Data_606_Lab_7_Inference _for_numerical_data.rmd

Enid Roman

2022-10-29

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

```
library(infer)
```

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

```
?yrbss
```

```
## starting httpd help server ... done
```

Exercise 1

```

glimpse(yrbss)
```

The cases in this dataset is CDC's Youth Risk Behavior Surveillance System. There are 13,583 cases, which is the same amount of the rows in this dataset on 13 variables, age, gender, grade, hispanic or not, race, height, weight, How often did they wear a helmet when biking in the last 12 months?, How many days did they text while driving in the last 30 days?, How many days were they physically active for 60+ minutes in the last 7 days?, How many hours of TV do they typically watch on a school night?, How many days did they do strength training (e.g. lift weights) in the last 7 days?, and How many hours of sleep do they typically get on a school night?

```
## 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"~
```

```
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

```
sum(is.na(yrbss$weight))
```

There are 1004 observations that are missing wights.

```
## [1] 1004
```

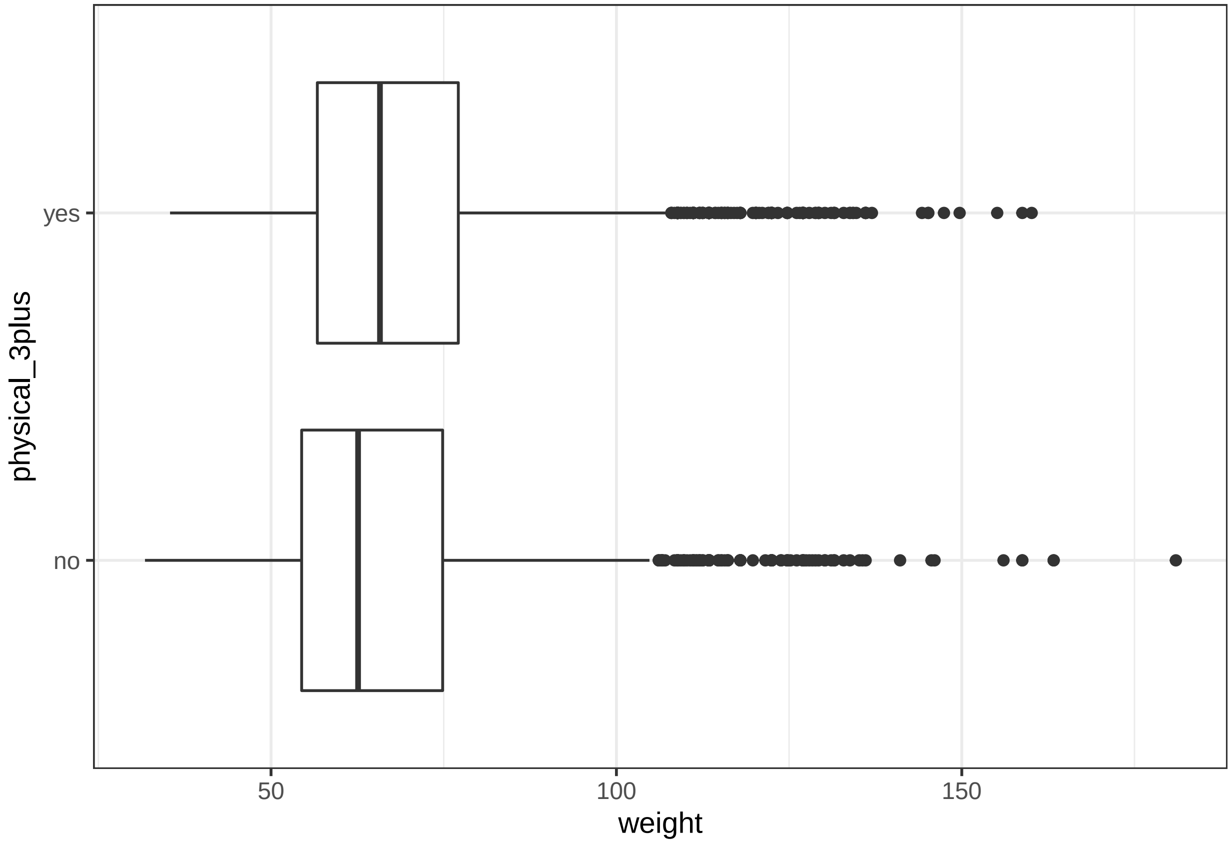
```
yrbss <- yrbss %>%
  mutate(physical_3plus = ifelse(yrbss$physically_active_7d > 2, "yes", "no"))
```

Exercise 3

