Chapter 7 - Inference for Numerical Data

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?

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?

```
cat('observations with missing weights are : ' ,1004)
```

```
## observations with missing weights are :  1004
```

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"))
```

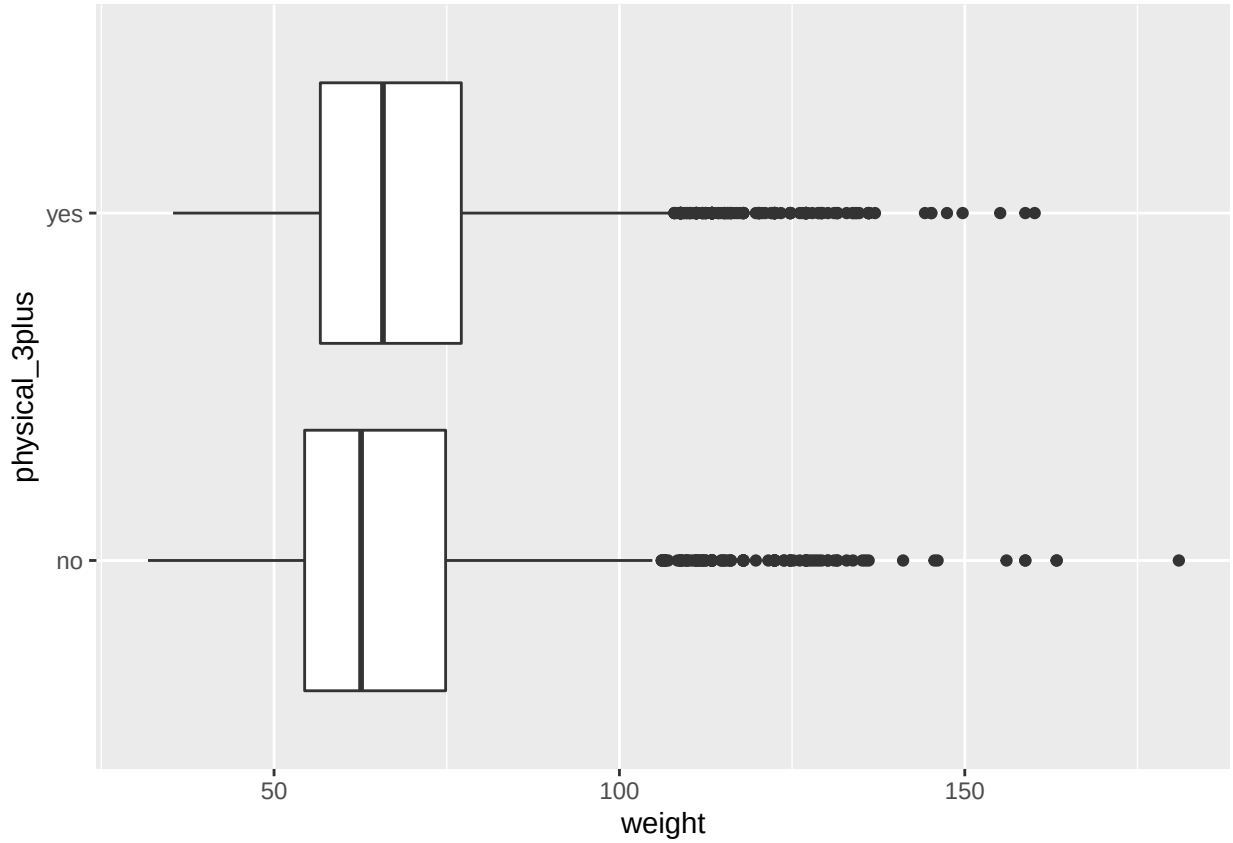
3. Make a side-by-side boxplot of `physical_3plus` and `weight`. Is there a relationship between these two variables? What did you expect and why?

Answer:

The relationship between student's weight and if they are physically active at least 3 times per week seems to show that those who are not physically active at least 3 times per week weigh less than those who are physically active at least 3 times per week. This is interesting as I would have expected those who were physically active at least 3 times per week to weigh less than those who were not physically active at least 3 times per week. These results are quite contrary to the way things are assume that people who exercise would weigh less.

```
yrbss2 <- yrbss %>%
  mutate(physical_3plus = ifelse(yrbss$physically_active_7d > 2, "yes", "no")) %>%
  na.exclude()

ggplot(yrbss2, aes(x=weight, y=physical_3plus)) + 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))
```

```
## # 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()`.

