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?

Insert your answer here

By using the glimpse function below, we can see there are 13,583 rows. Each represents one observation and case.

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?

Insert your answer here

By using the summary function above we can see there are 1,004 missing observations (NAs)

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"))
```

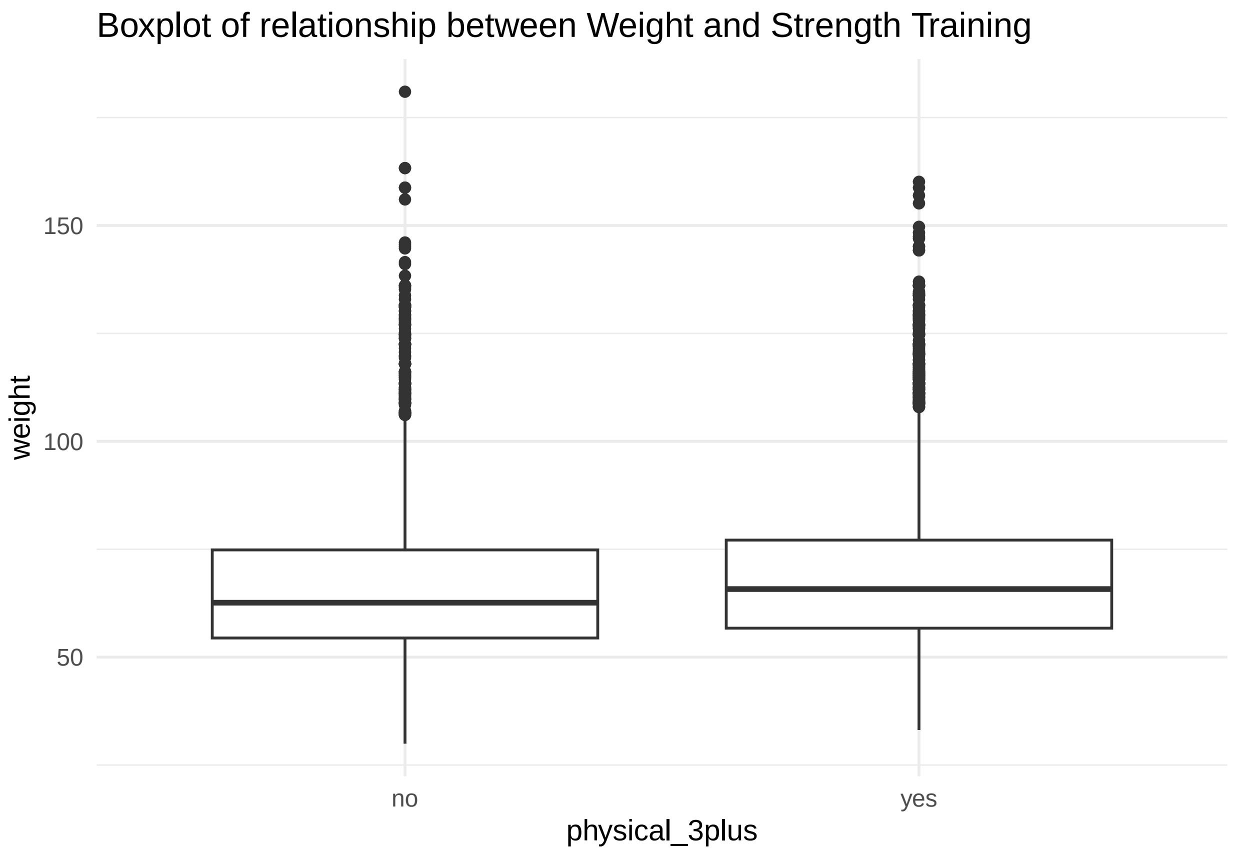
3. Make a side-by-side boxplot of `physical_3plus` and `weight`. Is there a relationship between these two variables? What did you expect and why?

Insert your answer here

The below boxplots show that the students who weighed more were slightly more likely to work out at least 3 days a week. This is due to increased muscle mass and in some cases overweight students looking to become fitter. This is the data I would expect to see. We do see more fluctuations in outliers among those who didn't work out, this is probably due to everweight people having a larger variance than fit people getting their weight from muscke mass.

```r
yrbss_x <- yrbss %>%
  mutate(physical_3plus = ifelse(yrbss$physically_active_7d > 2, "yes", "no")) |>
  filter(!is.na(physical_3plus))
```

```r
ggplot(yrbss_x, aes(x = physical_3plus, y = weight)) +
  geom_boxplot() +
  theme_minimal() +
  ggtitle("Boxplot of relationship between Weight and Strength Training")
```



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`.

```r
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()`.

Insert your answer here

Since there is greater than 10 successes and failures, we know we will have normal distribution when doing inference do to the CLT. I assume the data was randomly selected and that the datapoints are independent of each other.

```r
yrbss_x %>%
  drop_na() %>%
  group_by(physical_3plus) %>%
  summarise(freq = table(weight)) %>%
  summarise(n = sum(freq))
```

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

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.

