Inference for numerical data

John Cruz

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"~
```

There are 13,583 cases within the sample. Each case is a high school student responses to a survey.

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?

There are 1004 missing weights observations.

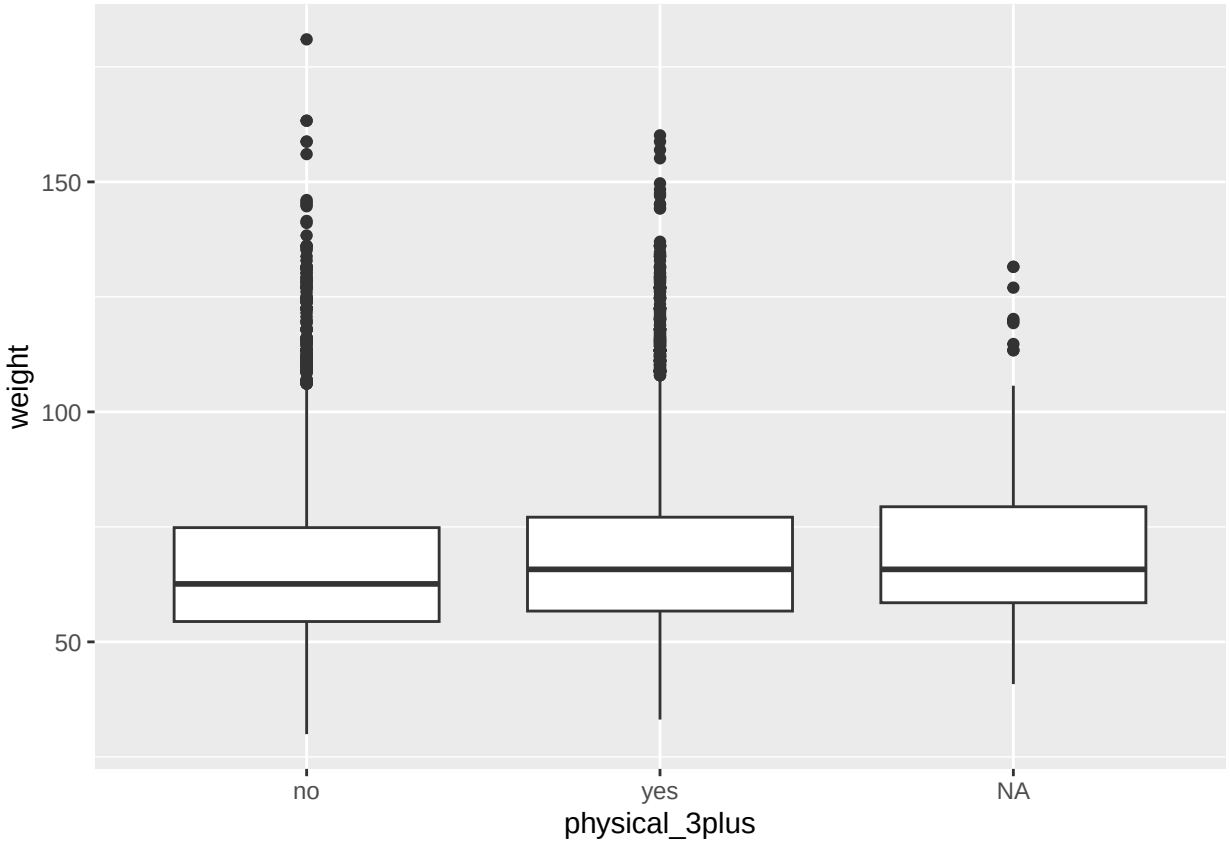
Next, consider the possible relationship between a high schooler's weight and their physical activity. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

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?

