Inference for numerical data

Gabriel Campos

Getting Started

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?
Categorical Data: Gender,grade, hispanic, race, helmet_12m, text_while_driving_30d
Numerical: age, height, weight, physically_active_7d, hours_tv_per_school_day, strenth_training_7d, school_night_hours_sleep
Outliers: 1004

ANSWER: There are 13,583 cases in our data set, representing frequency of texting and driving, and helmet use.

```
summary(yrbss)
```

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
```

2.

How many observations are we missing weights from?
ANSWER: 1004

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.

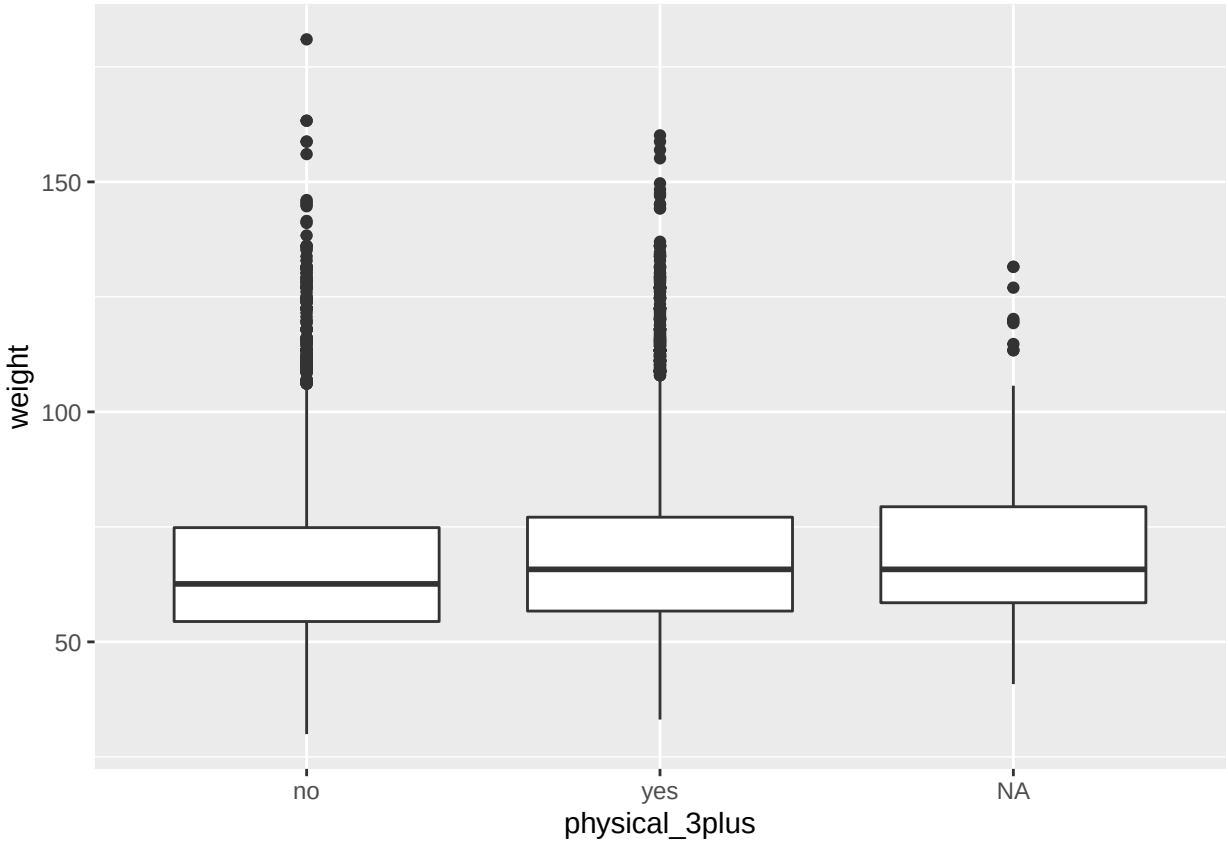
First, let's create a new variable `physical_3plus`, which will be coded as either "yes" if they are physically active for at least 3 days a week, and "no" if not.

```
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?
ANSWER:
I. The boxplot shows that the data is right skewed regardless if the person is physically active or not for 3 days.
II. There is more of a variance when not active and plenty of variance with the outliers.
III. The relationship shows weight **STARTS** to vary without activity,

This makes sense since 3 days of inactivity me begin to cause weight gain for some people, but not necessarily all, nor should the change be drastic.