```
yrbss_plot <- yrbss %>%
  mutate(physical_3plus = ifelse(yrbss$physically_active_7d > 2, "yes", "no")) %>%
  na.omit()
ggplot(yrbss_plot, aes(x=weight, y=physical_3plus)) + geom_boxplot() + theme_bw()
```

Yes, there seems to be a relationship between physical activity and weight. The weights are pretty similar for those that are physcial active in 3 days and those that are not. There is a higher concentration of weight measures clustered together for those who exercise than those who don't. There are more outliers in weight for those that don't exercise. The

data is more normally distribured for those who exercise than for those who don't exercise.



```
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
```

Exercise 4

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

Yes, all conditions is necessary for inference to be satisfied. The two conditions for inference are normality and independence. According to the data, we can see that it is a representative

sample of many students across national, state,and tribal territories. All students are independent of each other. According to the box plots above, we can also see that the data appears to be normally distributed. All three conditions for inference on the difference between two means are met.

```
## ‘summarise()‘ has grouped output by ’physical_3plus’. You can override using
## the ‘.groups‘ argument.
```

```
## # A tibble: 3 x 2
##   physical_3plus     n
##   <chr>          <int>
## 1 no              4022
## 2 yes             8342
## 3 <NA>             215
```

Exercise 5

H0: Students who are physically active 3 or more days per week have the same average weight as those who are not physically active 3 or more days per week.

```
set.seed(100)
obs_diff <- yrbss %>%
  filter(physical_3plus != "NA")%>%
  specify(weight ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

HA: Students who are physically active 3 or more days per week have a different average weight when compared to those who are not physically active 3 or more days per week.

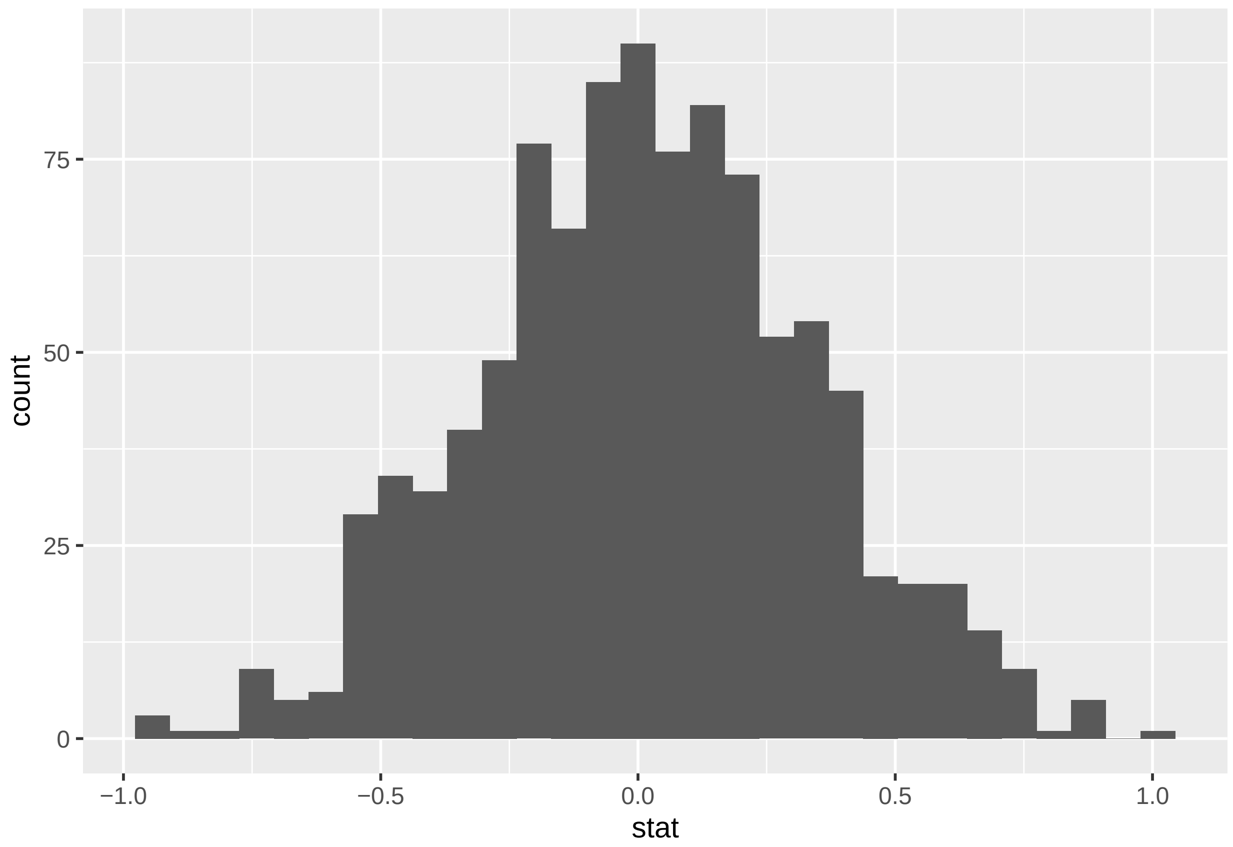
```
## Warning: Removed 946 rows containing missing values.
```

