Black …   1.65  132.  did not r… <NA>
##  8    14 male   9     not      Black …   1.88   71.2 never      <NA>
##  9    15 male   9     not      Black …   1.75   63.5 never      <NA>
## 10    15 male   10    not      Black …   1.37   97.1 did not r… <NA>
## # … with 13,573 more rows, and 5 more variables: physically_active_7d <int>,
## #   hours_tv_per_school_day <chr>, strength_training_7d <int>,
## #   school_night_hours_sleep <chr>, physical_3plus <chr>
```

Hide

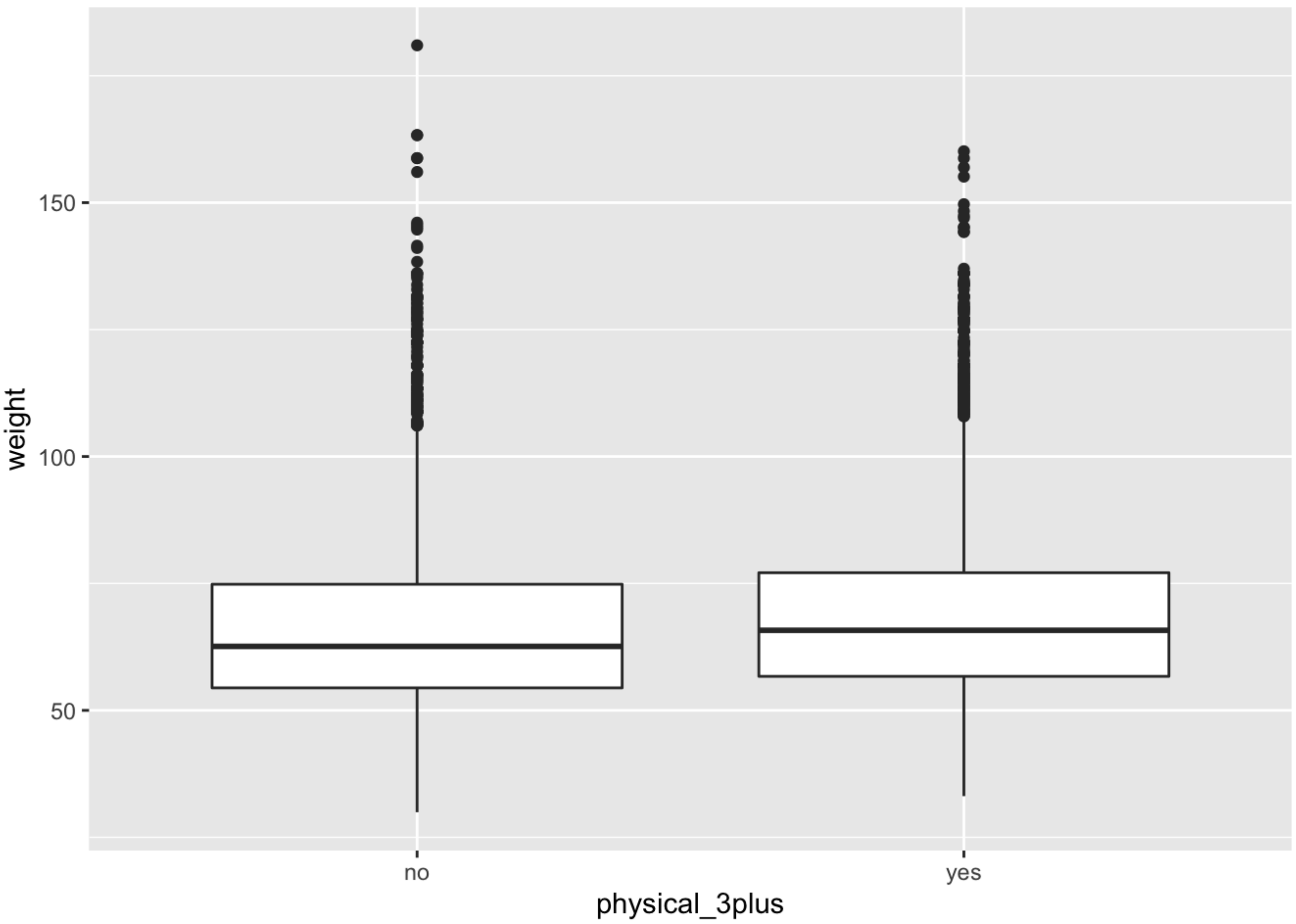
```
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(mean_weight = mean(weight, na.rm = TRUE))
```

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

Hide

```
yrbss$physical_3plus <- as.factor(yrbss$physical_3plus)

yrb_plot <- yrbss %>%
  filter(!is.na(weight),!is.na(physical_3plus))
p <- ggplot(yrb_plot, aes(x = physical_3plus,y = weight))+
  geom_boxplot()
p
```



Exercise 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

```
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 4

How many observations are we missing weights from?

