# Inference for numerical data

## 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?

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

ANSWER: - The cases in this data set are individual survey respondents to the YRBSS survey. - The are 13,583 cases in this sample.

```
nrow(yrbss)
```

```
## [1] 13583
```

```
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"~
```

```
head(yrbss)
```

```
## # A tibble: 6 x 13
##     age gender grade hispa~1 race  height weight helme~2 text_~3 physi~4 hours~5
##   <int> <chr>  <chr> <chr>   <chr>  <dbl>  <dbl> <chr>   <chr>     <int> <chr>
## 1    14 female 9     not     Blac~  NA      NA   never   0             4 5+
## 2    14 female 9     not     Blac~  NA      NA   never   <NA>          2 5+
## 3    15 female 9     hispan~ Nati~   1.73   84.4 never   30            7 5+
## 4    15 female 9     not     Blac~   1.6    55.8 never   0             0 2
## 5    15 female 9     not     Blac~   1.5    46.7 did no~ did no~       2 3
## 6    15 female 9     not     Blac~   1.57   67.1 did no~ did no~       1 5+
## # ... with 2 more variables: strength_training_7d <int>,
## #   school_night_hours_sleep <chr>, and abbreviated variable names 1: hispanic,
## #   2: helmet_12m, 3: text_while_driving_30d, 4: physically_active_7d,
## #   5: hours_tv_per_school_day
```

## 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?

ANSWER: - Missing weights from 1004 observations per the R chunk above.

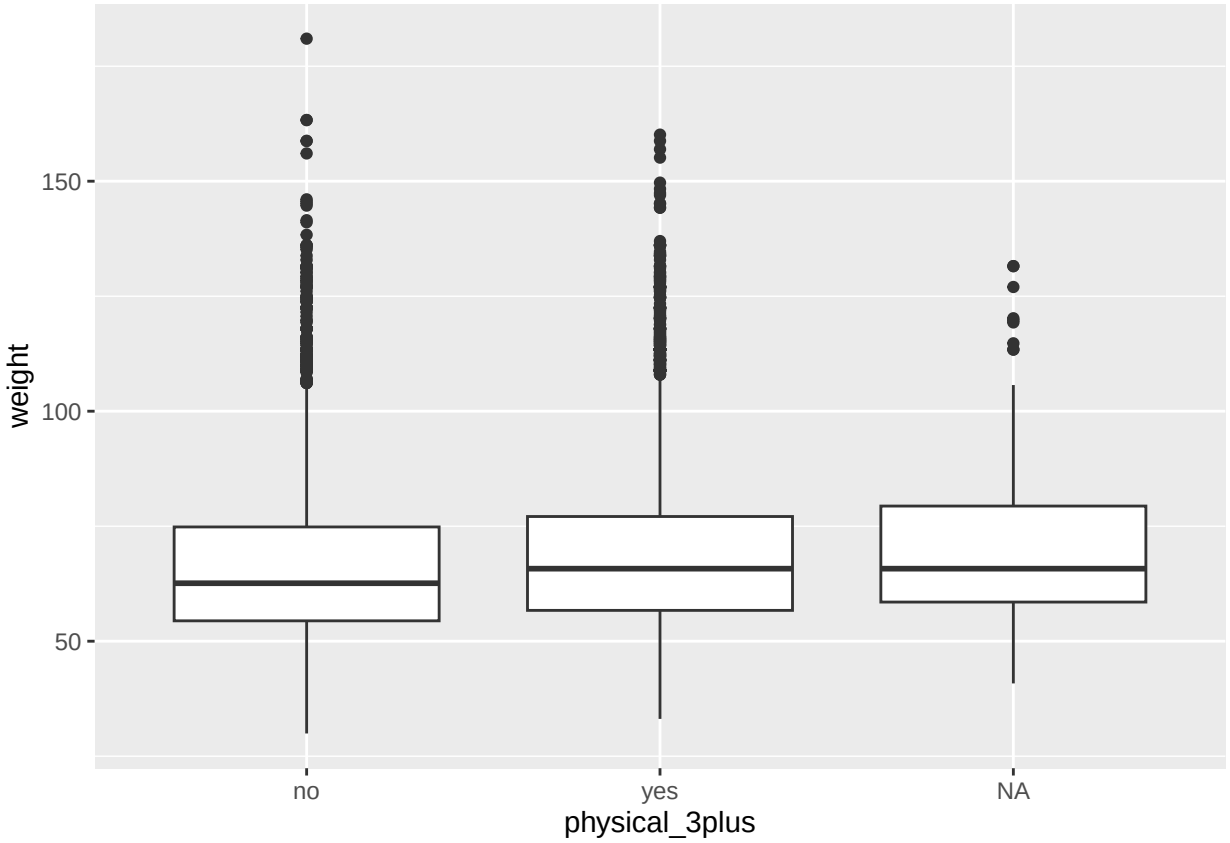
Next, consider the possible relationship between a high schooler's weight and their physical activity. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

