Lab 6: Number of Hours Worked per Week

t-Test: Inference and Confidence Intervals for a Single Mean

Team Captain: NAME HERE   Facilitator: NAME HERE
Recorder: NAME HERE   Resource Manager: NAME HERE

Invalid Date

The General Social Survey (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.

Research question: Does the mean hours worked in a week for all Americans spanning the years 1973-2018 differ from 40 hours?

1. Read the `gss.csv` data set into R and view the top 6 rows.

```
GSS <- read_csv("data/________")
```

```
Error: 'data/________' does not exist in current working directory ('C:/Users/erobin17/OneDri
```

```
# put code to view top 6 rows here.
```

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

3. Use the `favstats()` function to obtain the observed mean, standard deviation, and sample size. Make sure to identify them with the appropriate symbols/notation.

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

$\bar{x} =$

$s =$

$n =$

4. Fill in the blanks to create a histogram of the `hours` variable. You will want to play around with the `binwidth` argument until you get a distribution that looks "connected".

```
ggplot(data = GSS,
       mapping = aes(x = ____)) +
  geom_histogram(binwidth = ____, color = "white") +
  labs(title = "_____",
       x = "____",
       y = "____"
       )
```

```
Error: <text>:2:27: unexpected input
1: ggplot(data = GSS,
2:        mapping = aes(x = __
                             ^
```

5. Copy and modify the code above to create a boxplot of the `hours` variable.

```
# copy code and modify for boxplot
```

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

> **Distribution of Sample Means (Null Distribution)**
>
> - Recall, 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 do this using the "short-cut" of the *Central Limit Theorem* which leads to the t-distribution.

6. Predict what you think the Distribution of Sample Means would look like:

- Shape:
- Mean:
- Standard Deviation (aka Standard Error):

7. Check the normality assumption in order to use the t-distribution to test the research question. Explain.
8. Set up your null and alternative hypotheses (in words) to test the research question.

9. Complete the code below to conduct a t-test for our hypothesis test.

```
t_test(x = GSS,
       response = _____,
       mu = ____,
       alternative = "_________"
       )
```

```
Error: <text>:2:20: unexpected input
1: t_test(x = GSS,
2:        response = __
                      ^
```

10. Using the output above, make a decision regarding your hypothesis test.
11. Write a conclusion in context of the problem. Make sure to state your observed t-test statistic, degrees of freedom, and p-value in your conclusion.

Note that by default, the output from the `t_test()` function gave you a 95% confidence interval.

12. Interpret the 95% confidence interval in context of the data.
13. Why does it make sense that 40 does not fall within the 95% confidence interval? *Hint: think about what your confidence level and α cutoff are. How do these relate?*
14. Copy the code from your `t_test()` and adjust it to obtain a 99% confidence interval. *Hint: look at the `conf_level` argument.*

```
# Copy from above and adjust code for 99% confidence interval here.
```

15. Interpret, in the context of these data, the 99% confidence interval.
16. What do you notice about the 99% confidence interval compared to the 95% confidence interval? Is it wider? narrower? the same?
17. Based on your 99% confidence interval do you believe you would have evidence (at an $\alpha = 0.01$ cutoff) to conclude the mean hours worked in a week for all Americans differs from 40 hours? *Hint: think about (1) where does 40 fall in regards to the confidence interval? (2) what is your p-value from Q9?*