Inference for numerical data

Richie Rivera

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)
library(psych)
library(ggplot2)
set.seed(1994)
```

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

```
str(yrbss)
```

```
## tibble [13,583 x 13] (S3: tbl_df/tbl/data.frame)
##  $ age                   : int [1:13583] 14 14 15 15 15 15 15 14 15 15 ...
##  $ gender                : chr [1:13583] "female" "female" "female" "female" ...
```

```
##  $ grade                   : chr [1:13583] "9" "9" "9" "9" ...
##  $ hispanic                : chr [1:13583] "not" "not" "hispanic" "not" ...
##  $ race                    : chr [1:13583] "Black or African American" "Black or African American" "
##  $ height                  : num [1:13583] NA NA 1.73 1.6 1.5 1.57 1.65 1.88 1.75 1.37 ...
##  $ weight                  : num [1:13583] NA NA 84.4 55.8 46.7 ...
##  $ helmet_12m              : chr [1:13583] "never" "never" "never" "never" ...
##  $ text_while_driving_30d  : chr [1:13583] "0" NA "30" "0" ...
##  $ physically_active_7d    : int [1:13583] 4 2 7 0 2 1 4 4 5 0 ...
##  $ hours_tv_per_school_day : chr [1:13583] "5+" "5+" "5+" "2" ...
##  $ strength_training_7d    : int [1:13583] 0 0 0 0 1 0 2 0 3 0 ...
##  $ school_night_hours_sleep: chr [1:13583] "8" "6" "<5" "6" ...
```

```
case_count <- nrow(yrbss)
```

Interpreting a case to be an observation, each case is a highschool student along with some properties of that highschool student.

There are 13,583 cases in our sample.

End of your answer

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
```

```r
weight_na_count <- nrow(
  yrbss |>
    filter(is.na(weight))
)
```

2. How many observations are we missing weights from?

Insert your answer here

According to the line above, there are 1,004 observations missing weights.

End of your answer

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.

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

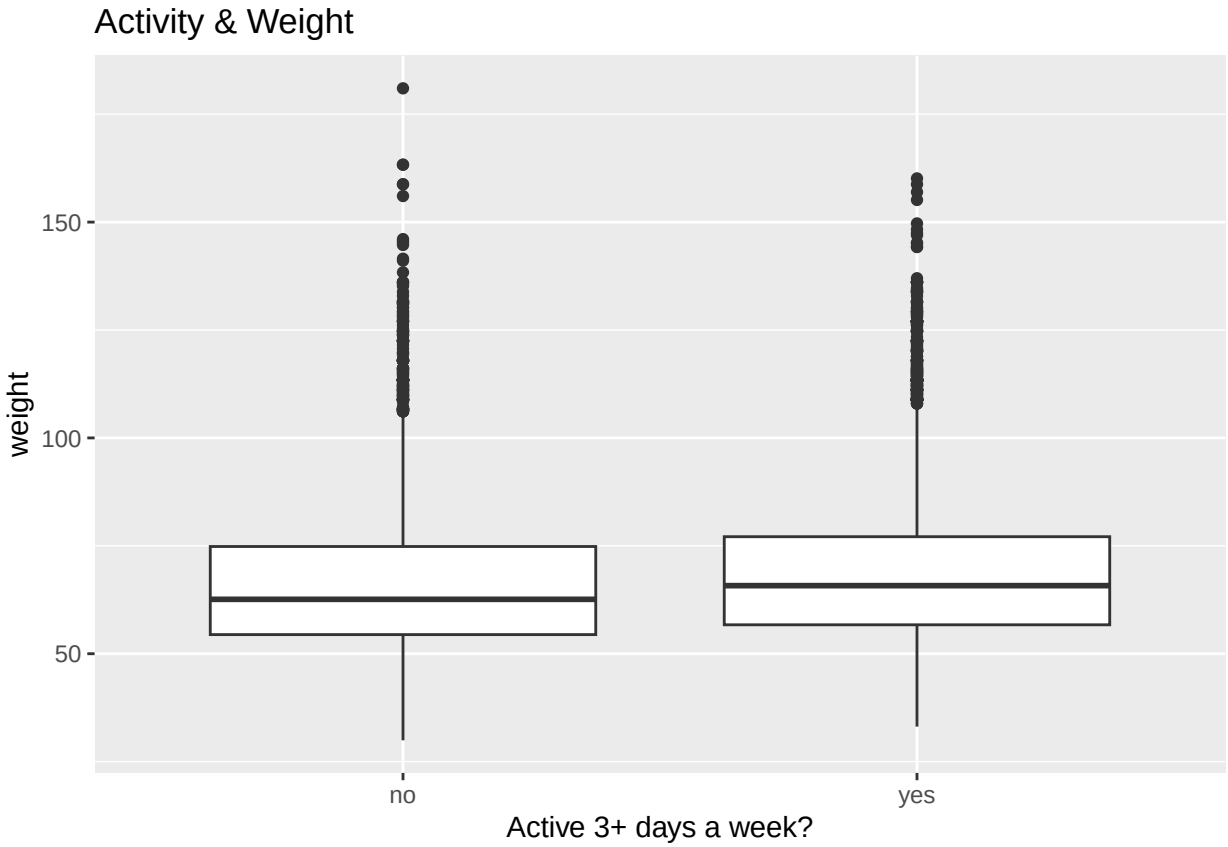
3. Make a side-by-side boxplot of `physical_3plus` and `weight`. Is there a relationship between these two variables? What did you expect and why?

