Inference for numerical data

Fomba Kassoh, Teamed up with Soulemane Ndoumbia on Lab 7

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 There are 13 cases in our sample. The cases are: 1. age: Age, in years. 2. gender: Gender. 3. grade: School grade. 4: hispanic: Hispanic or not. 5. race: Race / ethnicity. 6. height: Height, in meters (3.28 feet per meter). 7. weight: Weight, in kilograms (2.2 pounds per kilogram). 8. helmet_12m: How often did you wear a helmet when biking in the last 12 months? 9. text_while_driving_30d: How many days did you text while driving in the last 30 days? 10. physically_active_7d: How many days were you physically active for 60+ minutes in the last 7 days? 11. hours_tv_per_school_day: How many hours of TV do you typically watch on a school night? 12. strength_training_7d: How many days did you do strength training (e.g. lift weights) in the last 7 days? 13. school_night_hours_sleep: How many hours of sleep do you typically get on a school night?

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 There are 1004 missing observations in weights

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

```
## [1] 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?

Insert your answer here Relationship: There is a relationship. High schoolers who are not physically active for at least 3 days a week ("no") have the lowest median weight. High schoolers who are physically active for at least 3 days a week ("yes") have a higher median weight than the "no" group. The spread of weights for both groups is relatively similar, with outliers present in both categories. The observation from the box plot indicates that students who engage in physical activity for at least 3 days a week have a slightly higher median weight than those who don't.

Expectation: At first glance, I would expect physically active students to weigh less due to the calorie-burning nature of exercise. However, as previously mentioned, factors like increased muscle mass, dietary habits, and the nature of physical activities can influence this relationship.

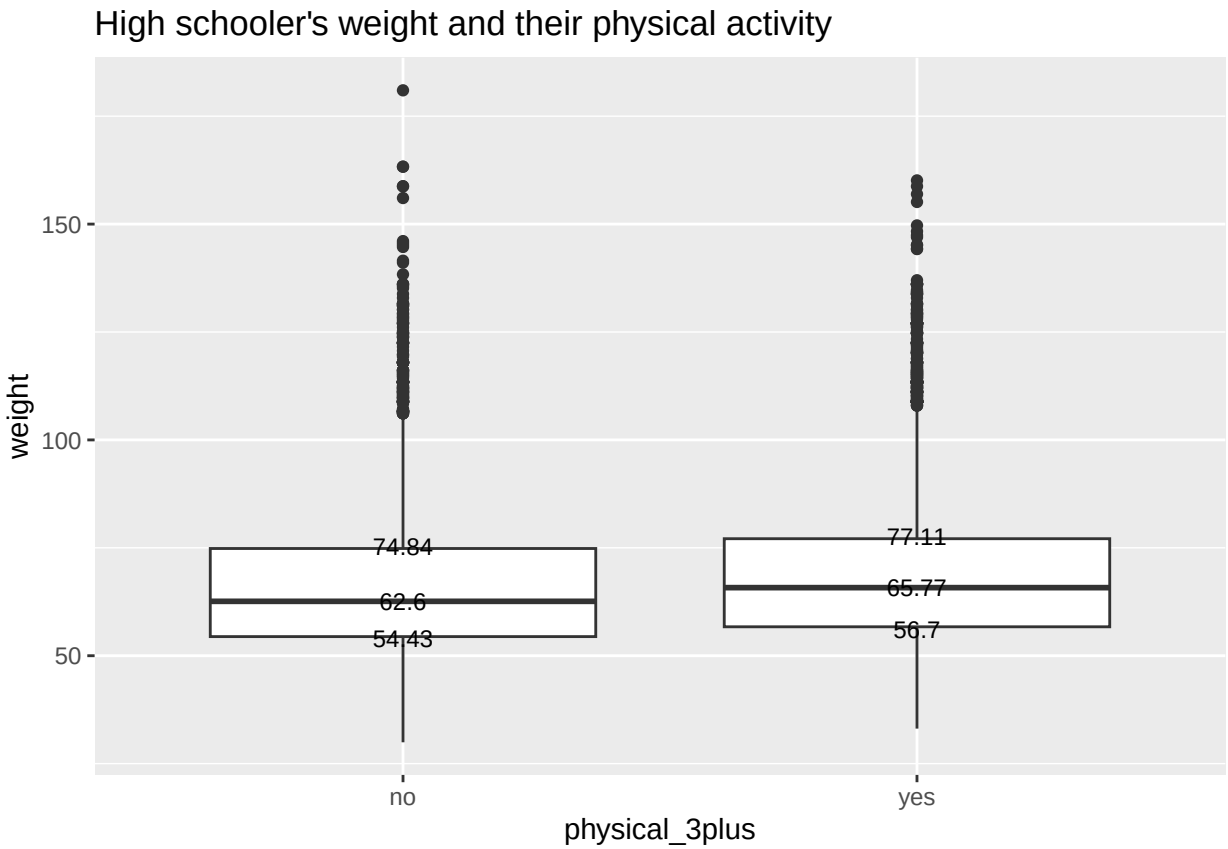
```
physical_3plus_data <- yrbss %>%
  filter(!is.na(physically_active_7d))

