Inference for numerical data

Nick Climaco

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?

```
colnames(yrbss)
```

```
##  [1] "age"                     "gender"
##  [3] "grade"                   "hispanic"
##  [5] "race"                    "height"
##  [7] "weight"                  "helmet_12m"
##  [9] "text_while_driving_30d"  "physically_active_7d"
## [11] "hours_tv_per_school_day" "strength_training_7d"
## [13] "school_night_hours_sleep"
```

There are 13583 observations and 13 variables columns where some hold categorical and some numerical data.

```
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

```
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?

```
yrbss |>
    count(is.na(weight))
```

```
## # A tibble: 2 x 2
##   `is.na(weight)`     n
##   <lgl>           <int>
## 1 FALSE           12579
## 2 TRUE             1004
```

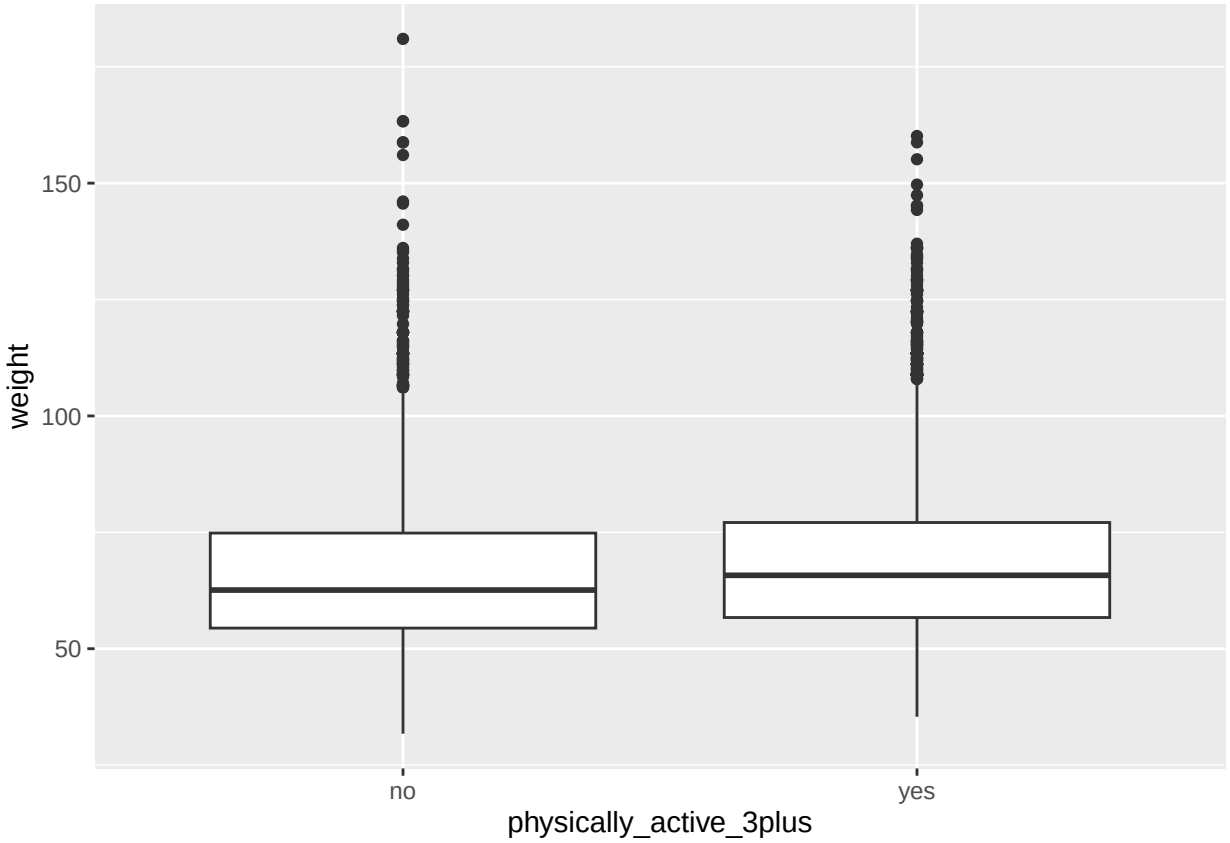
There are 1004 missing values in the weight column.