-There are 1004 NA values which correspond to missing weights.

…

Hide

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

yrbss
```

```
## # A tibble: 13,583 × 14
##      age gender grade hispanic race    height weight helmet_12m text_while_driv…
##    <int> <chr>  <chr> <chr>    <chr>    <dbl>  <dbl> <chr>      <chr>
##  1    14 female 9     not      Black …  NA      NA   never      0
##  2    14 female 9     not      Black …  NA      NA   never      <NA>
##  3    15 female 9     hispanic Native…   1.73   84.4 never      30
##  4    15 female 9     not      Black …   1.6    55.8 never      0
##  5    15 female 9     not      Black …   1.5    46.7 did not r… did not drive
##  6    15 female 9     not      Black …   1.57   67.1 did not r… did not drive
##  7    15 female 9     not      Black …   1.65  132.  did not r… <NA>
##  8    14 male   9     not      Black …   1.88   71.2 never      <NA>
##  9    15 male   9     not      Black …   1.75   63.5 never      <NA>
## 10    15 male   10    not      Black …   1.37   97.1 did not r… <NA>
## # … with 13,573 more rows, and 5 more variables: physically_active_7d <int>,
## #   hours_tv_per_school_day <chr>, strength_training_7d <int>,
## #   school_night_hours_sleep <chr>, physical_3plus <chr>
```

Exercise 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.

-Null: Students who are active 3 or more days per week have the SAME mean weight as students who are not active 3 or more days per week.

-Alternative: Students who are active 3 or more days per week have a different mean weight as students who are not active 3 or more days per 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

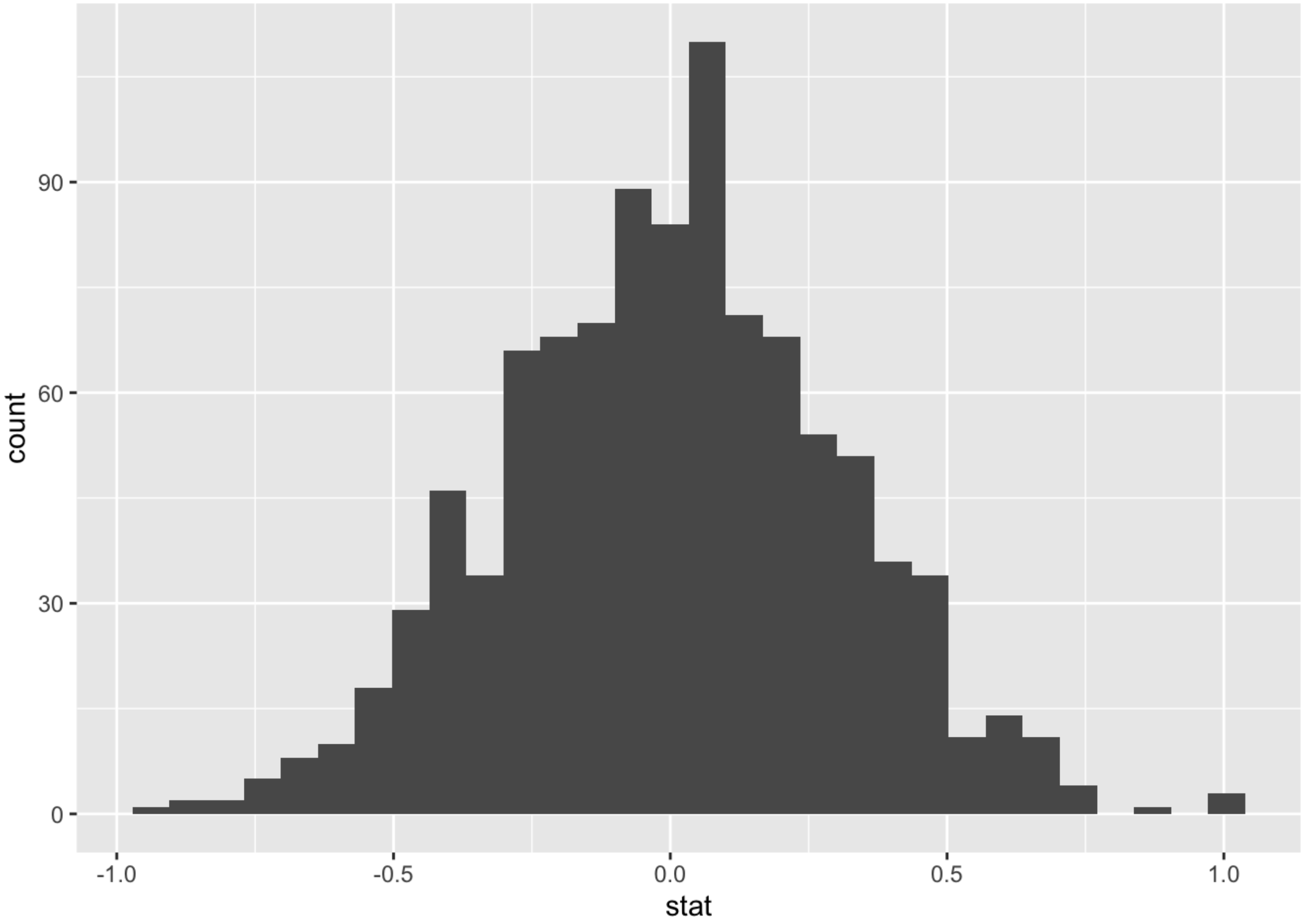
```
null_dist <- yrbss %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
## 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