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?

```
ggplot(yrbss, aes(x=physical_3plus, y=weight)) + geom_boxplot()
```




ANSWER: - The median weight value for high schoolers who are active 3 days or more per week is higher than for less active peers. - This could be due to more muscle, which weighs more than fat, among the physically active group. Or heavier people might be more likely to play sports (e.g. football/basketball). - However, less active high schoolers have more weight values on the extreme high end, which is what I would expect. Exercise is more difficult if you're extremely heavy. - In general I'm surprised at the median results but I feel I have reasonable hypotheses above.

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()`.

ANSWER: - Necessary conditions for inference: random (random sample); normal distribution; independent observations. - There isn't information from ?yrbss on whether the sample is random, but we assume so based on the distributions. - Sampling distribution is adequate (>100) and normally distributed for some fields (e.g. height, weight). - Observations are independent survey results from different people.

```
yrbss %>%
  group_by(physical_3plus) %>%
  summarise(size_physical_3plus = n())
```

```
## # A tibble: 3 x 2
##   physical_3plus size_physical_3plus
##   <chr>                        <int>
## 1 no                            4404
## 2 yes                           8906
## 3 <NA>                           273
```

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.

ANSWER: - Null Hypothesis: There *is no difference* in average weight between the population who exercise at least 3 days a week compared to those who exercise less. - Alternative Hypothesis: There *is a difference* in average weight between the population who exercise at least 3 days a week compared to those who exercise 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 %>%
  na.omit %>%
  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`.