# Create a basic boxplot
plot <- ggplot(physical_3plus_data, aes(x = physical_3plus, y = weight)) +
  geom_boxplot(na.rm = TRUE)

plot <- plot +
  labs(
    title = "High schooler's weight and their physical activity",
    y = "weight"
  )

quartiles <- physical_3plus_data %>%
  group_by(physical_3plus) %>%
  summarize(Q1 = quantile(weight, 0.25, na.rm = TRUE), Median = median(weight, na.rm = TRUE), Q3 = quan


plot +
  annotate("text", x = 1:2, y = quartiles$Q1 - 0.5, label = quartiles$Q1, size = 3) +
  annotate("text", x = 1:2, y = quartiles$Median, label = quartiles$Median, size = 3) +
  annotate("text", x = 1:2, y = quartiles$Q3 + 0.5, label = quartiles$Q3, size = 3)
```



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()`.

All conditions for inference are satisfied as follows: a. Sample size: The sample size is sufficiently large (n>30) for us to make inferences about the population from which the sample was drawn drawn. b. The yrbss dataset represents a random sample c. Assuming each student's response is independent of others; d. No significant extreme skewness observed from the box plot e. The mean and median are roughly the same.

```
yrbss %>%
  group_by(physical_3plus) %>%
  drop_na(physical_3plus) %>%
  summarise(n = n(), Mean = mean(weight, na.rm = TRUE), Median = median(weight, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   physical_3plus     n  Mean Median
##   <chr>          <int> <dbl>  <dbl>
## 1 no              4404  66.7   62.6
## 2 yes             8906  68.4   65.8
```

Insert your answer here

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.

Insert your answer here Null Hypothesis (H0): The average weights of those who exercise at least 3 times a week and those who don't are equal.

Alternative Hypothesis (H1): The average weights of those who exercise at least 3 times a week and those who don't are not equal.

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`.

```
set.seed(2000)
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"))
```

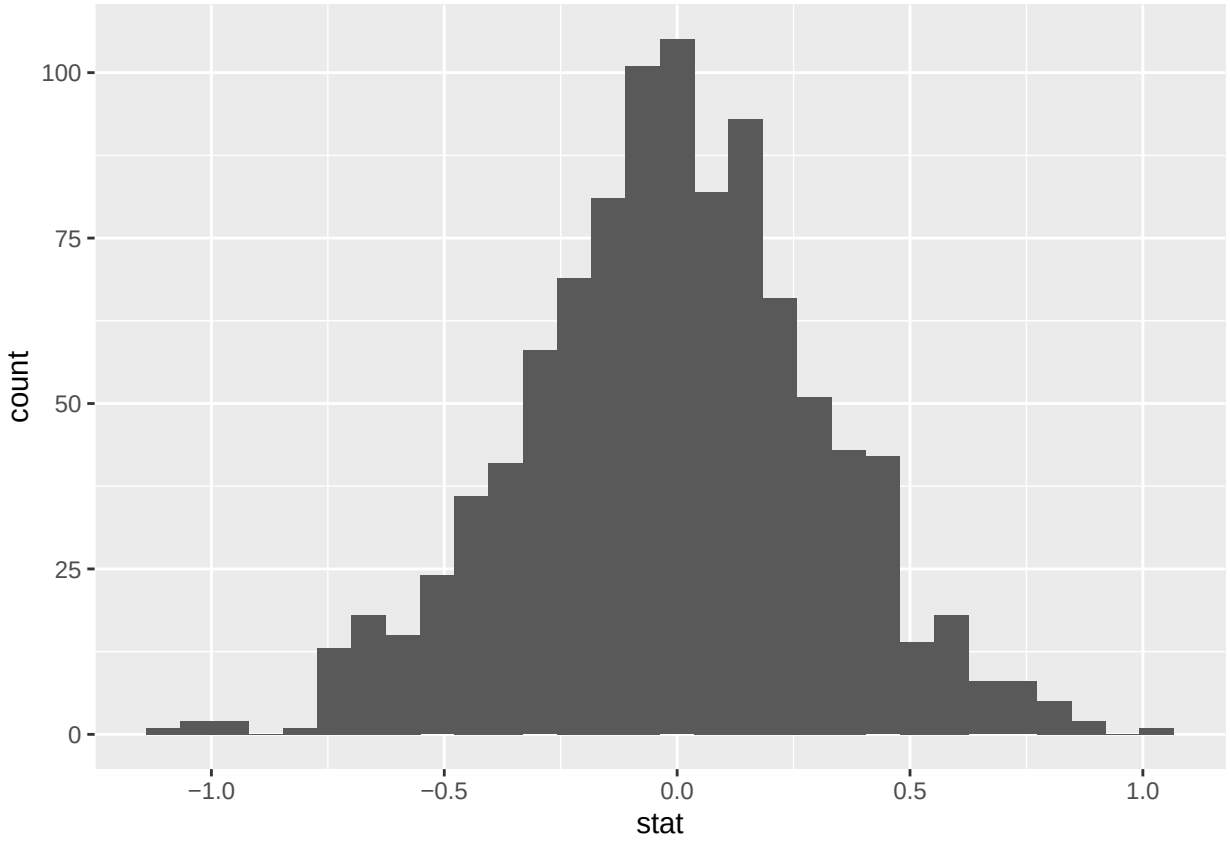
Here, `hypothesize` is used to set the null hypothesis as a test for independence. In one sample cases, the `null` argument can be set to "point" to test a hypothesis relative to a point estimate.

Also, note that the `type` argument within `generate` is set to `permute`, which is 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()
```