---

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?

```
yrbss <- yrbss |>
    mutate(physically_active_3plus = if_else(physically_active_7d > 2, "yes", "no")) |>
    na.omit()
```

```
yrbss |>
    ggplot(aes(x = physically_active_3plus, y = weight)) +
    geom_boxplot()
```



We expect that being physically active has relationship with weight due to the lower variance in weight for people who are physically active for at least 3 days in a week. Also, we notice that people who do not exercise regularly have wider range in weight. We need to conduct a hypothesis test to determine whether exercising regularly has a statistical signficant impact on one's weight.

Inference

```
yrbss |>
    group_by(physically_active_3plus) |>
    summarize(mean_weight = mean(weight, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##   physically_active_3plus mean_weight
##   <chr>                         <dbl>
## 1 no                             67.1
## 2 yes                            68.7
```

---

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()`.

The two conditions for inference are indepence and normality. So, looking at how the dataset was collected, a survey of randomly selected highschoolers from across the country. Each observation is independent. As for normality, we do not see any extreme outliers in the data so the condition is met.

```
yrbss |>
    group_by(physically_active_3plus) |>
    summarize(group_size = n())
```

```
## # A tibble: 2 x 2
##   physically_active_3plus group_size
##   <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.

$$h_0 : \mu_{diff} = 0 h_a : \mu_{diff} \neq 0$$

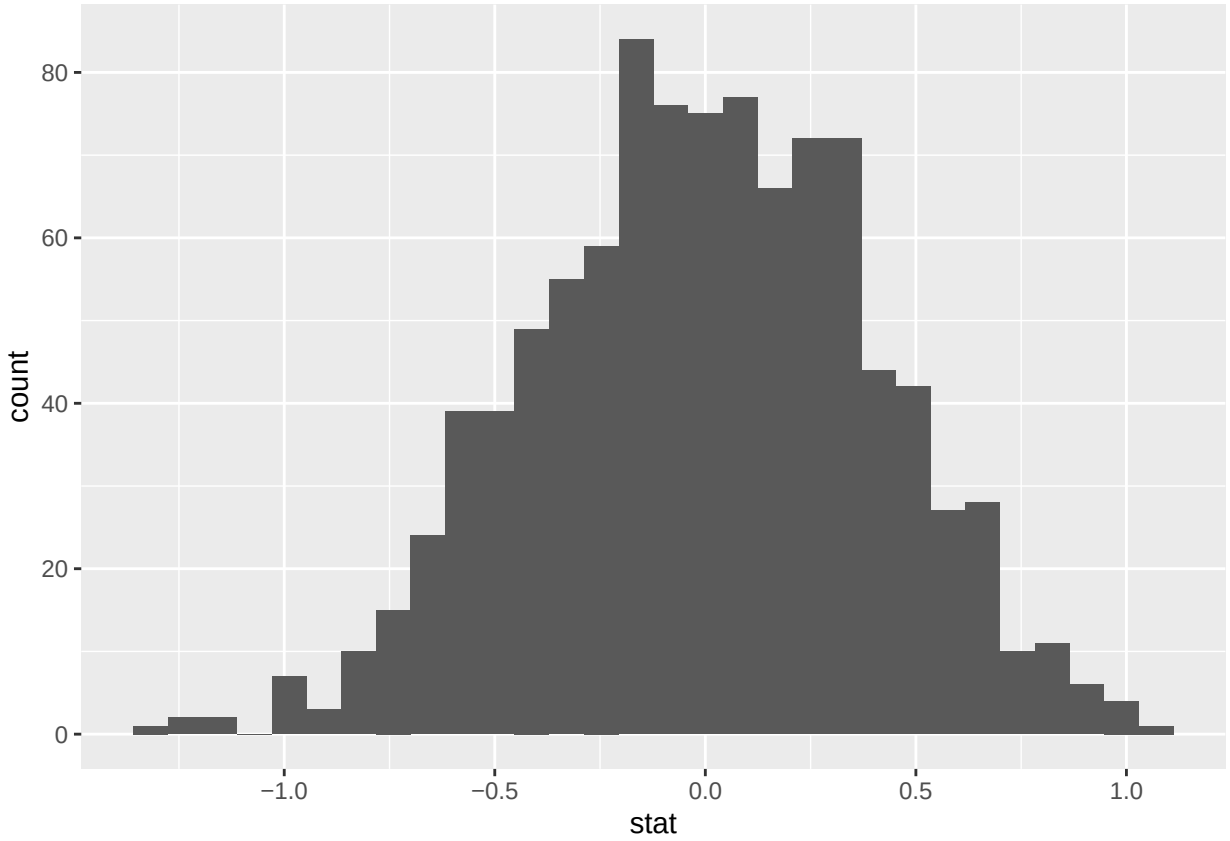
The null hypothesis where mean difference is equal to zero is the status que i.e. exercise has no impact on weight The alternative is where mean differecen is not zero suggests that exercise does have an impact on weight