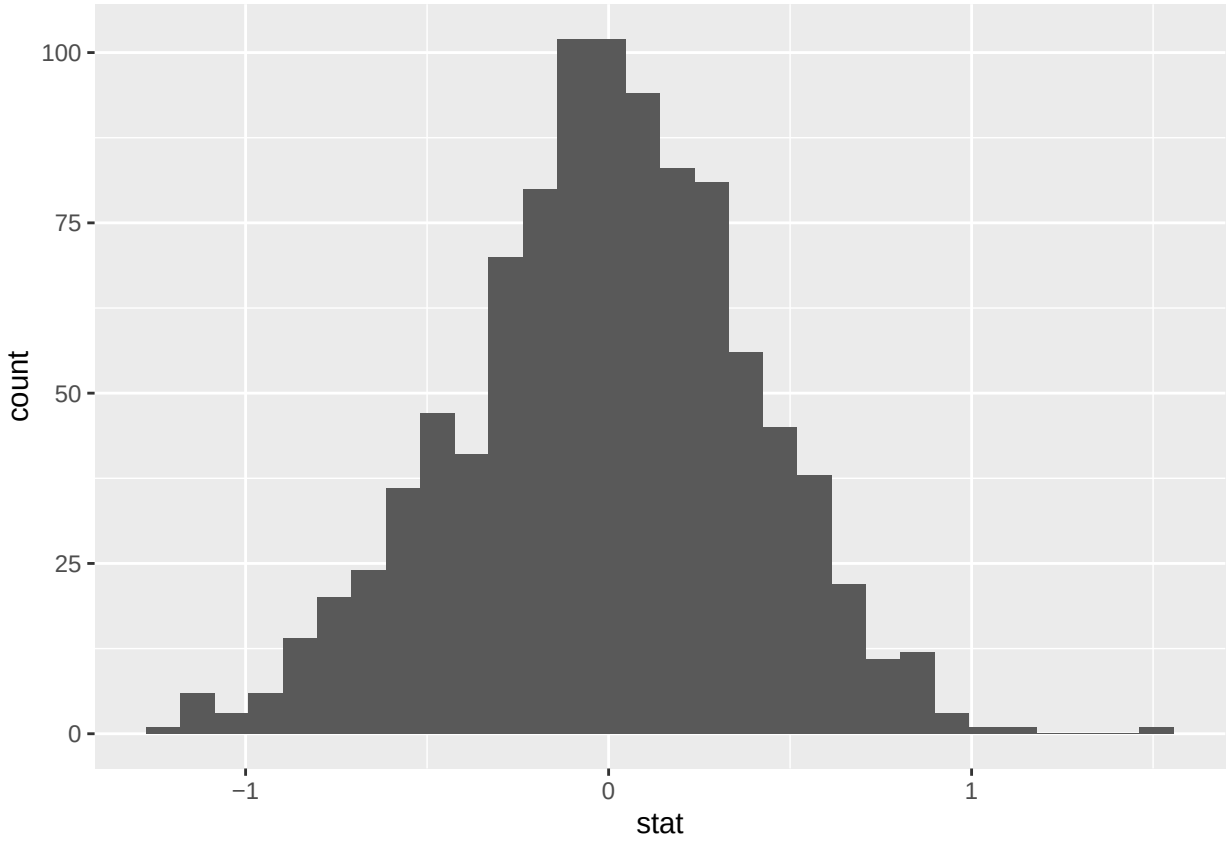
```
set.seed(1)
null_dist <- yrbss %>%
  na.omit %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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`?

ANSWER: - Per the R chunk below, 0 of these permutations have a difference of at least obs_stat (~1.53)

```
obs_diff_actual <- obs_diff$stat[1]

#Generates empty tibble, thus 0 observations meeting filter
null_dist%>%
  filter(abs(stat)>obs_diff_actual)
```

```
## Response: weight (numeric)
## Explanatory: physical_3plus (factor)
## Null Hypothesis: independence
## # A tibble: 0 x 2
## # ... with 2 variables: replicate <int>, stat <dbl>
```

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.

ANSWER: - The 95% confidence interval for the mean weight of the population who exercise fewer than 3 days per week is between 66.466 and 67.834 kilograms - The 95% confidence interval for the mean weight of the population who exercise 3 or more days per week is between 68.251 and 69.104 kilograms - Because the two confidence intervals don't overlap, this means there is a statistically significant difference between the two groups. - Therefore we reject the null hypothesis that there *is no difference* in average weight between population who exercise at least 3 days a week compared to those who exercise less. We accept the alt hypothesis that there *is a difference* in average weight between the population who exercise at least 3 days a week compared to those who exercise less.

```
ci_weight <- yrbss %>%
  na.omit %>%
  group_by(physical_3plus) %>%
  summarise(mean_weight = mean(weight, na.rm = TRUE),
            sd_weight = sd(weight, na.rm = TRUE),
            sample_size_weight = n())

inactive <- ci_weight %>% filter(physical_3plus == 'no')
inactive_mean <- inactive$mean_weight
inactive_sd <- inactive$sd_weight
inactive_n <- inactive$sample_size_weight
```