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



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

Insert your answer here To know how many of the permutations have a difference of at least `obs_stat`, we filter the null distribution for stat >= obs_diff. Base on the result below, none of the 'null permutations have a difference of at least `obs_diff`

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

```
## Warning: Using one column matrices in `filter()` was deprecated in dplyr 1.1.0.
## i Please use one dimensional logical vectors instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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

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")
```

```
## 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
```

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.

Interpretation: Based on the confidence interval below, We can state that at a 95% confidence level the true value of the difference between the average weight of those who exercise at least 3 times a week and those who don't falls within the range from approximately -0.6830667 to 0.606511.

```
# Calculate the 95% confidence interval
confidence_interval <- null_dist %>%
  get_ci(level = 0.95, type = "percentile")

# Print the confidence interval
confidence_interval
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.683    0.607
```

More Practice

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

Insert your answer here Interpretation: Based on the confidence interval below, I am reasonably confident that the true value of the difference between the average height of those who exercise at least 3 times a week and those who don't falls within the range from approximately -0.003787855 to 0.003959193 at a 95% confidence level.

This is almost not difference in height between those who exercise at least 3 times a week and those who don't. In context, height is not influenced by the number of exercise per week.

```
set.seed(3000)
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")) %>%
  get_ci(level = 0.95, type = "percentile")
```

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

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1 -0.00379  0.00396
```

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 The 90% confidence interval is narrower than the 95% confidence interval. This is because as you increase the confidence level from 90% to 95%, you need to widen the interval to be more certain that it contains the true population parameter. Conversely, if you decrease the confidence level from 95% to 90%, the interval becomes narrower because you are willing to accept less certainty.

```
set.seed(4000)
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")) %>%
  get_ci(level = 0.90, type = "percentile")
```

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

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1 -0.00329  0.00326
```

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 Null Hypothesis (H0): The average height of those who exercise at least 3 times a week and those who don't are equal.

Alternative Hypothesis (H1): The average height of those who exercise at least 3 times a week and those who don't are not equal.

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

```
set.seed(5000)
height_null_dist <- 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"))
```

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

```
height_null_dist %>%
  get_p_value(obs_stat = obs_height_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
```

Conclusion: Since the p-value of 0 is less than 0.5, we reject the null hypothesis.That is average height of those who exercise at least 3 times a week and those who don't are not equal.We make this conclusion bearing in mind that there is a 10% chance that our conclusion is wrong.

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 There 10 different options in the dataset for the `hours_tv_per_school_day`. These include att the variables except ade, race, and hispanic.

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: Is there a significant relationship between a person's weight (independent variable) and their sleep duration (dependent variable)?

Hypothesis Test:

Null Hypothesis (H0): There is no relationship between weight of people who sleep at least 8 hours a night and those who don't. Alternative Hypothesis (H1): There is a relationship between weight of people who sleep at least 8 hours a night and those who don't.

