Problem Set 2 Answer Key

ML

Set up

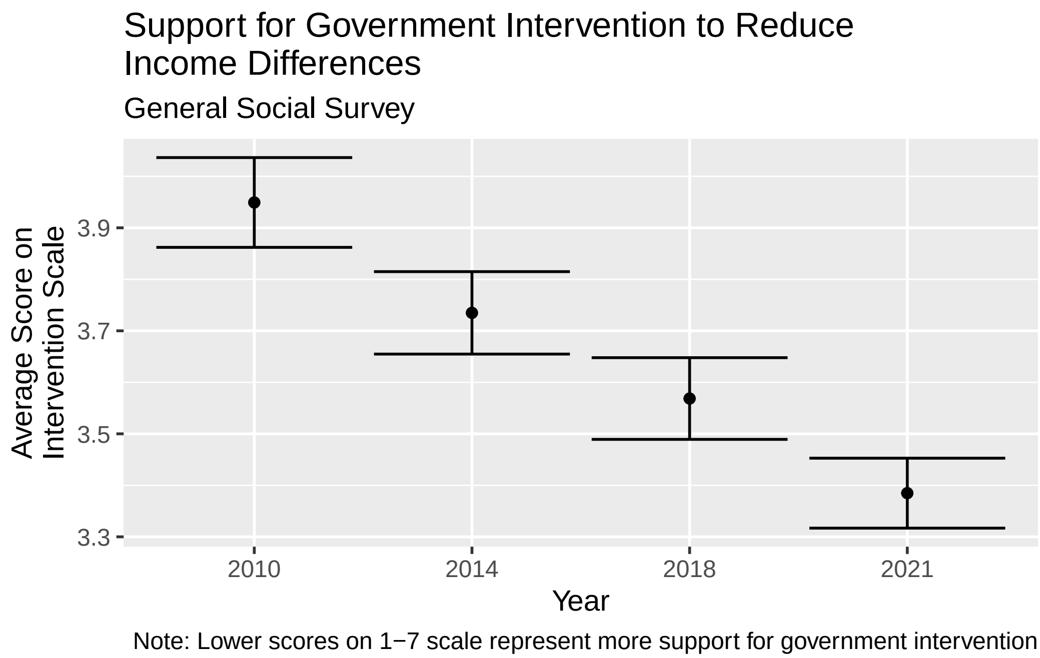
```
library(tidyverse)
ps2 <- read_csv("https://raw.githubusercontent.com/mjclawrence/soci385_f23/main/data/ps2.c
```

1. Without using any R shortcuts, find the 95% confidence interval for the mean of `eqwlth` in each of the years in the survey. Plot these intervals in a figure (with error bars), and use your figure to describe how the mean responses have changed over the survey years.

We did the basic set up of this one together in class. You only needed to add labels to the axes and the plot (the `labs()` function is the easiest way to do this). As for interpretation, recall what the confidence intervals represent: if the error bars do not overlap across years, then the means are significantly different.

```
ps2 |>
  filter(year %in% c(2010, 2014, 2018, 2021)) |>
  group_by(year) |>
  summarize(mean_eqwlth = mean(eqwlth, na.rm = TRUE),
            sd = sd(eqwlth, na.rm = TRUE),
            n = length(eqwlth),
            se = sd / sqrt(n),
            ll = mean_eqwlth - 1.96*se,
            ul = mean_eqwlth + 1.96*se) |>
  ggplot(aes(x = as.factor(year), y = mean_eqwlth,
             ymin = ll, ymax = ul)) +
  geom_point() + geom_errorbar() +
  labs(x = "Year",
       y = "Average Score on\nIntervention Scale",
       title = "Support for Government Intervention to Reduce\nIncome Differences",
```

```
        subtitle = "General Social Survey",
        caption = "Note: Lower scores on 1-7 scale represent more support for government in
```



2. Create a new variable grouping the `age` variable into the following categories: 18-24, 25-39, 40-54, 55-64, 65+. Which (if any) age categories showed significant differences in mean `eqwlth` scores between the 2018 and 2021 surveys? What is a sociological explanation for these differences?

