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?

There are 13,583 observations. Each contains information given by one youth on some classifying variables and values for risky behaviors.

```
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?

```
sum(!complete.cases(yrbss))
```

```
## [1] 5232
```

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.

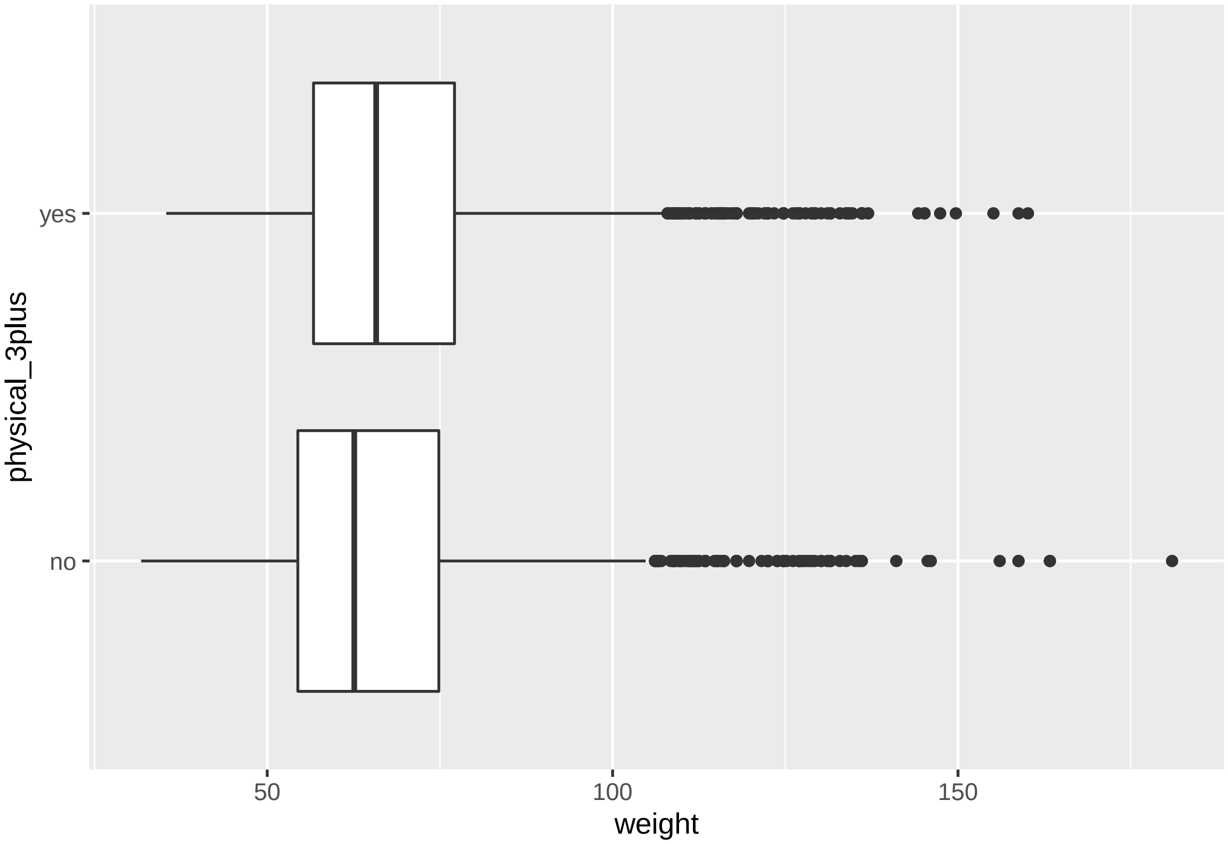
First, let's create a new variable `physical_3plus`, which will be coded as either "yes" if they are physically active for at least 3 days a week, and "no" if not.

```
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?

```
y_complete <- drop_na(yrbss)

y_complete %>%
  ggplot(aes(weight, physical_3plus)) +
  geom_boxplot()
```



There are some outlying heavy people who don't exercise much, but generally, people who exercise a little way a little more. Makes sense because muscle is heavy, but the difference seems too small to matter.

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()`.

```
y_complete %>%
  group_by(physical_3plus) %>%
  summarize(n = n())
```

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

The data are independent, and the number of observations pertaining to each level of the variable are sufficiently large. Thereis one very heavy outlier, but we don't mind.