Insert your answer here

```r
weight_data <- yrbss |>
  select(weight, physical_3plus) |>
  drop_na()

weight_box_plot <- ggplot(weight_data, aes(x = physical_3plus, y = weight)) +
  geom_boxplot() +
  labs(
    title = "Activity & Weight",
    x = "Active 3+ days a week?",
    y = "weight"
  )

weight_box_plot
```




```
weight_data |>
  group_by(physical_3plus) |>
  summarise(median_weight = median(weight))
```

```
## # A tibble: 2 x 2
##   physical_3plus median_weight
##   <chr>                  <dbl>
## 1 no                      62.6
## 2 yes                     65.8
```

According to the box plot, there doesn't seem to be a relationship between having 3+ days of activity and weight. Although I initially expected there would be a positive relationship, this analysis ignores height and other potential confounding factors.

End of your answer

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
```

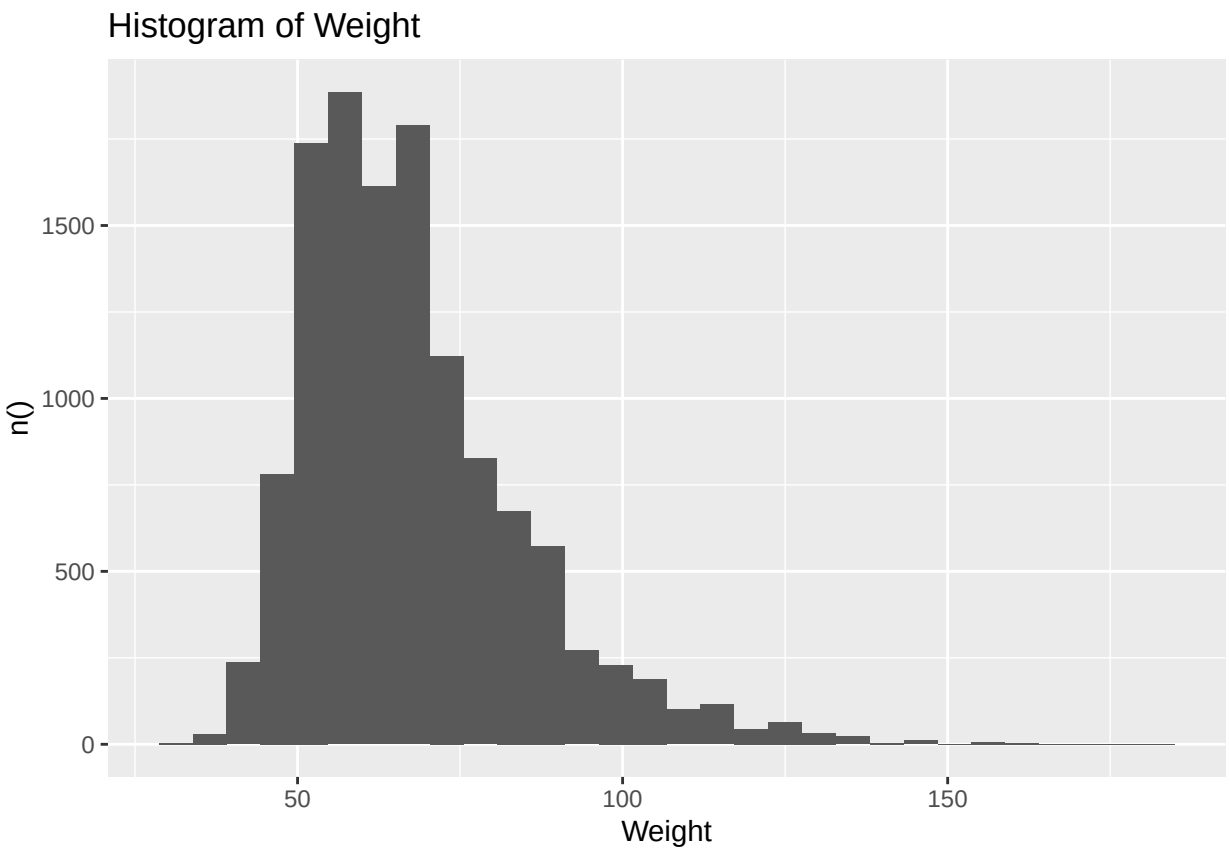
There is an observed difference, but is this difference statistically significant? In order to answer this question we will conduct a hypothesis test.

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

