Inference for numerical data

Getting Started

```
set.seed(606)
```

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

Answer: The cases in `yrbss` *are high school students. There are 13,583 cases in our sample.*

Remember that you can answer this question by viewing the data in the data viewer or by using the following command:

```r
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.

```r
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?

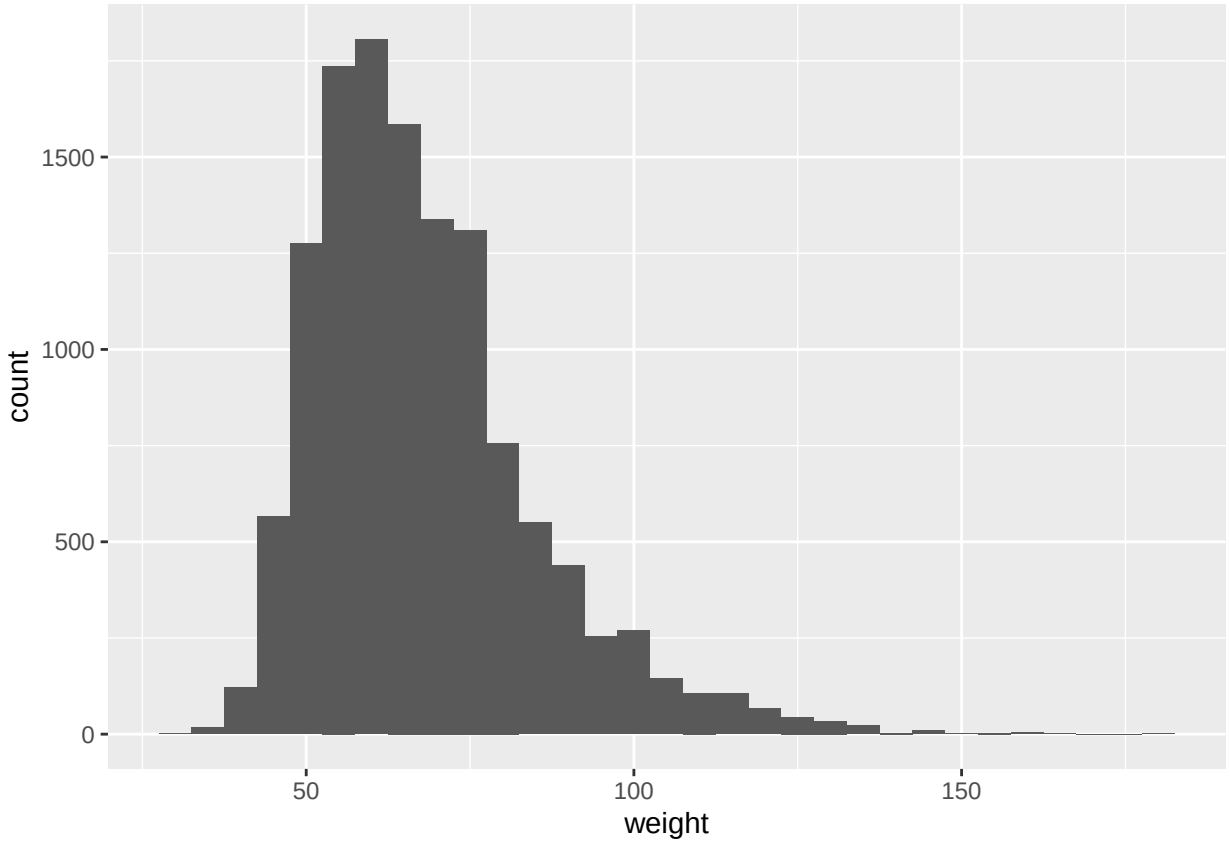
Insert your answer here

Answer: The distribution of ***weights*** *looks fairly normal, but right skewed. The mean is 67.9kg and the median fairly close at 64.4kg. Based on the histogram, the weights appear to have a wide spread. There are 1004 observations missing from the* ***weights*** *variable.*

```r
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
```

```r
yrbss |>
  ggplot(aes(x = weight)) +
  geom_histogram(binwidth = 5)
```



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.

