Inference for numerical data

Wilson Hernandez

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
head(yrbss)
```

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

This dataset has 13,583 observations. It spans children of high school age self-reporting several indicators of their health. These include their physical activity (strength-training frequency, physical activity frequency), media consumption (amount of time spent watching tv) and risk-taking activity (text messaging while driving, helmet usage) and sleep pattern.

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?

```
# Count NA values in the weight measurements column
na_count <- sum(is.na(yrbss$weight))

# Print the number of NA values
print(na_count)
```

```
## [1] 1004
```

This code above calculates the sum of logical values returned by `is.na()` function, where `TRUE` represents an NA value. The result gives you the count of NA values in the weight measurements column of your dataset. In this case, the value that is 1004 samples are missing a weight measurement (labeled "NA").

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.

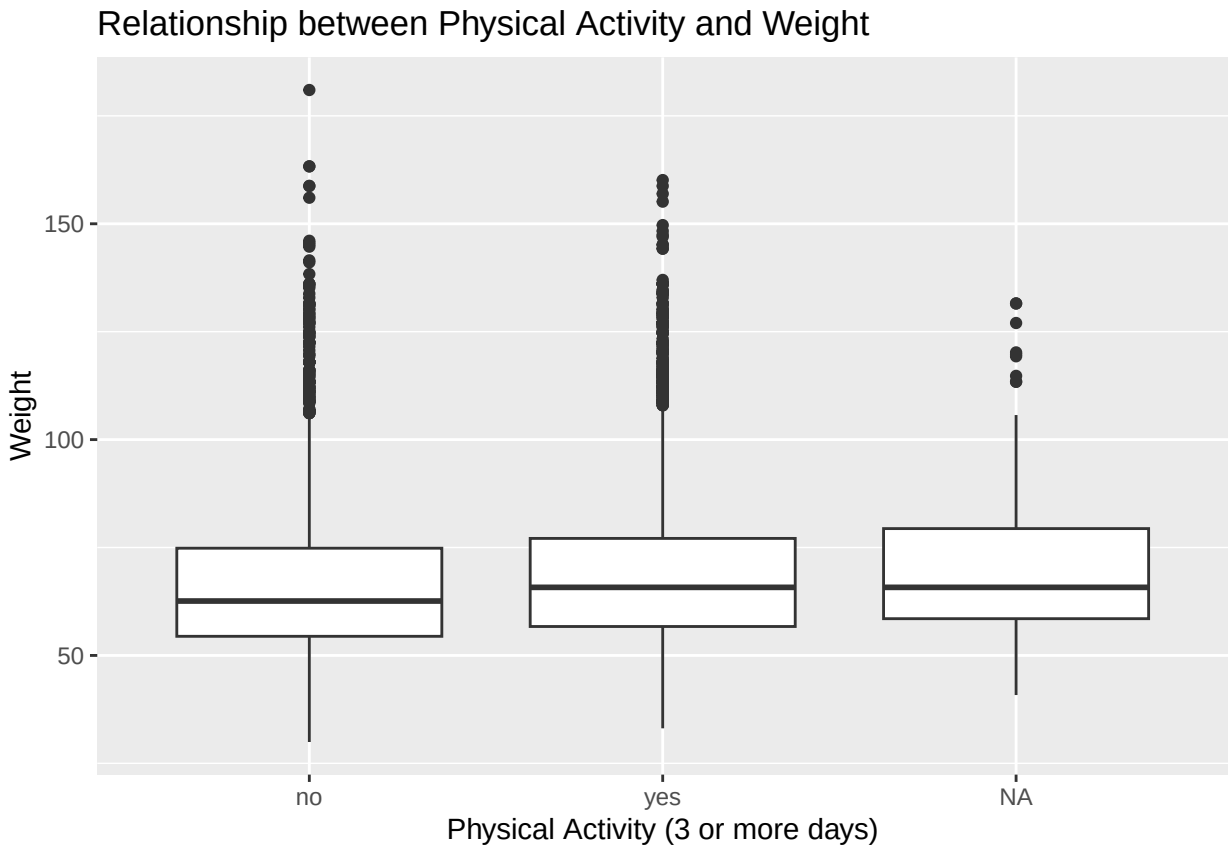
First, let's create a new variable `physical_3plus`, which will be coded as either "yes" if they are physically active for at least 3 days a week, and "no" if not.

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

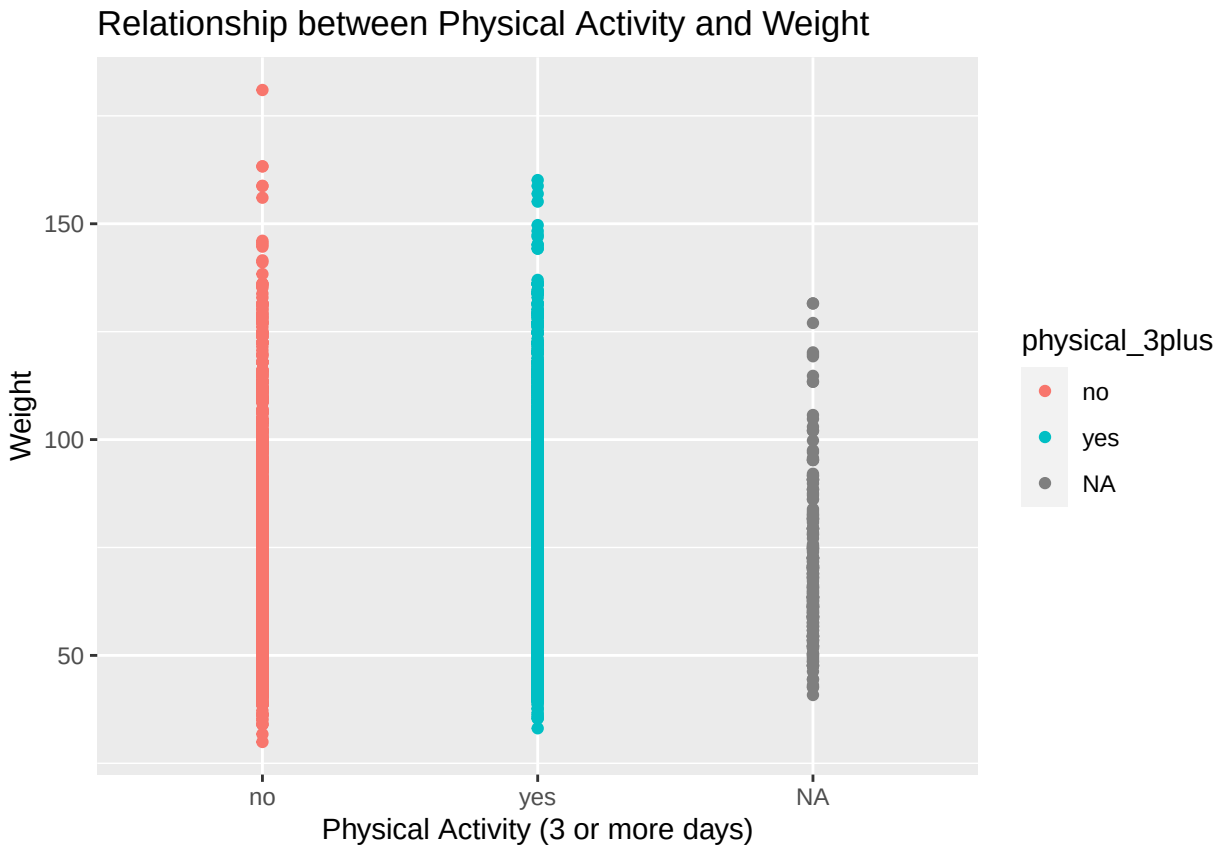
3. **Make a side-by-side boxplot of `physical_3plus` and `weight`. Is there a relationship between these two variables? What did you expect and why?**

I was a bit to surprised initially to find that the median weight for children that exercise is higher, but upon further reflection it makes sense given that children that exercise are more likely to develop muscle, which is heavy. The extreme outliers in weight in both directions seem to be in the category that do not exercise regularly, which does make sense.