Create age categories. Note that you want to use a new variable name so that you do not replace the existing values of the age variable.

```
ps2 <- ps2 |>
  mutate(age_cat = ifelse(age %in% c(18:24), "18-24",
                          ifelse(age %in% c(25:39), "25-39",
                                 ifelse(age %in% c(40:54), "40-54",
                                        ifelse(age %in% c(55:64), "55-64",
                                               "65+")))))
```

Want to check your work? Make a table with the old age variable and the new age_cat variable:

```r
table(ps2$age, ps2$age_cat)
```

```
   18-24 25-39 40-54 55-64 65+
18    42     0     0     0   0
19    89     0     0     0   0
20    83     0     0     0   0
21   110     0     0     0   0
22   118     0     0     0   0
23   127     0     0     0   0
24   138     0     0     0   0
25     0   185     0     0   0
26     0   151     0     0   0
27     0   157     0     0   0
28     0   165     0     0   0
29     0   202     0     0   0
30     0   208     0     0   0
31     0   181     0     0   0
32     0   190     0     0   0
33     0   207     0     0   0
34     0   212     0     0   0
35     0   187     0     0   0
36     0   171     0     0   0
37     0   210     0     0   0
38     0   175     0     0   0
39     0   207     0     0   0
40     0     0   186     0   0
41     0     0   208     0   0
42     0     0   186     0   0
43     0     0   191     0   0
44     0     0   178     0   0
45     0     0   160     0   0
46     0     0   177     0   0
47     0     0   156     0   0
48     0     0   139     0   0
49     0     0   169     0   0
50     0     0   185     0   0
51     0     0   194     0   0
52     0     0   156     0   0
53     0     0   210     0   0
54     0     0   181     0   0
55     0     0     0   197   0
```

```
56      0      0      0   213    0
57      0      0      0   192    0
58      0      0      0   195    0
59      0      0      0   215    0
60      0      0      0   206    0
61      0      0      0   180    0
62      0      0      0   184    0
63      0      0      0   191    0
64      0      0      0   158    0
65      0      0      0     0  175
66      0      0      0     0  149
67      0      0      0     0  194
68      0      0      0     0  159
69      0      0      0     0  163
70      0      0      0     0  172
71      0      0      0     0  131
72      0      0      0     0  112
73      0      0      0     0   98
74      0      0      0     0  138
75      0      0      0     0  116
76      0      0      0     0   94
77      0      0      0     0   96
78      0      0      0     0   79
79      0      0      0     0   86
80      0      0      0     0   75
81      0      0      0     0   64
82      0      0      0     0   51
83      0      0      0     0   47
84      0      0      0     0   56
85      0      0      0     0   40
86      0      0      0     0   44
87      0      0      0     0   22
88      0      0      0     0   33
89      0      0      0     0   94
```

We are testing differences in means, so use a t test. The most efficient way to do this is with a separate test for each age category. You are testing the mean of `eqwlth` so that should be the first variable included in the `t.test()` function. Brackets do the rest of the work here: each t.test should include a separate age category, and each line should include a different year (either 2018 or 2021).

```r
t.test(ps2$eqwlth[ps2$age_cat=="18-24" & ps2$year==2018],
       ps2$eqwlth[ps2$age_cat=="18-24" & ps2$year==2021])
```

```
    Welch Two Sample t-test

data:  ps2$eqwlth[ps2$age_cat == "18-24" & ps2$year == 2018] and ps2$eqwlth[ps2$age_cat == "1
t = 2.2681, df = 229.98, p-value = 0.02425
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0670278 0.9539347
sample estimates:
mean of x mean of y
 3.294964  2.784483
```

The difference for 18-24 year olds is significant since we can reject the null hypothesis. The test statistic of 2.2681 is more extreme than 1.96, the p-value of 0.025 is less than 0.05, and the confidence interval does not include the null hypothesis value of 0.

```r
t.test(ps2$eqwlth[ps2$age_cat=="25-39" & ps2$year==2018],
       ps2$eqwlth[ps2$age_cat=="25-39" & ps2$year==2021])
```

```
    Welch Two Sample t-test

data:  ps2$eqwlth[ps2$age_cat == "25-39" & ps2$year == 2018] and ps2$eqwlth[ps2$age_cat == "2
t = 2.8156, df = 954.91, p-value = 0.00497
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1053074 0.5897984
sample estimates:
mean of x mean of y
 3.252315  2.904762
```

