CUNY SPS DATA606 LAB7 -Inference for numerical data

Chinedu Onyeka

October 14th, 2021

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?

Solution 1:

```
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"~
```

There are 13,583 observations(cases) and 13 variables in this sample.

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?

Solution 2:

From checking the summary, we can see that there are 1004 NA's for weight. Hence, there are 1004 observations that are missing weights.

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.

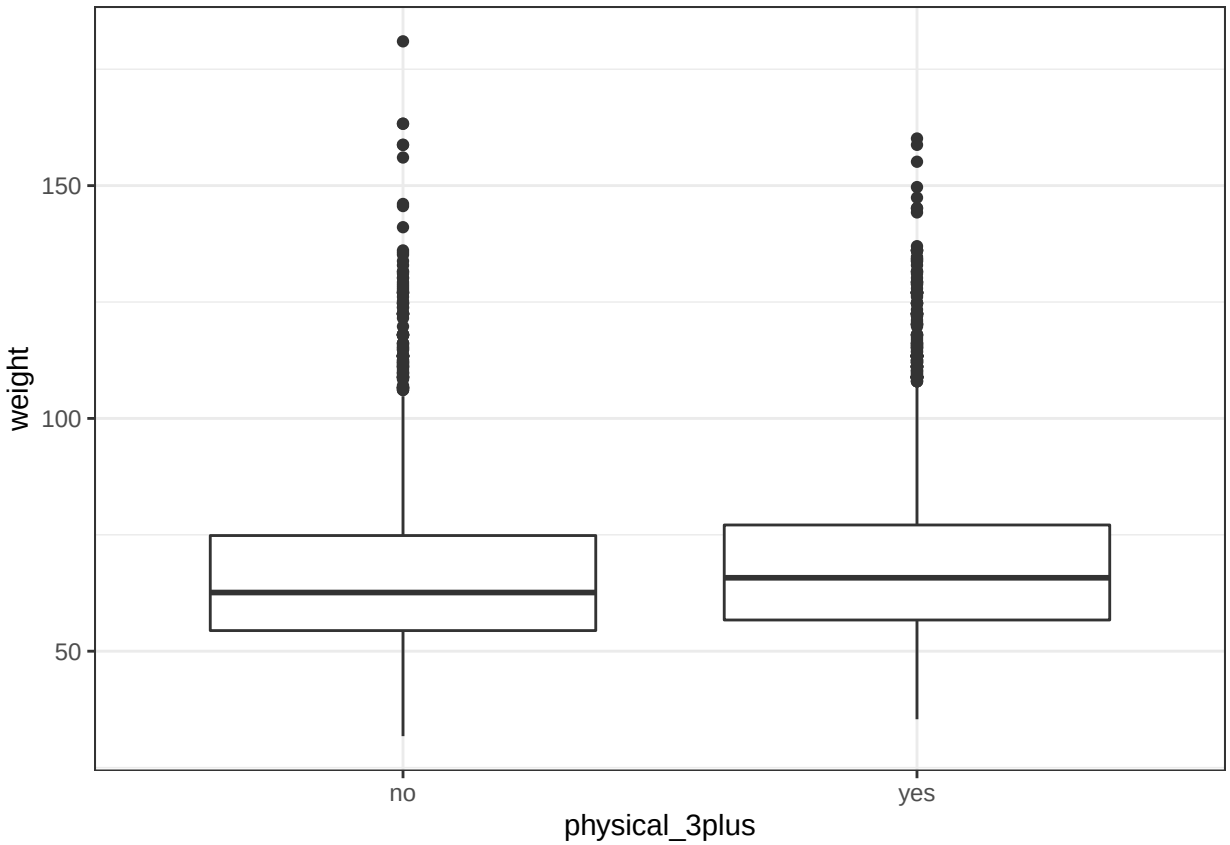
First, let's create a new variable `physical_3plus`, which will be coded as either "yes" if they are physically active for at least 3 days a week, and "no" if not.

```
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?

Solution 3:

```
yrbss %>% na.exclude() %>%
  ggplot(aes(x = weight, y = physical_3plus)) + geom_boxplot() + theme_bw() +coord_flip()
```



There seems to be a relationship. I expect that those who exercise frequently would have lesser weight compared to those who do not exercise frequently. There seems to be more outliers for those who do not exercise frequently.

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()`.

Solution 4:

```
yrbss %>% group_by(physical_3plus) %>% summarise(n = sum(table(weight)))
```

```
## # A tibble: 3 x 2
##   physical_3plus     n
##   <chr>          <int>
## 1 no              4022
## 2 yes             8342
## 3 <NA>             215
```

Independence: The data comes from a random sample of observations during one of the years the YRBSS was conducted. Hence, I can assume that the independence condition is satisfied.
Normality: The sample size is large enough and there seems to be no extreme outliers. Hence, I can assume that the sample can be modeled by a normal distribution.

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.

Solution 5:

Null Hypothesis, H_0: There is no difference in the average weight of those who exercise at least three times a week and those who don't.
Alternative Hypothesis, H_1: There is some difference in the average weight of those who exercise at least three times a week and those who don't.

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 %>%
  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`.

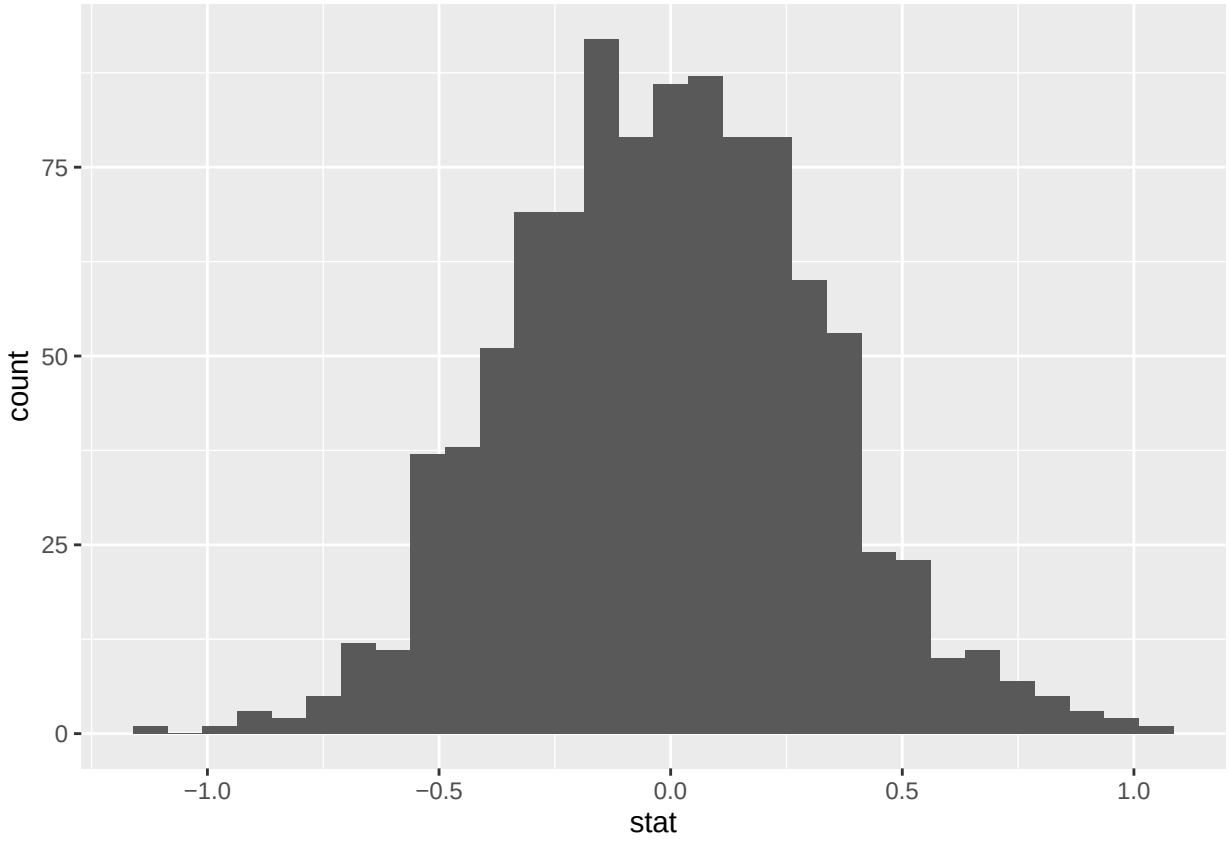
```
set.seed(110)
null_dist <- yrbss %>%
  specify(weight ~ physical_3plus) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