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.

H_0: The observed difference is due to random variance.

H_A: The two groups of students have distinct properties not due to random variance.

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`.

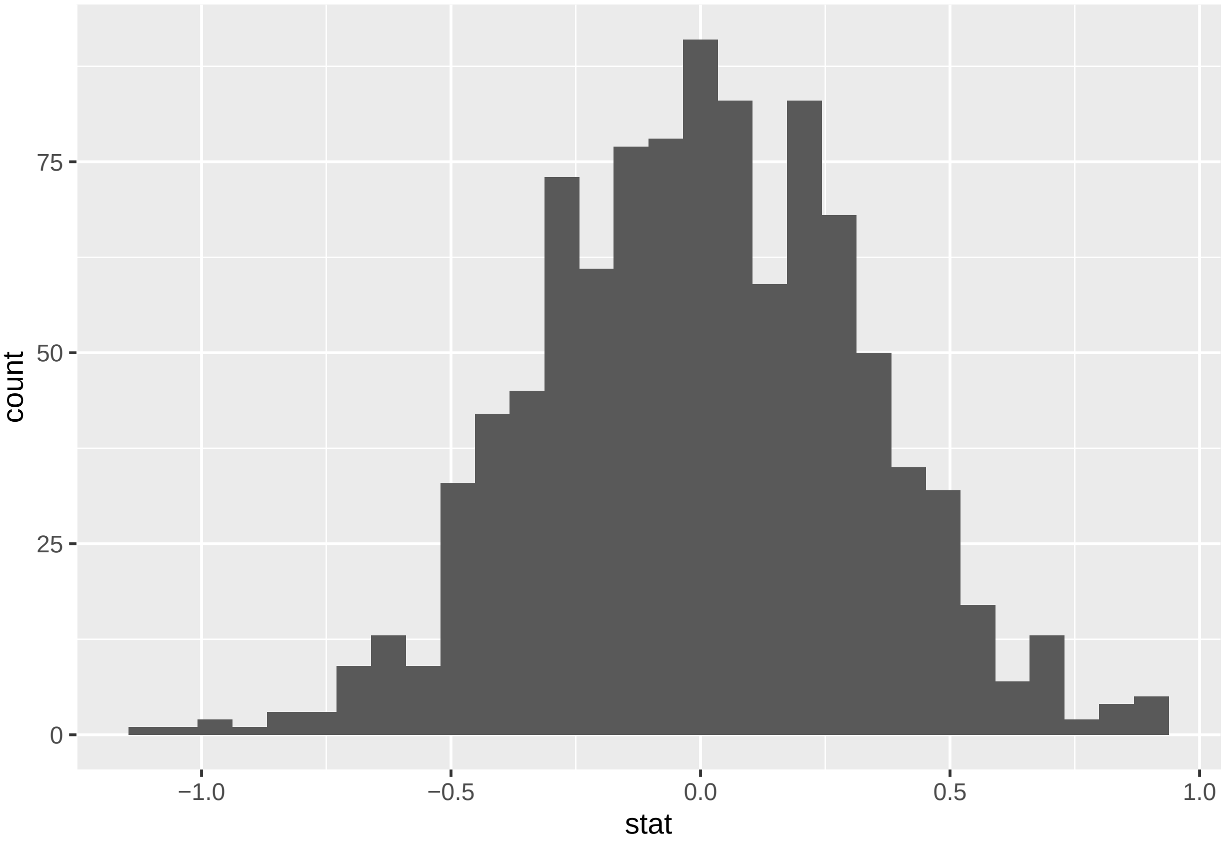
```
null_dist <- yrbss %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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:

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



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

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.

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

```
## [1] 0
```

Apparently zero??

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.

```
basics <- yrbss %>%
  group_by(physical_3plus) %>%
  summarize(count = n(),
            averages = mean(weight, na.rm = TRUE),
            deviations = sd(weight, na.rm = TRUE)) %>%
  drop_na()

x_bar <- diff(basics$averages)
SE <- sqrt(basics$deviations[1]^2 / basics$count[1] +
             basics$deviations[2]^2 / basics$count[2])

(interval <- c(x_bar - (1.96 * SE), x_bar + (1.96 * SE)))
```

```
## [1] 1.151287 2.397881
```

The real value for the difference among population means should fall within this interval 97.5% of the time.

More Practice

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

```
(y_complete %>%
   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     1.70
```

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.

