Inference for numerical data

Heleine Fouda

2023-10-22

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?

Insert your answer here

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?

Insert your answer here Weights is missing from 1004 observations. And there are a total of 9476 missing observations.

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

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

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

```
## [1] 9476
```

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.

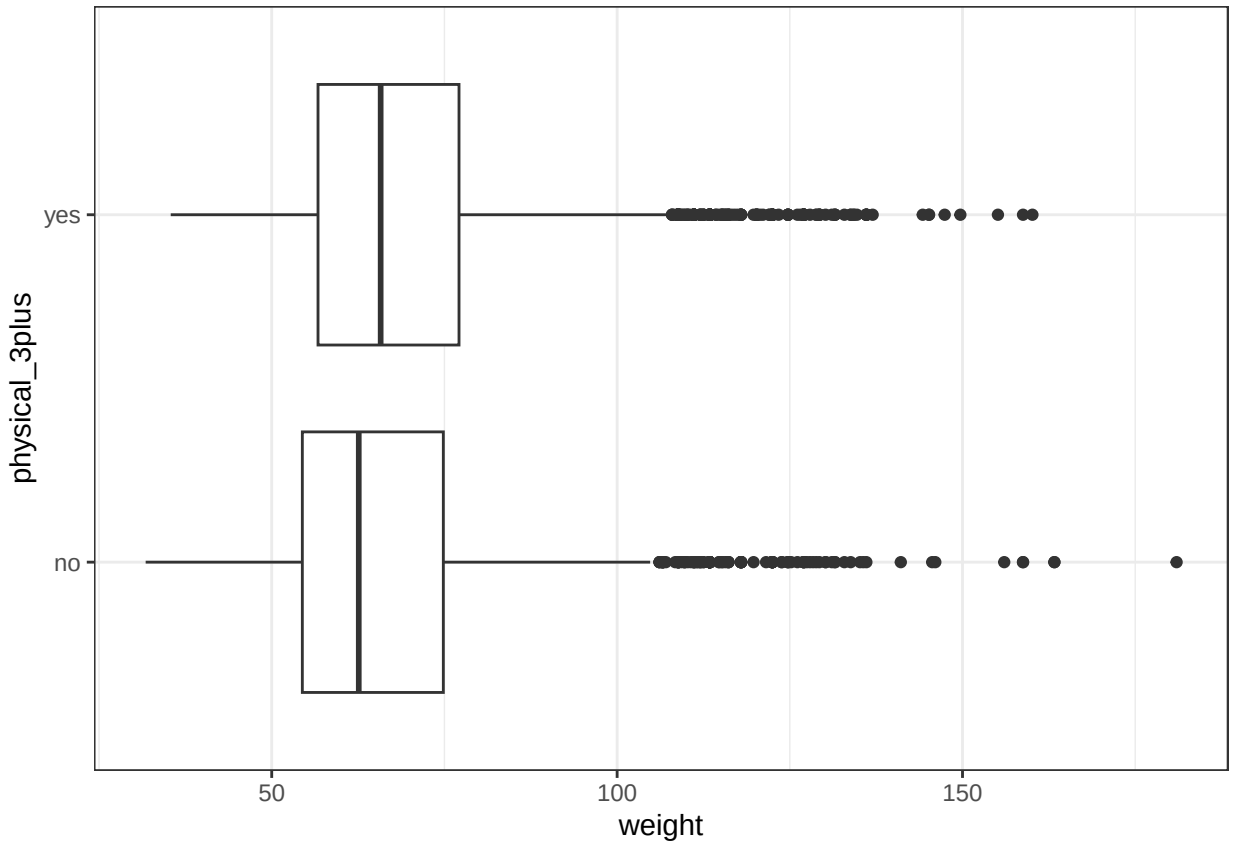
First, let's create a new variable `physical_3plus`, which will be coded as either "yes" if they are physically active for at least 3 days a week, and "no" if not.

```
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?

Insert your answer here Yes, there is a relationship between the two variables and the boxplot below reveals a comparatively small decrease in weight for those who exercise at least 3 days a week.

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



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()`.

Insert your answer here Inference is only possible when one deals with a normal distribution and when there's independence in the outcomes. These two conditions are met in the ybrss data base. The ybrss sample size counts thousands of american youths, which is clearly above the normal distribution's requirement of a sample size that is at least 30. Also, the youths have been randomly selected from all layers of the American society, thus meeting the randomization criterium. On the other hand, the test of independence is met by the

fact that the results of one youth does necessarily predict the result for another youth within the data set.

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.

Insert your answer here H0: There is no difference in the average weight for those who exercise at least 3 times a week and those who don't.

HA: There is a difference in the average weight for those who exercise at least 3 times a week and those who 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 %>%
  drop_na(physical_3plus) %>%
  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`.

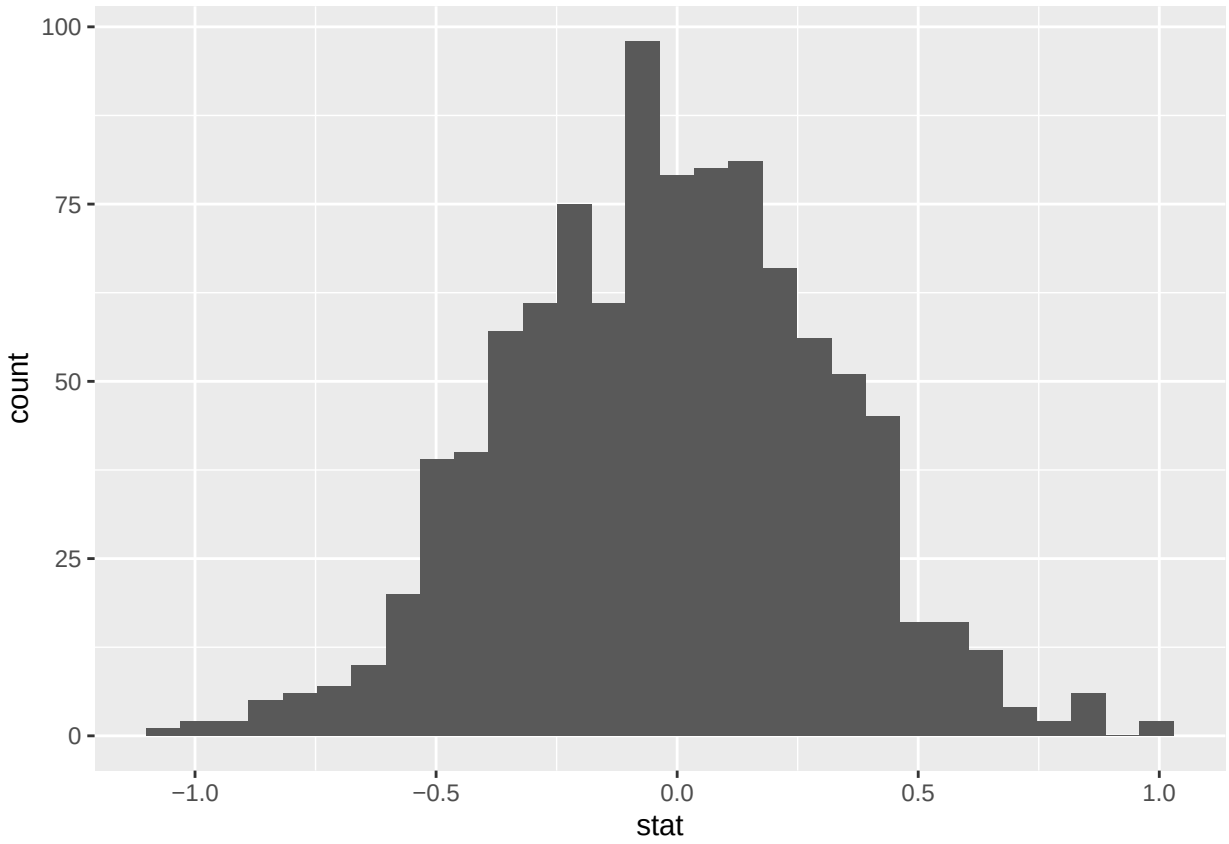
```
null_dist <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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:

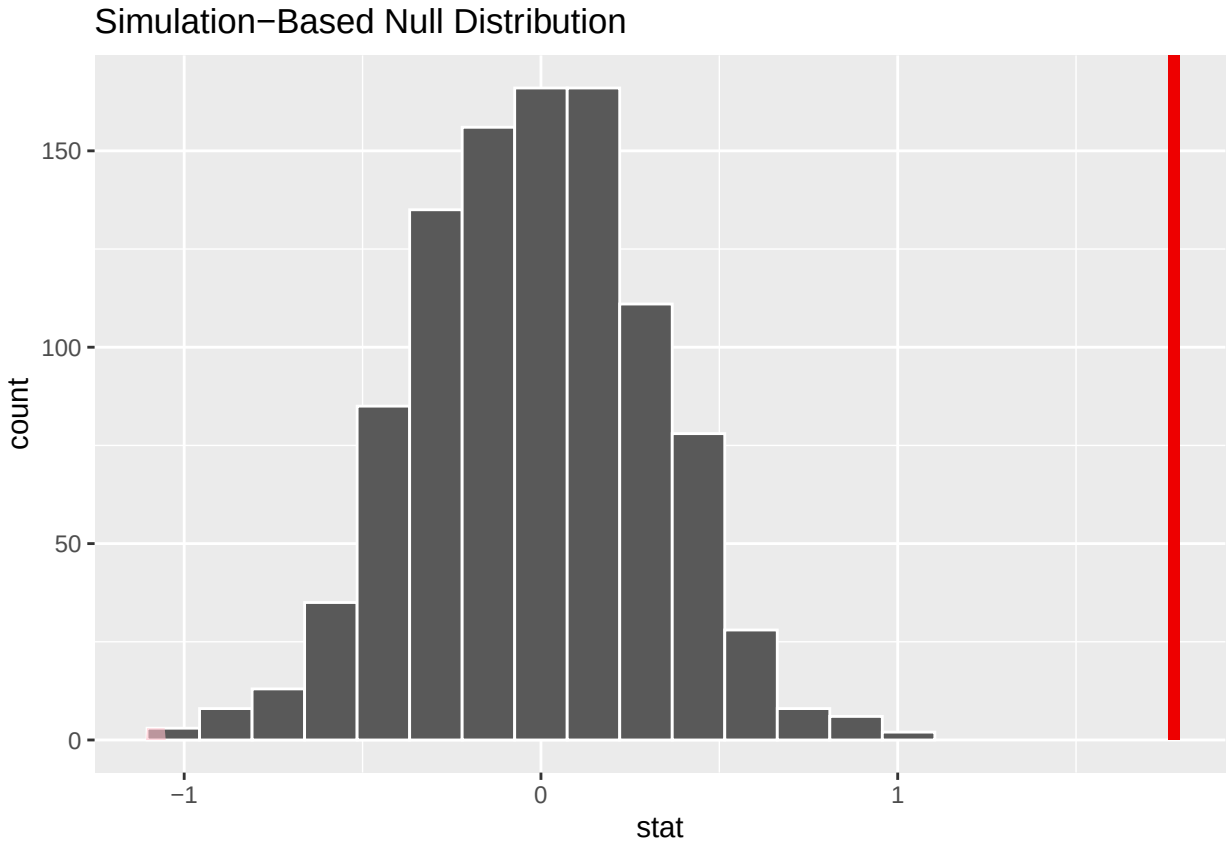
```
ggplot(data = null_dist, aes(x = stat)) +
  geom_histogram()
```