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:

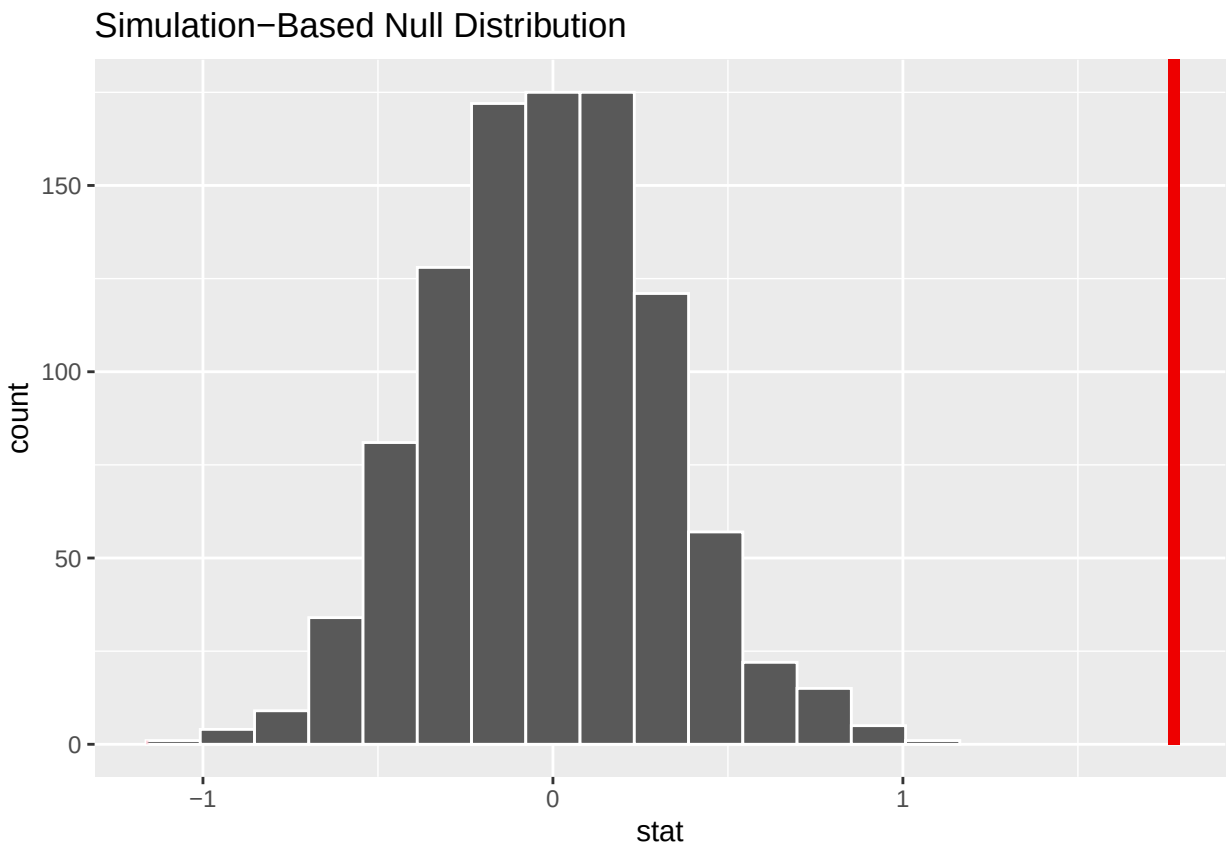
```
ggplot(data = null_dist, aes(x = stat)) +
  geom_histogram()
```



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

Solution 6:

```
visualize(null_dist) + shade_p_value(obs_stat = obs_diff, direction = "two-sided")
```



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 is 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.