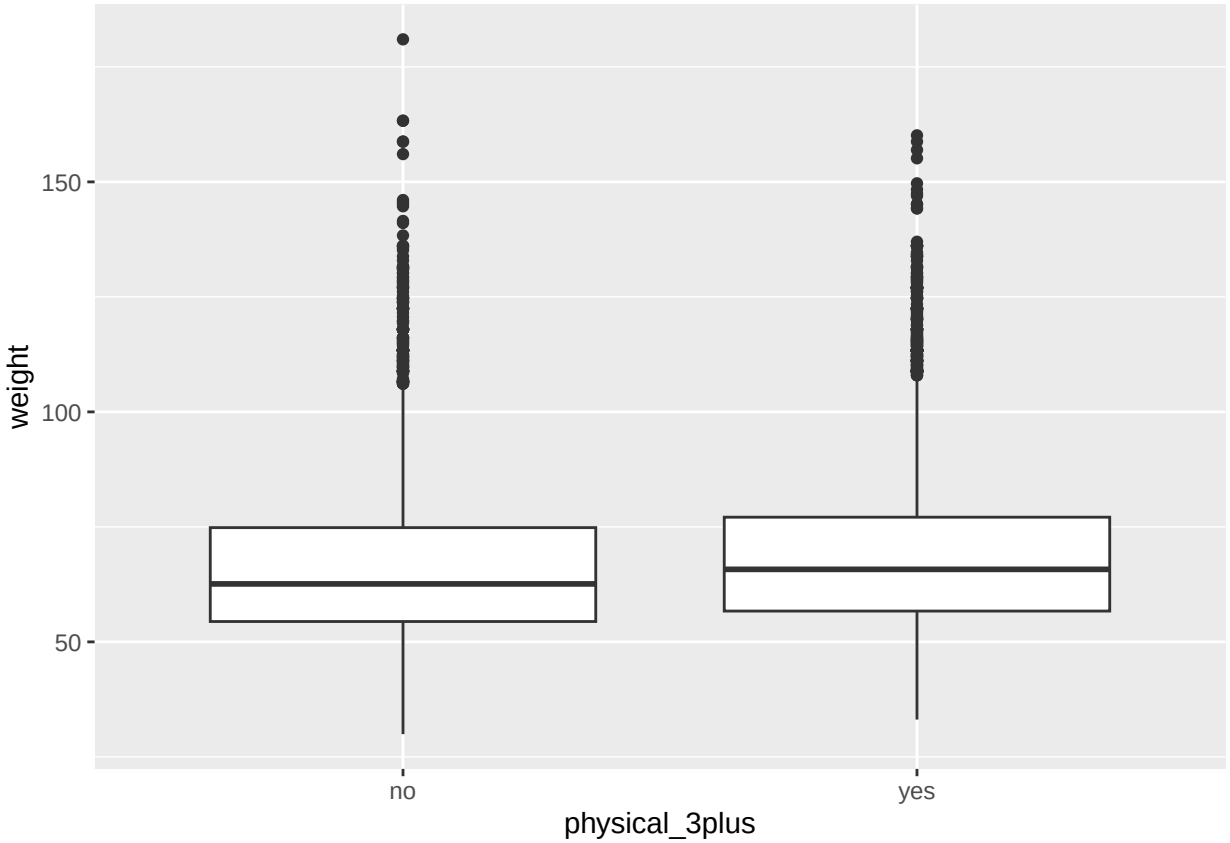
```
yrbss <- yrbss %>%
  mutate(physical_3plus = ifelse(yrbss$physically_active_7d > 2, "yes", "no"))
```

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

Answer: People who exercise at least three times a week seem to weigh slightly more than those who don't, based on the boxplot. I did not have a strong expectation for either scenario, because physical activity can either help people lose fat (lose weight) or help them gain muscle (add weight).

```
yrbss |>
  filter(physical_3plus != is.na(physical_3plus)) |>
  ggplot(aes(x = physical_3plus, y = weight)) +
  geom_boxplot()
```



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()`.