```
(y_complete %>%
   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     1.70     1.70
```

The confidence level is lower, so the interval is narrower.

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.

H_0: The difference is zero. H_A: The difference is not zero.

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

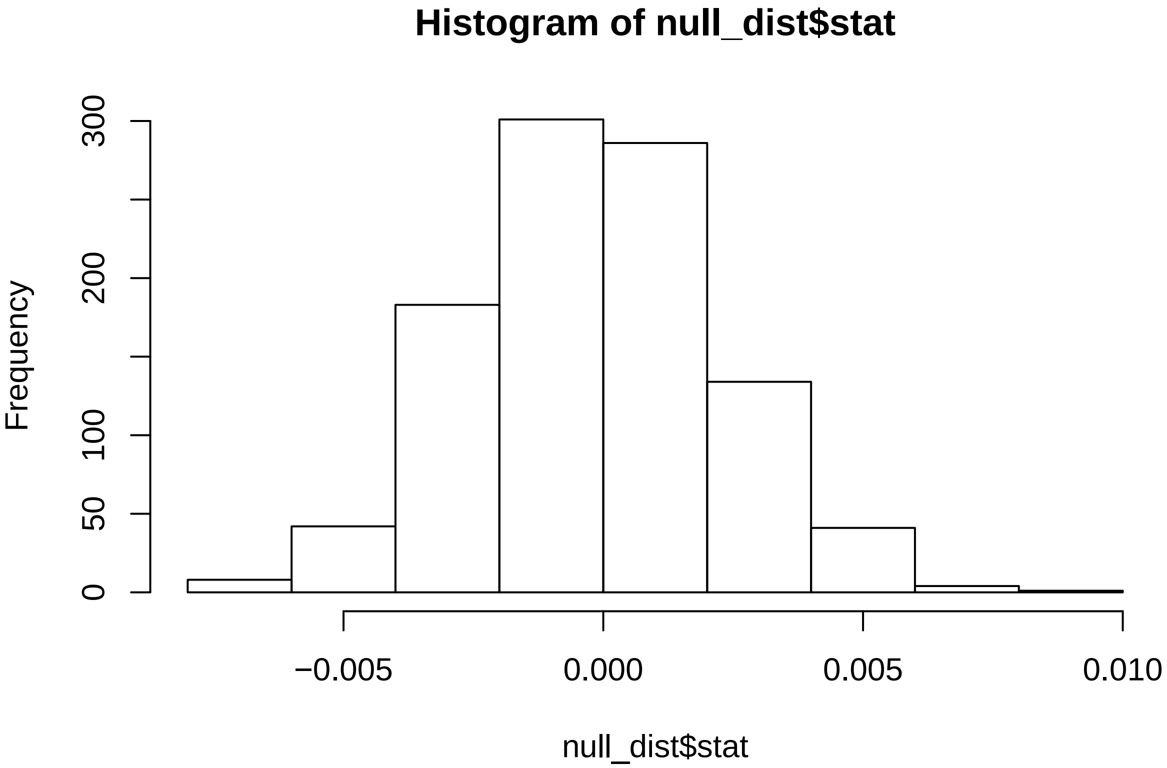
```
## # A tibble: 1 x 1
##     stat
##    <dbl>
## 1 0.0380
```

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

mean(null_dist$stat)
```

```
## [1] -0.0001709823
```

```
hist(null_dist$stat)
```



We fali to reject the null hypothesis. . . .

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.

```
unique(yrbss$hours_tv_per_school_day)
```

```
## [1] "5+"           "2"            "3"            "do not watch" "<1"
## [6] "4"            "1"            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.

Does getting enough sleep (7 or more hours) have an effect on wight?

H_0: The mean weight and the mean weight for people who sleep enough are the same. H_A: The two are not the same.

$\alpha = 0.05$

```
yrbss <- yrbss %>%
  mutate(enough_sleep = ifelse(school_night_hours_sleep %in% c("7","8","9","10+"),
                               "yes", "no"))
```

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

```
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 -1.27
```

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

mean(null_dist$stat)
```

```
## [1] 0.002564
```

We fail to reject the null hypothesis.