How many of these null permutations have a difference of at least obs_stat?

-None of the null permutations have a difference of 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 × 1
##   p_value
##     <dbl>
## 1       0
```

Hide

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

```
## # A tibble: 3 × 2
##   physical_3plus sd_weight
##   <chr>              <dbl>
## 1 no                  17.6
## 2 yes                 16.5
## 3 <NA>                17.6
```

Hide

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

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

Hide

```
#sample size n
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(freq = table(weight)) %>%
  summarise(n = sum(freq))
```

```
## `summarise()` has grouped output by 'physical_3plus'. You can override using the
## `.groups` argument.
## # A tibble: 3 × 2
##   physical_3plus     n
##   <chr>          <int>
## 1 no              4022
## 2 yes             8342
## 3 <NA>             215
```

Hide

```
x0_3 <- 66.67389
n0_3 <- 4022
s0_3 <- 17.63805
x_3 <- 68.44847
n_3 <- 8342
s_3 <- 16.47832

z = 1.96

ub_0 <- x0_3 + z*(s0_3/sqrt(n0_3))
lb_0 <- x0_3 - z*(s0_3/sqrt(n0_3))
ub_0
```

```
## [1] 67.219
```

Hide

```
lb_0
```

```
## [1] 66.12878
```

Hide

```
ub <- x_3 + z*(s_3/sqrt(n_3))
lb <- x_3 - z*(s_3/sqrt(n_3))

ub
```

```
## [1] 68.80209
```

Hide

```
lb
```

```
## [1] 68.09485
```

Exercise 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.

-The standard deviation is 17.63805 for students who do are not active at least 3 days per week and 16.47832 for those who are.

-The mean is 66.67389 for students who do are not active at least 3 days per week and 68.44847 for those who are.

-We are 95% confident that students who exercise at least three times a week have a mean weight between 68.09485 kg and 68.80209 kg. Students who don’t exercise at least three times a week have a mean weight between 66.12878 kg and 67.219 kg with 95% confidence.

Hide

```
height_data <- yrbss %>% select(height) %>% na.omit()

meanheight <- mean(height_data$height)
sd3 <- sd(height_data$height)
max3 <- max(height_data$height)
sdheight <- sd(height_data$height)
stderrorheight <- sdheight / sqrt(nrow(height_data))

max3
```

```
## [1] 2.11
```

Hide

```
meanheight + (2.5 * sd3)
```

```
## [1] 1.952984
```

Hide

```
tvalueheight <- qt(.05/2, nrow(height_data) - 1, lower.tail = FALSE)
rightintheight <- meanheight + tvalueheight * stderrorheight
leftintheight <- meanheight - tvalueheight * stderrorheight

leftintheight
```

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

Hide

```
rightintheight
```

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

Exercise 8

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

-The max of height_data is 2.11 and the mean plus 2.5 sd is 1.952984. -The 95% confidence interval is from 1.689411 meters to 1.693071 meters.

Hide

