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)
library(dplyr)

# my imports
library(ggplot2)
```

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"~
```

We have 13 variables (fields) in our sample dataset and 13,583 rows (cases/observations). Each case (row) corresponds to a teenager within the sample and the characteristics that apply to that teenager (age, gender, etc.).

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?

```
yrbss %>%
      filter(is.na(weight)) %>%
      nrow()
```

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

We are missing weights from 1004 rows/observations in our sample.

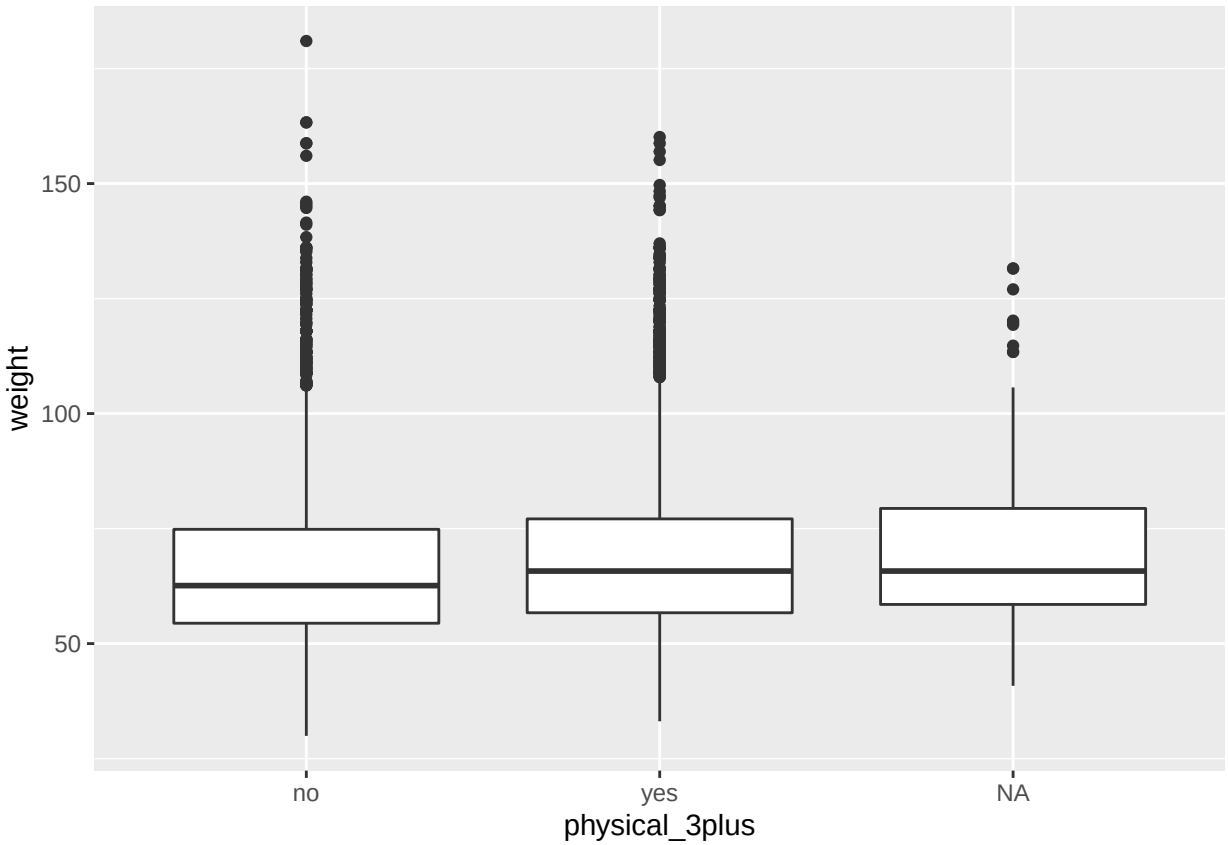
Next, consider the possible relationship between a high schooler's weight and their physical activity. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

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?

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



These variables do not appear to have a strong relatinoship, as the median and quartile weights appear to be in a similar range between those students who are and are not physically active at least 3 days a week. I did expect there to be a relationship between these two, as people who get more physical activity in tend to lose and keep off weight more easily.

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()`.

- *Random*: since this is a random sample of teenagers from the larger dataset, we can assume we meet

this condition, given that the sampling is truly random.
- *Normal*:Our sampling distribution of $\bar{x}$ needs to be approximately normal. This is true when our sample size is reasonably big ($n \geq 30$). each category (computed below) is greater than 30, so we can assume our sampling distribution will be normal.

```
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(group_size = n())
```

```
## # A tibble: 3 x 2
##   physical_3plus group_size
##   <chr>               <int>
## 1 no                   4404
## 2 yes                  8906
## 3 <NA>                  273
```

- *Independent*: Each observation in our sample dataset is a different teenager, and the observation of these characteristics for one teenager won't affect the next observation (assuming our sample is random)

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.

