Inference for numerical data

Peter Gatica

04/08/2021

```
# knitr::opts_chunk$set(eval = TRUE, results = True, fig.show = "show", message = FALSE, warning = FALS
```

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)
```

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?

   Answer: There are 6 cases in the dataset and sample and they are helmet_12m, text_while_driving_30d, physically_active_30d, hours_tv_per_school_day, strength_training_7d, school_night_hours_sleep.

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...
## $ gender                   <chr> "female", "female", "female", "female", "f...
## $ grade                    <chr> "9", "9", "9", "9", "9", "9", "9", "9", "9...
## $ hispanic                 <chr> "not", "not", "hispanic", "not", "not", "n...
## $ race                     <chr> "Black or African American", "Black or Afr...
## $ height                   <dbl> NA, NA, 1.73, 1.60, 1.50, 1.57, 1.65, 1.88...
## $ weight                   <dbl> NA, NA, 84.37, 55.79, 46.72, 67.13, 131.54...
## $ helmet_12m               <chr> "never", "never", "never", "never", "did n...
## $ text_while_driving_30d   <chr> "0", NA, "30", "0", "did not drive", "did ...
## $ physically_active_7d     <int> 4, 2, 7, 0, 2, 1, 4, 4, 5, 0, 0, 0, 4, 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, ...
## $ school_night_hours_sleep <chr> "8", "6", "<5", "6", "9", "8", "9", "6", "...
```

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
```

1. How many observations are we missing weights from?

   Answer: There are 1004 weights missing as signified by NA's.

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.

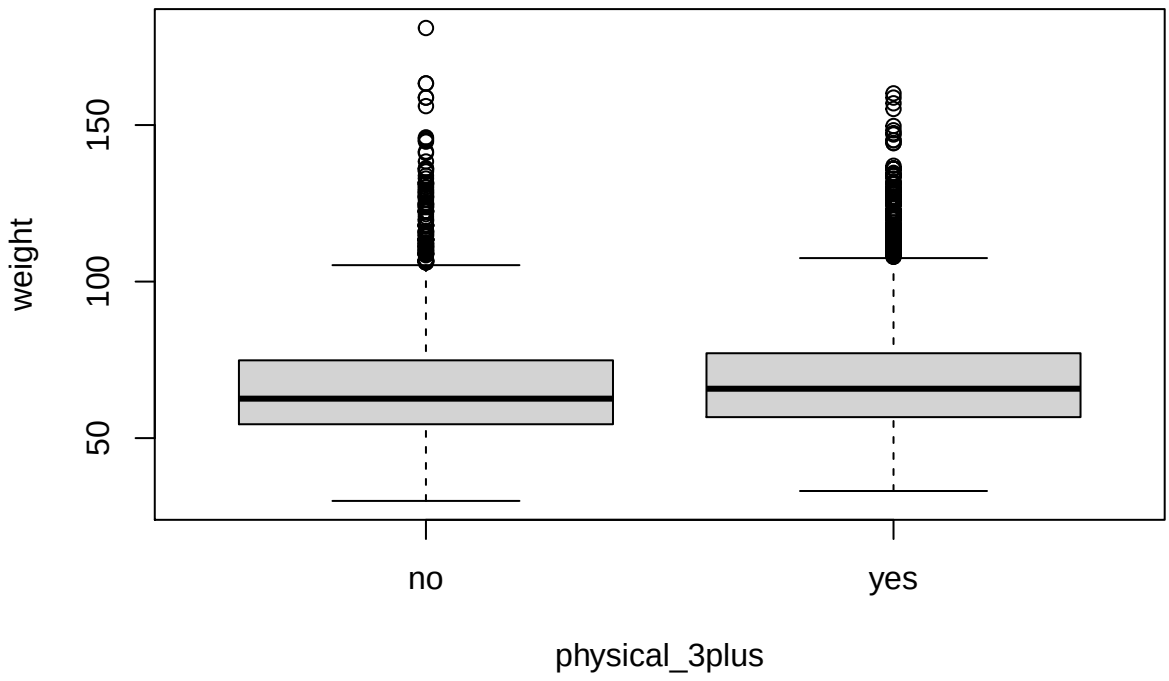
```
yrbss <- yrbss %>%
  mutate(physical_3plus = ifelse(yrbss$physically_active_7d > 2, "yes", "no"))
head(yrbss,20)
```

```
## # A tibble: 20 x 14
##      age gender grade hispanic race  height weight helmet_12m text_while_driv~
##    <int> <chr>  <chr> <chr>    <chr>  <dbl>  <dbl> <chr>      <chr>
##  1    14 female 9     not      Blac~  NA      NA   never      0
##  2    14 female 9     not      Blac~  NA      NA   never      <NA>
##  3    15 female 9     hispanic Nati~   1.73   84.4 never      30
##  4    15 female 9     not      Blac~   1.6    55.8 never      0
```