Assumptions:

The sample is representative of the population in terms of age, health, and other relevant factors. Sleep duration is measured accurately. The relationship between weight and sleep duration is approximately linear. The data follows a normal distribution. Significance Level (alpha): 0.05

```
yrbss %>%
  group_by(school_night_hours_sleep) %>%
  summarise(n = n())
```

```
## # A tibble: 8 x 2
##   school_night_hours_sleep     n
##   <chr>                    <int>
## 1 10+                        316
## 2 5                         1480
## 3 6                         2658
## 4 7                         3461
## 5 8                         2692
## 6 9                          763
## 7 <5                         965
## 8 <NA>                      1248
```

Create a new variable sleep_7plus for people who sleep at least 8 hours a night.

```
yrbss <- yrbss %>%
  mutate(sleep_7plus = ifelse(yrbss$hours_tv_per_school_day > 7, "yes", "no"))
```

Calculate the observed difference in the mean weight of people who sleep at least 8 hours a night and people who do not.

```
obs_sleep_diff <- yrbss %>%
  drop_na(sleep_7plus) %>%
  specify(weight ~ sleep_7plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

Generate a null distribution: Sample 1000 weight differences between people who sleep at least 8 hours a night and people who don't.

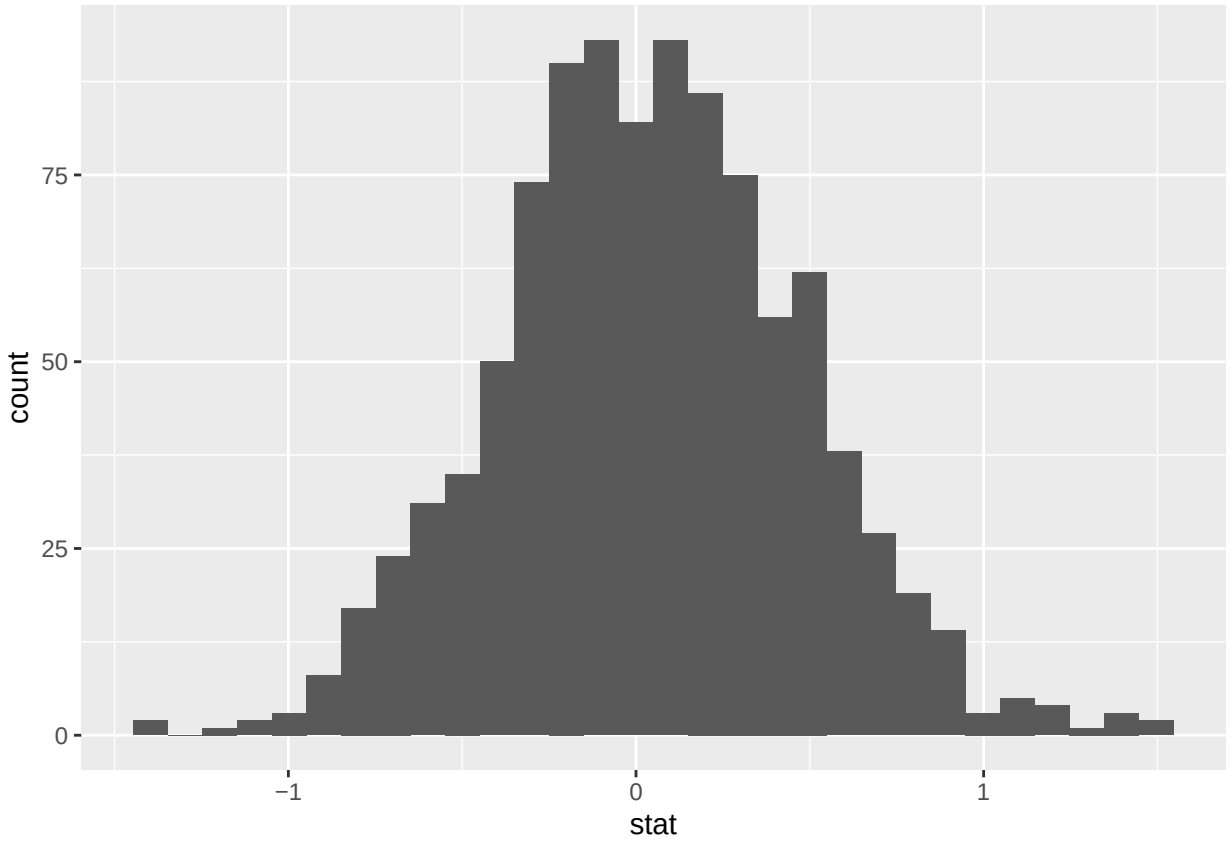
```
set.seed(6000)
sleep_null_dist <- yrbss %>%
  drop_na(sleep_7plus) %>%
  specify(weight ~ sleep_7plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
## Warning: Removed 937 rows containing missing values.
```

Visualize the null distribution to see if it follows a normal distribution

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

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



Calculate the p-value that corresponds to our test statistic.

```
sleep_null_dist %>%
  get_p_value(obs_stat = obs_sleep_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
```

Conclusion: Since p < alpha (0.5), we reject the null hypothesis. That is the probability that there is no relationship between sleep hours and weight seems to be wrong. That is, we go with the alternative hypothesis (H1) that there is a relationship between sleep hours and weight knowing that there is a 10% chance that the null hypothesis is true. * * *