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

Insert your answer here

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



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.

```
# Let's first explore the sample size
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(freq = table(weight)) %>%
  summarise(n = sum(freq))
```

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

No physical_3plus = 4022 Yes hysical_3plus = 8342

```r
# Let's find the mean of the distribution
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
```

Mean = 66.6738 for those not practicing physical activity at least 3 times a week Mean = 68.4484 for those having physical activity at least 3 times a week

```r
# Let's find the standard deviation
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(sd_weight = sd(weight, na.rm = TRUE))
```

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

sd = 17.6380 for No physical activity at least 3 times a week sd = 16.4783 for those having physical activity at least 3 times a week

```r
# Let's calculate confidence intervals

mean_nottraining <- 66.6738
n_nottraining <- 4022
sd_nottraining <- 17.6380

mean_training <- 68.4484
n_training<- 8342
sd_training <- 16.4783

z = 1.96

#CI for those not training
upper_ci_nottraining <- mean_nottraining + z*(sd_nottraining/sqrt(n_nottraining))

lower_ci_nottraining <- mean_nottraining - z*(sd_nottraining/sqrt(n_nottraining))

#CI for those training
upper_ci_training <- mean_training + z*(sd_training/sqrt(n_training))

lower_ci_training <- mean_training - z*(sd_training/sqrt(n_training))
```

```r
# View CI for those not training
c("Those not training:", lower_ci_nottraining, upper_ci_nottraining)
```

```
## [1] "Those not training:" "66.1286897147087"    "67.2189102852913"
```

```r
# View CI for those training
c("Those training:", lower_ci_training, upper_ci_training)
```

```
## [1] "Those training:" "68.094782797683" "68.802017202317"
```

More Practice

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

Insert your answer here One can say, within a 95% confidence interval that the average height of the students in the American population is between 1.689m and 1.693m.

```r
# Calculate CI for average height in meters
tab <- as.data.frame(table(yrbss$height))
freq <- sum(tab$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)
```

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

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.