```
set.seed(1998)
```

```
obs_diff <- yrbss %>%
  specify(weight ~ physically_active_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
null_dist <- yrbss %>%
  specify(weight ~ physically_active_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
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) |>
    nrow()
```

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

---

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.

Those who exercise at least 3 times a week

```
sub_yrbss <- yrbss |>
    na.omit() |>
    filter(physically_active_3plus %in% c('yes','no'))

t_test <- t.test(weight ~ physically_active_3plus, data = sub_yrbss, var.equal = FALSE, conf.int = TRUE)
t_test$conf.int
```

```
## [1] -2.3340891 -0.7213629
## attr(,"conf.level")
## [1] 0.95
```

The confidence interval (-2.33, -0.72) at alpha = 0.5, suggests that the true difference in mean falls inside the interval. Since, zero is not in the interval we say that we reject the null hypothesis. Thus, we can interpret the interval as evidence that there is a statistically significant difference between the two mean which then suggests that exercise does impact a person's weight.

More Practice

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

```
t.test(yrbss$height, conf.level = 0.95, conf.int = TRUE)
```

```
##
##  One Sample t-test
##
## data:  yrbss$height
## t = 1482.5, df = 8350, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.694810 1.699298
## sample estimates:
## mean of x
##  1.697054
```

The confidence interval (1.695, 1.699) at 95% confidence, the true average height is between these two values. So the average height is around 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.

```
t.test(yrbss$height, conf.level = 0.9, conf.int = TRUE)
```

```
##
##  One Sample t-test
##
## data:  yrbss$height
## t = 1482.5, df = 8350, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  1.695171 1.698937
## sample estimates:
## mean of x
##  1.697054
```

We expected the interval to narrow as we decreased the confidence level to 90% because we are less certain that the true mean in the interval.

---

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.

```
t_test <- t.test(height ~ physically_active_3plus, data = sub_yrbss, var.equal = FALSE, conf.int = TRUE)
t_test$conf.int
```

```
## [1] -0.04274834 -0.03329339
## attr(,"conf.level")
## [1] 0.95
```

---

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 |>
    count(hours_tv_per_school_day)
```

```
## # A tibble: 7 x 2
##   hours_tv_per_school_day     n
##   <chr>                   <int>
## 1 <1                       1407
## 2 1                        1172
## 3 2                        1738
## 4 3                        1309
## 5 4                         627
## 6 5+                        966
## 7 do not watch             1132
```

There are 7 unique categories including that student indicate how much time they watch tv.

---

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.

Question: Do the mean height of student who get enough sleep differ significantly to the mean height of student who lack sleep?

We will the compare means of the two groups similar to how we compared the means of the students who exercise at least 3 days a week in relation to thier weights. In this compare, we will compare the mean of the students who get enough sleep and students who lack sleep in relation to their height. Determining if sleep impact the height development of the students. In addition, we will alpha level 0.05 for this test.

```
subset_sleep <- yrbss |>
    mutate(enough_sleep = if_else(school_night_hours_sleep >= 8 | school_night_hours_sleep == "10+", "y
```

```r
t_test <- t.test(height ~ enough_sleep, data = subset_sleep, var.equal = FALSE, conf.int = TRUE)
t_test$conf.int
```

```
## [1] -0.005845808  0.004082679
## attr(,"conf.level")
## [1] 0.95
```

At 95% confidence, the confidence interval include zero which suggests that we fail to reject the null hypothesis. We conclude that getting enough sleep does not significantly hinder the height development of students, but might still hinder other aspects of the students development which can be examined in a different tests.
