Inference for numerical data

```
library(tidyverse)
library(openintro)
library(infer)
```

Exercise 1

```
data('yrbss', package='openintro')
```

```
?yrbss
```

```
glimpse(yrbss)
```

What are the cases in this data set? How many cases are there in our sample?

```
## 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 13,583 observations in this dataset and each row represents a student between 9th and 12th grade in public or private school in the US.

Exercise 2

```r
summary(yrbss$weight)
```

How many observations are we missing weights from?

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   29.94   56.25   64.41   67.91   76.20  180.99    1004
```

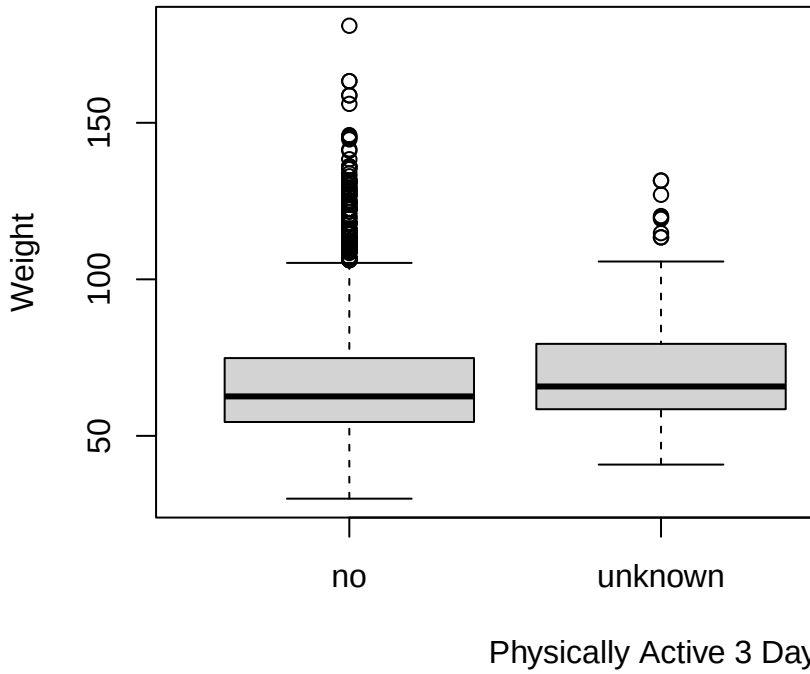
We have 1,004 NAs so that's how many we're missing.

Exercise 3

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

boxplot(yrbss$weight ~ yrbss$physical_3plus, data=yrbss, ylab="Weight", xlab="Physically Active 3 Days")
```

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?

I expected people who exercise more than 3 days to weigh less, but it looks like both groups are actually pretty similar so there's not a clear relationship at all between the two. I suppose that if the kids who are physically active do weigh more, it could be because of higher muscle mass. It does look like though that there aren't as many outliers in the `no` group though, so there aren't as many extremely obese observations.

Exercise 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 conditions necessary for inference are:

- Random: we can assume that the survey respondents were randomly selected.
- Normal: given the sample size is 13k+, that is more than 30 and so we can assume it's normal.
- Independent: there has not been resampling and our sample size is not more than 10% of the population

Exercise 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. H0: there is no difference in weight between those who exercise 3+ times a week and those who exercise less than 3 times a week, meaning the averages are the same. H1: there is a difference in weight between those who exercise 3+ times a week and those who do not exercise at all, meaning the averages are different.

Exercise 6

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

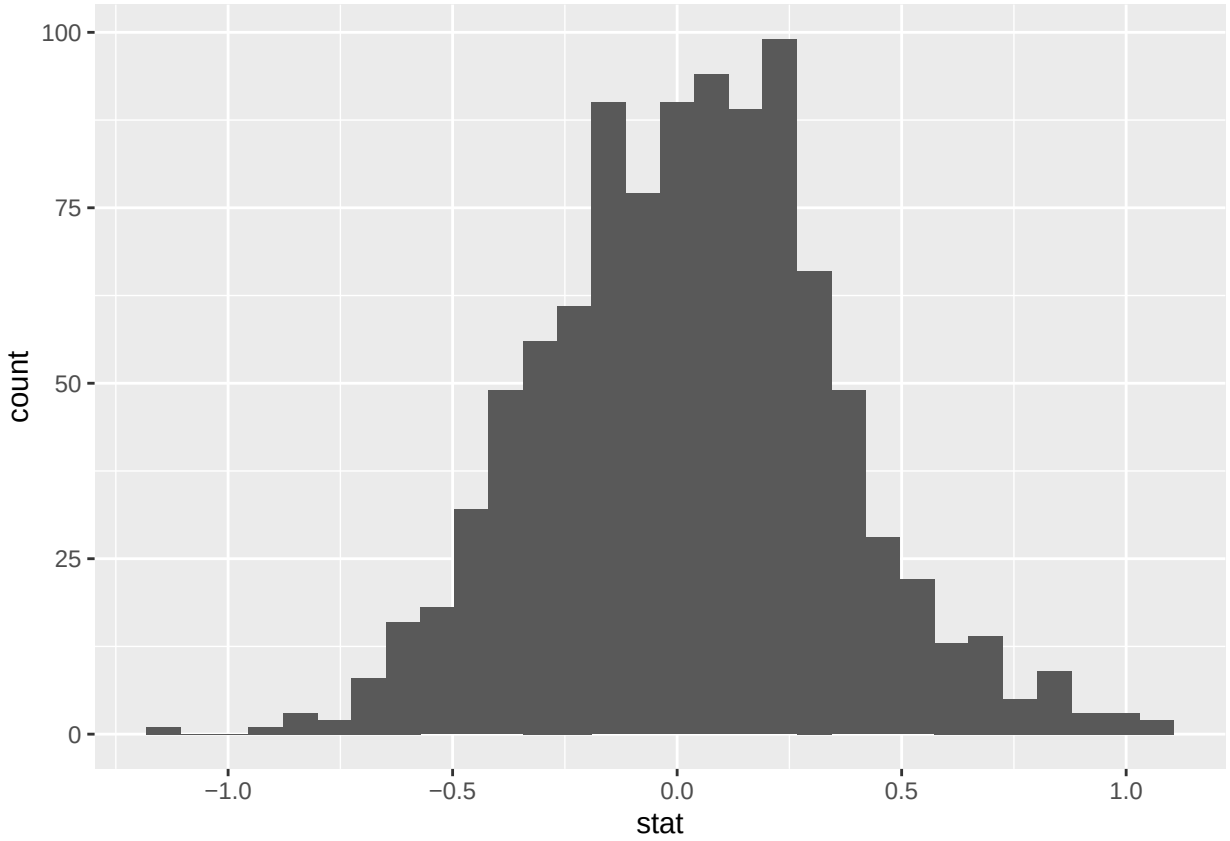
How many of these `null` permutations have a difference of at least `obs_stat`?

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

