Inference for numerical data

Sean Connin

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

Assuming "cases" refers to observations then there are 13,583 cases in this dataset.

```
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)
```

1. How many observations are we missing weights from?

There are 1004 missing observations for the weights variable.

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"))
```

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?

There does not appear to be a strong relationship between weight and physical activity (>= 3 days/wk). This doesn't surprise me given that other factors may influence this comparison (e.g., age, gender, hrs TV, etc.) and that 3 days with exercise may be enough to offset weight gain among teens.

```
ggplot(data = yrbss, mapping = aes(x=physical_3plus, y = weight))+
  geom_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

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()`.

The conditions for inference include:

1. Independence within and between the two groups; if sampling without replacement, n<10% of population.

This condition is satisfied. See below.

2. Normality - check outliers for each group separately, no extreme skew in either group.

This condition may not be satisfied. The interquartile ranges are similar for both groups. However, both groups appear to be right skewed – with a number of outliers between 100-150 lbs.

```
library(glue)

#obtain total obs for weight, do not consider NA cells.

tot<-sum(!is.na(yrbss$weight))

yrbss %>%
  group_by(physical_3plus) %>%
  summarise(group_num = n())

percent_yes <- round((8906/tot)*100, 2)
percent_no <- 100-percent_yes

glue("The total number of weight counts is {tot}.")
glue('Percent of yes responses: {percent_yes}%.')
glue('Percent of no responses: {percent_no}%.')
```

1. Write the hypotheses for testing if the average weights are different for those who exercise at least 3 times a week and those who don't.

H0 - There are no differences in the average weights of individuals who exercise at least 3 times a week and those that don't.

HA - The average weights of individuals who exercise at least 3 times a week are different from 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"))

(obs_diff) #-->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 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()
```

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

None

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")

#P --> 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.

the difference in weights does not straddle 0. As a result, we reject the null hypothesis.

```
null_dist%>%get_ci(point_estimate = obs_diff, level = .95, type = "se")

# lower ci = 1.13, upper ci = 2.42
```

More Practice

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

Since the confidence interval does not include 0 we reject the null hypothesis.

```
# Inference on a mean, start with the point statistic

x_bar <- yrbss %>%
  specify(response = weight) %>%
  calculate(stat = "mean") # --> 67.9065

glue("the sample mean is {x_bar}.")

#generate the null hypothesis

null <- yrbss %>%
   specify(response = weight) %>%
   generate(reps = 1000, type = "bootstrap") %>%
   calculate(stat = "mean")
```

```r
#get the CI using the null distribution and SE

null %>%get_ci(point_estimate = x_bar, level = 0.95, type = "se")

# lower ci = 67.6 and upper ci = 68.2

#get CI using p_value

null%>%get_p_value(obs_stat = x_bar, direction = "two-sided") #-->0.958

#visualize(null) +shade_confidence_interval(endpoints = ci)
```

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.

The widths of the two intervals are nearly identical. The width at 90% is one point narrower than at 95%. I'm not sure this is meaningful

```r
#get the CI using the null distribution and SE

null %>% get_ci(point_estimate = x_bar, level = 0.90, type = "se")

# lower ci = 67.7 and upper ci = 68.2

#visualize(null) +shade_confidence_interval(endpoints = ci)
```

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.

The confidence interval for the difference of means is -0.263 to 0.339. Since it includes null=0, we fail to reject the null hypothesis.

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

(ht_diff)  # --> 0.376

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

null %>% get_ci(point_estimate = ht_diff, level = 0.95, type = "se")  # --> -0.263, 0.339
```

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.

There are 8 different categories (options) for hours_tv_per_school_day.

```
cnt<-yrbss%>%distinct(hours_tv_per_school_day)
count(cnt)
```

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.

Ques: Do students who sleep less than 8 hrs a night weigh more than students who do not?

Ho - There is no difference in the average weight of students who sleep less than 8 hrs per day vs those who do.

Ha - The average weight of students who sleeps less that 8 hrs per day is different than those who do not

condition of independence holds, use bootstrapping to meet condition of normality.

Given a 95% confidence level, the confidence interval for the difference in average weights is -2.26 -to- -0.760. Therefore we do not reject the null hypothesis

Comparing boxplots for those who sleep < 8hrs/day ("no") to those who sleep 8 or more hrs/day ("yes"), it appears that the former students weigh more on the average than the latter students.

```
data <- yrbss%>%select(weight, school_night_hours_sleep)
data%>%distinct(school_night_hours_sleep)

data<-data%>%mutate(hrs_sleep = case_when(grepl("6", school_night_hours_sleep) ~ "no",
    grepl("<5", school_night_hours_sleep, ignore.case=TRUE) ~ "no", grepl("5", school_night_hours_sleep

#compare distributions

ggplot(data = data, mapping = aes(x=hrs_sleep, y = weight))+
  geom_boxplot()
```

```
# calculate point estimate for diff of means

o_diff <- data%>%
  specify(weight ~ hrs_sleep) %>%
  calculate(stat = "diff in means", order = c("yes", "no"))  #--> -1.51

# construct null distribution

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

# estimate confidence intervals for 95% confidence level

null_dist%>%get_ci(point_estimate = o_diff, level = .95, type = "se")

# the lower CI is -2.26 and the upper CI is -0.760, the point estimate is -1.51
```