Insert your answer here

Answer: To conduct a hypothesis test on the difference of means, our sample must meet two conditions: independence within and between between groups and normality for each group. Since we are assuming the data is an independent, random sample, independence is met. Normality for each group can be tested by looking for "extreme" outliers when n >= 30, which is the case for both "yes" and "no" groups. I have tested for "extreme" outliers by looking for values 3xIQR above or below the 3rd and 1st quartile. There appears to be at least one extreme outlier for each group for some very heavy observations. Based on the visualization of the distribution, there appears to be large right skew. Therefore, all the conditions necessary for inference are not met.

```
yrbss |>
  group_by(physical_3plus) |>
  summarise(n = n())
```

```
## # A tibble: 3 x 2
##   physical_3plus     n
##   <chr>          <int>
## 1 no              4404
## 2 yes             8906
## 3 <NA>             273
```

```
yrbss |>
  filter(physical_3plus == "yes") |>
  select(weight) |>
  summary()
```

```
##      weight
##  Min.   : 33.11
##  1st Qu.: 56.70
##  Median : 65.77
##  Mean   : 68.45
##  3rd Qu.: 77.11
##  Max.   :160.12
##  NA's   :564
```

```
#Checking for outliers for "yes" group
56.7 - (77.1-56.7)*3
```

```
## [1] -4.5
```

```
77.1 + (77.1-56.7)*3
```

```
## [1] 138.3
```

```
yrbss |>
  filter(physical_3plus == "no") |>
  select(weight) |>
  summary()
```

```
##      weight
##  Min.   : 29.94
##  1st Qu.: 54.43
```

```
##  Median : 62.60
##  Mean   : 66.67
##  3rd Qu.: 74.84
##  Max.   :180.99
##  NA's   :382
```

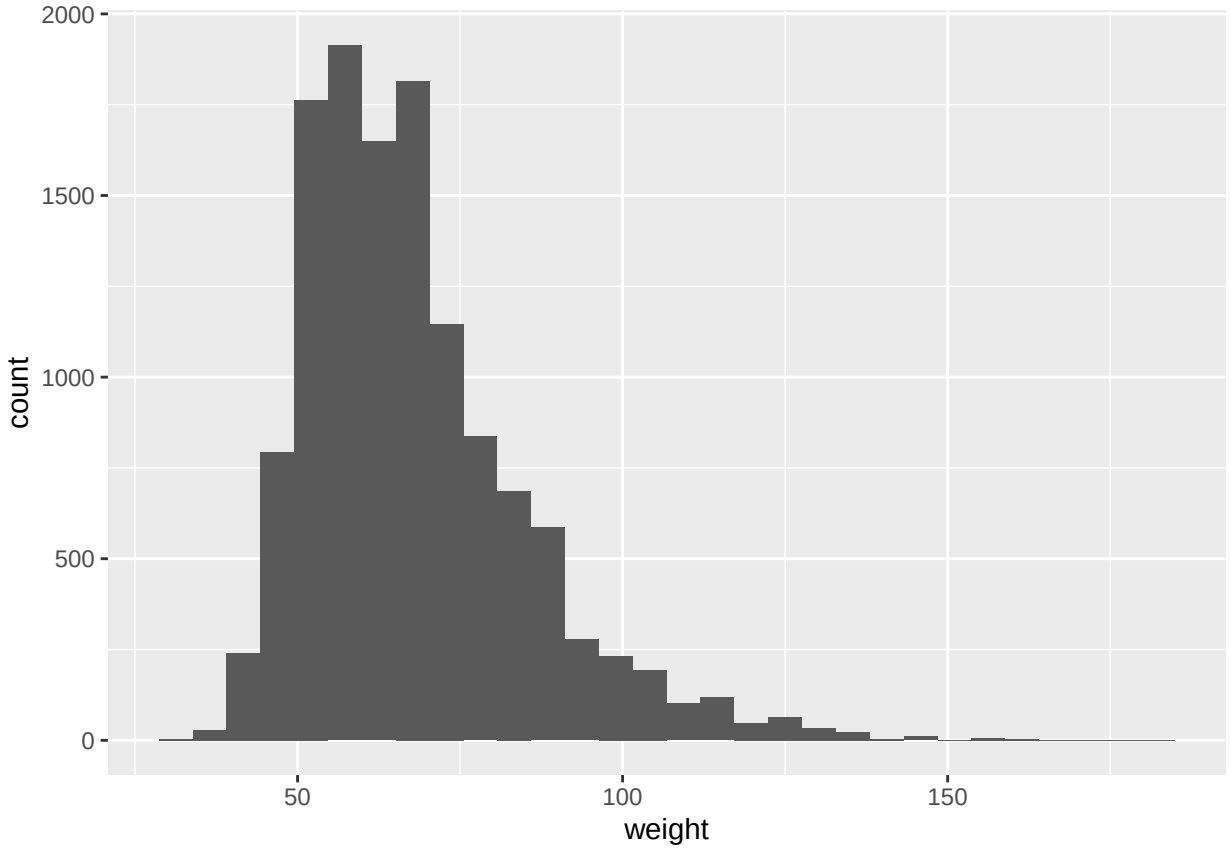
```
#Checking for outliers for "no" group
54.4 - (74.8-54.4)*3
```

```
## [1] -6.8
```

```
74.8 + (74.8-54.4)*3
```

```
## [1] 136
```

```
#Further investigation
ggplot(data = yrbss, aes(x = weight)) +
  geom_histogram()
```