```
##  5    15 female 9     not      Blac~   1.5    46.7 did not r~ did not drive
##  6    15 female 9     not      Blac~   1.57   67.1 did not r~ did not drive
##  7    15 female 9     not      Blac~   1.65  132.  did not r~ <NA>
##  8    14 male   9     not      Blac~   1.88   71.2 never      <NA>
##  9    15 male   9     not      Blac~   1.75   63.5 never      <NA>
## 10    15 male   10    not      Blac~   1.37   97.1 did not r~ <NA>
## 11    15 female 9     not      Blac~   1.68   69.8 did not r~ 0
## 12    15 female 9     not      Blac~   1.65   66.7 did not r~ did not drive
## 13    15 female 9     not      Blac~   1.63   67.1 did not r~ <NA>
## 14    16 male   9     not      Blac~   1.68   74.8 never      0
## 15    16 male   9     not      Blac~   1.85   74.4 did not r~ did not drive
## 16    15 male   9     not      Blac~   1.78   70.3 did not r~ 0
## 17    14 male   9     not      Blac~   1.73   73.5 never      did not drive
## 18    15 male   9     not      Blac~   1.83   67.6 never      0
## 19    14 male   9     not      Blac~   1.68   46.3 <NA>       0
## 20    16 male   9     not      Blac~   1.83   73.5 never      did not drive
## # ... with 5 more variables: physically_active_7d <int>,
## #   hours_tv_per_school_day <chr>, strength_training_7d <int>,
## #   school_night_hours_sleep <chr>, physical_3plus <chr>
```

1. 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?

```
boxplot(weight ~ physical_3plus, yrbss)
```



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))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # 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

1. 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()`.

```
no_group <- yrbss %>%
   filter(physical_3plus == "no")
no_group %>% summarise(n = n())
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  4404
```

```
yes_group <- yrbss %>%
   filter(physical_3plus == "yes")
yes_group %>% summarise(n = n())
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  8906
```

1. 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:

$$H_0 : \mu_{we} - \mu_{woe} = 0 H_A : \mu_{we} - \mu_{woe} \neq 0$$

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")))
```

```
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  1.77
```

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

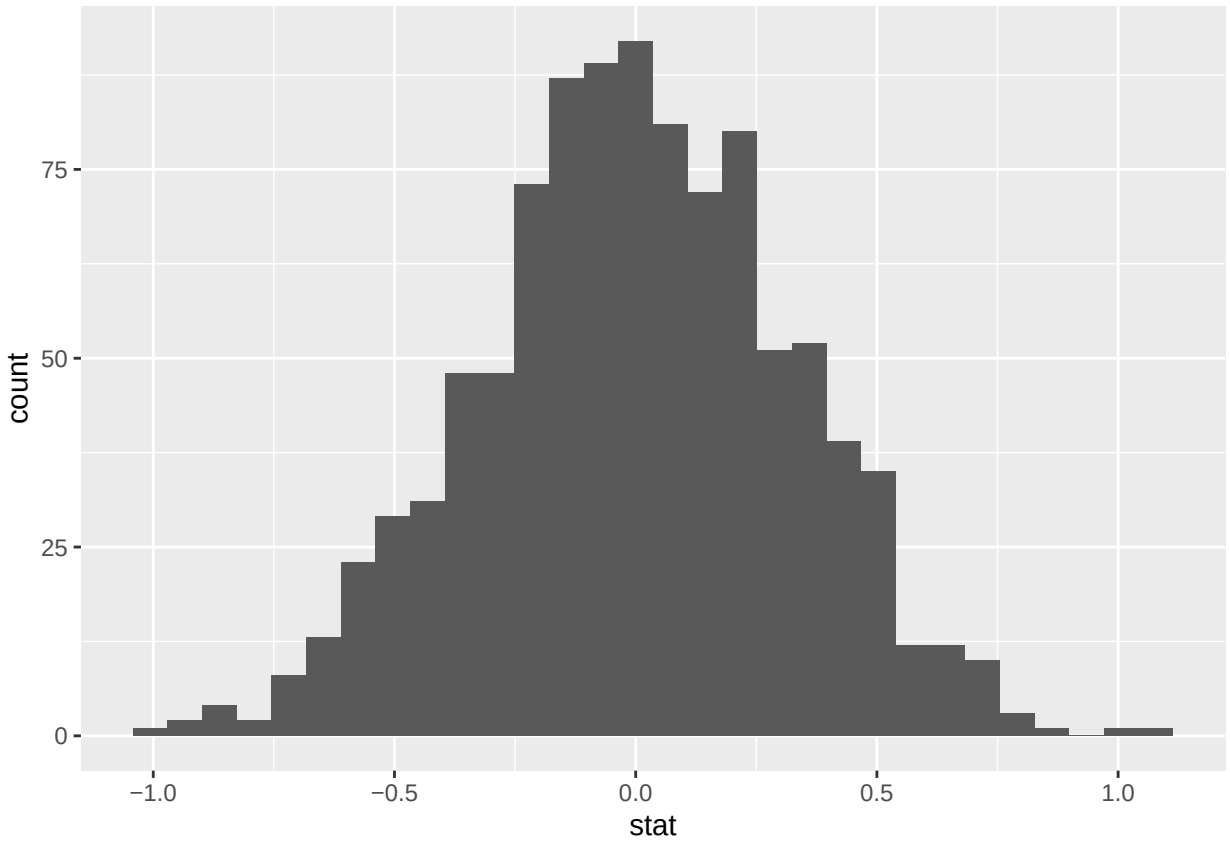
Here, `hypothesize` is used to set the null hypothesis as a test for independence. In one sample case, the `null` argument can be set to "point" to test a hypothesis relative to a point estimate.