Answer Main two conditions to be satisfied are- independence and normality. Based on the information given in the data set the data looks independent. We can assume With a sample size well over 1000 and no particularly outliers in the boxplot plotted earlier that the condition is satisfied.

```
yrbss2 %>%
  group_by(physical_3plus) %>%
  dplyr::summarise(n = n())
```

```
## # A tibble: 2 x 2
##   physical_3plus     n
##   <chr>          <int>
## 1 no              2656
## 2 yes             5695
```

```
#or another way
yrbss2 %>%
  group_by(physical_3plus) %>%
  summarise(freq = table(weight)) %>%
  summarise(n = sum(freq))
```

```
## # A tibble: 2 x 2
##   physical_3plus     n
##   <chr>          <int>
## 1 no              2656
## 2 yes             5695
```

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.

Answer: 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.

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.

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`.

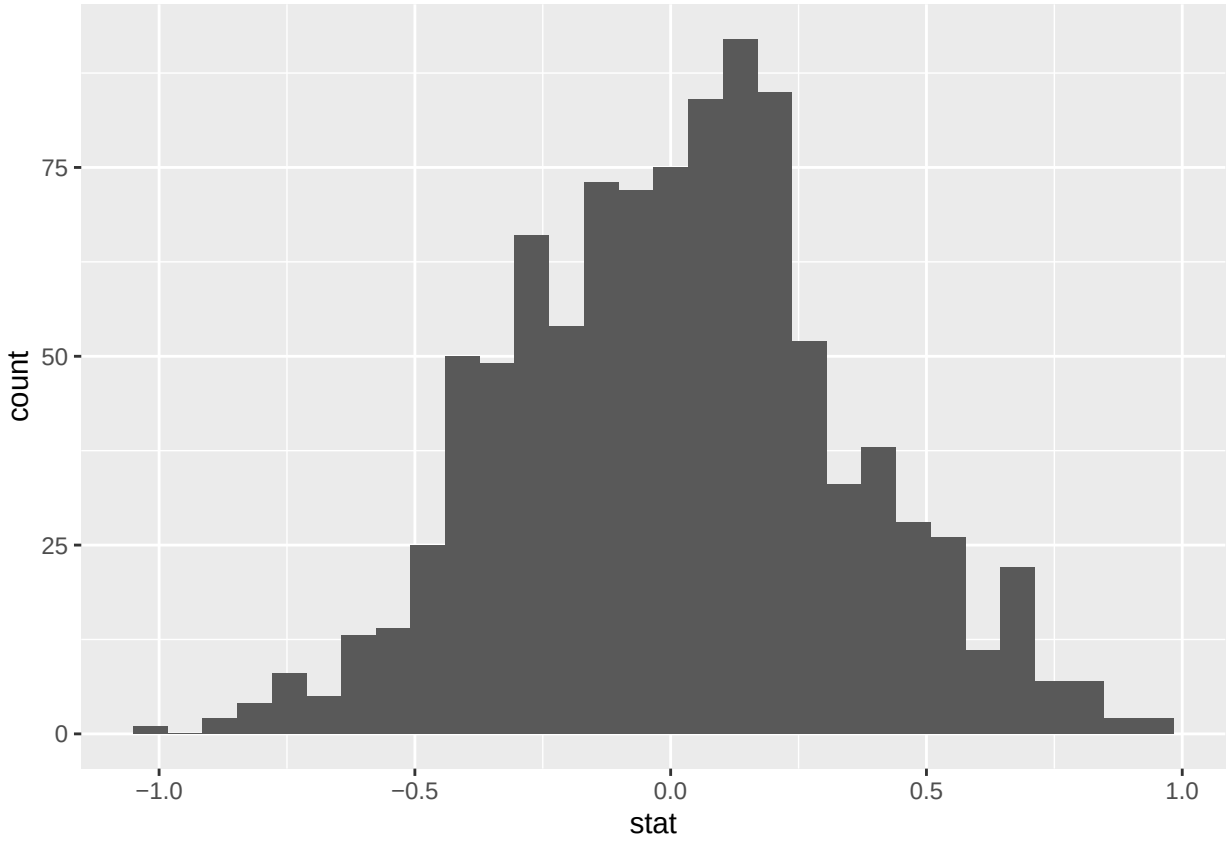
```r
null_dist <- yrbss %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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`, whichis the argument when generating a null distribution for a hypothesis test.

We can visualize this null distribution with the following code:

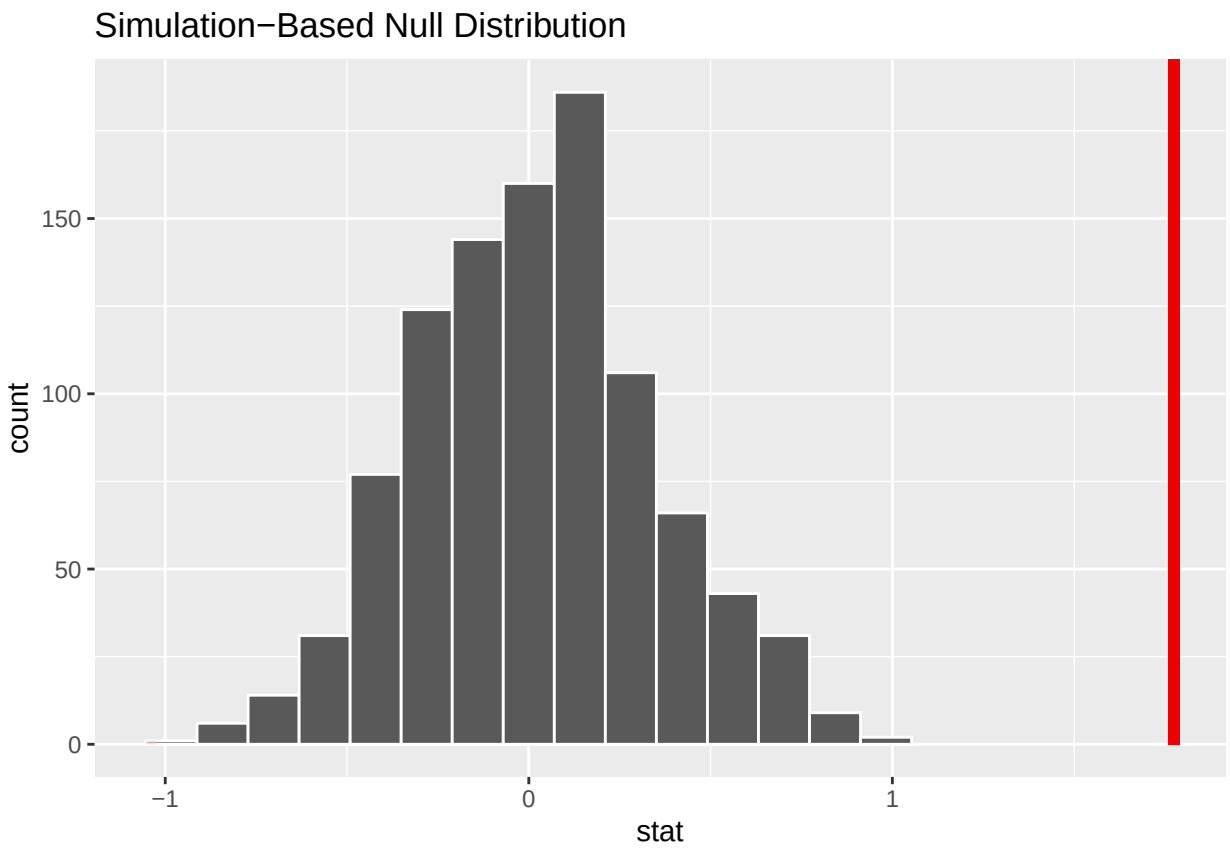
```r
ggplot(data = null_dist, aes(x = stat)) +
  geom_histogram()
```



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

Answer: The result is a very small number, close to zero

