Inference for numerical data

Mathew Katz

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 13853 cases in this data set.

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?

1004 NA's in this 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? **Before graphing I expected there to be a clear relationship between the two variables. It isn't super clear. These variables don't take into account the diet of the kids and therefore weight won't be directly correlated. If the variable included a section of 'works out and eats less than 2000 calories,' I think it would be an almost direct correlation.**

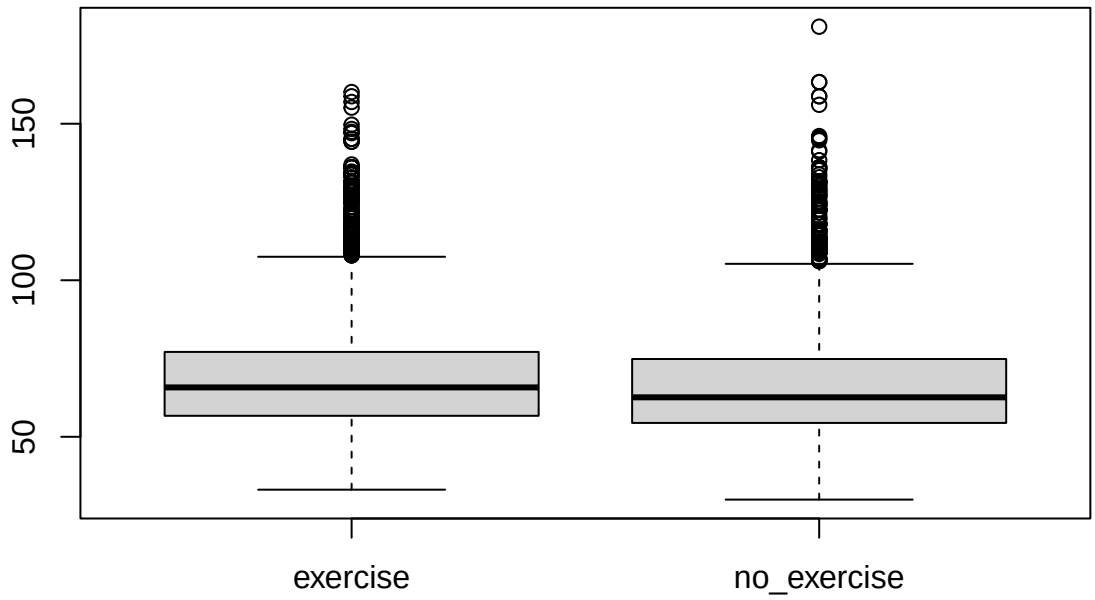
```
yrbss <- yrbss %>%
  mutate(physical_3plus = ifelse(yrbss$physically_active_7d > 2, "yes", "no"))

weight_exercise <- yrbss %>%
  filter(physical_3plus == "yes") %>%
  select(weight) %>%
```

```
  na.omit()

weight_noexercise <- yrbss %>%
  filter(physical_3plus == "no") %>%
  select(weight) %>%
  na.omit()

boxplot(weight_exercise$weight, weight_noexercise$weight,
        names = c("exercise", "no_exercise"))
```



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()`.

Yes;

1. Independant sample
2. Normality

```
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(mean_weight = mean(weight, na.rm = TRUE), count = n())
```

```
## # A tibble: 3 x 3
##   physical_3plus mean_weight count
##   <chr>                <dbl> <int>
## 1 no                    66.7  4404
## 2 yes                   68.4  8906
## 3 <NA>                  69.9   273
```

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.

The Null hypothesisis that there is no difference in average weights for those who exercise at least 3 times a week and those who don't.

An Alternative hypothesis is that tudents who are physically active 3 or more days per week have a different average weight when compared to those who are not physically active 3 or more days per week.