5. Write the hypotheses for testing if the average weights are different for those who exercise at least three times a week and those who don't.

Insert your answer here

Answer: Null Hypothesis: the average weight of people who exercise at least three times a week versus people who don't is the same. Alternative Hypothesis: the difference between the average weight of people who exercise at least three times a week versus people who don't is not zero.

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 %>%
  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`.

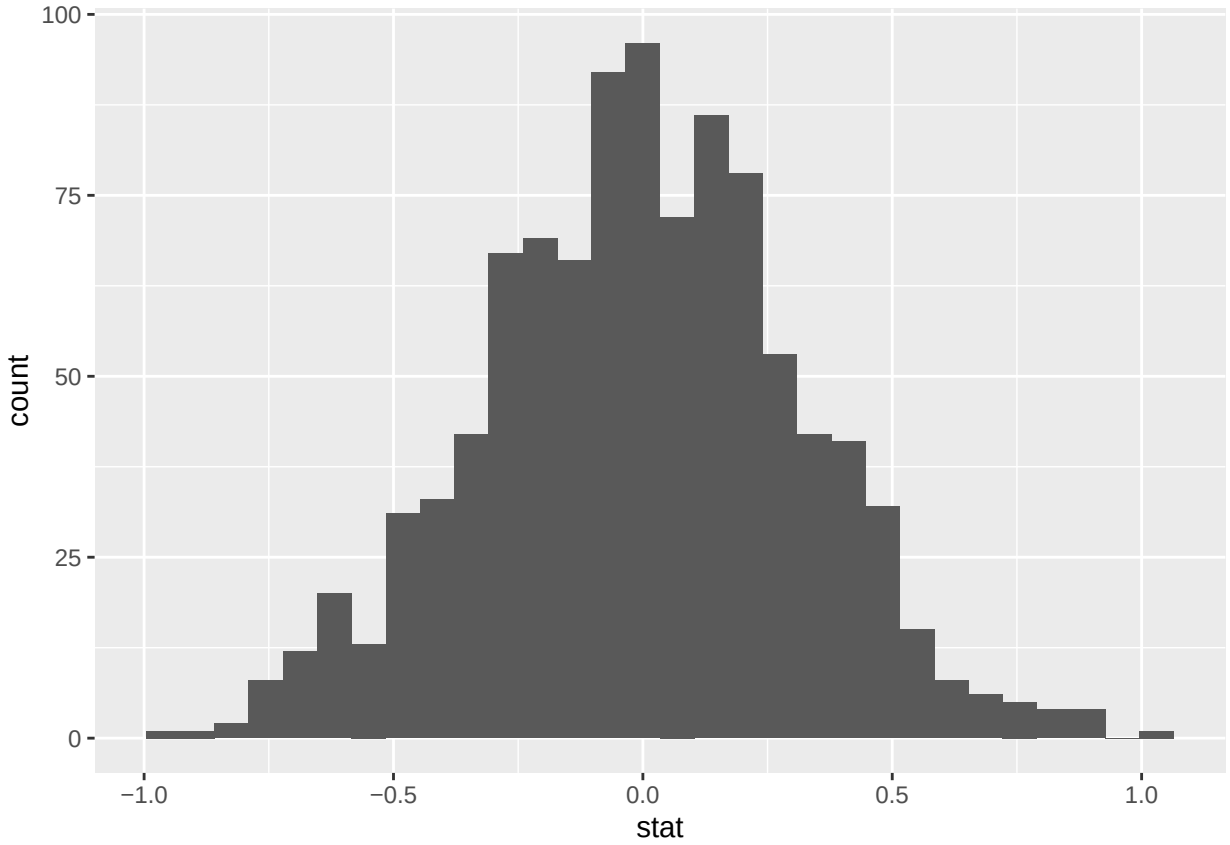
```
null_dist <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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

Answer: `obs_stat` is near 0. All (1,000) of the `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`.

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