Solution 7:

```
# create the table for mean, std deviation, and n for weights for those who exercise and those who don't
yrbss_table <- yrbss %>% group_by(physical_3plus) %>%
  summarise(
    mean_weight = mean(weight, na.rm = TRUE),
    std_weight = sd(weight, na.rm = TRUE),
```

```
    n = sum(table(weight)),
    SE = (std_weight)**2/n
  )
yrbss_table
```

```
## # A tibble: 3 x 5
##   physical_3plus mean_weight std_weight     n     SE
##   <chr>                <dbl>      <dbl> <int>  <dbl>
## 1 no                    66.7       17.6  4022 0.0773
## 2 yes                   68.4       16.5  8342 0.0326
## 3 <NA>                  69.9       17.6   215 1.44
```

The confidence interval is given by: $CI = pointestimate \mp Marginof Error$
$z_* = 1.96$ for 95% confidence interval.
$CI = \bar{x}_{yes} - \bar{x}_{no} \mp z_* \sqrt{\frac{(std_{yes})^2}{n} + \frac{(std_{no})^2}{n}}$

```
n1 <- yrbss_table$n[1]
n2 <- yrbss_table$n[2]
df = n1 + n2 - 2
t = abs(qt(0.025, df, lower.tail = FALSE))

point_estimate <- yrbss_table$mean_weight[2] - yrbss_table$mean_weight[1]
margin_error <- t*sqrt(yrbss_table$SE[1] + yrbss_table$SE[2])
CI <- round(c(point_estimate - margin_error, point_estimate + margin_error), 3)
paste0("The 95% confidence interval for the mean difference in weights of those",
" who exercise more than three times daily and those who do not is (", CI[1]," , ", CI[2], ")")
```

```
## [1] "The 95% confidence interval for the mean difference in weights of those who exercise more than
```

More Practice

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

Solution 8:

```
# 95% confidence interval for the height.
mean_height <- mean(yrbss$height, na.rm = TRUE)
sd_height <- sd(yrbss$height, na.rm = TRUE)
n <- sum(table(yrbss$height), na.rm = TRUE)
SE <- sd_height/sqrt(n)
point_estimate <- mean_height
z <- qnorm(0.025, lower.tail = FALSE)
ME <- z*SE
CI <- round(c(point_estimate - ME, point_estimate + ME), 3)
paste0("The 95% confidence interval for the average heights in meters is (", CI[1]," , ", CI[2], ")",
       ". That is, we are 95% confident that the height of a random individual selected from the sample
       " will be  between ", CI[1], " and ", CI[2], " meters")
```

```
## [1] "The 95% confidence interval for the average heights in meters is (1.689 , 1.693). That is, we a
```

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.

Solution 9:

```
# 90% confidence interval for the height.
z <- qnorm(0.05, lower.tail = FALSE)
ME <- z*SE
CI <- round(c(point_estimate - ME, point_estimate + ME), 3)
paste0("The 95% confidence interval for the average heights in meters is (", CI[1]," , ", CI[2], ")",
       ". That is, we are 95% confident that the height of a random individual selected from the sample
       " will be  between ", CI[1], " and ", CI[2], " meters")
```

```
## [1] "The 95% confidence interval for the average heights in meters is (1.69 , 1.693). That is, we ar
```

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.

Solution 10:

Null Hypothesis, H_0: There is no difference in the average height of those who exercise at least three times a week and those who don't.
Alternative Hypothesis, H_1: There is some difference in the average height of those who exercise at least three times a week and those who don't.

```
# create the table for mean, std deviation, and n for heights of those who excercise and those who don'
yrbss_table_h <- yrbss %>% group_by(physical_3plus) %>%
  summarise(
    mean_height = mean(height, na.rm = TRUE),
    std_height = sd(height, na.rm = TRUE),
    n = sum(table(height)),
    SE = (std_height)**2/n
  )
yrbss_table_h
```

```
## # A tibble: 3 x 5
##   physical_3plus mean_height std_height     n         SE
##   <chr>                <dbl>      <dbl> <int>      <dbl>
## 1 no                    1.67      0.103  4022 0.00000263
## 2 yes                   1.70      0.103  8342 0.00000128
## 3 <NA>                  1.71      0.107   215 0.0000530
```

```
n1 <- yrbss_table_h$n[1]
n2 <- yrbss_table_h$n[2]
df_h = n1 + n2 - 2
t = abs(qt(0.025, df, lower.tail = FALSE))
point_estimate_h <- yrbss_table_h$mean_height[2] - yrbss_table_h$mean_height[1]
Std_Error <- sqrt(yrbss_table_h$SE[1] + yrbss_table_h$SE[2])
margin_error <- t*Std_Error
Test_statistic <- (point_estimate_h - 0)/Std_Error
```

```
p_value <- pt(Test_statistic, df_h, lower.tail = FALSE)*2
CI <- round(c(point_estimate_h - margin_error, point_estimate_h + margin_error), 3)
paste0("The 95% confidence interval for the difference in heights of those who exercise",
" more than three times daily and those who do not is (", CI[1]," , ", CI[2], ")")
```

```
## [1] "The 95% confidence interval for the difference in heights of those who exercise more than three
```

```
paste0("The p_value is ", round(p_value, 5),
       ". Therefore, reject the null hypothesis at 0.05 level of significance")
```

```
## [1] "The p_value is 0. Therefore, reject the null hypothesis at 0.05 level of significance"
```