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"))
obs_diff
```

```
## Response: weight (numeric)
## Explanatory: physical_3plus (factor)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  1.77
```

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"))
null_dist
```

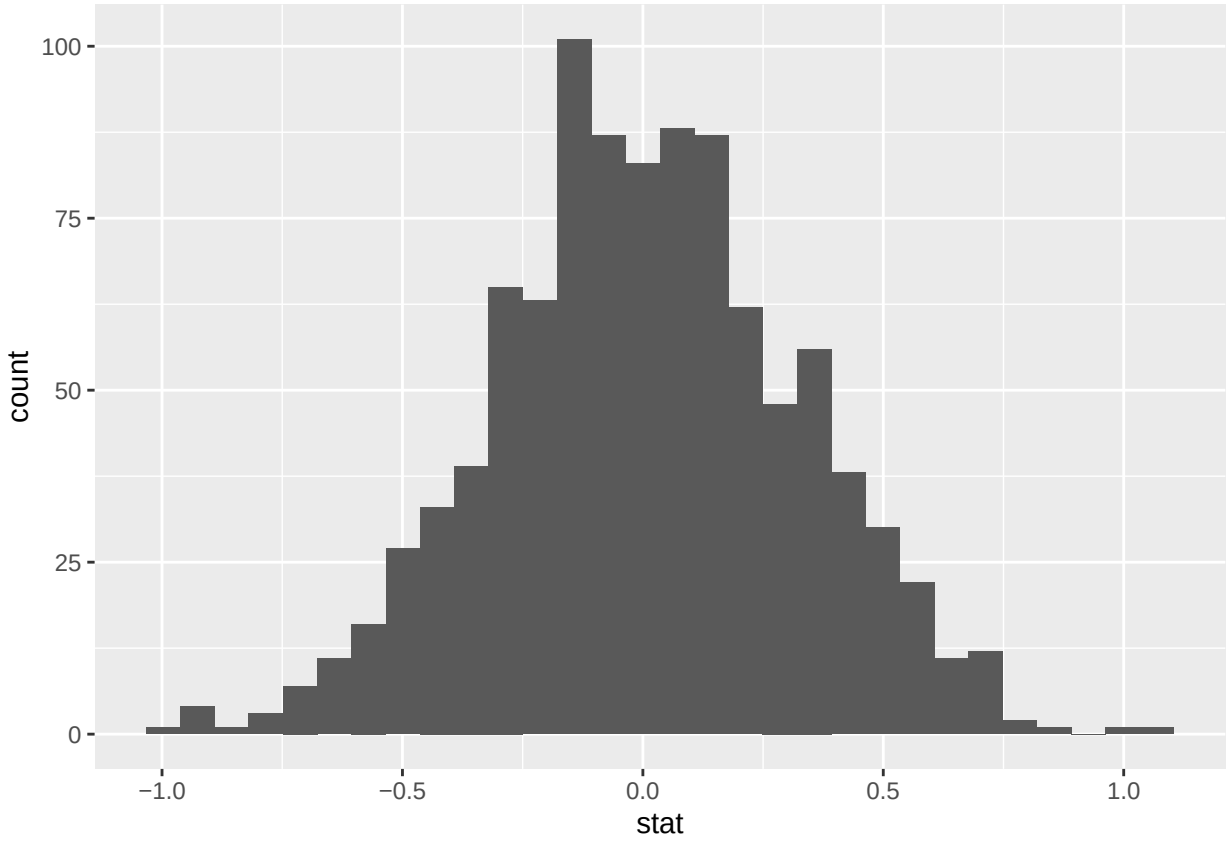
```
## Response: weight (numeric)
## Explanatory: physical_3plus (factor)
## Null Hypothesis: independence
## # A tibble: 1,000 x 2
##    replicate     stat
##        <int>    <dbl>
##  1         1 -0.0730
##  2         2 -0.364
##  3         3  0.279
##  4         4  0.600
##  5         5  0.264
##  6         6  0.0317
##  7         7  0.481
##  8         8 -0.0500
##  9         9  0.00306
## 10        10  0.0317
## # ... with 990 more rows
```

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`.