```
obs_stat
```

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

```
null_dist |>
  filter(stat >= obs_stat$p_value | stat <= obs_stat$p_value)
```

```
## Response: weight (numeric)
## Explanatory: physical_3plus (factor)
## Null Hypothesis: independence
```

```
## # A tibble: 1,000 x 2
##    replicate    stat
##        <int>   <dbl>
##  1         1 -0.180
##  2         2 -0.341
##  3         3  0.507
##  4         4 -0.0283
##  5         5 -0.221
##  6         6  0.196
##  7         7 -0.208
##  8         8 -0.442
##  9         9  0.0674
## 10        10  0.446
## # i 990 more rows
```

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.

Answer: We can have 95% confidence that the high school population difference between weights of those exercise at least three times a week and those who don't is between (1.17, 2.41).

```
yrbss |>
  drop_na(physical_3plus) |>
  specify(weight ~ physical_3plus) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "diff in means", order = c("yes", "no")) |>
  get_ci(level = .95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     1.17     2.41
```

More Practice

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

Insert your answer here

Answer: To construct a confidence interval for one sample mean, we can start by constructing a sampling t-distribution, which requires the sample to be independent and exhibit normality. Since ***yrbss*** *is a simple, random sample, independence is met. When n >= 30 (it is), we can check for any "extreme" outliers to test for normality. Based on the calculations below, there don't appear to be any "extreme" outliers within* ***height****. The 95% confidence interval is (1.689, 1.693). Therefore, we can have 95% confidence that the high school population mean height is within that interval.*

```
yrbss |>
  count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 13583
```

```
summary(yrbss$height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.270   1.600   1.680   1.691   1.780   2.110    1004
```

```
1.60 - (1.78 - 1.60)*3
```

```
## [1] 1.06
```

```
1.78 + (1.78 - 1.60)*3
```

```
## [1] 2.32
```

```
#Generate CI for height
yrbss |>
  mutate(height = height*1000) |>
  drop_na(height) |>
  specify(response = height) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "mean") |>
  get_ci(level = .95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    1689.    1693.
```

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

Answer: The width of the 90% CI is only .001m narrower than the 95% interval. I expected the interval to be narrower, because the lower the confidence level, the more narrow the interval. Since the difference seems small, this may suggest that the variance within ***height*** *is small.*

```
yrbss |>
  mutate(height = height*1000) |>
  drop_na(height) |>
  specify(response = height) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "mean") |>
  get_ci(level = .9)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    1690.    1693.
```

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

Answer: To conduct a hypothesis test for difference of means, we must first confirm independence and normality. Since ***yrbss*** *is a simple, random sample, independence is met. For normality we must check for "extreme" outliers. There does appear to be at least one potential "extreme" outlier in the "no" group. Upon further investigation with a histogram, the spread of the distribution looks fairly even, so normality is met. Our null hypothesis states that there is no difference in average height between high schoolers who exercise at least three times a week and those who don't. Our alternative hypothesis is that there is a non-zero difference between these means. Next, we calculate our difference of means statistic and generate a null distribution. We then find the p-value for the likelihood that the* ***obs_stat*** *falls within the null distribution. This p-value is near zero which means we can reject the null hypothesis.*

