Inference for numerical data

Kossi Akplaka

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 Open Intro resources, **open intro**.

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 work space.

```
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?

We have 13,583 rows in our data set.

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?

We are missing 1004 observations from the column "weights".

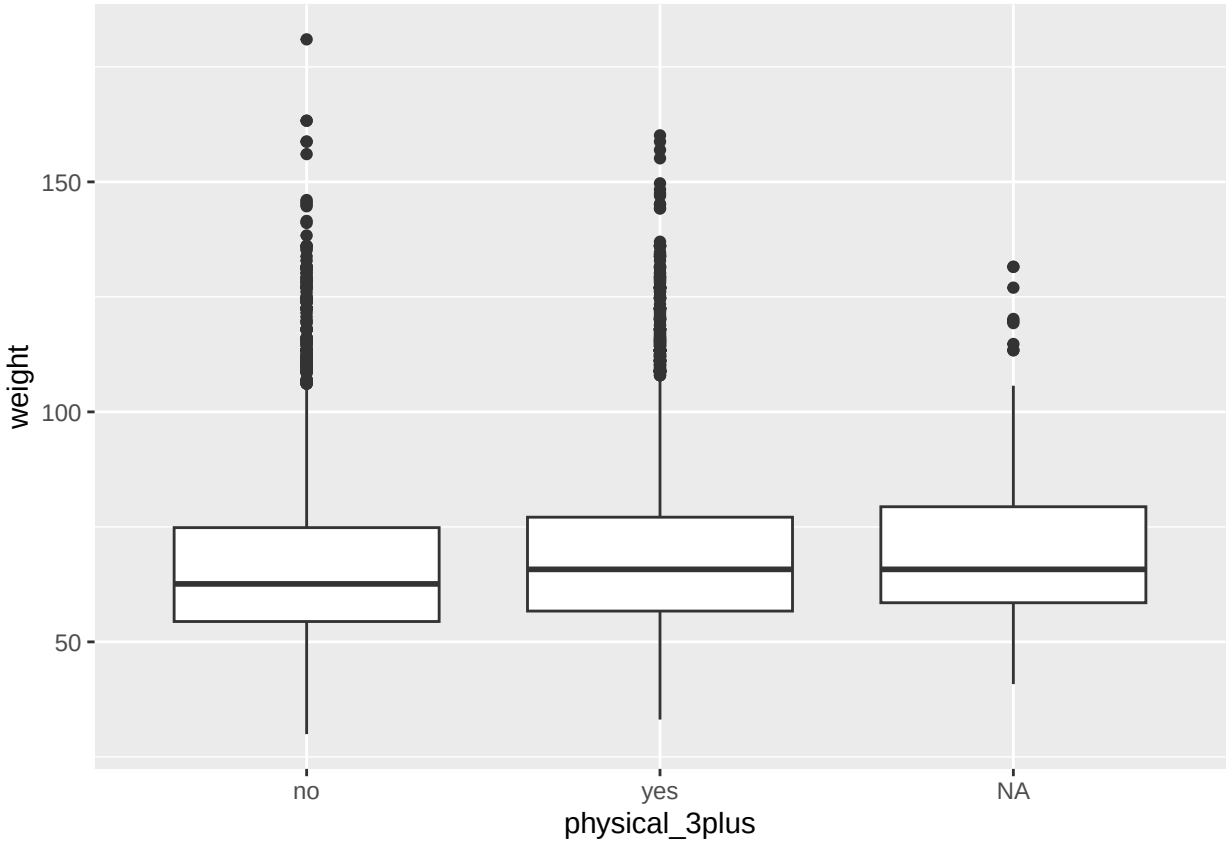
Next, consider the possible relationship between a high schooler's weight and their physical activity. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

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()
```



I would expect the median weight of people doing physical at least 3 days a week to be lower if they want to lose weight or higher if they are bulking. In contrary, the box plot shows that the median weight of people that are doing physical at least 3 times a week is greater. Nevertheless, there is an observed difference between these two groups.

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()`.

There are independence within the two groups because the high schooler that does are active at least 3 days a week are different from the people who doesn't. The sample size is less than 10% of all the high schooler in the united states.

```
yrbss %>%
  count(physical_3plus)
```

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

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.

Null Hypothesis (H0): The average weight is the same for those who exercise at least three times a week and those who don't. yes = no Alternative Hypothesis (Ha): The average weight is different for those who exercise at least three times a week and those who don't. yes != no

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`.