The difference for 25-39 year olds is significant since we can reject the null hypothesis. The test statistic of 2.8156 is more extreme than 1.96, the p-value of 0.005 is less than 0.05, and the confidence interval does not include the null hypothesis value of 0.

```r
t.test(ps2$eqwlth[ps2$age_cat=="40-54" & ps2$year==2018],
       ps2$eqwlth[ps2$age_cat=="40-54" & ps2$year==2021])
```

```
	Welch Two Sample t-test

data:  ps2$eqwlth[ps2$age_cat == "40-54" & ps2$year == 2018] and ps2$eqwlth[ps2$age_cat == "4
t = 1.8542, df = 796.46, p-value = 0.06408
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01497075  0.52536310
sample estimates:
mean of x mean of y
 3.584527  3.329331
```

The difference for 40-54 year olds is not significant since we cannot reject the null hypothesis. The test statistic of 1.8542 is not more extreme than 1.96, the p-value of 0.06408 is greater than 0.05, and the confidence interval does include the null hypothesis value of 0.

```
t.test(ps2$eqwlth[ps2$age_cat=="55-64" & ps2$year==2018],
       ps2$eqwlth[ps2$age_cat=="55-64" & ps2$year==2021])
```

```
	Welch Two Sample t-test

data:  ps2$eqwlth[ps2$age_cat == "55-64" & ps2$year == 2018] and ps2$eqwlth[ps2$age_cat == "5
t = 1.8845, df = 620.06, p-value = 0.05997
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01278959  0.62051612
sample estimates:
mean of x mean of y
 3.879433  3.575569
```

The difference for 55-64 year olds is not significant since we cannot reject the null hypothesis. The test statistic of 1.8845 is not more extreme than 1.96, the p-value of 0.05997 is greater than 0.05, and the confidence interval does include the null hypothesis value of 0.

```
t.test(ps2$eqwlth[ps2$age_cat=="65+" & ps2$year==2018],
       ps2$eqwlth[ps2$age_cat=="65+" & ps2$year==2021])
```

```
	Welch Two Sample t-test
```

```
data:  ps2$eqwlth[ps2$age_cat == "65+" & ps2$year == 2018] and ps2$eqwlth[ps2$age_cat == "65
t = 0.73563, df = 690.57, p-value = 0.4622
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1625736  0.3573869
sample estimates:
mean of x mean of y
 3.810089  3.712682
```

The difference for 65+ year olds is not significant since we cannot reject the null hypothesis. The test statistic of 0.73563 is not more extreme than 1.96, the p-value of 0.4622 is greater than 0.05, and the confidence interval does include the null hypothesis value of 0.

3. Does the proportion of respondents with "Hardly Any" confidence in congress differ between respondents at the lowest and highest extremes of the `eqwlth` scale? What is an additional variable you would want to explain your result in more detail?

There are two things to do to set up this one. First, create a binary variable distinguishing `conlegis` responses of "Hardly Any" from all other responses:

```
ps2 <- ps2 |>
  mutate(hardly_any = ifelse(conlegis == "Hardly Any", 1, 0))
```

Second, filter to only include responses with a 1 or 7 on the `eqwlth` scale. It makes sense to create a new data frame here so that you can use the full dataset for other questions.

```
q3 <- ps2 |>
  filter(eqwlth == 1 | eqwlth == 7)
```

We are testing differences in proportions for this question, so eventually we will use the `prop.test()` function. That function requires a table, so let's create it. Note here that since confidence in congress is the outcome, the `hardly_any` variable should be the column variable for this table, leaving support for government intervention as the row variable.

```
q3_table <- table(q3$eqwlth, q3$hardly_any)
```

With the table saved as an object, feed its name into `prop.test()`:

```
prop.test(q3_table)
```

```
    2-sample test for equality of proportions with continuity correction

data:  q3_table
X-squared = 122.95, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1760751 0.2488461
sample estimates:
   prop 1    prop 2 
0.4969596 0.2844991 
```

We can reject the null hypothesis here. The p-value is less than 0.05 and the confidence interval does not include the null hypothesis value of 0. For those reasons, we can conclude that the difference in proportions is significant.

4. summarize the results of the following three tests of association. In addition to offering sociological interpretations of your findings, describe why you chose which statistical tests to use.