```
ggplot(
  weight_data,
  aes(x = weight)
) +
  geom_histogram() +
  labs(
    title = "Histogram of Weight",
    x = "Weight",
    y = "n()",
  )
```



The conditions for inference are: 1. Our observations are a simple random sample from the population of interest. - This is true as yrbss is a random sample of high school youths 2. The variable being measured is normal - Since we just looked at weight, we can look at the histogram above to see that the data is fairly normal 3. The variables are independent - As each observation is a different random individual youth, we can conclude that the data is independent.

End of your answer

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

H_0: There *is no* difference in average weight between those who exercise at least times a week and those who don't

H_A: There *is* a difference in average weight between those who exercise at least times a week and those who don't

End of your answer

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`.

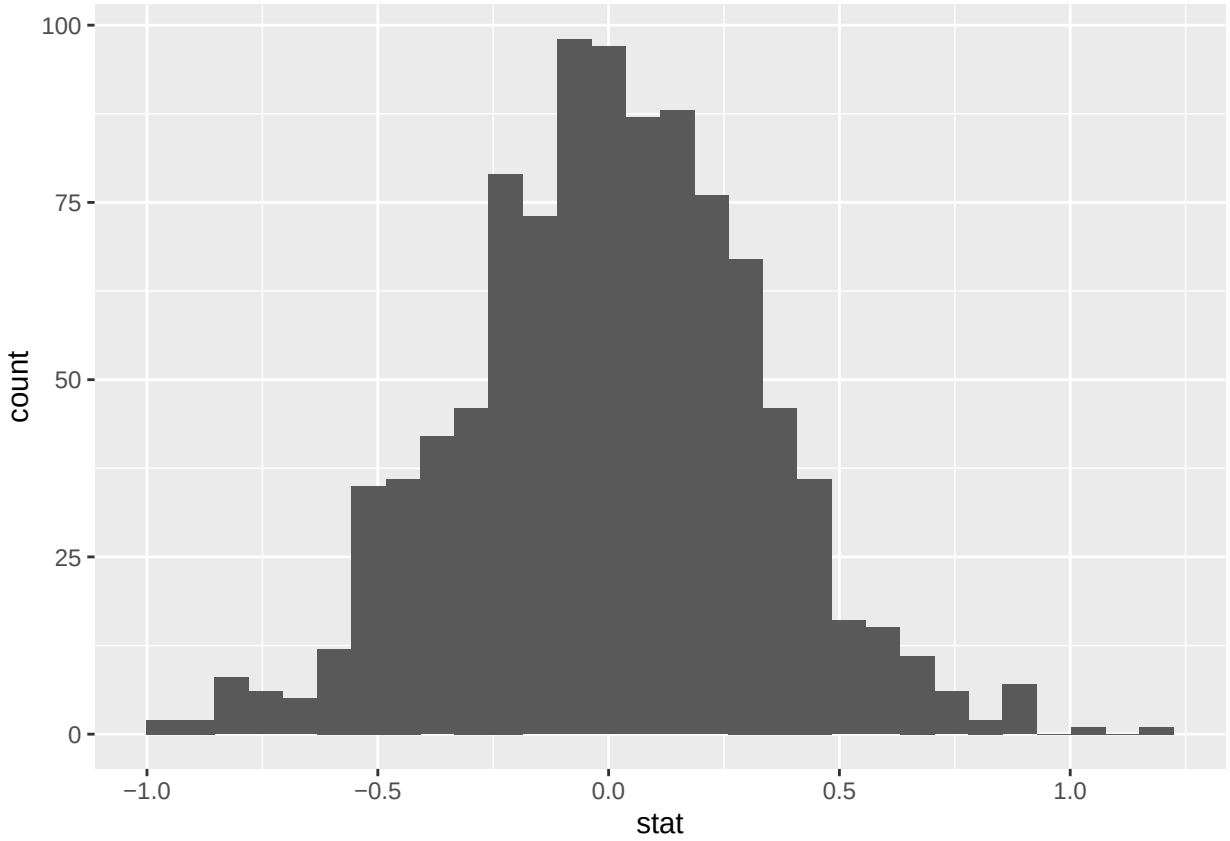
```
null_dist <- yrbss %>%
  drop_na(physical_3plus) %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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`, whichis 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

From the above, we know that `obs_diff` =1.77458426448825 and looking at the graph, there are no entries in obs_diff which are greater than 1.77458426448825

End of your answer

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.

Insert your answer here

We can get a confidence interval by using the get_ci() funciton on null_dist:

```
obs_diff_ci <- null_dist |>
  get_ci(level = 0.95)

obs_diff_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.607    0.640
```

Becasue 0 falls within the confidence interval range, we fail to reject the null hypothesis.

End of your answer

More Practice

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

Insert your answer here

We can do so with a t-test. We know we will have a large number of degrees of freedom but:

```
height_data <- yrbss |>
  select(height) |>
  drop_na(height)

height_95_result <- t.test(height_data$height, conf.level = 0.95)

height_95_lci <- height_95_result$conf.int[1]
height_95_uci <- height_95_result$conf.int[2]

height_95_result
```

```
##
##  One Sample t-test
##
## data:  height_data$height
## t = 1811.7, df = 12578, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.689411 1.693071
## sample estimates:
## mean of x
##  1.691241
```

From the above, we can see that the lower and upper bounds of the confidence interval is 1.6894112 and 1.6930708, respectively. This would mean that we are 95% confident that the full populations mean height to be between these two bounds.