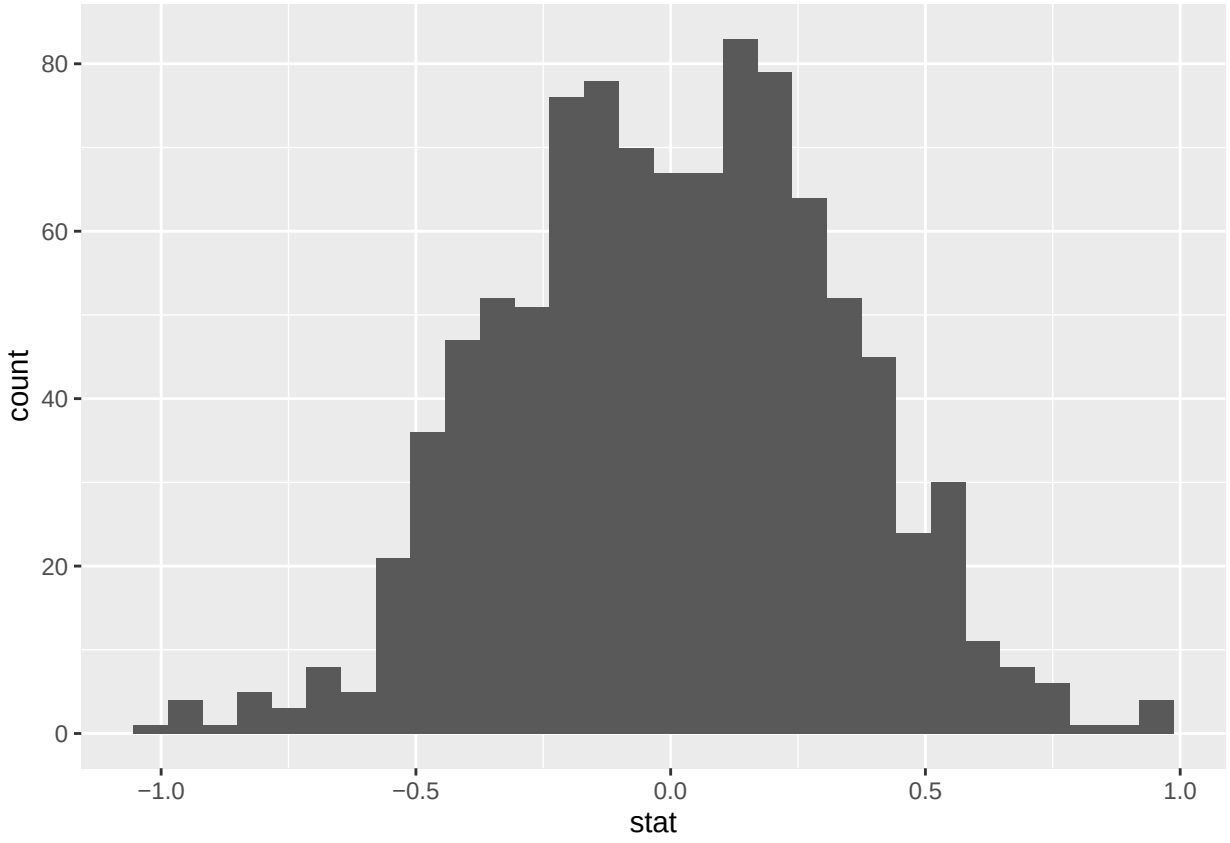
```
null_dist <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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`, 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()
```



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

```
null_dist %>%
  filter(stat>obs_diff$stat) %>%
  summarise(n())
```

```
## # A tibble: 1 x 1
##   `n()`
##   <int>
## 1     0
```

None of them. All the `null` permutations are lower than the `obs_stat` values.

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`.

```r
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 are 95% confident that the difference in mean between high schooler who exercise 3 times a week and the one who don't is between -0.663 and 0.637

```r
null_dist %>%
  get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.581    0.610
```

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 difference in average height is between 0.0039 and 0.0038m. This suggests that there is no difference in height between people who are active at least 3 times a week and people who are not.

```r
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 = .95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1 -0.00372  0.00355
```

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.

There is 90% of confidence that the difference of height is between -0.00325 and 0.00316. The 95% CI have a wider width than the 90% CI.

```
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 = .90)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1 -0.00341  0.00336
```

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 can write the hypothesis test:

Null Hypothesis (H0): The average height is the same for those who exercise at least three times a week and those who don't. yes = no

Alternative Hypothesis (Ha): The average height is different for those who exercise at least three times a week and those who don't. yes != no

Let's find the first difference in height

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

Now, let's simulate the test on the null distribution, which we will save as `null`.

```
null_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"))
```

Let's find how many of these `null` permutations have a difference of at least `obs_diff_height`?

```
null_height %>%
  filter(stat>obs_diff_height$stat) %>%
  summarise(n())
```

```
## # A tibble: 1 x 1
##   `n()`
##   <int>
## 1     0
```

None of the null permutation is at least obs_diff_height.

Now, we can find the p-value

```
null_height %>%
  get_p_value(obs_stat = obs_diff_height, 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 data set for the `hours_tv_per_school_day` there are.

```
table(yrbss$hours_tv_per_school_day)
```

```
##
##           <1            1            2            3            4           5+
##         2168         1750         2705         2139         1048         1595
## do not watch
##         1840
```

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: Conduct a hypothesis test evaluating whether the average weight is different for those who sleep at least ten hours(10+) and those who don't.

We can write the hypothesis test:

Null Hypothesis (H0): The average weight is the same for those who sleep at least 10 hours and those who don't. yes = no

Alternative Hypothesis (Ha): The average weight is different for those who sleep at least 10hours and those who don't. yes!=no

Let's create a new column sleep_10plus

```
yrbss <- yrbss %>%
  mutate(sleep_10plus = ifelse( yrbss$school_night_hours_sleep =='10+', 'yes', 'no'))
```

Let's find the first difference in weight

```
obs_weight <- yrbss %>%
  drop_na(sleep_10plus) %>%
  specify(weight ~ sleep_10plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

Now let's simulate the test on the null distribution, which we will save as `null`.

```
null_weight <- yrbss %>%
  drop_na(sleep_10plus) %>%
  specify(weight ~ sleep_10plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

Let's find how many of these `null` permutations have a difference of at least `obs_diff_height`?

```
null_weight %>%
  filter(stat>obs_weight$stat) %>%
  summarise(n())
```

```
## # A tibble: 1 x 1
##   `n()`
##   <int>
## 1    93
```

91 of the null permutation is at least obs_weight.

Now, we can find the p-value

```
null_weight %>%
  get_p_value(obs_stat = obs_weight, direction = "two_sided")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.186
```

The p value is 0.182 which is greater than .05, we fail to reject the H0. There is not enough evidence to show that there is a difference in weight between people who sleep 10_hours and those who don't.