```
# Boxplot to compare weight distribution for "yes" and "no" categories
ggplot(yrbss, aes(x = physical_3plus, y = weight)) +
  geom_boxplot() +
  xlab("Physical Activity (3 or more days)") +
  ylab("Weight") +
  ggtitle("Relationship between Physical Activity and Weight")
```




```
# Scatter plot to compare weight for different levels of physical activity
ggplot(yrbss, aes(x = physical_3plus, y = weight, color = physical_3plus)) +
  geom_point() +
  xlab("Physical Activity (3 or more days)") +
  ylab("Weight") +
  ggtitle("Relationship between Physical Activity and Weight")
```



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()`.**

Based on the context that I have available to me, it seems that we can proceed with our inference, although we could only begin to find correlation and not causation.

1. Independence: Because YRBSS is a survey, it's assumed that the responses are independent, meaning

one person's response doesn't influence another's.

2. Sample Size/Skew: We need enough data to achieve reliable results, a high enough sample size should ensure the central limit theorem applies and the sampling distribution should be approximately normal. Each subgroup (those physically active for more than 3 days and those not) has above 4,000 measurements, so although twice as many youth self-report having a minimum of 3 days of physical activity in a week compared to those that report that they don't, I would say that we have a sufficiently large dataset to proceed with our analysis.

```
group_sizes <- yrbss %>%
  group_by(physical_3plus) %>%
  summarise(group_size = n())

# Print group sizes
print(group_sizes)
```

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

3&4) Information regarding the methodology of this survey is available here: YRBSS Frequently Asked Questions [Source: CDC] This information seems to indicate that the requirements for Population Distribution and Random Sampling are comfortably met by this survey.

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.**

Null Hypothesis (H0): The average weight of students who exercise at least three times a week is equal to the average weight of students who don't. Mathematically represented as: 'mu'1 = 'mu'2.

Alternative Hypothesis (Ha): The average weight of students who exercise at least three times a week is not equal to the average weight of students who don't. Mathematically represented as: 'mu'1 \\neq 'mu'2.

Here, 'mu'1 represents the average weight of students who exercise at least three times a week, 'mu'2 represents the average weight of students who don't exercise as much.

This sets up a two-sided test as we are investigating whether the means are different, but not specifying in what direction (i.e., greater or less).

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) %>%
```

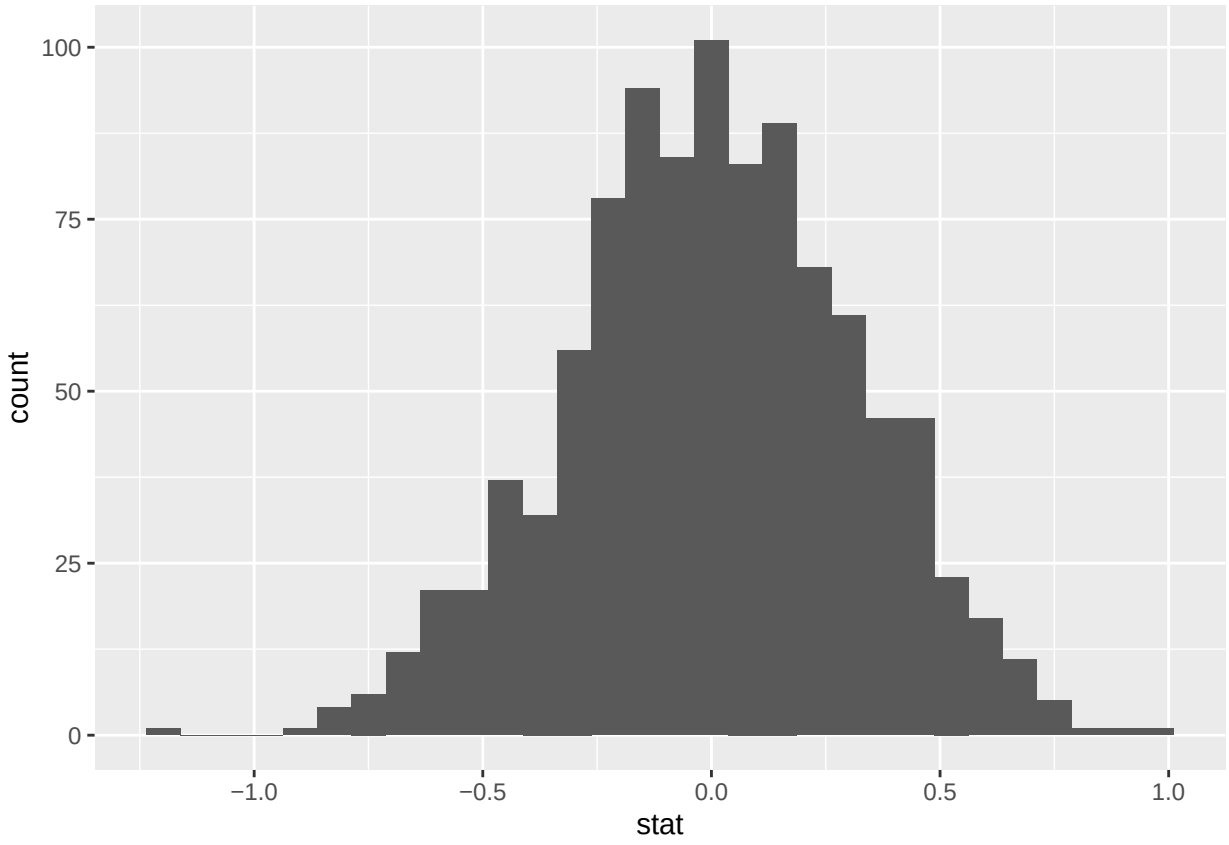
```
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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`?**

You can calculate the number of null permutations with a difference of at least `obs_stat` (the observed statistic or difference) using `get_pvalue()` function from the infer package. This function calculates the p-value for your observed statistic given your null distribution.