There is no sufficient statistical evidence to support the null hypothesis.

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.

Solution 11:

```
yrbss %>% group_by(hours_tv_per_school_day) %>% summarise(n())
```

```
## # A tibble: 8 x 2
##   hours_tv_per_school_day `n()`
##   <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 7 different options in this dataset. Also, there are 338 observations that had NA values.

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.

Solution 12:

Research Question: Do students who are heavier than the mean weight sleep less than those who are lighter than the mean weight?
Null Hypothesis, H_0: There is no difference in the sleep time of those who are heavier than the mean weight and those who are lighter than the mean weight.
Alternative Hypothesis, H_1: There is some difference in the sleep time of those who are heavier than the mean weight and those who are lighter than the mean weight.

Conditions:
Independence: The data comes from a random sample of observations during one of the years the YRBSS was conducted. Hence, I can assume that the independence condition is satisfied.
Normality: The sample size is large enough and there seems to be no extreme outliers. Hence, I can assume that the sample can be modeled by a normal distribution.

```
# subset the data
yrbss_sleep <- yrbss %>%
  mutate(sleep = ifelse(yrbss$school_night_hours_sleep < 6, "yes", "no"))
# Get the mean, standard deviation, and standard error
yrbss_table2 <- yrbss_sleep %>% group_by(sleep) %>%
  summarise(
    mean_weight = mean(weight, na.rm = TRUE),
    std_weight = sd(weight, na.rm = TRUE),
    n = sum(table(weight)),
    SE = (std_weight)**2/n
  )
yrbss_table2
```

```
## # A tibble: 3 x 5
##   sleep mean_weight std_weight     n     SE
##   <chr>       <dbl>      <dbl> <int>  <dbl>
## 1 no           67.5       16.5  8989 0.0303
## 2 yes          69.2       18.5  2492 0.137
## 3 <NA>         68.0       16.2  1098 0.238
```

```
# 95% confidence interval
n1 <- yrbss_table2$n[1]
n2 <- yrbss_table2$n[2]
df2 = n1 + n2 - 2
t = abs(qt(0.025, df, lower.tail = FALSE))
point_estimate2 <- yrbss_table2$mean_weight[2] - yrbss_table2$mean_weight[1]
Std_Error2 <- sqrt(yrbss_table2$SE[1] + yrbss_table2$SE[2])
margin_error2 <- t*Std_Error2
Test_statistic2 <- (point_estimate2 - 0)/Std_Error2
p_value2 <- pt(Test_statistic2, df2, lower.tail = FALSE)*2
CI2 <- round(c(point_estimate2 - margin_error2, point_estimate2 + margin_error2), 3)
paste0("The 95% confidence interval for the difference in heights of those who exercise",
" more than three times daily and those who do not is (", CI2[1]," , ", CI2[2], ")")
```

```
## [1] "The 95% confidence interval for the difference in heights of those who exercise more than three
```

```
paste0("The p_value is ", round(p_value2, 5),
       " which is less than 0.05. Therefore, reject the null hypothesis at 0.05 level of significance")
```

```
## [1] "The p_value is 8e-05 which is less than 0.05. Therefore, reject the null hypothesis at 0.05 leve
```