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()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
nrow(null_dist %>% filter(stat < obs_diff$stat))
```

```
## [1] 1000
```

```
Answer: There are no "null" permutations that have a difference of at least 1.77
```

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")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of ‘reps‘ chosen in the ‘generate()‘ step. See
## ‘?get_p_value()‘ for more information.
```

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

This the standard workflow for performing hypothesis tests.

1. 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.

```
xbar_1 <- mean(yes_group$weight, na.rm = TRUE)
s1 <- sd(yes_group$weight, na.rm = TRUE)
(n1 <- nrow(yes_group %>% filter(weight != 'NA')))
```

```
## [1] 8342
```

```
xbar_2 <- mean(no_group$weight, na.rm = TRUE)
s2 <- sd(no_group$weight, na.rm = TRUE)
(n2 <- nrow(no_group %>% filter(weight != 'NA')))
```

```
## [1] 4022
```

```
min_n <- n2
se <- (sqrt((s1^2 / n1) + (s2^2 / n2)))
me <- qt(0.95,df=min_n-1) * se
```

```
(low <- (xbar_1-xbar_2) - me)
```

```
## [1] 1.22917
```

```
(high <- (xbar_1-xbar_2) + me)
```

```
## [1] 2.319999
```

Answer: The 95% confidence interval for the difference between the weights of those who exercise at least three times a week and those who don't is between (1.23 kg, 2.32 kg).

More Practice

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

```
xbar <- mean(yrbss$height, na.rm = TRUE)
s <- sd(yrbss$height, na.rm = TRUE)
n <- nrow(yrbss %>% filter(height != 'NA'))
```

```
me <- qt(.95,df=n-1,) * s/sqrt(n)
```

```
(low <- xbar - me)
```

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

```
(high <- xbar + me)
```

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

Answer: The 95% confidence interval for the average height in meters is between (1.689m, 1.693m).

2. 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.

```
me <- qt(.90,df=n-1,) * s/sqrt(n)
```

```
(low <- xbar - me)
```

```
## [1] 1.690045
```

```
(high <- xbar + me)
```

```
## [1] 1.692437
```

Answer: The 90% confidence interval for the average height in meters is between (1.690m, 1.692m). The width seems to have become less at the 90% confidence level since it is a lower percentage.

3. 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.

$$H_0 : \mu_{hwe} - \mu_{hwoe} = 0 H_A : \mu_{hwe} - \mu_{hwoe} \neq 0$$

```
xbar_1 <- mean(yes_group$height, na.rm = TRUE)
s1 <- sd(yes_group$height, na.rm = TRUE)
(n1 <- nrow(yes_group %>% filter(height != 'NA')))
```

```
## [1] 8342
```

```
xbar_2 <- mean(no_group$height, na.rm = TRUE)
s2 <- sd(no_group$height, na.rm = TRUE)
(n2 <- nrow(no_group %>% filter(height != 'NA')))
```

```
## [1] 4022
```

```
min_n <- n2
(se <- (sqrt((s1^2 / n1) + (s2^2 / n2))))
```

```
## [1] 0.001977258
```

```r
me <- qt(0.95,df=min_n-1) * se
```

```r
(low <- round((xbar_1-xbar_2) - me,3))
```

```
## [1] 0.034
```

```r
(high <- round((xbar_1-xbar_2) + me,3))
```

```
## [1] 0.041
```

```r
(xbar_diff <- round((xbar_1 - xbar_2),3))
```

```
## [1] 0.038
```

```r
(t_score <- ((xbar_diff - 0) / se))
```

```
## [1] 19.21853
```

```r
(p_val <- 2 * pt(t_score,df=min_n - 1, lower.tail = FALSE))
```

```
## [1] 8.031991e-79
```

Answer: Reject the null hypothesis because the p_value falls below .05 and since we set the null value to 0, it does not fall with the confidence level. There is signicant evidence that there is a difference in the average height between those that exercise 3 or more days a week than those that do not.

4. 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.

5. 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.