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

```
## Warning: Removed 946 rows containing missing values.
```

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

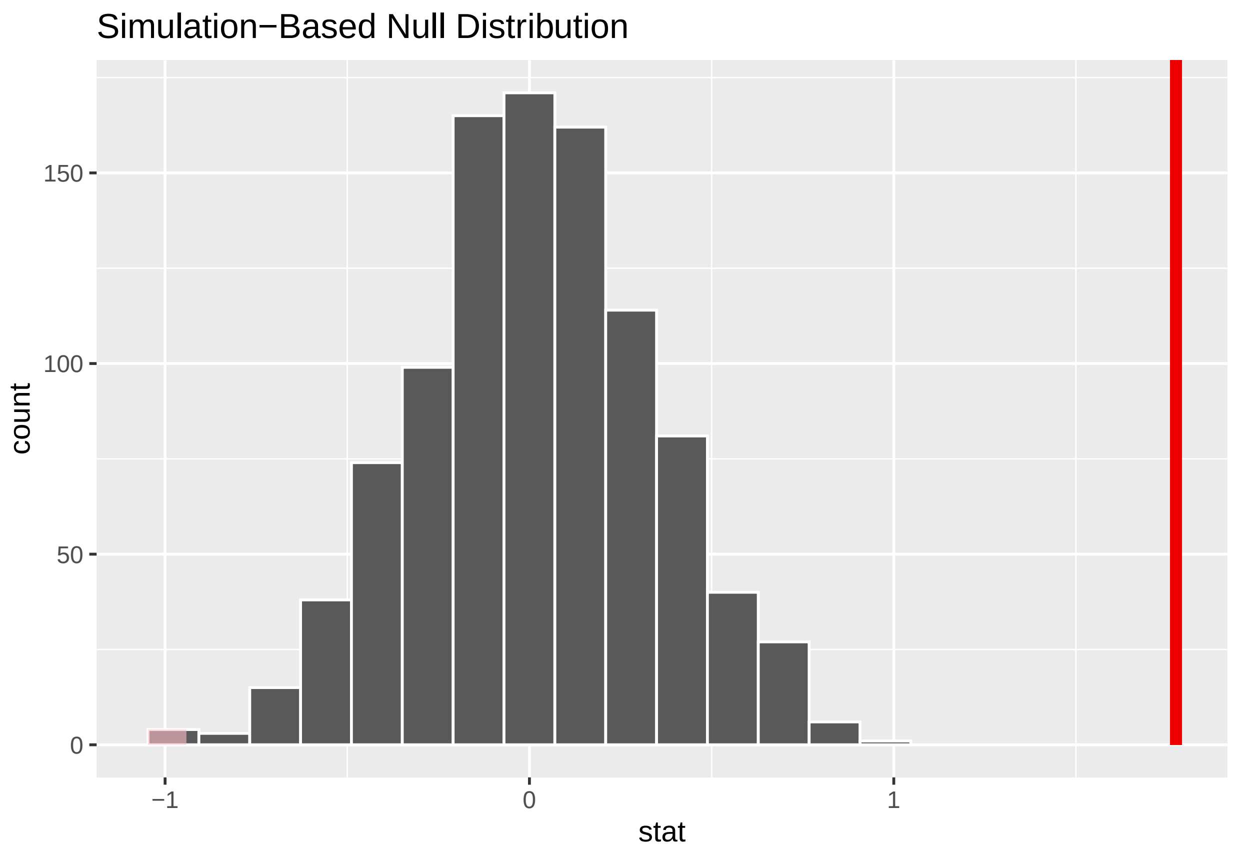
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



