Inference for numerical data

William Jasmine

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)

seed <- 1234
```

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)`.

WJ Response:

```
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"~
```

Using the `glimpse` command above, we see that there are 13,583 cases in the `yrbss` 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?

WJ Response:

Based on the above output of the `summary` function, the `weight` column in the `yrbss` column has 1,004 NA values. These comprise all the missing values, since it is a numeric column and cannot contain empty or blank strings.

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"))
```

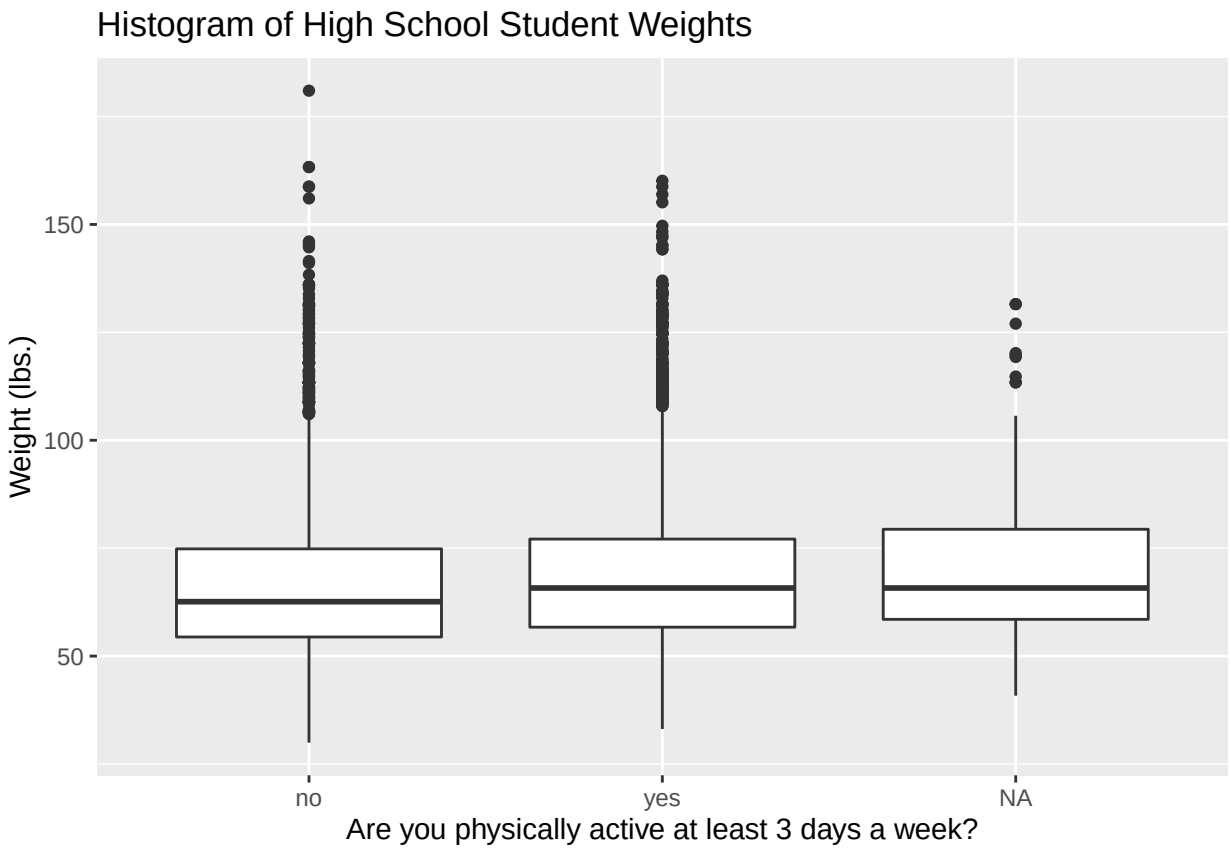
3. Make a side-by-side boxplot of `physical_3plus` and `weight`. Is there a relationship between these two variables? What did you expect and why?

WJ Response:

The cell below creates boxplots of two the `weight` column, grouped by the newly created `physical_3plus`:

```
ggplot(data = yrbss, aes(x = physical_3plus, y = weight)) +
  geom_boxplot() +
    labs(x = 'Are you physically active at least 3 days a week?',
         y = 'Weight (lbs.)',
```

```
        title = 'Histogram of High School Student Weights'
        )
```



The plot above shows that the two boxplots have a decent amount of overlap indicating that there might not be a very strong relationship between these two variables. This maybe goes against expectations (most people's first instinct might be people that are more physically active weigh less on average), but these ignore other factors such as this data only includes high school students, and that people who work out more might actually weigh more due to increased muscle mass. However, further testing will be required to prove this either way.

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()`.

WJ Response:

The conditions for inference when comparing the means of two groups are:

- Independence within groups
- Independence between groups
- The size of each group is large enough

The groups in this case are those students who work out at least 3 days a week, compared to those who don't. The first two conditions are easy to address: the description of the `yrbss` data states that the sample of high school students has been randomly selected; furthermore it cannot be the case that a high school student is physically active both at least 3 days a week and less than 3 days a week. Thus, we can say that there is independence within and between groups.