```
# Calculate p-value
p_value <- null_dist %>%
  get_pvalue(obs_stat = obs_diff, direction = "both")

# Print the p-value
print(p_value)
```

```
## # A tibble: 1 x 1
```

```
##   p_value
##     <dbl>
## 1       0
```

The `direction = "both"` argument is used because we are performing a two-tailed test (we're looking for a difference in either direction). The p-value is the proportion of the null distribution that is as or more extreme than our observed statistic.

A small p-value (typically less than 0.05) indicates strong evidence against the null hypothesis, so we reject the null hypothesis that the mean weights are the same between the two physical activity groups.

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.

```
#pre-processing by dropping the missing values

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


# Calculate the means and standard deviations for the two groups
group_stats <- yrbss_valid %>%
                group_by(physical_3plus) %>%
                summarise(mean_weight = mean(weight),
                          sd_weight = sd(weight),
                          n = n())

# Calculate the difference in mean weights
diff_mean <- diff(group_stats$mean_weight)

# Calculate the standard error of the difference in mean weights
se_diff <- sqrt(sum((group_stats$sd_weight)^2 / group_stats$n))

# Construct a 95% confidence interval
ci_lower <- diff_mean - 1.96* se_diff
ci_upper <- diff_mean + 1.96 * se_diff

# Print the confidence interval
cat("The 95% confidence interval for the difference in weights between the two groups is [", ci_lower,
```

```
## The 95% confidence interval for the difference in weights between the two groups is [ 1.124821 ,  2.4
```

In the context of the data, this means that we are 95% confident that individuals who exercise three times a week weigh on average between 1.124821 and 2.424348 kilograms more than individuals who do not work out as frequently.

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

```r
# Retain only the complete cases in 'height' column
valid_heights <- yrbss[complete.cases(yrbss$height), ]

# Now we can calculate the 95% confidence interval for the average height
height_t_test <- t.test(valid_heights$height)

# The mean height
mean_height <- mean(valid_heights$height)

# The 95% confidence interval
conf_interval <- height_t_test$conf.int

# Print the mean and the confidence interval
print(paste("The mean height is ", round(mean_height, 2),
            " meters. The 95% confidence interval is [",
            round(conf_interval[1], 2), ", ", round(conf_interval[2], 2),
            "] meters.", sep = ""))
```