```
yrbss |>
  ggplot(aes(x = physical_3plus, y = weight)) +
  geom_boxplot(na.rm = TRUE)
```



I would expect for there to be no relationship between weight and days physically active. This is because people can weigh more naturally, but also weigh more depending on the amount of body muscle obtained, which weighs heavier than fat. The box plot shows that there is not a definitive relationship with each other.

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()`.

It is not explicitly stated, but we have to assume that the CDC conducted a random sample, independent of bias. As for the sample size of 13,583, whenever a sample size is >30, we can use the Central Limit Theorem to let us assume it is normally distributed.

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 hypothesis is $\mu 1 = \mu 2$, where both groups of students who are active >3 days and students active <= 3 days are equal in average weights. The alternative hypothesis is where $\mu 1 = \mu 2$.

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(weight,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`.

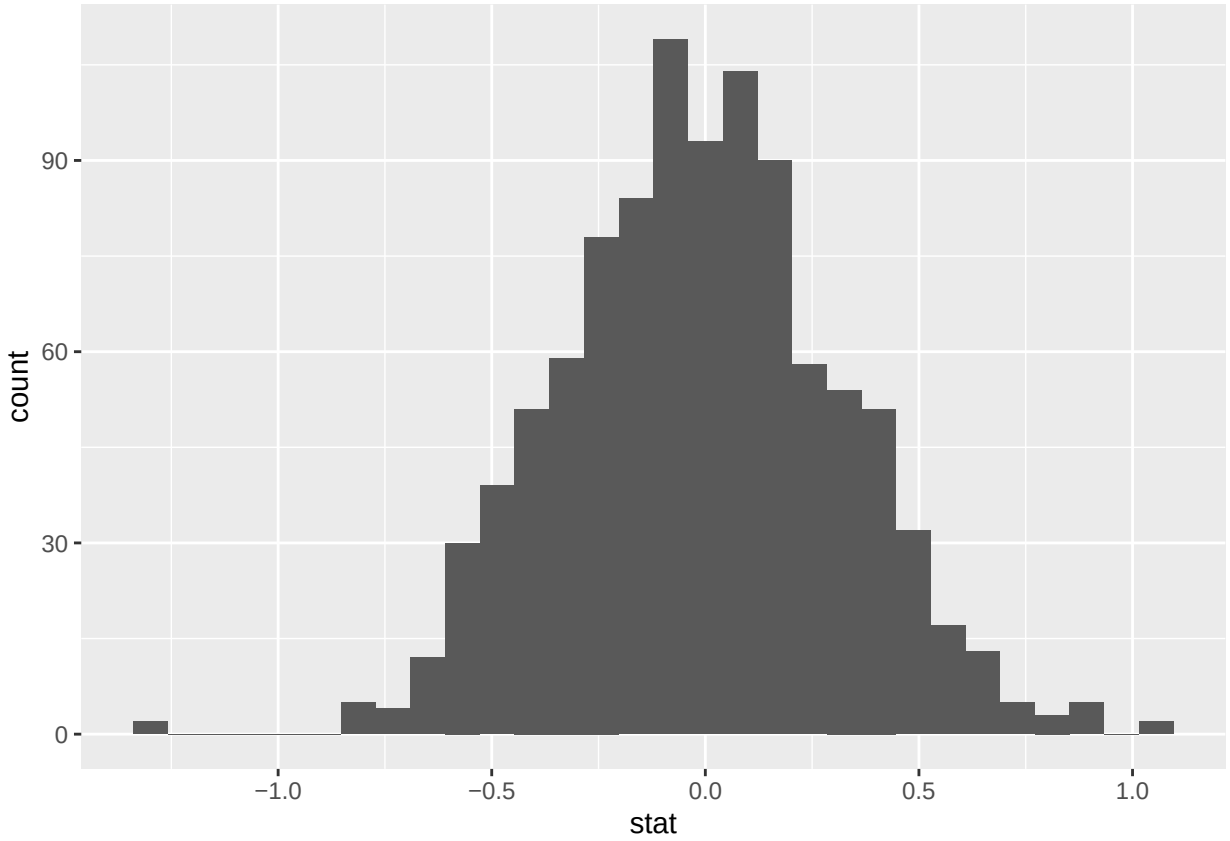
```
null_dist <- yrbss %>%
  drop_na(weight,physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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`?

Using the null_dist histogram, we can see that the obs_stat of 1.775 is well beyond the range of the graph, showing that there are zero occurrences.

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.

The confidence interval ranges between -0.662 to 0.666, inclusive of zero. This means that we fail to reject that there is a difference between btoh groups.

```
yrbss %>%
  drop_na(weight,physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no")) |>
  get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.626    0.549
```

More Practice

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

We are 95% confident, that the average height of students are between 1.69 and 1.69 meters. Rounding issues are not showing.

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

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     1.69     1.69
```

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 width is the same due to rounding, but it should be smaller because we are accounting for a smaller spread from the mean of the data.

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

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     1.69     1.69
```

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.

Null hypothesis is $\mu 1 = \mu 2$, where both groups of students who are active >3 days and students active <= 3 days are equal in average heights. The alternative hypothesis is where $\mu 1 = \mu 2$.

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 seven distinct groups not including NA

```
yrbss |>
distinct(hours_tv_per_school_day)
```

```
## # A tibble: 8 x 1
##   hours_tv_per_school_day
##   <chr>
## 1 5+
## 2 2
## 3 3
## 4 do not watch
## 5 <1
## 6 4
## 7 1
## 8 <NA>
```

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.

Are students on average taller than students who sleep less than 8 hours?

```
yrbss |>
distinct(school_night_hours_sleep)
```

```
## # A tibble: 8 x 1
##   school_night_hours_sleep
##   <chr>
## 1 8
## 2 6
## 3 <5
## 4 9
## 5 10+
## 6 7
## 7 5
## 8 <NA>
```

```
yrbss <- yrbss %>%
  mutate(sleep_less =
           ifelse(!school_night_hours_sleep %in% c('8', '9', '10+'), 'yes', 'no'))
```

```r
test_hyp <- yrbss %>%
  drop_na(sleep_less, height)

test_hyp %>%
  specify(height ~ sleep_less) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("yes", "no")) |>
  get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1 -0.00479  0.00303
```

We can see that the confidence interval between -0.00528 and 0.00250 include 0, so this means with a 95% confidence interval, we fail to rejec the null hypothesis. This means that students who sleep 8 or more hours compares to those that do not, there is no difference.