- Is there a significant association between `racehisp` and `eqwlth`?

A chi-squared test works for this question since only one of the variables (`eqwlth`) is ordered.

```
chisq.test(ps2$racehisp, ps2$eqwlth)
```

```
    Pearson's Chi-squared test

data:  ps2$racehisp and ps2$eqwlth
X-squared = 253.57, df = 18, p-value < 2.2e-16
```

We can reject the null hypothesis because the p-value is less than 0.05. There is a significant association between race and support for government intervention to reduce income differences.

- Among respondents with less than a high school diploma, is there a significant association between `racehisp` and `eqwlth`?

We know from the question above that a chi-squared test could work for these variables. Let's try it:

```r
chisq.test(ps2$racehisp[ps2$degree=="Less Than High School"],
           ps2$eqwlth[ps2$degree=="Less Than High School"])
```

```
Warning in chisq.test(ps2$racehisp[ps2$degree == "Less Than High School"], :
Chi-squared approximation may be incorrect


    Pearson's Chi-squared test

data:  ps2$racehisp[ps2$degree == "Less Than High School"] and ps2$eqwlth[ps2$degree == "Les
X-squared = 26.503, df = 18, p-value = 0.08879
```

The "Warning: Chi-squared approximation may be incorrect" is alerting us to the fact that some cells in the table of these variables have small expected frequencies. Let's check them.

```r
chisq.test(ps2$racehisp[ps2$degree=="Less Than High School"],
           ps2$eqwlth[ps2$degree=="Less Than High School"])$expected
```

```
Warning in chisq.test(ps2$racehisp[ps2$degree == "Less Than High School"], :
Chi-squared approximation may be incorrect


                  1          2         3         4         5          6
  Black    34.838608  6.5522152 12.465190 20.615506 10.547468  4.9541139
  Hispanic 85.199367 16.0237342 30.484177 50.416139 25.794304 12.1155063
  Other     5.174051  0.9731013  1.851266  3.061709  1.566456  0.7357595
  White    92.787975 17.4509494 33.199367 54.906646 28.091772 13.1946203

                  7
  Black    11.026899
  Hispanic 26.966772
  Other     1.637658
  White    29.368671
```

Since several cells have expected frequencies below five, we need to use Fisher's Test instead of a regular chi-squared test:

```r
fisher.test(ps2$racehisp[ps2$degree=="Less Than High School"],
            ps2$eqwlth[ps2$degree=="Less Than High School"],
```

```
                simulate.p.value = TRUE)
```

```
    Fisher's Exact Test for Count Data with simulated p-value (based on
    2000 replicates)

data:  ps2$racehisp[ps2$degree == "Less Than High School"] and ps2$eqwlth[ps2$degree == "Les
p-value = 0.07746
alternative hypothesis: two.sided
```

Note: you probably have a slightly different p-value since we are using different simulations. But the p-value should still be greater than 0.05. For that reason, we cannot reject the null hypothesis.

- Is there a significant association between age (using the categories you created in #2) and confidence in congress?

Both variables are ordered here so we will want to use the Goodman Kruskal Gamma test. Before doing so, put the `conlegis` values in order from least to most confidence. The age categories are already ordered; we would also need to assert the order for those values if they were not.

```
ps2 <- ps2 |>
  mutate(conlegis = factor(conlegis,
                           levels = c("Hardly Any",
                                      "Only Some",
                                      "A Great Deal")))
```

The GKgamma test is in the vcdExtra package. Load it:

```
library(vcdExtra)
```

The `GKgamma()` function reads a saved table:

```
q4_table <- table(ps2$age_cat,
                  ps2$conlegis)

GKgamma(q4_table)
```

```
gamma         : -0.199
std. error    : 0.015
CI            : -0.229 -0.169
```

The negative gamma value tells us that the association is negative. That means that higher values of age tend to have lower values on the conlegis scale. In more substantive terms, that means that older respondents tend to have less confidence in congress.

To find out if that negative association is significant, calculate the test statistic by dividing gamma by its standard error:

```
-.199/.015
```

```
[1] -13.26667
```

This is the value we can compare to -1.96. Since our test statistic of -13.267 is more extreme than -1.96, we can reject the null hypothesis. The negative association between age and confidence in congress is significant.