```
tvalueheight <- qt(.1/2, nrow(height_data) - 1, lower.tail = FALSE)
rightintheight <- meanheight + tvalueheight * stderrorheight
leftintheight <- meanheight - tvalueheight * stderrorheight

leftintheight
```

```
## [1] 1.689705
```

Hide

```
rightintheight
```

```
## [1] 1.692777
```

Exercise 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.

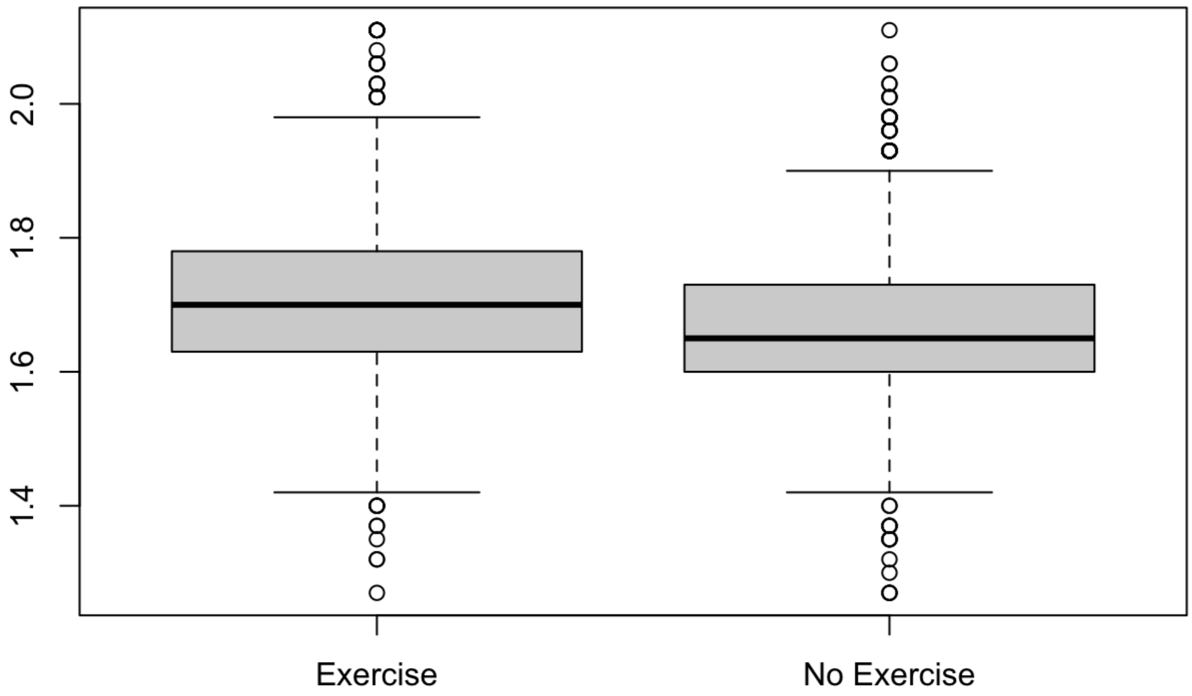
-The 90% confidence interval is from 1.69 meters to 1.693 meters

Hide

```
height_exercise <- yrbss %>%
  filter(physical_3plus == "yes") %>%
  select(height) %>%
  na.omit()

height_noexercise <- yrbss %>%
  filter(physical_3plus == "no") %>%
  select(height) %>%
  na.omit()

# Box Plot
boxplot(height_exercise$height, height_noexercise$height,
        names = c("Exercise", "No Exercise"))
```



Hide

```
mean4 <- mean(height_noexercise$height)
sd4<- sd(height_noexercise$height)
max4 <- max(height_noexercise$height)

mean5 <- mean(height_exercise$height)
sd5 <- sd(height_exercise$height)
max5 <- max(height_exercise$height)

max4
```

```
## [1] 2.11
```

Hide

```
mean4 + (2.5 * sd4)
```

```
## [1] 1.922732
```

Hide

```
max5
```

```
## [1] 2.11
```

Hide

```
mean5 + (2.5 * sd5)
```

```
## [1] 1.961452
```

Hide

```
# Standard Error
meandiff <- mean5 - mean4
stderror <-
  sqrt(
  ((mean5^2) / nrow(height_exercise)) +
  ((mean4^2) / nrow(height_noexercise))
  )

# T-Value and CI
degfreedomht2 <- 4022-1
tvalueht2 <- qt(.05/2, degfreedomht2, lower.tail = FALSE)
rightintervalht <- meandiff + tvalueht2 * stderror
leftintervalht <- meandiff - tvalueht2 * stderror

leftintervalht
```

```
## [1] -0.02552409
```

Hide

```
rightintervalht
```

```
## [1] 0.1007759
```

Hide

```
# P-Value
pvalueht2 <- 2*pt(tvalueht2,degfreedomht2, lower.tail = FALSE)

pvalueht2
```

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

Exercise 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.

-The max of height_noexercise is 2.11 and the mean plus 2.5 sd is 1.922732.

-The max of height_exercise is 2.11 and the mean plus 2.5 sd is 1.961452.

-The 95% confidence interval is from -0.02552409 to 0.1007759.

-The P-Value is 0.05….”If P is high, null will fly,” so we fail to reject the null hypothesis.

Hide

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