- H_0: there *is no difference* in the average weights between those who exercise at least twice a week and those who don't
- H_a: there *is a difference* in the average weights between those who exercise at least twice a week and those who do not.

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 %>%
  filter(!(is.na(physical_3plus) | is.na(weight))) %>%
  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`.

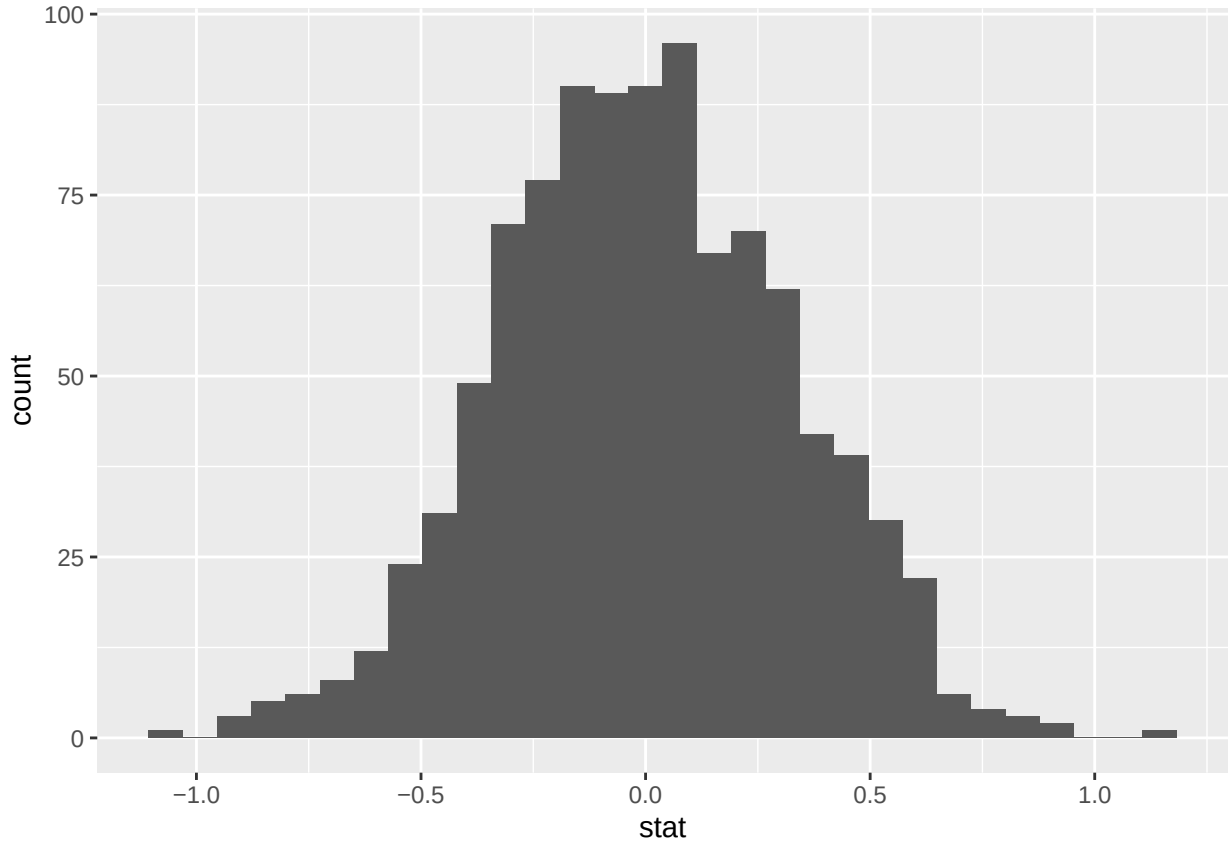
```
null_dist <- yrbss %>%
  filter(!(is.na(physical_3plus) | is.na(weight))) %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = 'diff in means', order = c("yes", "no"))
```

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`?

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

null_dist %>%
          filter(stat >= obs_stat) %>%
          summarise(count_greater = n())
```

```
## # A tibble: 1 x 1
##   count_greater
##           <int>
## 1             0
```

There are no `null_dist` permutations with a difference greater than that specified in `obs_diff`.

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.

We can accomplish this using R' built-in `t.test` function, which constructs a confidence interval of mean differences between two groups

```
active_teens <- yrbss %>%
                    filter(physical_3plus == "yes")
inactive_teens <- yrbss %>%
                    filter(physical_3plus == "no")
t.test(active_teens$height, inactive_teens$height, alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  active_teens$height and inactive_teens$height
## t = 19.029, df = 7973.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03374994 0.04150183
## sample estimates:
## mean of x mean of y
##  1.703213  1.665587
```

The 95% confidence interval $(0.03374994, 0.04150183)$ means we can be 95% sure the true difference in means lies within that interval. Since 0 (which means no true difference in the averages) does not lie in the confidence interval, we can say that there is a true difference in the means of these two groups, and that there is a difference in the average weights between those who exercise regularly and those who don't.

More Practice

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