#### Excercise 6

None of the values are greater than the obs_diff_stat.

```
visualize(null_dist) +
  shade_p_value(obs_stat = obs_diff, direction = "two_sided")
```



```
obs_diff_val<-obs_diff$stat[1]

null_list<-as.list(null_dist$stat)
null_abs<-lapply(null_list, FUN=function(x){abs(x)})

null_dist%>%
  summarise(mean= mean(stat, na.rm=TRUE))
```

```
## # A tibble: 1 x 1
##     mean
##    <dbl>
## 1 0.0170
```

```
null_dist%>%
  filter(stat>obs_diff_val)
```

```
## Response: weight (numeric)
## Explanatory: physical_3plus (factor)
## Null Hypothesis: independence
## # A tibble: 0 x 2
## # ... with 2 variables: replicate <int>, stat <dbl>
```

```
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
```

Exercise 7

```
# SD

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

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.

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

```
# Mean

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
```

```
# 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 x 2
##   physical_3plus     n
##   <chr>          <int>
## 1 no              4022
## 2 yes             8342
## 3 <NA>             215
```

```
mean_not_active <- 66.67389
n_not_active <- 4022
sd_not_active <- 17.63805

mean_active <- 68.44847
n_active <- 8342
sd_active <- 16.47832

z = 1.96

#CI for those not active
upper_ci_not_act <- mean_not_active + z*(sd_not_active/sqrt(n_not_active))

lower_ci_not_act <- mean_not_active - z*(sd_not_active/sqrt(n_not_active))

#CI for those active
upper_ci_act <- mean_active + z*(sd_active/sqrt(n_active))

lower_ci_act <- mean_active - z*(sd_active/sqrt(n_active))
```

```
c("Those not active:", lower_ci_not_act, upper_ci_not_act)
```

```
## [1] "Those not active:" "66.1287781694363"  "67.2190018305637"
```

```
c("Those active:", lower_ci_act, upper_ci_act)
```

```
## [1] "Those active:"    "68.0948523684916" "68.8020876315084"
```

Exercise 8

```


tb <- as.data.frame(table(yrbss$height))
freq <- sum(tb$Freq)

mean_height <- mean(yrbss$height, na.rm = TRUE)
sd_height <- sd(yrbss$height, na.rm = TRUE)
sample_height <- yrbss %>%
  summarise(freq = table(height)) %>%
  summarise(n = sum(freq, na.rm = TRUE))

height_upper <- mean_height + z*(sd_height/sqrt(sample_height))

height_lower <- mean_height - z*(sd_height/sqrt(sample_height))

c(height_lower,height_upper)
```

At a 95% Confidence Interval, we can say the average height of the students of the population is between 1.689m and 1.693m.

```
## $n
## [1] 1.689411
##
## $n
## [1] 1.693071
```

Exercise 9 (see below)

```
z2 <- 1.645

height_upper_2 <- mean_height + z2*(sd_height/sqrt(sample_height))

height_lower_2 <- mean_height - z2*(sd_height/sqrt(sample_height))

c(height_lower_2 ,height_upper_2)
```

```
## $n
## [1] 1.689705
##
## $n
## [1] 1.692777
```

```
x <- abs(height_lower_2 - height_lower)
y <- abs(height_upper_2 - height_upper)

c(x,y)
```

Difference between both Conference Intervals:

```
## $n
## [1] 0.0002940511
##
## $n
## [1] 0.0002940511
```

Exercise 10

H0: Students who are physically active 3 or more days per week have the same average height as those who are not physically active 3 or more days per week.

HA: Students who are physically active 3 or more days per week have a different average height when compared to those who are not physically active 3 or more days per week.

At a confidence level of 95%, the average height of students who are physically active at least 3 days/week, is between ~1.701m and 1.705m. The average height of students who are not physically active is between ~1.663m and ~1.670m.

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

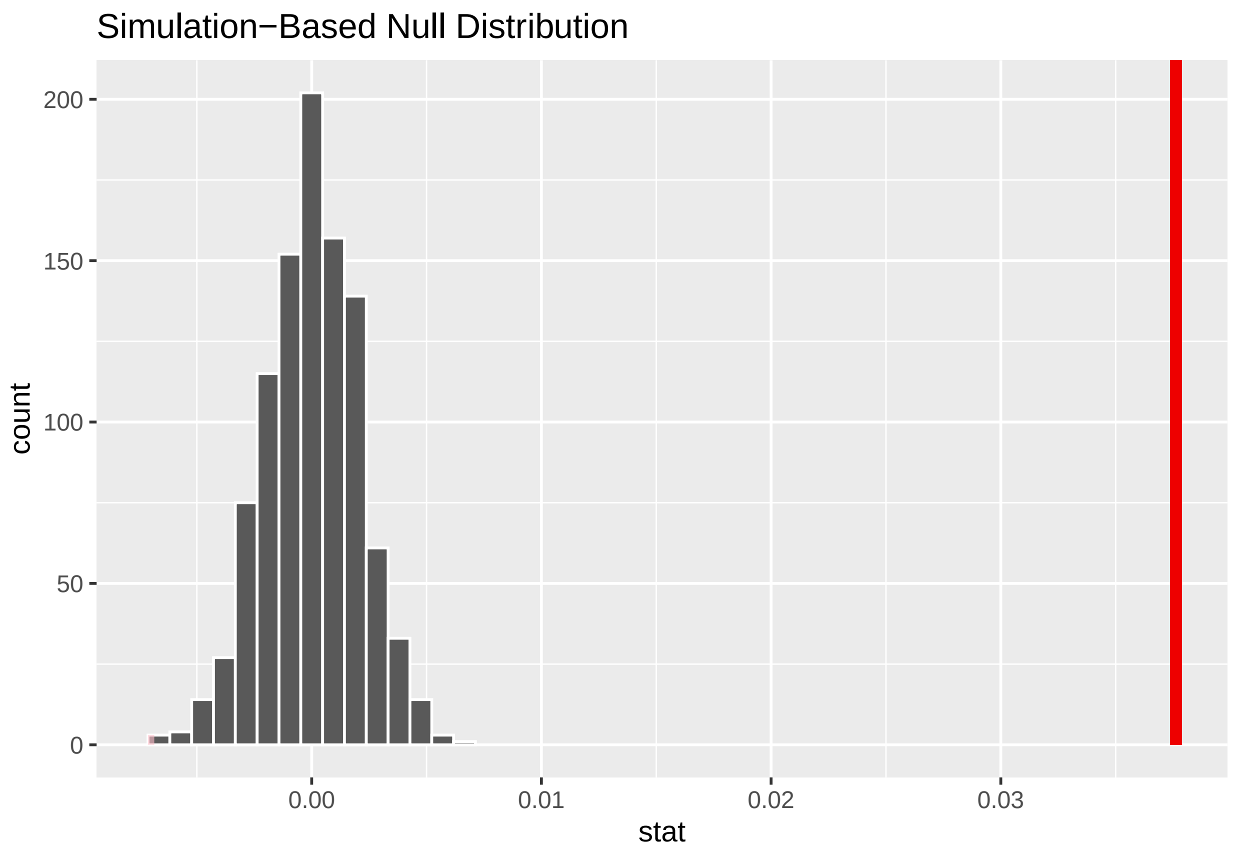
The P value is less than 0.05, we reject the null hypothesis. Thus, there is a difference of those who are physically active at least 3x/week.