End of your answer

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

Continuing with a t-test:

```
height_data <- yrbss |>
  select(height) |>
  drop_na(height)

height_90_result <- t.test(height_data$height, conf.level = 0.90)

height_90_lci <- height_90_result$conf.int[1]
height_90_uci <- height_90_result$conf.int[2]

height_90_result
```

```
##
##  One Sample t-test
##
## data:  height_data$height
## t = 1811.7, df = 12578, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  1.689705 1.692777
## sample estimates:
## mean of x
##  1.691241
```

Here we find that we are 90% confident that the lower and upper bounds of the confidence interval is 1.6897054 and 1.6927765, respectively. This is a change from the 95% confidence of $-2.9421708 \times 10^{-4}$ in the lower bound and 2.9421708×10^{-4} for the upper bound.

The very small difference between the upper and lower bounds of the confidence interval indicates a high degree of precision in our estimate of the parameter.

End of your answer

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

I'll assume that exercise means those who are physically active. With that:

H_0: There is no difference in average heights between those who are physically active 3+ days a week and those who are not.

H_A: There is a difference in average heights between those who are physically active 3+ days a week and those who are not.

```r
height_exersise <- yrbss |>
  mutate(active_3plus = ifelse(physically_active_7d >= 3, "yes", "no")) |>
  select(active_3plus, height) |>
  drop_na()

height_exersise_summary <- describeBy(
  height_exersise$height,
  group = height_exersise$active_3plus,
  mat = TRUE,
  skew = FALSE
)

height_exersise_summary[, c(2, 4:7)]
```

```
##     group1    n     mean        sd  min
## X11     no 4022 1.665587 0.1028581 1.27
## X12    yes 8342 1.703213 0.1032956 1.27
```

```r
h_e_se_yes  <- (0.1032956) / (8342)
h_e_se_no   <- (0.1028581) / (4022)

h_e_se_tot  <- sqrt(h_e_se_yes + h_e_se_no)

h_e_mean_yes  <- 1.703213
h_e_mean_no   <- 1.665587

h_e_mean_diff <- h_e_mean_yes - h_e_mean_no

h_e_lci <- (h_e_mean_diff) - (1.96 * h_e_se_tot)
h_e_uci <- (h_e_mean_diff) + (1.96 * h_e_se_tot)

h_e_reject <- ifelse(
  (
    (0 >= h_e_lci) & (0 <= h_e_uci)
  ), "fail to reject", "reject"
)

h_e_result <- ifelse(
  (
    (0 >= h_e_lci) & (0 <= h_e_uci)
  ),
  "is not",
  "is"
)
```

Because our interval is between (0.0255507,0.0497013), we reject our null hypothesis. Which means that there is a difference in average heights between those who are physically active 3+ days a week and those who are not.

End of your answer

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

There are 8 unique options for hours_tv_per_school_day. These are: 5+, 2, 3, do not watch, <1, 4, 1, NA

End of your answer

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

Is there a difference in height and how much someone sleeps?

H_0: There is no difference between the heights of each sleep group.