```
null_dist <- yrbss %>% filter(physical_3plus != "unknown") |>
  specify(weight ~ physical_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()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



So `obs_diff`'s stat is 1.77 – none of these permutations get up to that, it looks like the max in the distribution is a little over 1.0.

Exercise 7

```
ex7 <- yrbss %>%
  group_by(physical_3plus) %>%
  summarise(mean_weight = mean(weight, na.rm = TRUE),
            sd_weight = sd(weight, na.rm = TRUE),
            count = n())
z <- 1.96

not_active <- ex7 |> filter(physical_3plus == "no")

not_active_mean <- not_active |> select(mean_weight)
not_active_sd <- not_active |> select(sd_weight)
not_active_count <- not_active |> select(count)

active <- ex7 |> filter(physical_3plus == "yes")

active_mean <- active |> select(mean_weight)
active_sd <- active |> select(sd_weight)
active_count <- active |> select(count)
```

```r
# confidence interval for not active
upper_not_active <- not_active_mean$mean_weight + z * (not_active_sd$sd_weight / sqrt(not_active_count$
lower_not_active <- not_active_mean$mean_weight - z * (not_active_sd$sd_weight / sqrt(not_active_count$
sprintf("The 95 percent confidence interval for not physically active students' weight is %f to %f", lo
```

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.

```
## [1] "The 95 percent confidence interval for not physically active students' weight is 66.152952 to 6
```

```r
# confidence interval for active
upper_active <- active_mean$mean_weight + z * (active_sd$sd_weight / sqrt(active_count$count))
lower_active <- active_mean$mean_weight - z * (active_sd$sd_weight / sqrt(active_count$count))
sprintf("The 95 percent confidence interval for physically active students' weight is %f to %f", lower_
```

```
## [1] "The 95 percent confidence interval for physically active students' weight is 68.106233 to 68.79
```

Seeing as the two confidence intervals do not overlap, we are 95% confident that there is a difference in weight between these two groups; we reject the null hypothesis.

Exercise 8

```r
ex8 <- yrbss %>%
  summarise(mean_height = mean(height, na.rm = TRUE),
            sd_height = sd(height, na.rm = TRUE),
            count = n())

z <- 1.96

height_mean <- ex8 |> select(mean_height)
height_sd <- ex8 |> select(sd_height)
height_count <- ex8 |> select(count)

# confidence interval for height
upper_height <- height_mean$mean_height + z * (height_sd$sd_height / sqrt(height_count$count))
lower_height <- height_mean$mean_height - z * (height_sd$sd_height / sqrt(height_count$count))
sprintf("The 95 percent confidence interval for the height of students is %f to %f", lower_height, upper
```

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

```
## [1] "The 95 percent confidence interval for the height of students is 1.689480 to 1.693002"
```

We are 95% confident that the average height of a student is between 1.689 and 1.693 meters.