```
active <- ci_weight %>% filter(physical_3plus == 'yes')
active_mean <- active$mean_weight
active_sd <- active$sd_weight
active_n <- active$sample_size_weight


ci_inactive_bottom <- inactive_mean - 1.96 * inactive_sd / sqrt(inactive_n)
ci_inactive_top <- inactive_mean + 1.96 * inactive_sd / sqrt(inactive_n)

ci_active_bottom <- active_mean - 1.96 * active_sd / sqrt(active_n)
ci_active_top <- active_mean + 1.96 * active_sd / sqrt(active_n)

ci_inactive_bottom
```

```
## [1] 66.46554
```

```
ci_inactive_top
```

```
## [1] 67.83394
```

```
ci_active_bottom
```

```
## [1] 68.2511
```

```
ci_active_top
```

```
## [1] 69.10383
```

---

## More Practice

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

ANSWER: - We are 95% confident that the average height for the population which the yrbss sample was based on (assuming the sample was taken randomly from the population) resides between 1.6894m and 1.6931m. - See R chunk below for calculations.

```
height_df <- yrbss %>%
  filter(!is.na(height)) %>%
  select(height)

#Stats necessary to calculate confidence interval
sample_mean_height <- mean(height_df$height)
standard_dev_height <- sd(height_df$height)
sample_size_height <- nrow(height_df)
standard_error_height <- standard_dev_height/sqrt(sample_size_height)
list_of_heights <- height_df$height
```

```
t_statistic_95_height <- qt(.025,sample_size_height-1)

#Calculating confidence interval min and max
bottom_interval_95_height <- sample_mean_height - abs(t_statistic_95_height*standard_error_height)
top_interval_95_height <- sample_mean_height + abs(t_statistic_95_height*standard_error_height)

bottom_interval_95_height
```

```
## [1] 1.689411
```

```
top_interval_95_height
```

```
## [1] 1.693071
```

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.

ANSWER: - We are 90% confident that the average height for the population which the yrbss sample was based on (assuming the sample was taken randomly from the population) resides between 1.6897m and 1.6928m. - The width of the 90% confidence interval is larger (by about 9%), because if we express less confidence, we can make a wider estimate.

```
t_statistic_90_height <- qt(.05,sample_size_height-1)

bottom_interval_90_height <- sample_mean_height - abs(t_statistic_90_height*standard_error_height)
top_interval_90_height <- sample_mean_height + abs(t_statistic_90_height*standard_error_height)

bottom_interval_90_height
```

```
## [1] 1.689705
```

```
top_interval_90_height
```

```
## [1] 1.692777
```

```
width_95_height <- top_interval_95_height - bottom_interval_95_height
width_90_height <- top_interval_90_height - bottom_interval_95_height

width_95_height
```

```
## [1] 0.003659587
```

```
width_90_height
```

```
## [1] 0.00336537
```

```
width_95_height/width_90_height
```

```
## [1] 1.087425
```

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.

ANSWER: - Null Hypothesis: There *is no difference* in average height between the population who exercise at least 3 days a week compared to those who exercise less. - Alternative Hypothesis: There *is a difference* in average height between the population who exercise at least 3 days a week compared to those who exercise less.

- The 95% confidence interval for the mean height of the population who exercise fewer than 3 days per week is between 1.667 and 1.675 meters.
- The 95% confidence interval for the mean height of the population who exercise 3 or more days per week is between 1.706 and 1.712 meters.
- This means there is a statistically significant difference between the two groups.
- Therefore we reject the null hypothesis that there *is no difference* in average height between the population who exercise at least 3 days a week compared to those who exercise less. We accept the alt hypothesis that there *is a difference* in average height between the population who exercise at least 3 days a week compared to those who exercise less.

```
ci_height <- yrbss %>%
  na.omit %>%
  group_by(physical_3plus) %>%
  summarise(mean_height = mean(height, na.rm = TRUE),
            sd_height = sd(height, na.rm = TRUE),
            sample_size_height = n())


inactive_height <- ci_height %>% filter(physical_3plus == 'no')
inactive_mean_height <- inactive_height$mean_height
inactive_sd_height <- inactive_height$sd_height
inactive_n_height <- inactive_height$sample_size_height

active_height <- ci_height %>% filter(physical_3plus == 'yes')
active_mean_height <- active_height$mean_height
active_sd_height <- active_height$sd_height
active_n_height <- active_height$sample_size_height