The last condition states that the size of each group is large enough. In this case, we will consider a sample size of 30 as sufficiently large given is the size required for the CLT to be applicable and the distribution of sample means from each group to be approximately normal. The code cell below checks to make sure this is indeed the case:

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

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

As is clear in the output above, each group has is of sufficiently large size ($n > 30$).

---

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.

WJ Response:

Given μ_1 = the mean weight of those students who are physically active at least 3 days a week and μ_2 = the mean weight of those students who are physically active less than 3 days a week, we can define the following null and alternative hypotheses (H_0 and H_A, respectively):

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

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`.

```r
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`.

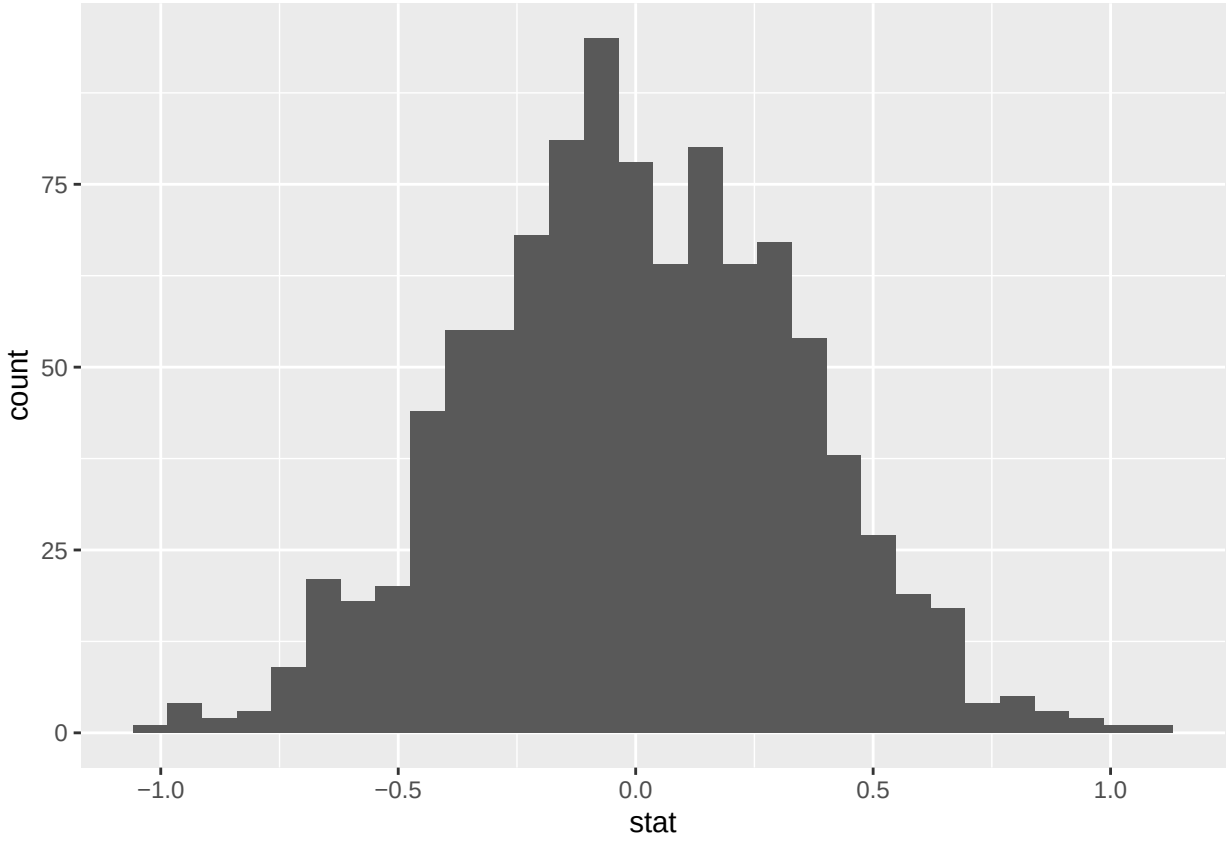
```r
set.seed(seed)

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`, which is the argument when generating a null distribution for a hypothesis test.

We can visualize this null distribution with the following code:

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



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

WJ Response:

The code chunk below checks to see if any of the permutations in the `null_dist` data frame have a difference in means greater than what is observed in the `yrbss` data set (=1.77).

```
(null_dist$stat > obs_diff[[1]]) %>%
  table()
```

```
## .
## FALSE
##  1000
```

The output above reveals that all permutations have a difference below what was observed. This is also clearly seen in the above histogram, as the mean difference values span only from about -1 to 1.

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.