```
## [1] "The mean height is 1.69 meters. The 95% confidence interval is [1.69, 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.

```r
# Calculate 90% Confidence Interval
height_t_test_90 <- t.test(valid_heights$height, conf.level = 0.90)

# The 90% confidence interval
conf_interval_90 <- height_t_test_90$conf.int

# Print the 90% confidence interval
print(paste("The 90% confidence interval for the average height is [",
            round(conf_interval_90[1], 2), ", ", round(conf_interval_90[2], 2),
            "] meters.", sep = ""))
```

```
## [1] "The 90% confidence interval for the average height is [1.69, 1.69] meters."
```

```r
# Calculate the widths of the confidence intervals
width_95 <- conf_interval[2] - conf_interval[1]
width_90 <- conf_interval_90[2] - conf_interval_90[1]

# Print the comparison
print(paste("The width of the 95% confidence interval is ", round(width_95, 2),
            " meters and the width of the 90% confidence interval is ",
            round(width_90, 2), " meters.", sep = ""))
```

```
## [1] "The width of the 95% confidence interval is 0 meters and the width of the 90% confidence interva
```

Overall, we would expect the 90% confidence interval to be narrower than the 95% confidence interval. This is because a 90% confidence level implies that we're willing to accept a larger probability (10% vs. 5%) that the true population parameter falls outside our calculated confidence interval. Therefore, we don't need the interval to be as wide as we would for a higher confidence level.

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.**

The mean height for group1 (those who exercise at least three times a week) is 1.703213 and for group2 (those who don't exercise three times a week) is 1.665587. This means that, on average, those who exercise at least three times a week tend to be taller.

The 95% confidence interval for the difference in means ranges from 0.03374994 to 0.04150183. Since this interval does not include 0, we can conclude that there's a significant difference in the average height between the two groups.

```
# Separate the data into two groups
group1 <- yrbss$height[yrbss$physical_3plus == "yes"]
group2 <- yrbss$height[yrbss$physical_3plus == "no"]

# Conduct a two-sample t-test
test_result <- t.test(group1, group2)

# Print the results
print(test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  group1 and group2
## t = 19.029, df = 7973.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03374994 0.04150183
## sample estimates:
## mean of x mean of y
##  1.703213  1.665587
```

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.**

```
# Identify unique values in the hours_tv_per_school_day column
unique_values <- unique(yrbss$hours_tv_per_school_day)

# Print the unique values
print(unique_values)
```

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

The options are as follows: "do not watch", '1', '2', '3', '4', '5+', 'NA'.

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.**

"Is there a significant correlation between a student's height and the amount of sleep they get?" In this case, we are examining whether there is a relationship between these two variables: 'school_night_hours_sleep' and 'height'

I always grew up believing that you would grow if you slept more, but I never actually bother to look this up.

From the research question, we can formulate the following two hypotheses:

The null hypothesis H0: There is no correlation between a student's weight and their sleep duration. The alternative hypothesis Ha: There is a significant correlation between a student's weight and their sleep

duration.

From the research question, we can formulate the following two hypotheses:

The null hypothesis H0: There is no correlation between a student's weight and their sleep duration. The alternative hypothesis Ha: There is a significant correlation between a student's weight and their sleep duration. I will go with the standard α of .05 - this means that we are willing to accept a 5% chance of rejecting the null hypothesis even if it is true.

```
#preprocessing data- looking at unique values first

unique(yrbss$school_night_hours_sleep)
```

```
## [1] "8"   "6"   "<5"  "9"   "10+" "7"   "5"   NA
```