Zero of the 'null' permutations have a difference of at least obs_stat.

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

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

```
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.

```
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(sd_weight = sd(weight, na.rm = TRUE))
```

```
## # A tibble: 3 x 2
##   physical_3plus sd_weight
##   <chr>              <dbl>
## 1 no                  17.6
## 2 yes                 16.5
## 3 <NA>                17.6
```

```
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
```

```
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(freq = table(weight)) %>%
  summarise(n = sum(freq))
```

```
## # A tibble: 3 x 2
##   physical_3plus     n
##   <chr>          <int>
## 1 no              4022
## 2 yes             8342
## 3 <NA>             215
```

```
x_3 <- 66.67389
n_3 <- 4022
s_3 <- 17.63805
x3 <- 68.44847
n3 <- 8342
s3 <- 16.47832

z = 1.96

uci_not <- x_3 + z*(s_3/sqrt(n_3))
lci_not <- x_3 - z*(s_3/sqrt(n_3))
uci_not
```

```
## [1] 67.219
```

```
lci_not
```

```
## [1] 66.12878
```

```
u_ci <- x3 + z*(s3/sqrt(n3))
l_ci <- x3 - z*(s3/sqrt(n3))

u_ci
```

```
## [1] 68.80209
```

```
l_ci
```

```
## [1] 68.09485
```

With 95% confident that students who exercise at least three times a week have an average weight between 68.09 kg and 68.8 kg. Also those students who do not exercise at least three times a week have an average weight between 66.13 kg and 67.22 kg with 95% confident.

More Practice

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

```
x_h <- mean(yrbss$height, na.rm = TRUE)
sd_h <- sd(yrbss$height, na.rm = TRUE)
n_h <- yrbss %>%
  summarise(freq = table(height)) %>%
  summarise(n = sum(freq, na.rm = TRUE))

u_h <- x_h + z*(sd_h/sqrt(n_h))
l_h <- x_h - z*(sd_h/sqrt(n_h))
u_h
```

```
##          n
## 1 1.693071
```

```
l_h
```

```
##          n
## 1 1.689411
```

The average height of the students in this population is between 1.689m and 1.693m.

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.

```
t_90 <- 1.645
upper_ci_height_90 <- x_h + t_90*(sd_h/sqrt(n_h))
lower_ci_height_90 <- x_h - t_90*(sd_h/sqrt(n_h))
upper_ci_height_90
```

```
##          n
## 1 1.692777
```

```
lower_ci_height_90
```

```
##          n
## 1 1.689705
```

The new confidence interval is 1.689705 to 1.692777. Our intervals at a 95% confidence level were 1.689411 and 1.693071. The difference in these two confidence intervals are below:

```
dif1 <- (u_h - l_h)
dif2 <- (upper_ci_height_90 - lower_ci_height_90)
dif1
```

