Pre-Lab 2: Inference for One Mean

Professor I. Johnson

Goals for this Pre-Lab.

- To check Validity Conditions for Theory-Based methods for Inference with One Mean
- To apply Theory-Based methods for Inference with One Mean and draw appropriate conclusions.
- To apply calculation techniques using tools from Lab 1.
- To learn to display quantitative data in a histogram.

Setup and packages

As usual, we start by loading our two packages: `mosaic` and `ggformula`. 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. Add the other package you need.

```
        library(mosaic)
        library(ggformula)
```

Loading in data

We'll load the example data, `GSS22clean.csv`. It is available at this Url:

https://raw.githubusercontent.com/lJohnson-math/Math138/main/GSS22clean.csv (https://raw.githubusercontent.com/lJohnson-math/Math138/main/GSS22clean.csv)

We'll use the `read.csv()` function to read in the data.

```
        #load data
        GSS22 <- read.csv("https://raw.githubusercontent.com/IJohnson-math/Math138/ma
in/GSS22clean.csv")
```

This dataset comes from the 2022 General Social Survey (GSS), which is collected by NORC at the University of Chicago. It is a random sample of households from the United States, and has been running since 1972, so it is very useful for studying trends in American life. The data I've given you is a subset of the questions asked in the survey, and the data has been cleaned to make it easier to use. But, there are still some messy aspects (which we'll discover as we analyze it further throughout our class!).

Research question

Suppose we wanted to know for all U.S. workers if the mean number of hours worked in a week is different than 40. We could write our null and alternative hypotheses as

$$H_0 : \mu = 40$$

$$H_a : \mu \neq 40$$

Basic commands to view the data.

In Lab 1, we used the following commands to view parts of the GSS22 data: `glimpse`, `head`, `tail`. You can also view the data in another tab by clicking on 'GSS22' in the **Environment** pane. This second option allows you to scroll up and down, and left and right to view the data.

You may use these commands on the GSS22 data to review the data.

```
        head(GSS22)
        #glimpse(GSS22)
```

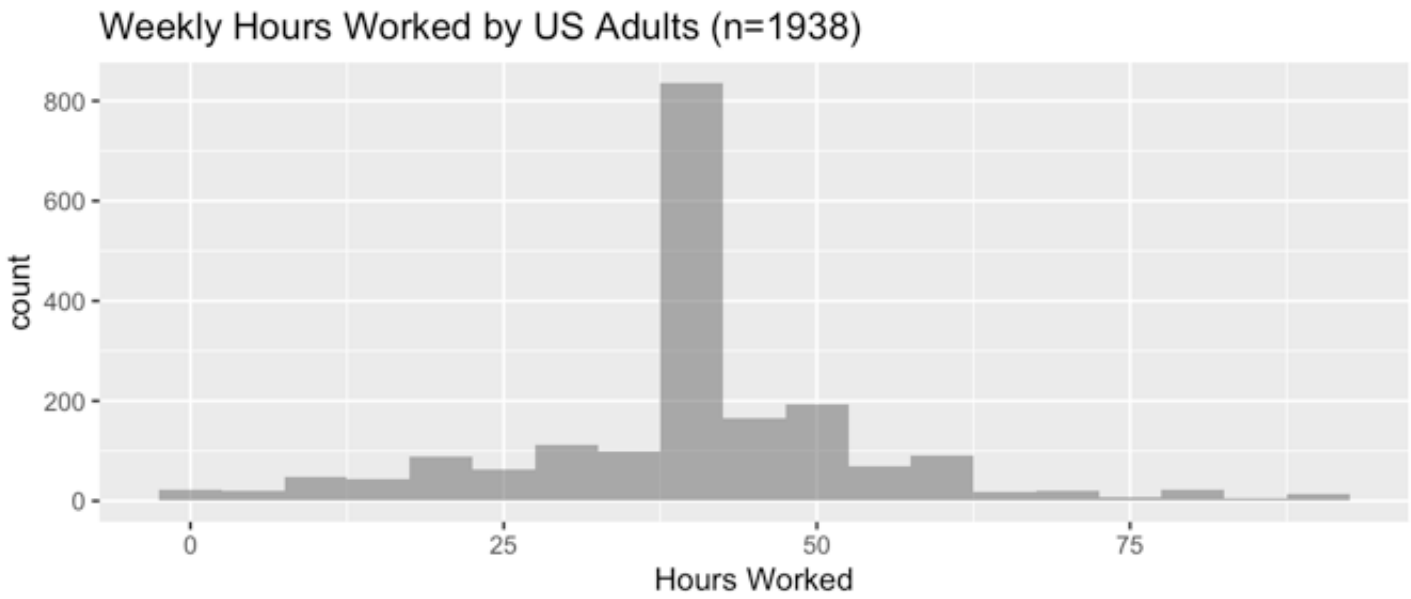
Inference for One Mean

In this lab we will consider the mean of the number of `hours_worked_last_week` as our statistic of interest. Looking within the GSS22 data we see many NA values for the variable `hours_worked_last_week`. Let's start by filtering out the NA values. The command `filter` is used to keep the observational units that satisfy a given property. In this example the property is `!is.na(hours_worked_last_week)`; here, the exclamation point, `!`, is read as "not", so this command keeps the observational units that do *not* have NA as an entry for the variable `hours_worked_last_week`.

```
        #NEW command to filter out NA values from data
        GSS22 <- filter(GSS22, !is.na(hours_worked_last_week))
```