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))
paste0("With # of no's = ",length(which(yrbss$physical_3plus == "no")),
       " and # of yes' = ", length(which(yrbss$physical_3plus == "yes")))
```

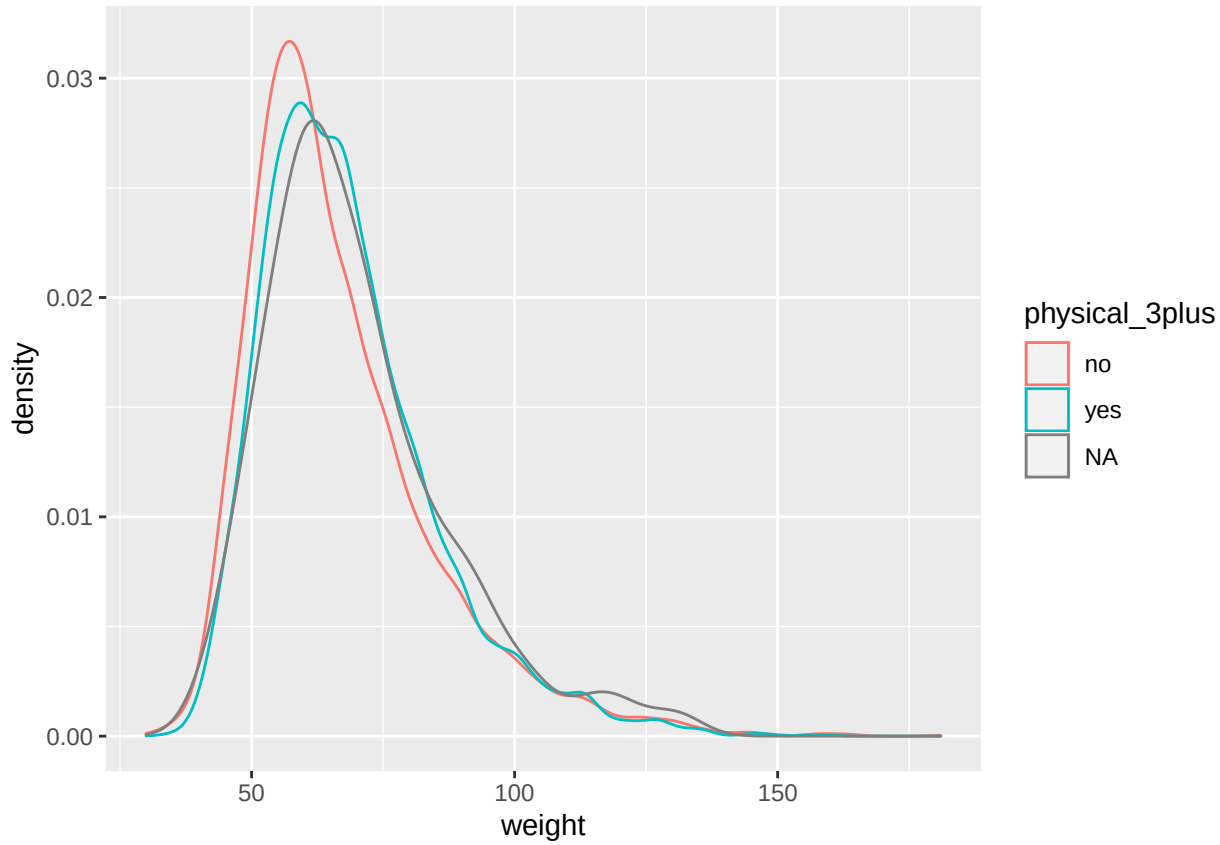
There is an observed difference, but is this difference statistically significant? In order to answer this question we will conduct a hypothesis test.

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: Number of candidates inactive for 3 days is 4044 and active within at 8906 making sample size sufficient since $n<30$. Each event is independent, so despite the skewed data, conditions for inference is satisfied


0.03
0.02
0.01
0.00
density
50
100
150
weight
physical_3plus
no
yes
NA

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
H_0:$\mu_{exercise \neq 3day} = \mu_{exercise=3days}$
H_A:$\mu_{non-exercise \neq 3day} > \mu_{exercise=3day}$

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`.

```
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`, 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`?
ANSWER:

The # of `null` permutations have a difference of at least `obs_stat` is 0

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`.

```r
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.
```

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: Using Inference function in **library(DATA606)** with parameters *type* = *ci* and *method* = *theoretical* for the **CLT** method, for the below output. We see that we cannot reject the NULL hypothesis that 3days of inactivity does not necessarily result in weight gain, since the data does not support our H_A