```r
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.

```
#get Standard deviation for the group
yrbss2 %>%
  group_by(physical_3plus) %>%
  summarise(sd_weight = sd(weight, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##   physical_3plus sd_weight
##   <chr>              <dbl>
## 1 no                  18.0
## 2 yes                 16.4
```

```r
#means of the weights
yrbss2 %>%
  group_by(physical_3plus) %>%
  summarise(mean_weight = mean(weight, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##   physical_3plus mean_weight
##   <chr>                <dbl>
## 1 no                    67.1
## 2 yes                   68.7
```

```r
#sample size for the group

yrbss2 %>%
  group_by(physical_3plus) %>%
  dplyr::summarise(n = n())
```

```
## # A tibble: 2 x 2
##   physical_3plus     n
##   <chr>          <int>
## 1 no              2656
## 2 yes             5695
```

```r
xnot3 <- 67.1
nnot3 <- 2656
snot3 <- 18.0
x3 <- 68.7
n3 <- 5695
s3 <- 16.4

z = 1.96

uci_not <- xnot3 + z*(snot3/sqrt(nnot3))
lci_not <- xnot3 - z*(snot3/sqrt(nnot3))


cat('With 95% confident that students who did not exercise at least three times a week have an average
```

```
## With 95% confident that students who did not exercise at least three times a week have an average wei
```

```r
u_ci <- x3 + z*(s3/sqrt(n3))
l_ci <- x3 - z*(s3/sqrt(n3))


cat('With 95% confident that students who exercise at least three times a week have an average weight b
```

```
## With 95% confident that students who exercise at least three times a week have an average weight betw
```

More Practice

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

```
calculate_height_ci_95 <- yrbss %>%
  specify(response = height) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
  get_ci(level = 0.95)


cat('With 95% confident the average height in meters between :',calculate_height_ci_95$lower_ci, 'and '
```

```
## With 95% confident the average height in meters between : 1.689454 and  1.693082
```

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.

