Problem Set 2 Solutions

1. Without using any R shortcuts, find the 95% confidence interval for the mean of `eqwlth` in each of the following years: 2010, 2014, and 2018. 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.

Students will probably do this separately for observations from each year.

For example, from 2010:

```
year_2010 <- filter(ps2, year == 2010)
mean_2010 <- mean(year_2010$eqwlth)
se_2010 <- sd(year_2010$eqwlth) /
  sqrt(length(year_2010$eqwlth))
ll_2010 <- mean_2010 - 1.96*se_2010
ul_2010 <- mean_2010 + 1.96*se_2010
ci_2010 <- c(ll_2010, mean_2010, ul_2010)
ci_2010
```

```
## [1] 3.835497 3.942435 4.049374
```

And from 2014:

```
year_2014 <- filter(ps2, year == 2014)
mean_2014 <- mean(year_2014$eqwlth)
se_2014 <- sd(year_2014$eqwlth) /
  sqrt(length(year_2014$eqwlth))
ll_2014 <- mean_2014 - 1.96*se_2014
ul_2014 <- mean_2014 + 1.96*se_2014
ci_2014 <- c(ll_2014, mean_2014, ul_2014)
ci_2014
```

```
## [1] 3.635316 3.734094 3.832871
```

And from 2018:

```
year_2018 <- filter(ps2, year == 2018)
mean_2018 <- mean(year_2018$eqwlth)
se_2018 <- sd(year_2018$eqwlth) /
  sqrt(length(year_2018$eqwlth))
ll_2018 <- mean_2018 - 1.96*se_2018
ul_2018 <- mean_2018 + 1.96*se_2018
ci_2018 <- c(ll_2018, mean_2018, ul_2018)
ci_2018
```

```
## [1] 3.466194 3.564421 3.662648
```

Once you have a confidence interval for each year, combine them into a table:

```
ci_years <- round(rbind(ci_2010, ci_2014, ci_2018),3)
ci_years
```

```
##          [,1]  [,2]  [,3]
## ci_2010 3.835 3.942 4.049
## ci_2014 3.635 3.734 3.833
## ci_2018 3.466 3.564 3.663
```

