Inference for numerical data

Your reproducible lab report: Before you get started, download the R Markdown template for this lab. Remember all of your code and answers go in this document:

```
download.file("https://dyurovsky.github.io/85309/post/rmd/lab8.Rmd",
              destfile = "lab8.Rmd")
```

American salaries

Since 2005, the American Community Survey polls ~$3.5 million households yearly. We will work with a randon sample of 2000 observations from the 2012 ACS.

The data

As always, let's get the data from the course website

```
acs <- read_csv("https://dyurovsky.github.io/85309/data/lab8/acs.csv")
```

Below is the *codebook* for this dataset:

- `income` : Yearly income (wages and salaries)
- `employment` : Employment status, not in labor force, unemployed, or employed
- `hrs_work` : Weekly hours worked
- `race` : Race, White, Black, Asian, or other
- `age` : Age
- `gender` : gender, male or female
- `citizens` : Whether respondent is a US citizen or not
- `time_to_work` : Travel time to work
- `lang` : Language spoken at home, English or other
- `married` : Whether respondent is married or not
- `edu` : Education level, hs or lower, college, or grad
- `disability` : Whether respondent is disabled or not
- `birth_qrtr` : Quarter in which respondent is born, jan thru mar, apr thru jun, jul thru sep, or oct thru dec

Note that this dataset contains some people who are not in the labor force or not employed. First, let's subset the dataset for those who are employed. We will call this new dataset `acs_emp`, short for "employed". Remember that we use the `filter` function for subsetting the data based on attributes stored in a variable.

```
acs_emp <- acs %>%
  filter(employment == "employed")
```

Exercise 1 What percent of the original sample (`acs`) are employed?

Next let's take a look at the income distribution by gender. The first step would be to create a visualization:

```
ggplot(acs_emp, aes(x = gender, y = income)) +
  geom_boxplot()
```

We can also obtain summary statistics such as means and standard deviations and sample sizes.

```
acs_emp %>%
  group_by(gender) %>%
  summarise(xbar = mean(income),
            s = sd(income),
            n = n())
```

Exercise 2 At a first glance how do the average incomes of males and females compare? Make sure to include the visualization and the summary statistics in your answer, and discuss/interpret them.

We can use R's `t.test` function to make statistical inferences about differences like this using the t-distribution.

Exercise 3 Use `t.test` to find the 95% confidence interval for the difference between the average incomes of males and females using, and interpret this interval. HINT: You might want to make two separate tibbles using the `filter` function–one that has just the males and one that has just the females

Exercise 4 Based on this interval is there a statistically significant difference between the average incomes of men and women? Why, or why not?

Confounding variables

There is a clear difference between the average salaries of men and women, but could some, or all, of this difference be attributed to a variable other than gender? Remember that we call such variables confounding variables. We will evaluate whether `hrs_work` is a confounder for the relationship between gender and income.

Let's start by just looking at the how many hours the average employee works.

```
ggplot(acs_emp, aes(x = hrs_work)) +
  geom_histogram(binwidth = 10)
```

Exercise 5 Well it sure looks like many of the employees work 40 hours. But not all. Describe the shape and spread of this distribution.

Exercise 6 Do we have a reason to think that the people in our sample come from a population of workers that isn't just full time employees? Perform a one-sample `t-test` to determine if the mean number of hours worked is different from what you would expect if everyone were a full time employee. *Hint:* Think about what your Null Hypothesis should be when you call `t.test` !

Ok, since it looks like we might have some part time workers, let's see if this matters for the conclusion we drew earlier. First convert the `hrs_work` variable to a categorical variable (with levels `"full time"` or `"part time"`) so that we can use methods we have learned so far in the course to run the analysis. (Later in the course we will learn how to work with numerical explanatory variables in a regression model setting.)

Recoding variables

We want to create a new variable, say `emp_type` , with levels `"full time"` or `"part time"` depending on whether the employee works 40 hours or more per week or less than 40 hours, respectively. Remember, we create a new variable with the mutate function.

```
acs_type <- acs_emp %>%
  mutate(emp_type = if_else(hrs_work >= 40, "full time", "part time"))
```

The `if_else()` function has three arguments: a logical test, return values for TRUE elements of test, and return values for FALSE elements of test. In this case, `emp_type` will be coded as `"full time"` for observations where `hrs_work` is greater than or equal to 40, and as `"part time"` otherwise.

To find out what percent of the sample is full vs. part time, we turn to summary statistics:

```
acs_type %>%
  group_by(emp_type) %>%
  summarise(total_type = n()) %>%
  ungroup() %>%
  mutate(prop_type = total_type/sum(total_type))
```

Here we first grouped the data by the new `emp_type` variable, and then we calculated proportions of full and part time employees by first counting how many there are in each group (`n()`), and then `ungroup` ing the data and dividing the total in each group by the total in all groups.

Exercise 7 Create a bar plot of the distribution of the `emp_type` variable, and also include the summary statistics you calculated above in your answer. What percent of the sample are full time and what percent are part time employees?

Exercise 8 Are women more heavily represented among full time employees or part time employees? Answer this question using summary statistics (code provided below) and a visualization.

```
acs_type %>%
  group_by(emp_type, gender) %>%
  summarise(n = n())
```

More Practice

Exercise 9 Create two subsets of the `acs_emp` dataset: one for full time employees and one for part time employees. No interpretation is needed for this question, just the code is sufficient.

Exercise 10 Use a hypothesis test to evaluate whether there is a difference in average incomes of **full time** male and female employees, and also include a confidence interval (at the equivalent confidence level) estimating the magnitude of the average income difference.

Exercise 11 Use a hypothesis test to evaluate whether there is a difference in average incomes of **part time** male and female employees, and also include a confidence interval (at the equivalent confidence level) estimating the magnitude of the average income difference.

Exercise 12 What do your findings from these hypothesis test suggest about whether or not working full or part time might be a confounding variable in the relationship between gender and income?

Exercise 13 Pick **another** numerical variable from the dataset to be your response variable, and also pick a categorical explanatory variable (can be one we used before). Conduct the appropriate hypothesis test to compare means of the response variable across two levels of the explanatory variable. Make sure to state your research question, and interpret your conclusion in context of the dataset. Note that you can use the complete `acs` dataset, the subsetted `acs_emp` dataset, or another subset that you create.