```
## Warning: Removed 946 rows containing missing values.
```

```r
set.seed(100)
null_dist_hgt <- yrbss %>%
  filter(physical_3plus != "NA")%>%
  specify(height ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
## Warning: Removed 946 rows containing missing values.
```

```r
visualize(null_dist_hgt) +
  shade_p_value(obs_stat = obs_diff_hgt, direction = "two_sided")
```



```
null_dist_hgt %>%
  get_p_value(obs_stat = obs_diff_hgt, 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
```

```
# Non active
mean_height_2 <- 1.6665
samples_2 <- 4022
sd_height_2 <- 0.1029

# Active
mean_height_2a <- 1.7032
samples_2a <- 8342
sd_height_2a <- 0.1033

z_2 = 1.96
```

```
#Non Active
upper_non_active <- mean_height_2 + z*(sd_height_2/sqrt(samples_2))

lower_non_active <- mean_height_2 - z*(sd_height_2/sqrt(samples_2))

c("Non-active heights:", lower_non_active, upper_non_active)
```

```
## [1] "Non-active heights:" "1.66331982943891"    "1.66968017056109"
```

```
#Active
upper_active <- mean_height_2a + z*(sd_height_2a/sqrt(samples_2a))

lower_active <- mean_height_2a - z*(sd_height_2a/sqrt(samples_2a))

c("Active heights:", lower_active, upper_active)
```

```
## [1] "Active heights:"  "1.70098322660715" "1.70541677339285"
```

Exercise 11

```


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

There are 7 different options for the 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
```

Exercise 12

Is there evidence that students who are heavier than the mean weight sleep more than students who weight less than the mean weight?

HO: There is a relationship between weight and sleep

HA: There is no relationship between weight and sleep

95% confident level

Conditions: ####-Independent sample-yes - Normality - yes

Results:

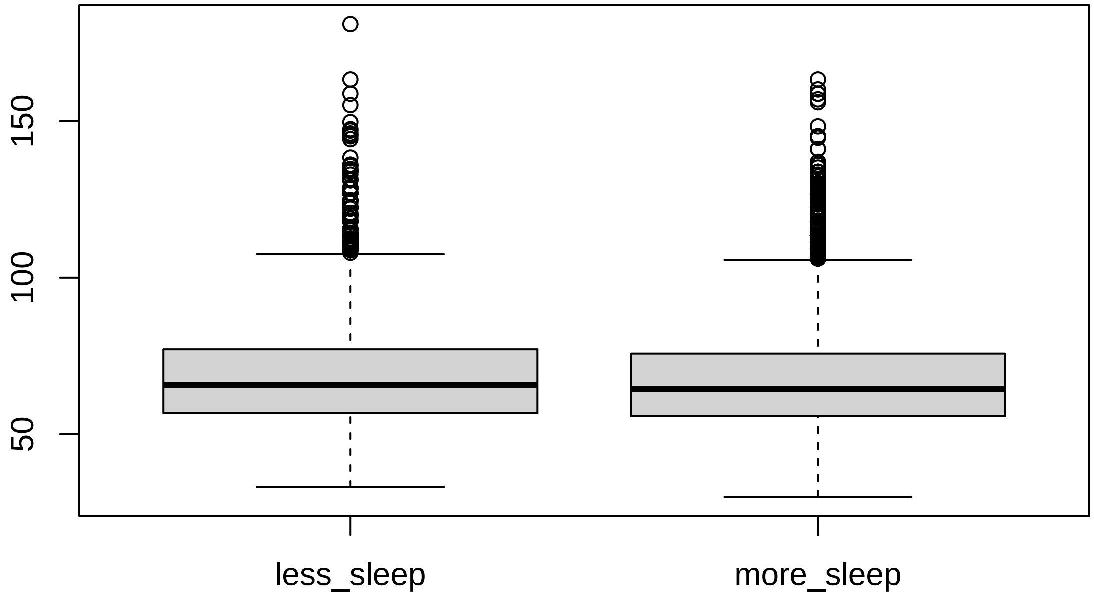
```
yrbss <- yrbss %>%
  mutate(sleep_less = ifelse(yrbss$school_night_hours_sleep < 6, "yes", "no"))

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"))
```

Because our P-value is equal to our alpha, 0.05, we cannot reject the null hypothesis. Therefore, we cannot determine there exists a relationship between weight and sleep.



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

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

```
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
```

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

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