WJ Response:

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

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.656    0.656
```

The 95% confidence interval using the null distribution for the difference of means when comparing the weights of those students who are physically active at least 3 time a day and those who are not, is far from the range of the observed difference. In fact, the upper range of the confidence interval would need to be almost three times larger to encapsulate the observed difference of 1.77.

This confidence interval supports the fact that our p-value is so low, and that we can reject the null hypothesis. In other words, there is a significant difference on weight when comparing those that are very physically active to those who are not. For even further proof of this, we can calculate the confidence interval of the difference between means from the actual data (as opposed to the null distribution):

```
set.seed(seed)

yrbss %>%
  filter(!(is.na(physical_3plus) | is.na(weight))) %>%
    specify(weight ~ physical_3plus) %>%
    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     1.16     2.46
```

Note that now the confidence interval does not span 0, providing a third way to prove significance.

More Practice

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

WJ Response:

The code below calculates the confidence interval at 95%:

```
set.seed(seed)

yrbss %>%
```

```
    filter(!is.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
```

We can see that this is not a helpful confidence interval, given that the numbers do not include enough decimal points to see any noticeable difference between them. As such, the following code chunk recalculates the confidence interval using the actual formula:

$$CI = \mu \pm Z \cdot \frac{\sigma}{\sqrt{n}}$$

```
height_stats <-
  yrbss %>%
    filter(!is.na(height)) %>%
      summarise(avg = mean(height),
                sigma = sd(height),
                size = n())

avg = height_stats$avg[[1]][1]
sigma = height_stats$sigma[[1]][1]
size = height_stats$size[[1]][1]
cl = 0.95
crit_val = abs(qnorm((1-cl)/2))

ci_lower = avg - (crit_val * (sigma / sqrt(size)))
ci_upper = avg + (crit_val * (sigma / sqrt(size)))

cat(ci_lower, ci_upper)
```

```
## 1.689411 1.693071
```

The number of the left represents the lower part of the confidence interval while the right represents the upper part. With this higher level of detail we can now see that there is an actual difference between the upper and lower confidence intervals. This interval means we are 95% confident that the average student height falls between those two values.

---

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.

WJ Response:

The code below calculates the confidence interval at 90%:

```
height_stats <-
  yrbss %>%
    filter(!is.na(height)) %>%
      summarise(avg = mean(height),
```

```
                sigma = sd(height),
                size = n())

avg = height_stats$avg[[1]][1]
sigma = height_stats$sigma[[1]][1]
size = height_stats$size[[1]][1]
cl = 0.90
crit_val = abs(qnorm((1-cl)/2))

ci_lower = avg - (crit_val * (sigma / sqrt(size)))
ci_upper = avg + (crit_val * (sigma / sqrt(size)))

cat(ci_lower, ci_upper)
```

```
## 1.689705 1.692776
```

In this case, this confidence interval means we are 90% confident that the mean of student height falls between those two values. Note that this interval is slightly smaller than the previous because the more the a confidence interval shrinks, the less sure you are that the actual value falls within it.

---

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.

WJ Response:

First, we define our null and alternative hypotheses. Given that μ_1 = the mean height of someone who exercises at least 3 times a week and μ_2 = the mean height of someone who exercises less that 3 times a week:

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

Now, for conducting the actual test:

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

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

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

The above value represents the approximate p value, which is 0. Given this result, we can conclude that there is a statistically significant difference in the mean height when comparing those students who exercise at least 3 times a week and those who don't. This would hold true even for values of α extremely close to 0.

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.

WJ Response:

```
(yrbss %>%
   filter(!is.na(hours_tv_per_school_day)))$hours_tv_per_school_day %>%
  unique() %>%
    length()
```

```
## [1] 7
```

There are 7 options for the values present in the `hours_tv_per_school_day` column (8 if you consider not responding to the question as an option).

---

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.

WJ Response:

Question: Does the average weight of a high school student differ for those students who get less 5 or less hours of sleep compared to those who don't?

Hypotheses: Given μ_1 = mean weight of students who get 5 or less hours of sleep and μ_2 = mean weight of students who get more than 5 hours of sleep a night:

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

Assumptions: There exists independence both within and between these two groups given the data collection methodology and the phrasing of the research question. The last condition requires that we check the size of each group:

```
yrbss <- yrbss %>%
  mutate(less5_sleep =
           ifelse(school_night_hours_sleep %in% c('5', '<5'), 'yes', 'no'))

yrbss %>%
  filter(!(is.na(less5_sleep) | is.na(weight))) %>%
    count(less5_sleep)
```

```
## # A tibble: 2 x 2
##   less5_sleep     n
##   <chr>       <int>
## 1 no          10342
## 2 yes          2237
```

In our sample there are 2,237 students who get 5 or less hours of sleep, and 10,342 that do not. Given the large sample size we can assume that the central limit theorem holds and that all conditions for inference are met.

Performing the test:

The hypotheses test is carried out below using confidence intervals and an $\alpha = 0.01$:

```
set.seed(seed)

yrbss %>%
  filter(!(is.na(less5_sleep) | is.na(weight))) %>%
    specify(weight ~ less5_sleep) %>%
    generate(reps = 1000, type = "bootstrap") %>%
    calculate(stat = "diff in means", order = c('yes', 'no')) %>%
    get_ci(level = 0.99)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.435     2.76
```

We see that even using the 99% confidence interval, the range does not span 0 (the expected difference in means if H_0 were true). Thus, we reject H_0 and conclude that there is a statistically significant difference in the average weight when comparing those students who get 5 or less hours of sleep to those who don't.