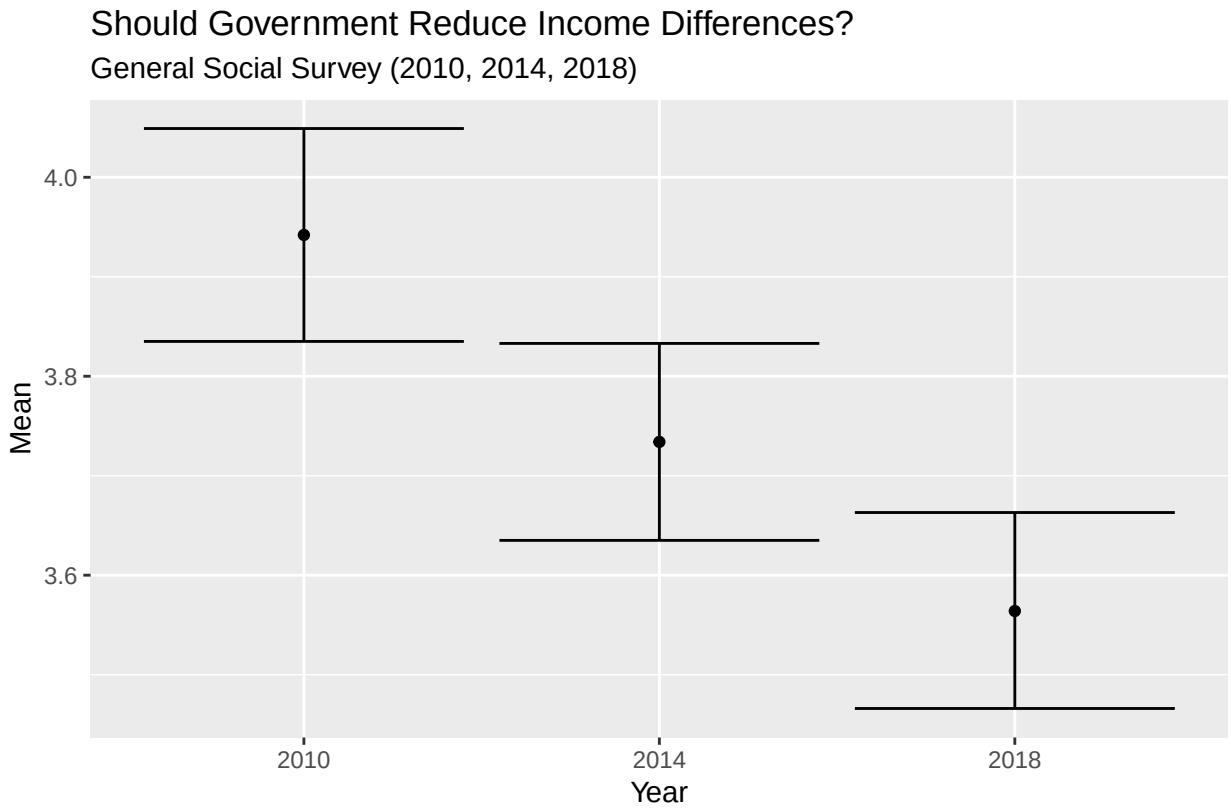
Then clean up the table

```
ci_years_rownames <- as.factor(c(2010, 2014, 2018))
ci_years_df <- cbind.data.frame(ci_years_rownames, ci_years)
colnames(ci_years_df) <- c("Year", "LL", "Mean", "UL")
ci_years_df
```

```
##         Year    LL  Mean    UL
## ci_2010 2010 3.835 3.942 4.049
## ci_2014 2014 3.635 3.734 3.833
## ci_2018 2018 3.466 3.564 3.663
```

And use the new data frame for your plot

```
years_plot <- ggplot(ci_years_df, aes(x = Year, y = Mean,
                                      ymin = LL, ymax = UL))
years_plot + geom_point() + geom_errorbar() +
     labs(title = "Should Government Reduce Income Differences?",
          subtitle = "General Social Survey (2010, 2014, 2018)",
          caption = "Note: Lower values indicate more support for reducing income differences") + # not
  theme(plot.caption = element_text(hjust = 0))
```



Here's a shortcut using group_by and summarize:

```
q1 <- ps2 |>
  filter(year == 2010 | year == 2014 | year == 2018) |>
  group_by(year) %>%
     summarize(mean = mean(eqwlth),
          sd = sd(eqwlth),
          n = length(eqwlth),
          se = sd / sqrt(n),
          ll = mean - 1.96*se,
          ul = mean + 1.96*se)
q1
```