Insert your answer here

```r
# Calculating a new confidence interval at 90% confidence interval
z <- 1.645

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

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

c(height_lower_ci ,height_upper_ci)
```

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

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.

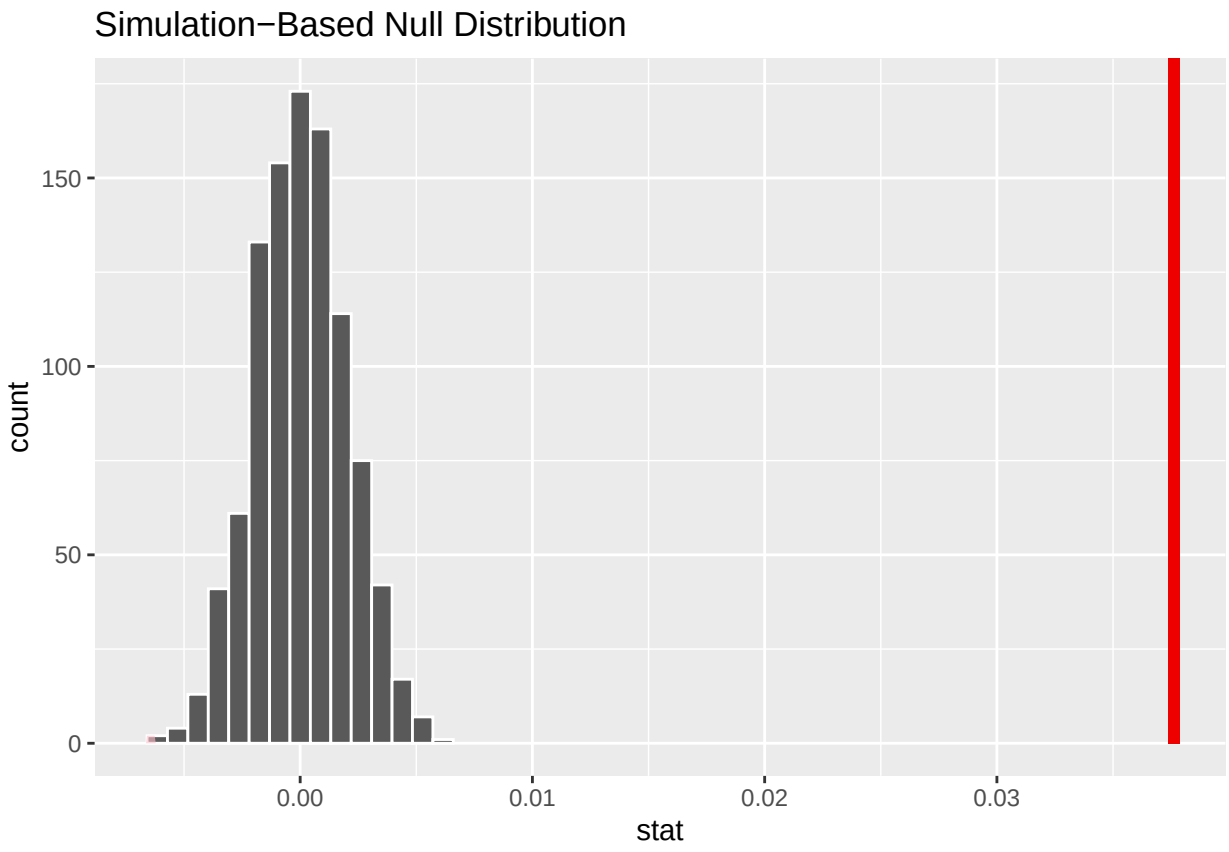
Insert your answer here HO: There is no difference in the average height of those who exercise at least three times a week and those who don't.