```
#Check normality for "yes" group
yrbss |>
  filter(physical_3plus == "yes") |>
  select(height) |>
  summary()
```

```
##      height
##  Min.   :1.270
##  1st Qu.:1.630
##  Median :1.700
##  Mean   :1.703
##  3rd Qu.:1.780
##  Max.   :2.110
##  NA's   :564
```

```
1.63 - (1.78-1.63)*3
```

```
## [1] 1.18
```

```
1.78 + (1.78-1.63)*3
```

```
## [1] 2.23
```

```
#Check normality for "no" group
yrbss |>
  filter(physical_3plus == "no") |>
  select(height) |>
  summary()
```

```
##      height
##  Min.   :1.270
```

```
##  1st Qu.:1.600
##  Median :1.650
##  Mean   :1.666
##  3rd Qu.:1.730
##  Max.   :2.110
##  NA's   :382
```

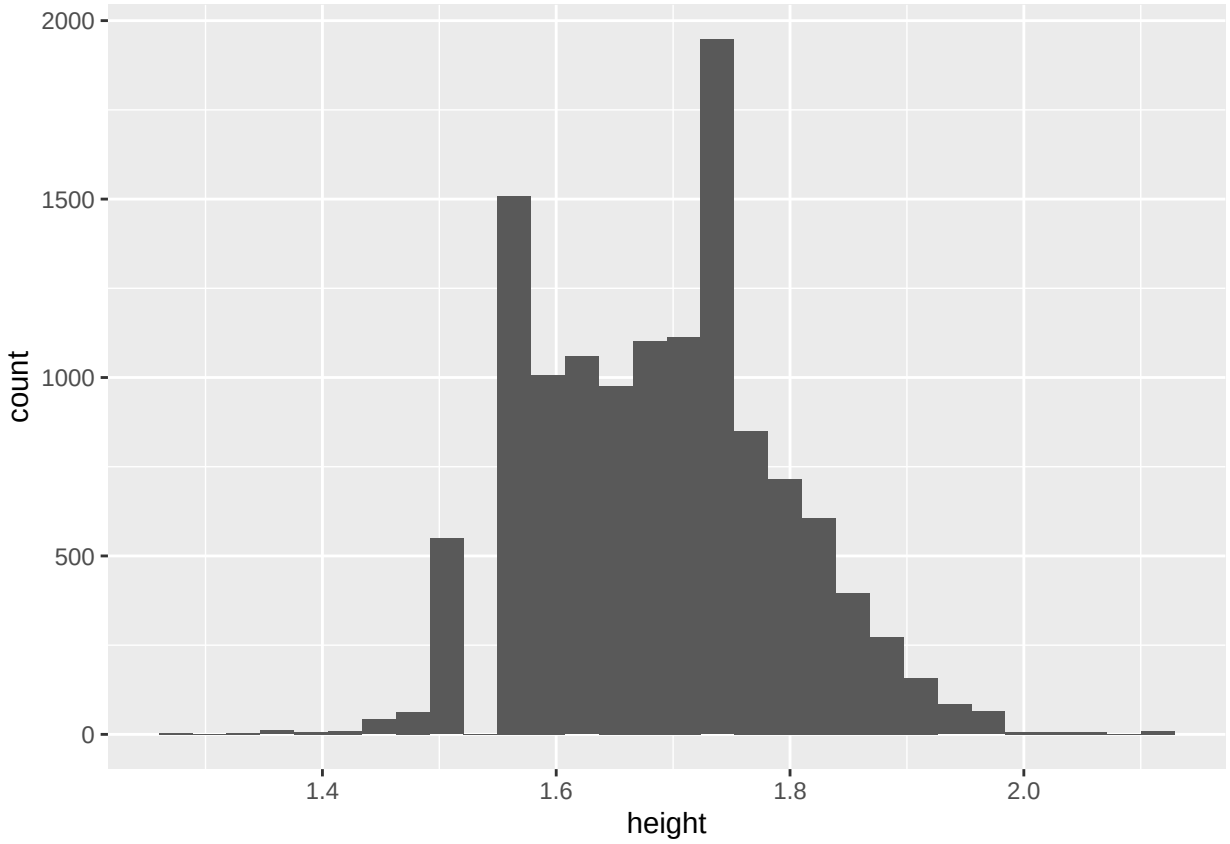
```
1.60 - (1.73-1.60)*3
```