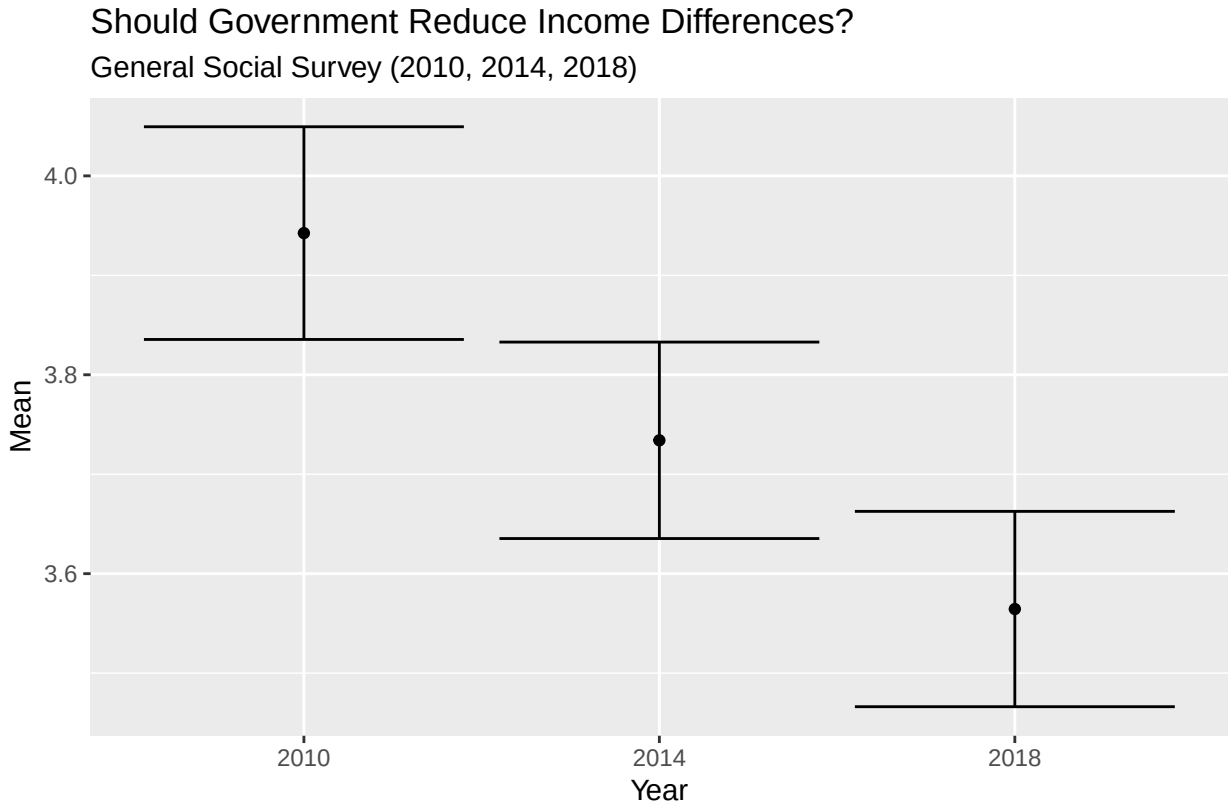
```
## # A tibble: 3 x 7
##    year  mean    sd     n     se    ll    ul
##   <int> <dbl> <dbl> <int>  <dbl> <dbl> <dbl>
## 1  2010  3.94  2.01  1355 0.0546  3.84  4.05
## 2  2014  3.73  2.06  1666 0.0504  3.64  3.83
## 3  2018  3.56  1.96  1529 0.0501  3.47  3.66
```

And the plot (note that with this approach you need to assert the `year` variable as a factor variable):

```
plot_q1 <- ggplot(q1, aes(x = factor(year), y = mean,
                          ymin = ll, ymax = ul))

plot_q1 + geom_point() + geom_errorbar() +
```

```
     labs(x = "Year", y = "Mean",
          title = "Should Government Reduce Income Differences?",
          subtitle = "General Social Survey (2010, 2014, 2018)",
          caption = "Note: Lower values indicate more support for reducing income differences") + theme
```



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 2010 and 2018 surveys?

New variable using mutate and ifelse:

```
ps2 <- ps2 %>%
     mutate(agecat =
          ifelse(age %in% 18:24, 1,
             ifelse(age %in% 25:39, 2,
               ifelse(age %in% 40:54, 3,
                 ifelse(age %in% 55:64, 4, 5)))))
```

Tests for each age category:

```
t.test(ps2$eqwlth[ps2$agecat==1&ps2$year==2010],
       ps2$eqwlth[ps2$agecat==1&ps2$year==2018])
```