Exercise 9

```r
z <- 1.65

# confidence interval for height
upper_height <- height_mean$mean_height + z * (height_sd$sd_height / sqrt(height_count$count))
lower_height <- height_mean$mean_height - z * (height_sd$sd_height / sqrt(height_count$count))
sprintf("The 90 percent confidence interval for the height of students is %f to %f", lower_height, upper
```

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.

```
## [1] "The 90 percent confidence interval for the height of students is 1.689759 to 1.692723"
```

We are 90% confident that the average height of a student is between 1.690 and 1.693 meters. Since all numbers except the critical value are the same, the confidence interval is smaller in width as the confidence level goes down.

Exercise 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. H0: The average height between students who exercise 3+ times a week and students who don't are the same. H1: The average height between students who exercise 3+ times a week and students who don't are not the same.

```r
ex10 <- yrbss %>%
  group_by(physical_3plus) %>%
  summarise(mean_height = mean(height, na.rm = TRUE),
            sd_height = sd(height, na.rm = TRUE),
            count = n())
z <- 1.96

not_active <- ex10 |> filter(physical_3plus == "no")

not_active_mean <- not_active |> select(mean_height)
not_active_sd <- not_active |> select(sd_height)
not_active_count <- not_active |> select(count)

active <- ex10 |> filter(physical_3plus == "yes")

active_mean <- active |> select(mean_height)
active_sd <- active |> select(sd_height)
active_count <- active |> select(count)

# confidence interval for not active
upper_not_active <- not_active_mean$mean_height + z * (not_active_sd$sd_height / sqrt(not_active_count$
lower_not_active <- not_active_mean$mean_height - z * (not_active_sd$sd_height / sqrt(not_active_count$
sprintf("The 95 percent confidence interval for not physically active students' height is %f to %f", low
```

```
## [1] "The 95 percent confidence interval for not physically active students' height is 1.662549 to 1.
```

```r
# confidence interval for active
upper_active <- active_mean$mean_height + z * (active_sd$sd_height / sqrt(active_count$count))
lower_active <- active_mean$mean_height - z * (active_sd$sd_height / sqrt(active_count$count))
sprintf("The 95 percent confidence interval for physically active students' height is %f to %f", lower_a
```

```
## [1] "The 95 percent confidence interval for physically active students' height is 1.701067 to 1.70535
```

As these do not overlap, we reject the null hypothesis and are 95% confident that students who exercise 3+ times a week have a taller height than students who don't.

Exercise 11

```r
ex11 <- yrbss %>%
  group_by(hours_tv_per_school_day) %>%
  summarise(count = n())
ex11
```

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.

```
## # A tibble: 8 x 2
##   hours_tv_per_school_day count
##   <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 8 options here: <1, 1-4, 5+, do not watch, and then people who didn't respond.

Exercise 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. Are students who get at least 8 hours of sleep taller?

We are assuming that all conditions for inference are satisfied:

- Random: we can assume that the survey respondents were randomly selected.
- Normal: given the sample size is 13k+, that is more than 30 and so we can assume it's normal.
- Independent: there has not been resampling and our sample size is not more than 10% of the population

H0: There is no difference in average height for students depending on whether they get 8+ hours of sleep or not. H1: There is a difference in average height for students depending on whether they get 8+ hours of sleep or not.

```
yrbss <- yrbss %>%
  mutate(sleep_8plus = replace_na(ifelse(yrbss$school_night_hours_sleep >= 8, "yes", ifelse(yrbss$schoo

ex12 <- yrbss %>%
  group_by(sleep_8plus) %>%
  summarise(mean_height = mean(height, na.rm = TRUE),
            sd_height = sd(height, na.rm = TRUE),
            count = n())
z <- 1.96

no_sleep <- ex12 |> filter(sleep_8plus == "no")

no_sleep_mean <- no_sleep |> select(mean_height)
no_sleep_sd <- no_sleep |> select(sd_height)
no_sleep_count <- no_sleep |> select(count)

sleep <- ex12 |> filter(sleep_8plus == "yes")

sleep_mean <- sleep |> select(mean_height)
sleep_sd <- sleep |> select(sd_height)
sleep_count <- sleep |> select(count)

# confidence interval for not active
upper_no_sleep <- no_sleep_mean$mean_height + z * (no_sleep_sd$sd_height / sqrt(no_sleep_count$count))
lower_no_sleep <- no_sleep_mean$mean_height - z * (no_sleep_sd$sd_height / sqrt(no_sleep_count$count))
sprintf("The 95 percent confidence interval for students who don't sleep 8+ hours' height is %f to %f",
```

```
## [1] "The 95 percent confidence interval for students who don't sleep 8+ hours' height is 1.687895 to
```

```
# confidence interval for active
upper_sleep <- sleep_mean$mean_height + z * (sleep_sd$sd_height / sqrt(sleep_count$count))
lower_sleep <- sleep_mean$mean_height - z * (sleep_sd$sd_height / sqrt(sleep_count$count))
sprintf("The 95 percent confidence interval for students who sleep 8+ hours' height is %f to %f", lower_
```

```
## [1] "The 95 percent confidence interval for students who sleep 8+ hours' height is 1.688781 to 1.695
```

Given these confidence intervals do overlap, we cannot reject the null hypothesis at 95% confidence.