ci_inactive_bottom_height <- inactive_mean_height - 1.96 * inactive_sd_height / sqrt(inactive_n_height)
ci_inactive_top_height <- inactive_mean_height + 1.96 * inactive_sd_height / sqrt(inactive_n_height)

ci_active_bottom_height <- active_mean_height - 1.96 * active_sd_height / sqrt(active_n_height)
ci_active_top_height <- active_mean_height + 1.96 * active_sd_height / sqrt(active_n_height)

ci_inactive_bottom_height
```

```
## [1] 1.667239
```

```
ci_inactive_top_height
```

```
## [1] 1.675013
```

```
ci_active_bottom_height
```

```
## [1] 1.706458
```

```
ci_active_top_height
```

```
## [1] 1.711835
```

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.

ANSWER: - 8 different options

```
answer_11 <- yrbss %>%
  group_by(hours_tv_per_school_day)%>%
  summarise(n())

nrow(answer_11)
```

```
## [1] 8
```

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 $\alpha$ level, and conclude in context.

ANSWER: - Research Question: Is there a statistically significant relationship between height and sleep time in the population surveyed in the sample yrbss? Based on .05 alpha level. - Assumptions: No multicollinearity between height and another factor that could be skewing results; Assume yrbss table is representative sample of population that's random, normally distributed, with independent observations, and an adequate sample.

- Null Hypothesis: There *is no difference* in average height between the the population who sleeps 7 hours or more and those who sleep less.
- Alternative Hypothesis: There *is a difference* in average height between the the population who sleeps 7 hours or more and those who sleep less.
- The 95% confidence interval for the mean height of the population who exercise sleeps 7 hours or more is between 1.692 and 1.698 meters.
- The 95% confidence interval for the mean height of the population who exercise sleeps less than 7 hours is between 1.696 and 1.702 meters.
- Since the confidence intervals overlap, this means there is *no* statistically significant difference between the two groups.
- Therefore we fail to reject the null hypothesis that there *is no difference* in average height between the population who exercise at least 3 days a week compared to those who exercise less. Assumes .05 alpha level.
- Explanation in plain words: the population height estimates for different levels of sleeping were too similar to think there's some meaningful relationship between height and sleeping.

```r
yrbss <- yrbss %>%
  mutate(sleep_7plus = ifelse(yrbss$school_night_hours_sleep >= 7, "yes", "no"))

ci_heightsleep <- yrbss %>%
  na.omit %>%
  group_by(sleep_7plus) %>%
  summarise(mean_height = mean(height, na.rm = TRUE),
            sd_height = sd(height, na.rm = TRUE),
            sample_size_height = n())

inactive_hs <- ci_heightsleep %>% filter(sleep_7plus == 'no')
inactive_mean_hs <- inactive_hs$mean_height
inactive_sd_hs <- inactive_hs$sd_height
inactive_n_hs <- inactive_hs$sample_size_height

active_hs <- ci_heightsleep %>% filter(sleep_7plus == 'yes')
active_mean_hs <- active_hs$mean_height
active_sd_hs <- active_hs$sd_height
active_n_hs <- active_hs$sample_size_height


ci_inactive_bottom_heightsleep <- inactive_mean_hs - 1.96 * inactive_sd_hs / sqrt(inactive_n_hs)
ci_inactive_top_heightsleep <- inactive_mean_hs + 1.96 * inactive_sd_hs / sqrt(inactive_n_hs)

ci_active_bottom_heightsleep <- active_mean_hs - 1.96 * active_sd_hs / sqrt(active_n_hs)
ci_active_top_heightsleep <- active_mean_hs + 1.96 * active_sd_hs / sqrt(active_n_hs)

ci_inactive_bottom_heightsleep
```

```
## [1] 1.691231
```

```r
ci_inactive_top_heightsleep
```

```
## [1] 1.698057
```

```r
ci_active_bottom_heightsleep
```

```
## [1] 1.695939
```

```r
ci_active_top_heightsleep
```

```
## [1] 1.701892
```

---