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

```
## data:  ps2$eqwlth[ps2$agecat == 1 & ps2$year == 2010] and ps2$eqwlth[ps2$agecat == 1 & ps2$year == 20
## t = 1.2668, df = 259.45, p-value = 0.2064
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1487366  0.6853056
## sample estimates:
## mean of x mean of y
##  3.558140  3.289855
```

```
t.test(ps2$eqwlth[ps2$agecat==2&ps2$year==2010],
       ps2$eqwlth[ps2$agecat==2&ps2$year==2018])
```

```
##
##  Welch Two Sample t-test
##
## data:  ps2$eqwlth[ps2$agecat == 2 & ps2$year == 2010] and ps2$eqwlth[ps2$agecat == 2 & ps2$year == 20
## t = 2.7597, df = 763.67, p-value = 0.005924
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1080838 0.6407684
## sample estimates:
## mean of x mean of y
##  3.626741  3.252315
```

```
t.test(ps2$eqwlth[ps2$agecat==3&ps2$year==2010],
       ps2$eqwlth[ps2$agecat==3&ps2$year==2018])
```

```
##
##  Welch Two Sample t-test
##
## data:  ps2$eqwlth[ps2$agecat == 3 & ps2$year == 2010] and ps2$eqwlth[ps2$agecat == 3 & ps2$year == 20
## t = 2.928, df = 719.81, p-value = 0.003519
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1424466 0.7222289
## sample estimates:
## mean of x mean of y
##  4.013263  3.580925
```

```
t.test(ps2$eqwlth[ps2$agecat==4&ps2$year==2010],
       ps2$eqwlth[ps2$agecat==4&ps2$year==2018])
```

```
##
##  Welch Two Sample t-test
##
## data:  ps2$eqwlth[ps2$agecat == 4 & ps2$year == 2010] and ps2$eqwlth[ps2$agecat == 4 & ps2$year == 20
## t = 1.4141, df = 491.63, p-value = 0.158
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1017774  0.6244483
## sample estimates:
## mean of x mean of y
##  4.143478  3.882143
```

```
t.test(ps2$eqwlth[ps2$agecat==5&ps2$year==2010],
       ps2$eqwlth[ps2$agecat==5&ps2$year==2018])
```

```
##
##  Welch Two Sample t-test
##
## data:  ps2$eqwlth[ps2$agecat == 5 & ps2$year == 2010] and ps2$eqwlth[ps2$agecat == 5 & ps2$year == 20
## t = 2.8857, df = 541.27, p-value = 0.004062
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1563343 0.8229912
## sample estimates:
## mean of x mean of y
##  4.288462  3.798799
```

Can see there are significant differences in means for groups 2, 3, and 5: p-values are less than .05, test statistics are more extreme than 1.96, and zero is not in the confidence intervals.

We could also do this much more efficiently using group_by and summarise_each (but we have not seen how to do so in class so no expectation that you know how to do this).

```
multiple_ttests <- ps2 |>
  group_by(agecat) |>
  filter(year == 2010 | year == 2018) %>%
  summarise_each(list(~t.test(.[year == 2010], .[year == 2018])$statistic,
                      ~t.test(.[year == 2010], .[year == 2018])$p.value,
                      ~t.test(.[year == 2010], .[year == 2018])$estimate[[1]],
                      ~t.test(.[year == 2010], .[year == 2018])$estimate[[2]],
                      ~t.test(.[year == 2010], .[year == 2018])$conf.int[[1]],
                      ~t.test(.[year == 2010], .[year == 2018])$conf.int[[2]]),
                 vars = eqwlth) |>
  mutate(difference = `vars_[[..3` - `vars_[[..4`) |>
  select(-c(4:5))
```

```
## Warning: ‘summarise_each_()‘ was deprecated in dplyr 0.7.0.
## Please use ‘across()‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.
```

```
round(multiple_ttests,3)
```

```
## # A tibble: 5 x 6
##   agecat ‘vars_$..1‘ ‘vars_$..2‘ ‘vars_[[..5‘ ‘vars_[[..6‘ difference
##    <dbl>       <dbl>       <dbl>        <dbl>        <dbl>      <dbl>
## 1      1        1.27       0.206       -0.149        0.685      0.268
## 2      2        2.76       0.006        0.108        0.641      0.374
## 3      3        2.93       0.004        0.142        0.722      0.432
## 4      4        1.41       0.158       -0.102        0.624      0.261
## 5      5        2.89       0.004        0.156        0.823      0.49
```