HA: There is a difference in the average height of those who exercise at least three times a week and those who don't.

```
# Calculating the observed difference
obs_diff_height <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(height ~ physical_3plus) %>%
  calculate(.,stat = "diff in means", order = c("yes", "no"))
```

```
# Simulating data under the assumption of independence through permutation
set.seed(131017)
null_dist_height <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(height ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
# visualizing the null distribution
visualize(null_dist_height) +
  shade_p_value(obs_stat = obs_diff_height, direction = "two_sided")
```




```
 # calculating the p-value
null_dist_height %>%
  get_p_value(obs_stat = obs_diff_height, direction = "two_sided")
```

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

Since the p-value is less than 0.05, we need to reject the null hypothesis and accept the alternative hypothesis that there is indeed a difference in average weight for those who are physically active at least 3x/week compared to those who are not.

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.

Insert your answer here

```
# Calculating the number of options for the variable "hours_tv_per_school_day"
yrbss %>%group_by(hours_tv_per_school_day)%>% summarise(n())
```

```
## # A tibble: 8 x 2
##   hours_tv_per_school_day `n()`
##   <chr>                   <int>
## 1 1                        1750
## 2 2                        2705
## 3 3                        2139
## 4 4                        1048
## 5 5+                       1595
## 6 <1                       2168
## 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.

Insert your answer here Research Question: Do students who weight more than the average weight sleep more than students who weight less than the average weight?

HO: There is relationship between weight and sleep

HA: There is a no relationship between weight and sleep

Chosen confidence level:95%

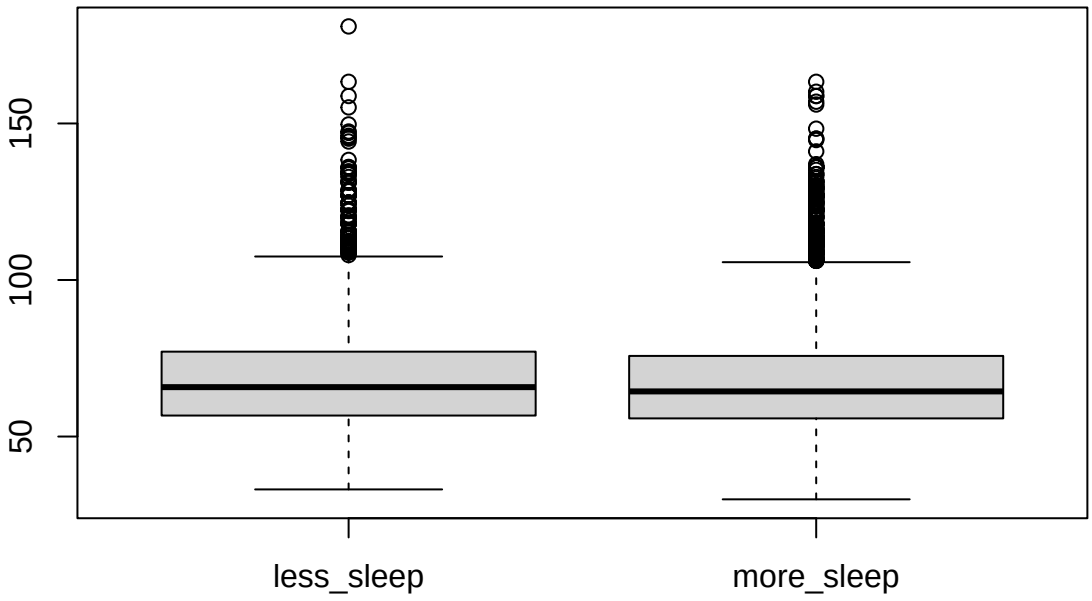
Normal distribution & Independence of the sample size conditions met.

```
# Preparing our data set
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()
```

```r
# Visualizing our data set
boxplot(weight_less$weight, weight_more$weight,
        names = c("less_sleep", "more_sleep"))
```



Results:

```r
mean1 <- mean(weight_less$weight)
sd1<- sd(weight_less$weight)
max1 <- max(weight_less$weight)
max1
```

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

```r
mean2 <- mean(weight_more$weight)
sd2 <- sd(weight_more$weight)
max2 <- max(weight_more$weight)
max2
```

```
## [1] 163.3
```

```r
mean_diff <- mean2 - mean1
sd <-
  sqrt(
  ((mean2^2) / nrow(weight_more)) +
  ((mean1^2) / nrow(weight_less))
  )
```

```r
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
```

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

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

p- value found is at exactly 0.05 making it rather difficult to determine whether there is in fact a relationship between weight and sleep. * * *