```
## # A tibble: 8 × 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
```

Exercise 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 different options in the dataset hours_tv_per_school_day, not including those labeled “NA”.

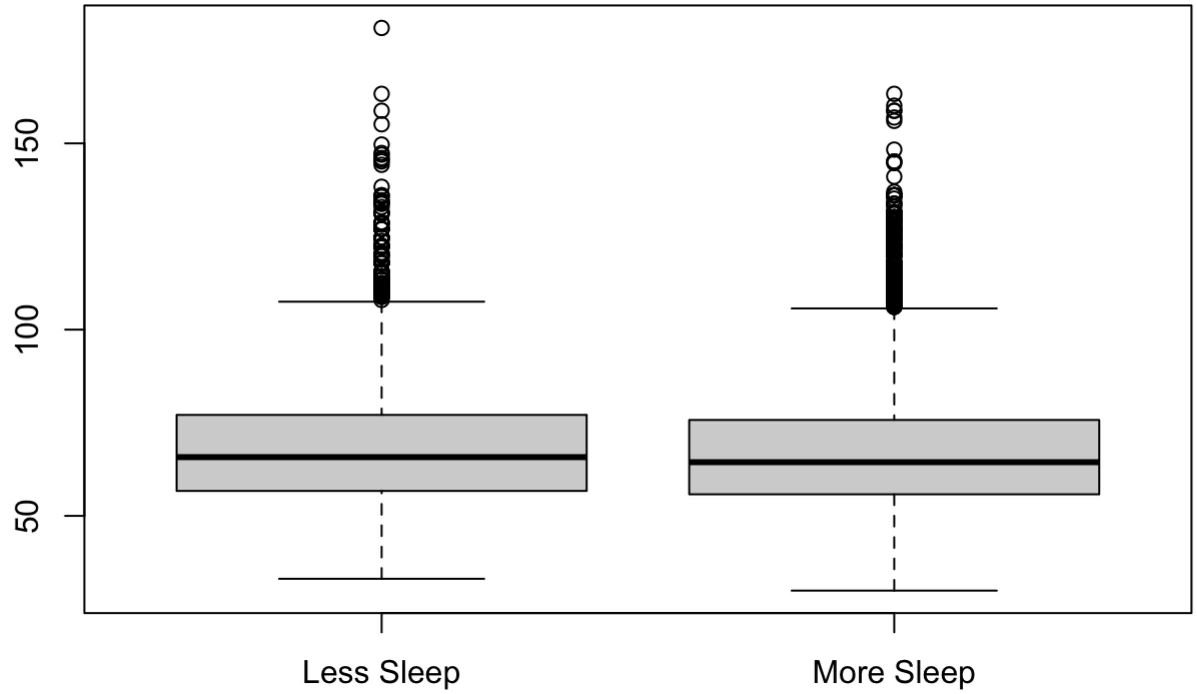
Hide

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

weight_less <- yrbss %>%
  select(weight, sleep_less) %>%
  filter(sleep_less == "yes") %>%
  na.omit()

weight_more <- yrbss %>%
  select(weight, sleep_less) %>%
  filter(sleep_less == "no") %>%
  na.omit()

boxplot(weight_less$weight, weight_more$weight,
        names = c("Less Sleep", "More Sleep"))
```



Hide

```
mn <- mean(weight_less$weight)
sd <- sd(weight_less$weight)
max <- max(weight_less$weight)
max
```

```
## [1] 180.99
```

Hide

```
mn1 <- mean(weight_more$weight)
sd2 <- sd(weight_more$weight)
max2 <- max(weight_more$weight)

mean_diff <- mn1 - mn
sd <-
  sqrt(
  ((mn1^2) / nrow(weight_more)) +
  ((mn^2) / nrow(weight_less))
  )

df <- 2492-1
t <- qt(.05/2, df, lower.tail = FALSE)

upper_ci <- mean_diff + t * sd
lower_ci <- mean_diff - t * sd

c(lower_ci ,upper_ci)
```

```
## [1] -4.666506  1.442799
```

Hide

```
p_value <- 2*pt(t,df, lower.tail = FALSE)
p_value
```

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

Exercise 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 that students who weigh more than the mean weight sleep more than students who are lighter than the mean weight?

Null: There is a relationship between weight and sleep

Alternative: There is no relationship between weight and sleep

95% confident level

Since the P-value equals alpha, we fail to reject the null hypothesis. We can’t determine that a relationship exists between weight and sleep.