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?

```r
# Create a binary variable
q3 <- ps2 |>
  mutate(conf_hi = ifelse(conlegis == "Hardly any", 1, 0)) |>
  filter(eqwlth==1 | eqwlth==7)

# Create a table
q3_table <- table(q3$eqwlth, q3$conf_hi)
round(prop.table(q3_table,1),3)
```

```
##
##         0     1
##   1 0.449 0.551
##   7 0.357 0.643
```

```r
# Run prop.test on the table
prop.test(q3_table)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  q3_table
## X-squared = 22.042, df = 1, p-value = 2.668e-06
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.0536743 0.1304696
## sample estimates:
##    prop 1    prop 2
## 0.4488704 0.3567985
```

Yes, the proportions are significantly different. We can reject the null hypothesis that the difference is zero because the p-value is less than .05 and zero is not in the confidence interval. Don't look at the x-squared statistic for the proportion test.

4a. Is there an association between racehisp and eqwlth?

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

```
##
##  Pearson's Chi-squared test
##
## data:  ps2$racehisp and ps2$eqwlth
## X-squared = 349.55, df = 18, p-value < 2.2e-16
```

Use a chi-squared test here because both variables are categorical but at least one (racehisp) is not ordered. Yes, there is a significant association because we can reject the null hypothesis since the p-value is less than .05. That means the two variables are dependent.

4b. Among respondents with less than a high school diploma, is there an association between race/Hispanic status and eqwlth?

```
q4b <- filter(ps2, degree=="Less than HS")
chisq.test(q4b$racehisp, q4b$eqwlth)
```

```
## Warning in chisq.test(q4b$racehisp, q4b$eqwlth): Chi-squared approximation may
## be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  q4b$racehisp and q4b$eqwlth
## X-squared = 23.22, df = 18, p-value = 0.1823
```

We get an error message, so let's see if the expected frequency in each cell is at least 5.

```
chisq.test(q4b$racehisp, q4b$eqwlth)$expected
```

```
## Warning in chisq.test(q4b$racehisp, q4b$eqwlth): Chi-squared approximation may
## be incorrect
```

```
##             q4b$eqwlth
## q4b$racehisp         1         2        3         4        5         6
##     Black     59.01484 12.708210 21.41246  38.64688 17.23442  7.311573
##     Hispanic 115.34718 24.838773 41.85163  75.53709 33.68546 14.290801
##     Other     11.73591  2.527201  4.25816   7.68546  3.42730  1.454006
##     White    152.90208 32.925816 55.47774 100.13056 44.65282 18.943620
##             q4b$eqwlth
## q4b$racehisp         7
##     Black    19.671612
##     Hispanic 38.449060
##     Other     3.911968
##     White    50.967359
```

Some cells have fewer than five expected observations, so we need to use Fisher's Test

```
fisher.test(q4b$racehisp, q4b$eqwlth, simulate.p.value = TRUE)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  q4b$racehisp and q4b$eqwlth
## p-value = 0.1799
## alternative hypothesis: two.sided
```

Cannot reject the null hypothesis because the p-value is greater than .05. That means there is not a significant association between these variables.

4c. Is there an association between age (use categories) and confidence in Congress?

These variables are ordered, so use the GK Gamma test

```r
library(vcdExtra)

# Put conlegis in the right order
ps2$conlegis <- factor(ps2$conlegis,
     levels = c("Hardly any", "Only some", "A great deal"))

# Make a table
q4c_table <- table(ps2$agecat, ps2$conlegis)

# Run the test on the table
GKgamma(q4c_table)
```

```
## gamma        : -0.215
## std. error   : 0.015
## CI           : -0.245 -0.186
```

Then find the test statistic:

```r
-.215 / .015
```

```
## [1] -14.33333
```

There is a significant negative association: respondents in the youngest age categories tend to have more confidence in congress than respondents in the older age categories.