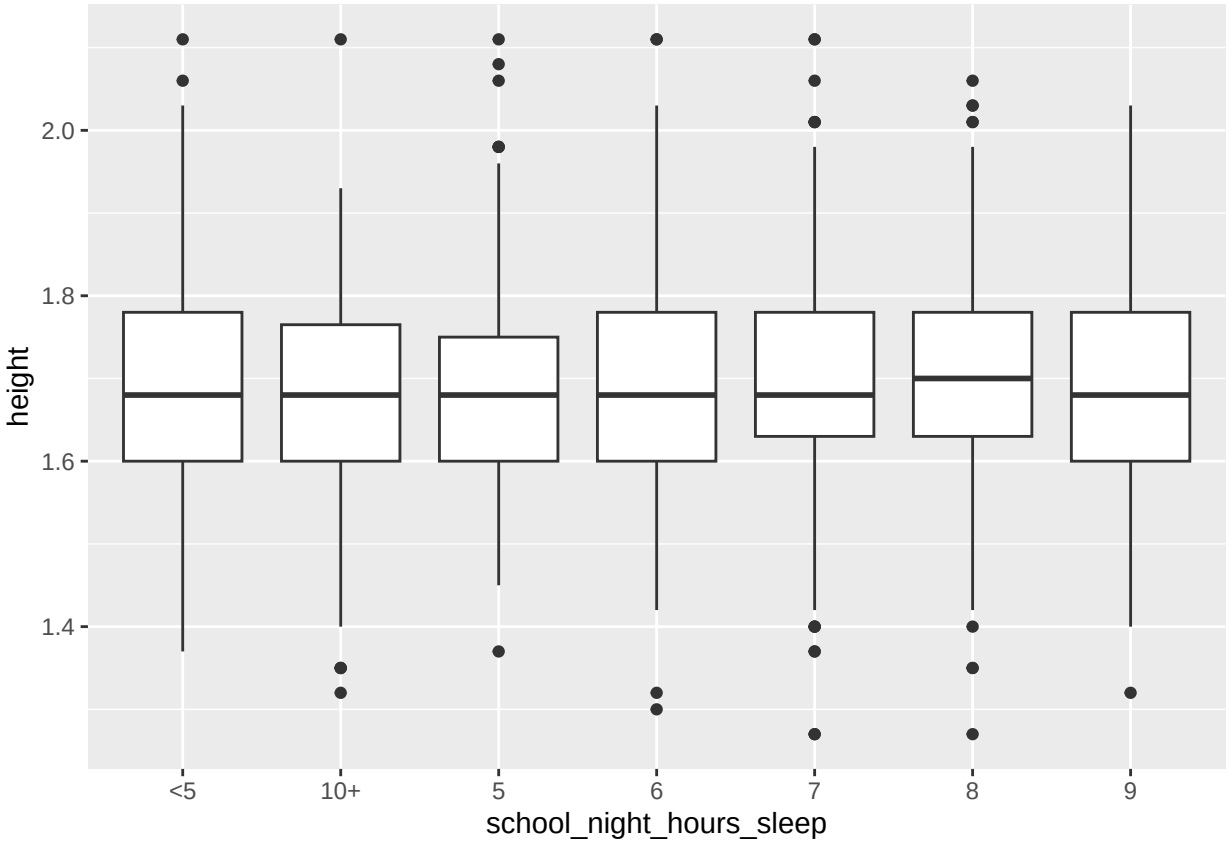
H_A: There is difference between the heights of each sleep group.

```
sleep_height <- yrbss |>
  select(height, school_night_hours_sleep) |>
  drop_na()

sleep_height
```

```
## # A tibble: 11,481 x 2
##    height school_night_hours_sleep
##     <dbl> <chr>
##  1   1.73 <5
##  2   1.6  6
##  3   1.5  9
##  4   1.57 8
##  5   1.65 9
##  6   1.88 6
##  7   1.75 <5
##  8   1.37 <5
##  9   1.68 10+
## 10   1.65 6
## # i 11,471 more rows
```

```
ggplot(
  sleep_height,
  aes(
    x = school_night_hours_sleep,
    y = height
  )
) +
  geom_boxplot() +
  theme(legend.position = "none")
```



Although we can employ ANOVA here, we can visually see that each mean completely overlaps with the mean of all the others, meaning that we can visually determine that there is no difference in height based on sleeping habits.

Let's employ ANOVA anyway and determine our alpha to be 0.05:

```
one_way_anova_height_sleep <- aov(
  height ~ school_night_hours_sleep,
  data = sleep_height
)

summary(one_way_anova_height_sleep)
```

```
##                              Df Sum Sq Mean Sq F value Pr(>F)
## school_night_hours_sleep      6   0.17 0.02776   2.538 0.0186 *
## Residuals                 11474 125.49 0.01094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Opposing our visual results, we can see that our F value is 2.538 and our Pr(>F) is 0.0186. Interpreting these results, the F score is suggesting that there is some evidence that mean heights vary between groups based on sleep and the p value is showing that the F score result is statisitcally significant.

We have sufficient evidence to reject the null hypothesis. This would mean that there is a statisitcally significant difference in height between at least one of the groups.

End of your answer