```
## [1] 1.21
```

```
1.60 + (1.73-1.60)*3
```

```
## [1] 1.99
```

```
#Further investigate
ggplot(data = yrbss, aes(x = height)) +
  geom_histogram()
```




```
#Calculate obs_stat
height_obs_diff <- yrbss |>
                drop_na(physical_3plus) |>
                specify(height ~ physical_3plus) |>
                calculate(stat = "diff in means", order = c("yes", "no"))
```

```r
#Generate null distribution
null_height_dist <- 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"))

#Get p-value
null_height_dist |>
  get_p_value(obs_stat = height_obs_diff, direction = "two_sided")
```

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

Insert your answer here

Answer: There are seven options for ***hours_tv_per_school_day****.*

```r
yrbss |>
  drop_na(hours_tv_per_school_day) |>
  distinct(hours_tv_per_school_day) |>
  count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     7
```

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

Answer: I would like to test whether there is a significant difference in height between high schoolers who sleep the recommended eight hours a night versus those who don't, using an alpha level of 0.05. The null hypothesis is that those who sleep at least eight hours a night and those who don't have the same average height. The alternative hypothesis is that there is a non-zero difference in average height between these two groups. Independence has already been established in the above problems, but now we must assess normality of the height variable within each sleep group. Based on the calculations below, there does not appear to be any "extreme" outliers, so normality is met. After generating our difference in means statistic and null distribution, we calculate the likelihood of our statistic falls within the null distribution. The resulting p-value of 0.34 is much greater than our alpha level of 0.05, so we fail to reject the null hypothesis.

```
yrbss <- yrbss |>
  mutate(school_sleep2 = str_remove(school_night_hours_sleep, "\\+"),
         school_sleep2 = str_remove(school_sleep2, "<"),
         school_sleep2 = parse_number(school_sleep2),
         recommended_sleep = ifelse(school_sleep2 >=8, "yes", "no"))


yrbss |>
  filter(recommended_sleep == "yes") |>
  select(height) |>
  summary()
```

```
##      height
##  Min.   :1.270
##  1st Qu.:1.600
##  Median :1.700
##  Mean   :1.692
##  3rd Qu.:1.780
##  Max.   :2.110
##  NA's   :306
```

```
1.60 - (1.78-1.60)*3
```

```
## [1] 1.06
```

```
1.78 + (1.78-1.60)*3
```

```
## [1] 2.32
```

```
yrbss |>
  filter(recommended_sleep == "no") |>
  select(height) |>
  summary()
```

```
##      height
##  Min.   :1.27
##  1st Qu.:1.60
##  Median :1.68
##  Mean   :1.69
##  3rd Qu.:1.75
##  Max.   :2.11
##  NA's   :548
```

```
1.60 - (1.75-1.60)*3
```

```
## [1] 1.15
```

```
1.75 + (1.75-1.60)*3
```

```
## [1] 2.2
```

```r
height_sleep_stat <- yrbss |>
                      drop_na(recommended_sleep) |>
                      specify(height ~ recommended_sleep) |>
                      calculate(stat = "diff in means", order = c("yes", "no"))

height_sleep_null_dist <- yrbss |>
  drop_na(recommended_sleep) |>
  specify(height ~ recommended_sleep) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "diff in means", order = c("yes", "no"))

height_sleep_null_dist |>
  get_p_value(obs_stat = height_sleep_stat, direction = "two_sided")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1    0.34
```
