Lab 3: Inference for a Single Mean

Group member names here!

Reminders

- Bolded questions will be answered in the Canvas quiz (*Lab 3: One Mean Inference*) AS A GROUP. I will ask for group member names on the quiz.
- Your group only needs to submit ONE final pdf/html lab write up to canvas (*Lab 3: Completion*); **every** question included in the lab should be answered.

Getting set-up

We use three packages in this course: `tidyverse`, `mosaic`, and `infer`. To load a package, you use the `library()` function, wrapped around the name of a package. I've put the code to load one package into the chunk below.

1. Load in the other two packages you need.

Loading in data

As usual, we'll load in the data we are going to work with: `GSS_clean.csv`. The data should be inside the `data` folder in your RStudio Cloud. Just like before, we'll use the `read_csv()` function to read in the data. *Note: we named our data set `GSS`.*

```
GSS <- read_csv("data/gss.csv")
```

Investigating hours worked

For this lab, we'll be working with data from the General Social Survey (GSS). The GSS is a high-quality survey which gathers data on American society and opinions, conducted since 1972.This data set is a sample of 500 entries from the GSS, spanning years 1973-2018, including demographic markers and some economic variables.

Specifically, we will consider the variable `hours`. Say we wanted to know if the true mean number of hours worked in a week is different than 40. We could write out our null and alternative hypotheses like so:

$$H_o : \mu = 40$$

$$H_a : \mu \neq 40$$

2. In words and context of this problem, what does the parameter, μ, represent?

Then, we can find our point estimate (otherwise known as our sample statistic, or sample mean).

3. Use the favstats() function to obtain the point estimate for μ.

```
# your code goes here!
```

Okay, so that's not exactly 40. But, it would be nice to know what the distribution of hours worked looks like!

Visualizing hours worked

4. Fill in the blanks to create a histogram of the `hours` variable.

```
ggplot(data = GSS,
       mapping = aes(x = ____)) +
  geom_histogram(binwidth = ____) +
  labs(x = "____")
```

5. How would you describe the distribution of hours worked? *Keep in mind you should address: center, spread, shape, and outliers!*

Sampling distribution

When we're doing inference, we are interested in knowing what other statistics we might have gotten from other samples. If we were to record the mean hours worked for thousands of other samples of 500 respondents, we would be able to create the sampling distribution for the mean.

However, we only have **one** sample. So, we will need to approximate the sampling distribution. We have two options for how to do this:

1. using the t-distribution or
2. using a bootstrap distribution.

Today we will focus on using the t-distribution and save the bootstrap distribution for another day!

The t-distribution

We can use the (Student's) t-distribution to approximate the sampling distribution, but the accuracy of this approximation depends on two key conditions.

6. What are the two conditions we need to check so we know the t-distribution is a good approximation for the sampling distribution?

7. Are these conditions reasonable to assume? Why or why not?

Confidence interval for one mean

As I said in class, the fortunate aspect of using the t-distribution is that it makes calculations much easier. In *Activity 3: Sleepless Nights* we found our confidence interval by handing using:

- the point estimate (sample mean)
- the standard error
- a t* multiplier

8. How many degrees of freedom does the t-distribution we should use have?

Today, since we are programming in R, we will use a function to get this interval for us! The `t_test()` function is what we will use. I've filled in the necessary pieces for you, but I'll talk you through them.

- `x` tells `t_test()` what data set to use
- `response` tells `t_test()` what variable (from the data set) to look at
- `conf_int` tells `t_test()` YES (TRUE) I want a confidence interval
- `conf_level` tells `t_test()` what percentage confidence interval you want

Run the code below and see what you get!

```
t_test(x = GSS,
       response = hours,
       conf_int = TRUE,
       conf_level = 0.95)
```

```
## # A tibble: 1 x 7
##   statistic  t_df   p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>     <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1      62.4   499 5.95e-238 two.sided       41.4     40.1     42.7
```

9. Adjust the code chunk below to obtain a 90% confidence interval.

```
t_test(x = GSS,
       response = hours,
       conf_int = TRUE,
       conf_level = 0.90)
```

```
## # A tibble: 1 x 7
##   statistic  t_df   p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>     <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1      62.4   499 5.95e-238 two.sided       41.4     40.3     42.5
```

10. Interpret, in the context of these data, the 90% confidence interval you obtained in question 9.

Hypothesis test for one mean

Our goal in a hypothesis test is to decide if our sample provides sufficient evidence to support the claim that the mean hours worked per week in the population is different from 40.

This is a different goal from a confidence interval! A confidence interval provides a range of values for the population parameter, whereas a hypothesis test tells you if it is or is not a certain value.

11. Based on your 90% confidence interval do you believe you would reject or fail to reject the null hypothesis? *Hint: Where does the mean null value fall in relation to the confidence interval you found?*

Alright, let's see! We can use the `t_test()` to obtain the p-value for our hypothesis test. We can keep the code we had before, but we do need to give `t_test()` two more pieces of information:

- `alternative`: the direction of the alternative hypothesis (two-sided $\neq$, greater $>$, or less $<$)
- `mu`: the hypothesized null mean value

The code looks like this:

```
t_test(x = GSS,
       response = hours,
       conf_int = FALSE,
       conf_level = 0.90,
       alternative = "two-sided",
       mu = 40)
```

```
## # A tibble: 1 x 5
##   statistic  t_df p_value alternative estimate
##       <dbl> <dbl>   <dbl> <chr>          <dbl>
## 1      2.09   499  0.0376 two.sided       41.4
```

12. Based on the p-value what would you decide to do regarding your null hypothesis? Why?

13. Write a conclusion in context of the scenario.