```
##             n
## 1 0.003659302
```

```
dif2
```

```
##           n
## 1 0.0030712
```

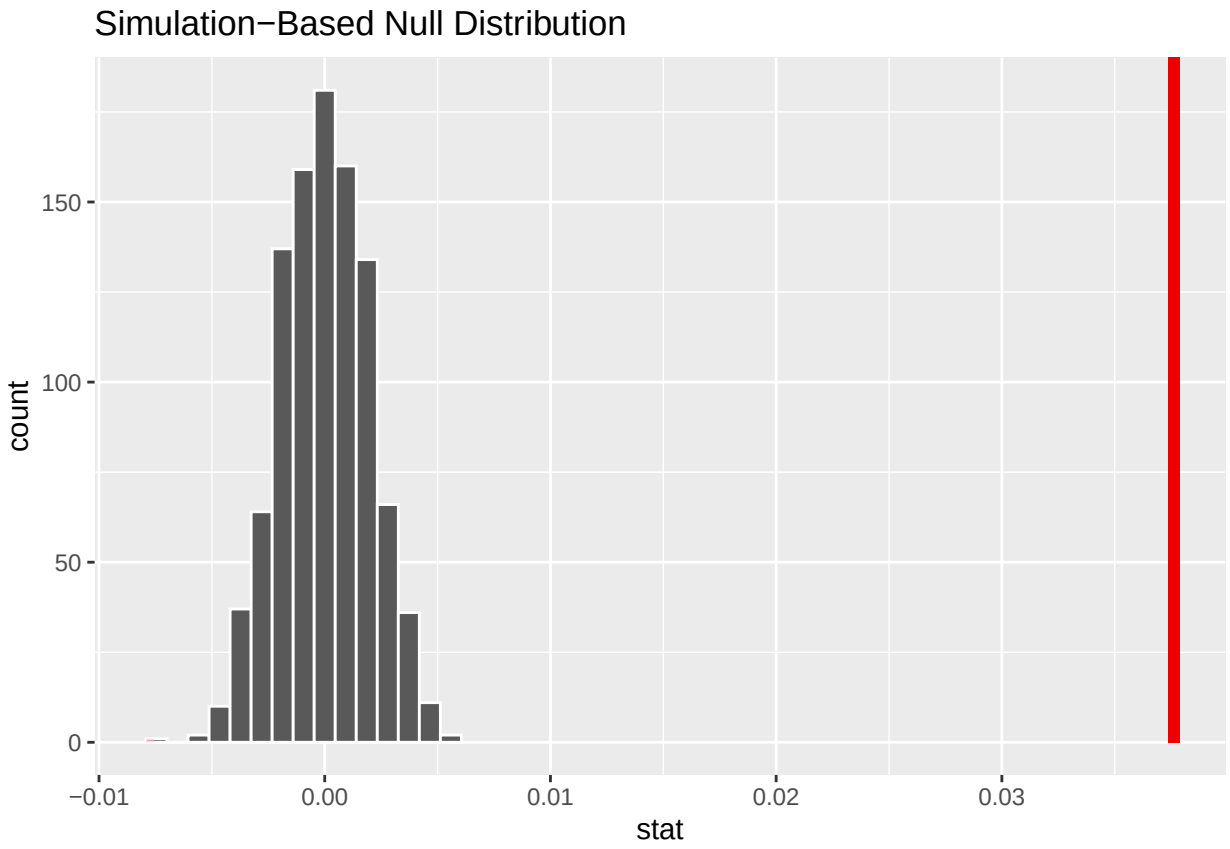
The 95% confidence interval has a slightly larger range than the confidence interval 90%

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.

```
obs_diff_hgt <- yrbss %>%
  filter(!(is.na(physical_3plus) | is.na(height))) %>%
  specify(height ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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

```
visualize(null_dist_hgt) +
  shade_p_value(obs_stat = obs_diff_hgt, direction = "two_sided")
```




```
null_dist_hgt %>%
  get_p_value(obs_stat = obs_diff_hgt, direction = "two_sided")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

```
x_t <- 1.6665
n_t <- 4022
s_t <- 0.1029
x_yt <- 1.7032
n_yt <- 8342
s_yt <- 0.1033

z = 1.96


ut <- x_t + z*(s_t/sqrt(n_t))
lt <- x_t - z*(s_t/sqrt(n_t))
ut
```

```
## [1] 1.66968
```

```
lt
```

```
## [1] 1.66332
```

```
uyt <- x_yt + z*(s_yt/sqrt(n_yt))
lyt <- x_yt - z*(s_yt/sqrt(n_yt))
uyt
```

```
## [1] 1.705417
```

```
lyt
```

```
## [1] 1.700983
```

With 95% confident that the average height of students who are physically active at least 3 days per week is between 1.705 and 1.701 and the average height of students who are not physically active at least 3 days per week is between 1.670 and 1.663. 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.

```
yrbss %>%group_by(hours_tv_per_school_day)%>% summarise(n())
```

```
## # A tibble: 8 x 2
##   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
```

There are 7 different options.

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.