Insert your answer here

H0: There is no difference in average weights for those who exercise at least 3 times a week and those who don't. Ha: There is a difference in average weights for those who exercise at least 3 times a week and those who don't.

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 %>%
  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`.

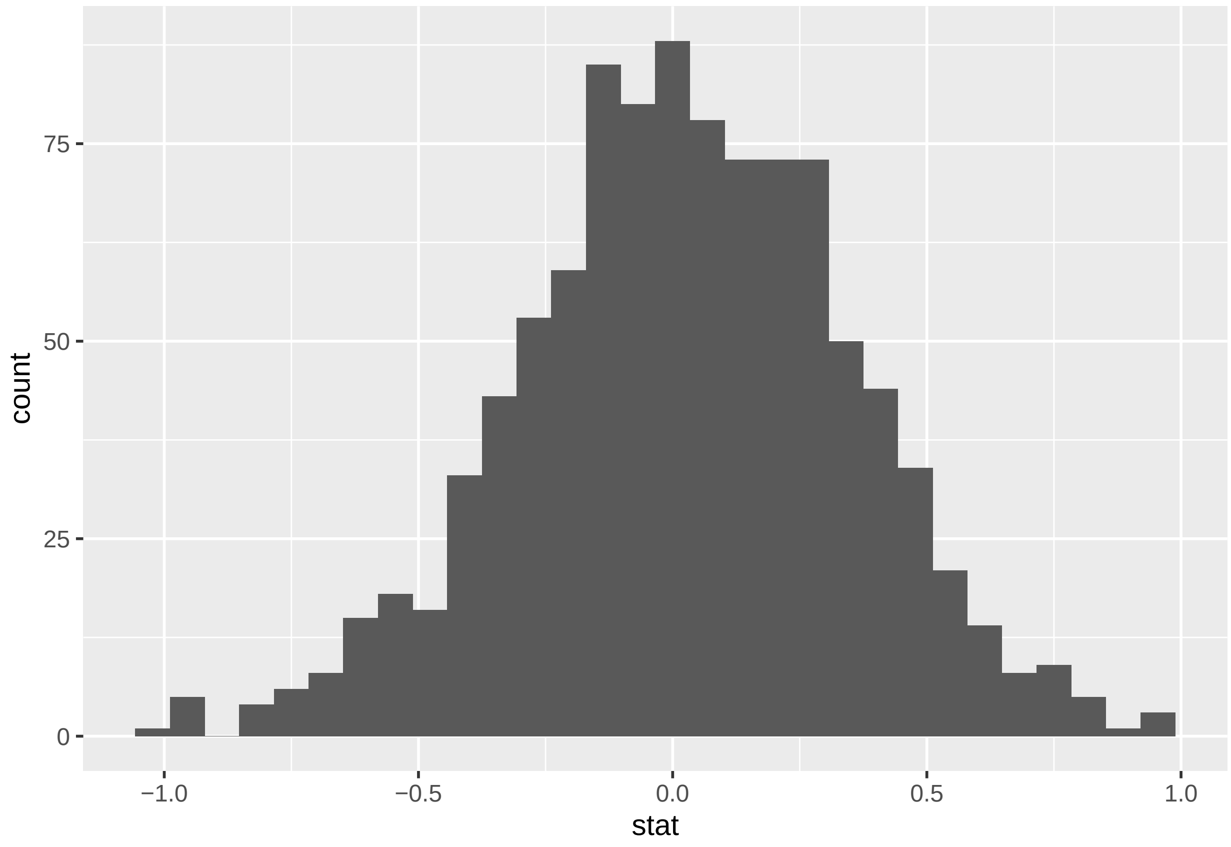
```
null_dist <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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`?

Insert your answer here

The below code shows that zero of the null permutations had a difference of at least the obs_diff.

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

```
## Response: weight (numeric)
## Explanatory: physical_3plus (factor)
## Null Hypothesis: independence
## # A tibble: 0 x 2
## # i 2 variables: replicate <int>, stat <dbl>
```

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.

```
#(confidence_interval <- c(mean(null_dist$stat) - 1.96 * (sd(null_dist$stat) / #sqrt(length(null_dist$s
```

```
yrbss_weight_physical <- yrbss[complete.cases(yrbss$weight, yrbss$physical_3plus), ]

group_stats <- yrbss_weight_physical %>%
  group_by(physical_3plus) %>%
  summarise(
    mean_weight = mean(weight),
    sd_weight = sd(weight),
    n = n()
  )

diff_mean <- diff(group_stats$mean_weight)

diff_se <- sqrt(sum((group_stats$sd_weight)^2 / group_stats$n))

(confidence_interval <- c(diff_mean - 1.96 * diff_se, diff_mean + 1.96 * diff_se))
```

```
## [1] 1.124821 2.424348
```

The confidence interval ranges from 1.125 to 2.424. Since the ci spans the observed differential we can reject the null hypothesis and accept the alternative that there is a correlation between these two variables. This fits with our p value being 0.

More Practice

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

Insert your answer here