Let's look at the data. Create a histogram of `hours_worked_last_week`. Add a title and units along the horizontal axis.

```
        #NEW command to create a histogram for a quantitative variable
        gf_histogram(~hours_worked_last_week, data=GSS22, binwidth = 5, title="Weekly
Hours Worked by US Adults (n=1938)",xlab="Hours Worked" )
```



Now we can compute and display the mean, the value of our point estimate. We name the statistic `xbar` (in place of the symbol $\bar{x}$).

```
        #NEW command for our favorite statistics
        favstats(~hours_worked_last_week, data=GSS22)
```

```
##  min Q1 median Q3 max     mean       sd    n missing
##    0 37     40 45  89 40.18163 14.10734 1938       0
```

```
        xbar <- mean(~hours_worked_last_week, data=GSS22)
        xbar
```

```
## [1] 40.18163
```

Validity Conditions for a One-sample t-test

The quantitative variable should have a symmetric distribution, or you should have at least 20 observations and the sample distribution should not be strongly skewed.

When these conditions are met we can use the t-distribution to approximate the p-value for our hypothesis test. It's important to keep in mind that these conditions are rough guidelines and not a guarantee. All theory-based methods are approximations. They will work best when the sample distribution is symmetric, the sample size is large, and there are no large outliers. When in doubt, use a simulation-based method as a cross-check.

Check Validity Conditions: In this example we have $n = 1938$ observations, which is much larger than 20, and our sample distribution is symmetric as seen above in the histogram. Thus the validity conditions for theory-based inference with one mean are satisfied.

Calculating the standardized statistic, the t-statistic

The standardized statistic, t, is found using the formula

$$t = \frac{\bar{x} - \mu}{SE(\bar{x})}$$

and standard error for the null distribution is given by

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}.$$

Calculate and display the standardized t- statistic

```
        #calculate the standard deviation of the sample, s
        s <- sd(~hours_worked_last_week, data=GSS22)
        s
```

```
## [1] 14.10734
```

```
        # n is the number of observational units (after filtering)
        n <- 1938

        #calculate standard error
        SE <- s/sqrt(n)
        SE
```

```
## [1] 0.3204559
```

```
        #mu is the mean of the null distribution
        mu <- 40

        #now we can calculate (and display) the standardized statistic
        t <- (xbar - mu)/SE
        t
```

```
## [1] 0.566788
```

We use the command `t.test( )` to calculate a p-value. We must input the variable, the data, the alternative hypothesis, and the mean of the null distribution into `t.test( )`. As we saw in Lab 1, the options for alternative are “two.sided”, “greater”, “less” depending on the inequality in the alternative hypotheses. We must also enter the null-hypothesis parameter as `mu = 40`.

```
        t.test(~hours_worked_last_week, data=GSS22, alternative="two.sided", mu=40)
```

```
##
##  One Sample t-test
##
## data:  hours_worked_last_week
## t = 0.56679, df = 1937, p-value = 0.5709
## alternative hypothesis: true mean is not equal to 40
## 95 percent confidence interval:
##  39.55316 40.81011
## sample estimates:
## mean of x
##  40.18163
```

Conclusions

Our data is from a random sample of n=1938 US workers collected through the General Social Survey. We consider the number of hours worked last week, a quantitative variable, and investigated whether or not the mean number of hours worked last week by US workers is equal to 40. Since our sample is random, if our findings are significant we may generalize our conclusions the larger population of US workers.

What can be concluded from the t-statistic and p-value?

Our statistic, the sample mean of $\bar{x} = 40.18$ hours worked last week, is 0.566 of a standard deviation above the hypothesized mean of 40 hours worked last week. A t-statistic between -1 to 1 standard deviations away from the mean of the null is weak evidence against the null hypothesis. From our random sample of n=1938 households, we obtained a statistic that is typical if the average number hours worked last week is indeed 40 hours. Therefore will accept the null as plausible, since our evidence agrees with the supposition that people in the US work an average of 40 hours per week.

Similarly, the p-value of 0.5709 is very large. When a p-value is greater than 0.05, as ours is here, we have no evidence against the null hypothesis. Thus, we will accept the null hypothesis as plausible. The mean number of hours worked by US workers is quite plausibly equal to 40 hours per week.

Notice that the t-statistic and p-value give the same conclusions, as expected. Both imply that the null-hypothesis is plausible, so we cannot reject the null. Thus we conclude that the US population of workers, in 2022, is likely to have worked about 40 hours per week on average.

Instructions to Print Your Lab!

- Knit to html and you should see the html file in the **Files** pane.
- Open the .html file in a Web Browser window.
- Print from the browser, BUT FIRST select the layout as 2 Pages per sheet. You may also print double-sided to save paper.
- Bring your printed file to class to turn it in. It is due one week after the Pre-Lab is completed in class.