Inference for numerical data

Shana Green

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(ggplot2)
library(DATA606)
```

Creating a reproducible lab report

To create your new lab report, in RStudio, go to New File -> R Markdown... Then, choose From Template and then choose `Lab Report for OpenIntro Statistics Labs` from the list of templates.

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?

The cases in this data set analyzes the behaviors of the high school students. There are 13,583 cases in our sample.

```
colnames(yrbss)
```

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

```
glimpse(yrbss)
```

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)
```

2. How many observations are we missing weights from?

After running the summary for yrbss$weight, there are 1004 observations missing weights.

```
sum(is.na(yrbss$weight))
```

After running the summary for yrbss$weight, there are 1004 observations 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?

```
ggplot(yrbss, aes(x=weight, y=physical_3plus)) +  geom_boxplot(fill=’ blue’,color="black") +theme_class
```

```
## Warning: Removed 1004 rows containing non-finite values (stat_boxplot).
```

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))
```

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()`.

We can assume an independent sample and since the normality has at least 30 samples, then yes.

```r
yrbss %>%
  group_by(physical_3plus) %>%
  summarize(mean_weight = mean(weight, na.rm = TRUE), count = n())
```

5. Write the hypotheses for testing if the average weights are different for those who exercise at least three times a week and those who don't.

H_0 : Students who are physically active 3 or more days per week have the same average weight as those who are not physically active 3 or more days per week.

H_A : Students who are physically active 3 or more days per week have a different average weight when compared to those who are not physically active 3 or more days per week.

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`.

```r
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`.

```r
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:

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

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

```
obs_diff <- yrbss %>%
  specify(weight ~ physical_3plus) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
## Warning: Removed 1219 rows containing missing values.
```

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

```
## Warning: Removed 1219 rows containing missing values.
```

```
null_dist %>% filter(stat >=    obs_diff) %>% nrow()
```

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.
```

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.

```
inference(yrbss, y = yrbss$weight, x = yrbss$physical_3plus, est = "mean", type = "ci", null = NULL,
         alternative = "twosided", method = "theoretical")
```

```
## Warning: Ignoring success since y are numerical.
```

More Practice

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

```
heightdf <- yrbss %>% select(height) %>% na.omit()

mean_height <- mean(heightdf$height)
sd_df <- sd(heightdf$height)
max_df <- max(heightdf$height)
sd_height <- sd(heightdf$height)
error_height <- sd_height / sqrt(nrow(heightdf))
```

```
tv_df<- qt(.05/2, nrow(heightdf) - 1, lower.tail = FALSE)
right <- mean_height + tv_df * error_height
left <- mean_height - tv_df * error_height
left
right
```

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.

```
tv_df<- qt(.1/2, nrow(heightdf) - 1, lower.tail = FALSE)
right <- mean_height + tv_df * error_height
left <- mean_height - tv_df * error_height
left
right
```

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.

H_0 : There is no difference in the average height of those who are exercise active at least 3 days per week and those who don't.

H_A : There is no difference in the average height of those who are exercise active at least 3 days per week and those who don't.

```
height_exercise <- yrbss %>%
  filter(physical_3plus == "yes") %>%
  select(height) %>%
  na.omit()

height_noexercise <- yrbss %>%
  filter(physical_3plus == "no") %>%
  select(height) %>%
  na.omit()

# Starting with a box plot for an initial idea
boxplot(height_exercise$height, height_noexercise$height,
        names = c("exercise", "no_exercise"))
```

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.

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

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.

“‘