```
(confidence_interval95 <- c(mean(yrbss$height, na.rm = TRUE) - 1.96 * (sd(yrbss$height, na.rm = TRUE) /
```

```
## [1] 1.689480 1.693002
```

The confidence interval is 1.689480-1.693002. this means that we are 95% sure that the mean of the height is between those numbers nd is therefore roughly 1.69 meters.

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.

Insert your answer here

```
z <- qnorm((1 + 0.90) / 2)

(confidence_interval90 <- c(mean(yrbss$height, na.rm = TRUE) - z * (sd(yrbss$height, na.rm = TRUE) / sq
```

```
## [1] 1.689763 1.692719
```

The confidence interval is the same 1.689763-1.692719. This is means we are 90% confident that the mean is between those numbers. Normally we expect the 90% ci to be a narrower range than the 95%ci since we are less certain about it. However, in this instance the difference is so small to be almost completely irrelevant (yet still there), this is due to the range being so small already.

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.

Insert your answer here

H0: There is no difference in average heights for those who exercise at least 3 times a week and those who don't. Ha: There is a difference in average heights for those who exercise at least 3 times a week and those who don't.

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

```
null_dist_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"))
```

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

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

```
#(confidence_interval_test <- c(mean(null_dist_height$stat) - 1.96 * (sd(null_dist_height$stat) / #sqrt

yrbss_height_physical <- yrbss[complete.cases(yrbss$height, yrbss$physical_3plus), ]

group_stats2 <- yrbss_height_physical %>%
  group_by(physical_3plus) %>%
  summarise(
    mean_height = mean(height),
    sd_height = sd(height),
    n = n()
  )

diff_mean2 <- diff(group_stats2$mean_height)

diff_se2 <- sqrt(sum((group_stats2$sd_height)^2 / group_stats2$n))

(confidence_interval2 <- c(diff_mean2 - 1.96 * diff_se2, diff_mean2 + 1.96 * diff_se2))
```

```
## [1] 0.03375046 0.04150131
```

I got a p value of zero again. This means we should reject the null hypothesis and accept the alternative, there is a difference in average height between those who are physically active 3+ times a week and those who aren't. The observed differential falls within the CI spanning 0.0337-0.0415. This also means that we should 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.

Insert your answer here

```
(unique_options <- unique(yrbss$hours_tv_per_school_day))
```

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

There are 8 different options in the selected column including NAs.

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.

Insert your answer here

Question

Does the average weight of a student who sleeps more than 8 hours a night differ from the average weight of a student who does not sleep at least 8 hours a night?

Null hypothesis is that there is no correlation between the average weight of a student and the amount of hours they sleep. Alternative hypothesis is that there is indeed a relationship. We will use an α of 0.05. That means we will use a threshold that ensures we are 95% confident in our final conclusion.

```
#add the column with the success or failure and drop NAs
yrbss_z <- yrbss |>
  mutate(sleep_8plus = ifelse((yrbss$school_night_hours_sleep == "8") | (yrbss$school_night_hours_sleep
  filter(!is.na(sleep_8plus))
```

```
#manually check the observed means
yrbss_z %>%
  group_by(sleep_8plus) %>%
  summarise(mean_weight = mean(weight, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##   sleep_8plus mean_weight
##   <chr>             <dbl>
## 1 no                 68.2
## 2 yes                67.2
```

```
#save the observed differential
obs_diff_weight <- yrbss_z %>%
  specify(weight ~ sleep_8plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
null_dist_weight <- yrbss_z %>%
  specify(weight ~ sleep_8plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.004
```

The p value is 0.006. Since this is lower than the 0.05 value we decided to use as our α we reject the null hypothesis and accept the alternative that there is a relationship between weight and sleeping 8+ hours a night.

We can further illustrate this by calculating confidence intervals

```
#get complete cases
yrbss_weight_sleep <- yrbss_z[complete.cases(yrbss_z$weight, yrbss_z$sleep_8plus), ]

group_stats3 <- yrbss_weight_sleep %>%
  group_by(sleep_8plus) %>%
  summarise(
    mean_weight = mean(weight),
    sd_weight = sd(weight),
    n = n()
  )

diff_mean3 <- diff(group_stats3$mean_weight)

diff_se3 <- sqrt(sum((group_stats3$sd_weight)^2 / group_stats3$n))

(confidence_interval3 <- c(diff_mean3 - 1.96 * diff_se3, diff_mean3 + 1.96 * diff_se3))
```

```
## [1] -1.6537692 -0.3141333
```

The CI ranges from -1.65 to -0.31. This means that we are 95% that sleeping more than 8 hours will have a correlation with being somewhere between those numbers lighter than someone who does not sleep 8 hours. Since those numbers are all negative, that means we are 95% sure that someone who sleep 8+ hours will be heavier than someone who doesn't. The observed differential in this sample was indeed within this range at -0.98, meaning those who slept 8+ hours were on average 0.98 kilograms heavier than those who did not.