```
# filtering out
height_mean <- yrbss %>%
                    filter(!is.na(height)) %>%
                    summarise(mean_height = mean(height))


height_sd <- yrbss %>%
                    filter(!is.na(height)) %>%
                    summarise(height_sd = sd(height))
n_height <- yrbss %>%
                    filter(!is.na(height)) %>%
                    summarise(n_height = n())

# calculate margin of error (z = 1.96 for 95% CI)
margin <- qt(0.025, n_height[[1]] -1) * (height_sd[[1]] / sqrt(n_height[[1]]))
margin
```

```
## [1] -0.001829794
```

```
ci <- c(height_mean - margin, height_mean + margin)
```

This confidence interval menas we can be 95% sure that the true average height lies within the interval $(1.689411, 1.693071)$.

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.

```
# calculate margin of error (z = 1.96 for 90% CI)
margin <- qt(0.05, n_height[[1]] -1) * (height_sd[[1]] / sqrt(n_height[[1]]))
margin
```

```
## [1] -0.001535577
```

```
ci_90 <- c(height_mean - margin, height_mean + margin)
ci_90
```

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

This interval will be less wide than the 95% interval, as it will contain fewer "possible" values for the true mean. In other words, the 90% confidence interval contains less area under the curve of the distribution, as we're taking a smaller probability around the point estimate. We see this in the bounds of our 90% interval $(1.692777, 1.689705)$.

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.

We will complete a t-test between the two groups

H_0: There is no difference in average height between those that exercise 3 times a week and those who don't
H_a: There is a difference in avg height between those who exercise 3 times a week and those who don't

Since we're comparing averages between two groups, a t-test with a $\alpha = 5$ significance level should suffice.

```
active_teens <- yrbss %>%
                    filter(physical_3plus == "yes")
inactive_teens <- yrbss %>%
                    filter(physical_3plus == "no")
t.test(active_teens$height, inactive_teens$height, alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  active_teens$height and inactive_teens$height
## t = 19.029, df = 7973.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03374994 0.04150183
## sample estimates:
## mean of x mean of y
##  1.703213  1.665587
```

Since $p < \alpha = 0.05$, we can reject the null hypothesis that there is no difference in the average height between those who exercise at least 3 days a week and those who don't.

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.

```
# Need to find unique values in our hours of tv vector
yrbss %>%
```

```
      group_by(hours_tv_per_school_day) %>%
      count(hours_tv_per_school_day)
```

```
## # A tibble: 8 x 2
## # Groups:   hours_tv_per_school_day [8]
##   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.

Research question: is there a difference in the average weights between those who sleep 8 hours or more a night and those who don't?

Clean the data first:

```
library(stringr)
# cleaning up sleep column so we can get all students who
yrbss$school_night_hours_sleep = str_replace(yrbss$school_night_hours_sleep, "10+", "10")
yrbss$school_night_hours_sleep = as.numeric(yrbss$school_night_hours_sleep)

no_sleep <- yrbss %>%

                mutate(enough_sleep = ifelse(yrbss$school_night_hours_sleep> 8, "yes", "no"))
```

Hypotheses H_0: there is *no difference* in the average weight of students who sleep at least 8 hours a day and those who don't H_a: there *is a difference* in the average weight of students who sleep at least 8 hours a day and those who don't

We'll conduct this test with a 5% significance level ($\alpha = 0.05$\$)

Since we're checking if the group means are equal or not, we'll needa two-sided t-test. We are dealing with groups that are independent (one student's sleep doesn't impoact another's). As well as groups that are sufficiently large enough (n $>$ 30 for successess and failures).

```
no_sleep %>%
      group_by(enough_sleep) %>%
      count(enough_sleep)
```

```
## # A tibble: 3 x 2
## # Groups:   enough_sleep [3]
##   enough_sleep     n
##   <chr>        <int>
## 1 no           10291
## 2 yes            763
## 3 <NA>          2529
```

```r
sleepy_teens <- no_sleep %>%
                    filter(enough_sleep == "no")
awake_teens <- no_sleep %>%
                    filter(enough_sleep == "yes")
t.test(sleepy_teens$weight, awake_teens$weight, alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  sleepy_teens$weight and awake_teens$weight
## t = 3.6259, df = 821.71, p-value = 0.0003057
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.033783 3.474072
## sample estimates:
## mean of x mean of y
##  67.81291  65.55898
```

Our p-value of 0.0003057 is less than our significance level of 5%, we can reject the null hypotheses and state that there is a difference in the average weight between students who sleep at least 8 hours a night and those who don't. This makes sense because the health benefits of getting enough sleep are fairly well-documented, so people who sleep enough are more likely to be healthy and at a healthy weight.