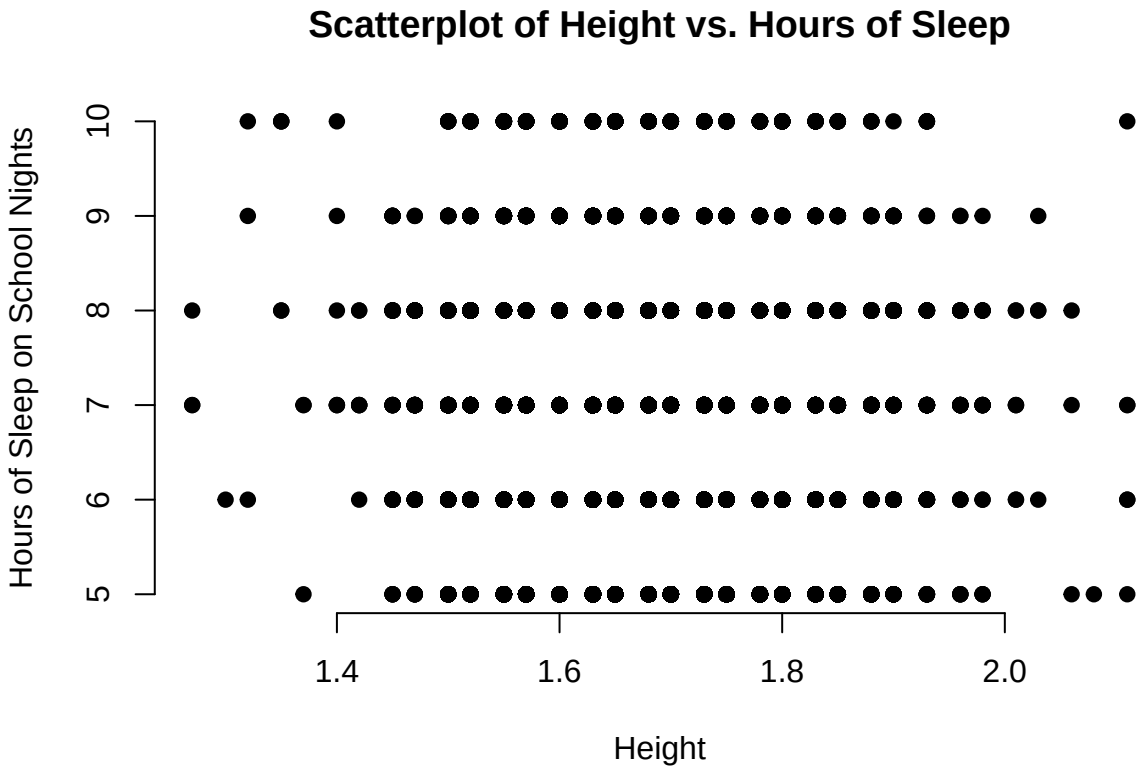
```
unique(yrbss$height)
```

```
##  [1]   NA 1.73 1.60 1.50 1.57 1.65 1.88 1.75 1.37 1.68 1.63 1.85 1.78 1.83 1.55
## [16] 1.52 1.42 1.45 1.70 1.80 1.90 1.93 1.35 1.98 1.96 1.47 1.40 2.11 2.01 2.03
## [31] 2.06 1.27 1.32 1.30 2.08
```

```
#creating a subset of data

data_subset <- yrbss[, c('school_night_hours_sleep', 'height')]

#using complete case to drop the nas

complete_data <- data_subset[complete.cases(data_subset), ]

#the unique values for sleep include the value '10+' which is not numeric and can't be accounted for in

complete_data <- complete_data %>%
  mutate(school_night_hours_sleep = ifelse(school_night_hours_sleep == "10+", 10, as.numeric(school_nigh

#generating summary statistics

summary(complete_data)
```

```
##  school_night_hours_sleep     height
##  Min.   : 5.000           Min.   :1.270
##  1st Qu.: 6.000           1st Qu.:1.600
##  Median : 7.000           Median :1.680
##  Mean   : 6.946           Mean   :1.691
##  3rd Qu.: 8.000           3rd Qu.:1.780
##  Max.   :10.000           Max.   :2.110
##  NA's   :859
```

```
#generating scatter plot

plot(complete_data$height, complete_data$school_night_hours_sleep,
     main="Scatterplot of Height vs. Hours of Sleep",
     xlab="Height",
     ylab="Hours of Sleep on School Nights",
     pch=19, frame=FALSE)
```




```
#using corr test to determine whether or not there is a relationship

correlation <- cor.test(complete_data$height, complete_data$school_night_hours_sleep)

print(correlation)
```

```
##
##  Pearson's product-moment correlation
##
## data:  complete_data$height and complete_data$school_night_hours_sleep
## t = 1.3653, df = 10620, p-value = 0.1722
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.005771911  0.032256460
## sample estimates:
##       cor
## 0.01324706
```

There appears to be no significant correlation between height and the amount of sleep in school nights among the students in our sample data. This means that a student's height doesn't predict how much they sleep, and vice versa, according to our data and chosen significance level.