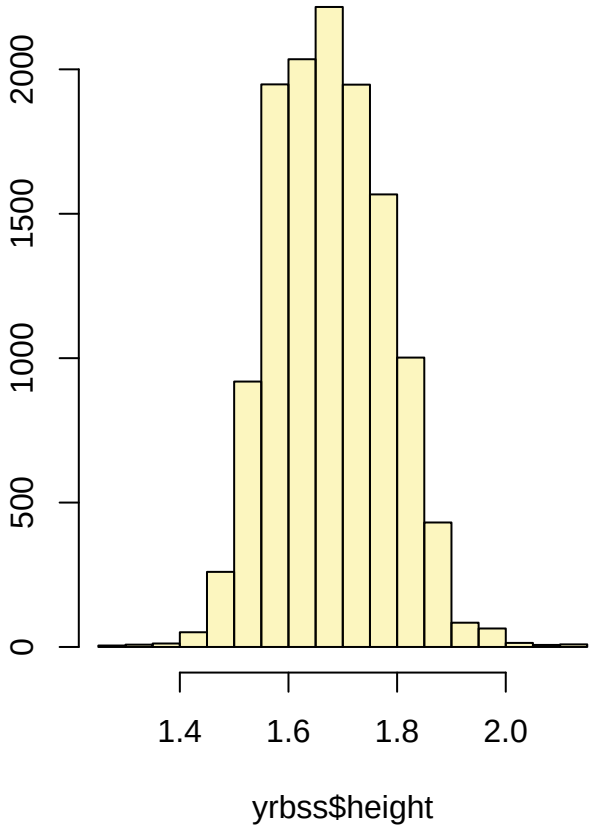
```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_no = 4022, mean_no = 66.6739, sd_no = 17.6381
## n_yes = 8342, mean_yes = 68.4485, sd_yes = 16.4783

## Observed difference between means (no-yes) = -1.7746
##
## Standard error = 0.3315
## 95 % Confidence interval = ( -2.4243 , -1.1248 )
```

More Practice

8.

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

```
## Single mean
## Summary statistics:
```

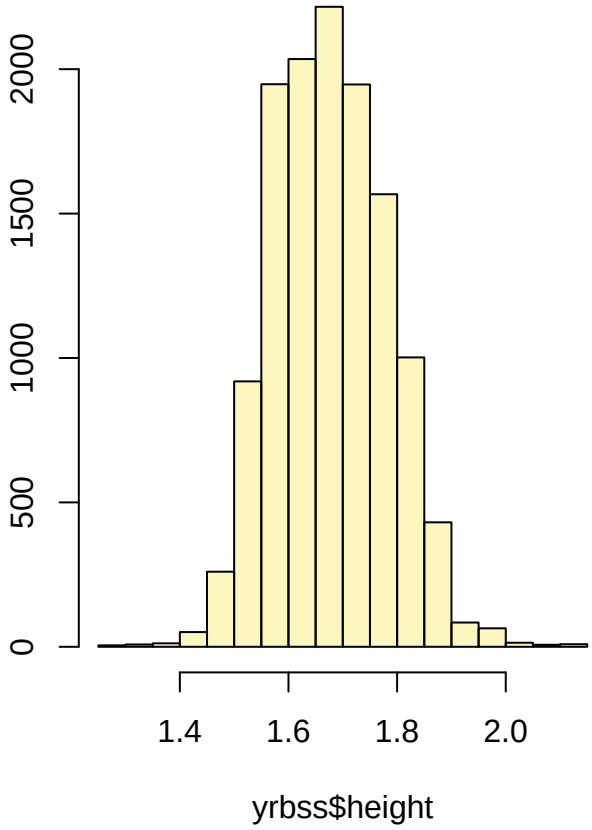



```
## mean = 1.6912 ;  sd = 0.1047 ;  n = 12579
## Standard error = 9e-04
## 95 % Confidence interval = ( 1.6894 , 1.6931 )
```

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.

```
## Single mean
## Summary statistics:
```




```
## mean = 1.6912 ;  sd = 0.1047 ;  n = 12579
## Standard error = 9e-04
## 90 % Confidence interval = ( 1.6897 , 1.6928 )
```

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.

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

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

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 There are 8 different options **<1,1,2,3,4,5+,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.

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_no = 6493, mean_no = 67.2397, sd_no = 16.2602
## n_yes = 4988, mean_yes = 68.7439, sd_yes = 17.8082

## Observed difference between means (no-yes) = -1.5041
##
## Standard error = 0.323
## 95 % Confidence interval = ( -2.1371 , -0.8711 )
```

ANSWER:

H_0**:**$\mu_{weight \neq sleep7less} = \mu_{weight=sleep7less}$
H_A**:**$\mu_{weight \neq sleep7less} > \mu_{weight=sleep7less}$

Using Inference function in **library(DATA606)** with parameters $type = ci$ and $method = theoretical$ for the **CLT** method, for the below output. We see that we cannot reject the NULL hypothesis that Less than 7 hours of sleep does not necessarily result in weight gain, since the data does not support our H_A