Answer: As expected, the 95% confidence interval has a slightly larger range than the confidence interval 90%. This larger range is necessary to be more certain about the population parameter.

```
calculate_height_ci_90 <- yrbss %>%
  specify(response = height) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
  get_ci(level = 0.90)


cat('With 90% confident the average height in meters between :',calculate_height_ci_90$lower_ci, 'and '
```

```
## With 90% confident the average height in meters between : 1.689684 and  1.692661
```

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. **Answer** HO: There is no difference in the average height of those who are physically active at least 3 days per week and those who are not. HA: There is a difference in the average height of those who are physically active at least 3 days per week and those who are not

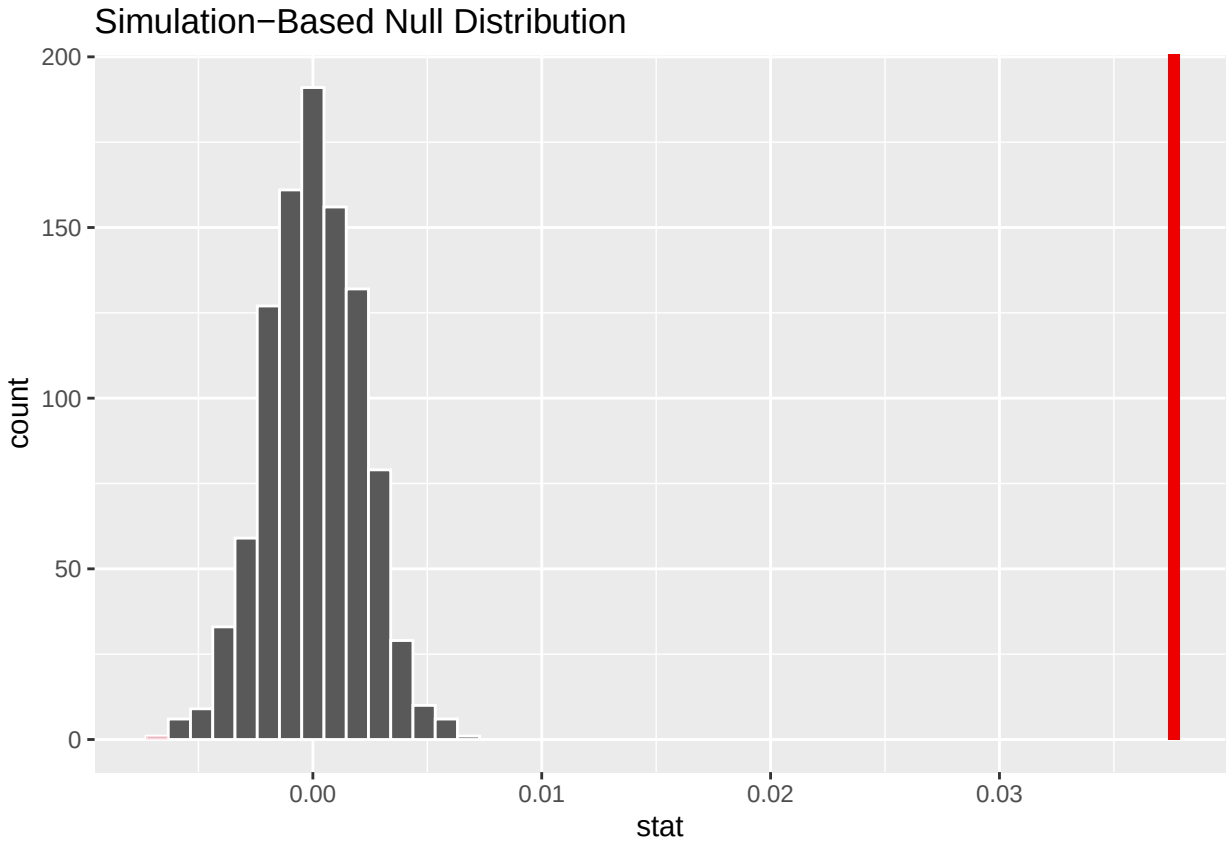
Because p-value is very small, smaller than 0.05, we should reject the null hypothesis.

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


set.seed(455678)
null_dist_hgt <- yrbss %>%
  specify(height ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
```

```
  calculate(stat = "diff in means", order = c("yes", "no"))

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")
```

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

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

```
## # 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
```

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.
