Problem Set 2. Stalled progress toward gender equality

Info 3370. Studying Social Inequality with Data Science. Spring 2023

Due: 5pm on 10 Mar 2023. Submit on Canvas.

Welcome to the second problem set!

- Use this .Rmd template to complete the problem set
- If you want to print the assignment, here is a pdf
- In Canvas, you will upload the PDF produced by your .Rmd file
- Don't put your name on the problem set. We want anonymous grading to be possible
- We're here to help! Reach out using Ed Discussion or office hours

This problem set involves both reading and data analysis. The data analysis involves the data wrangling skills we learned in class.

Reading for this the problem set

This problem set is based on an exemplar of high-quality descriptive research.

> England, Paula, Andrew Levine, and Emma Mishel. 2020. Progress toward gender equality in the United States has slowed or stalled, PNAS 117(13):6990–6997 https://doi.org/10.1073/pnas.1918891117.

Because the authors have provided a replication package, we can exactly recreate the analysis! We will learn from what they have done.

We will use two data files from the replication package for the paper.

1. `PNAS_CPS_Complete_Labeled.dta` contains data from the 1970–2018 Current Population Survey Annual Social and Economic Supplement (CPS-ASEC), accessed by the authors through IPUMS
2. `PNAS_Degree_Ratios.csv` contains data on college completion from the National Center for Education Statistics (NCES)

A note about sex and gender

As we have discussed in class, sex typically refers to categories assigned at birth (e.g., female, male). Gender is a multifaceted social construct enacted in social life; there are many genders of which man, woman, and nonbinary are only some. In this problem set, we will follow the reading by referring to gender inequality between men and women even though the categories coded in the data are male and female. However, we recognize that this survey instrument is an imperfect tool that surely misses forms of gender inequality beyond the measured categories.

Prepare your environment

```
library(tidyverse)
library(haven)
# Note: If not doing the bonus, you won't need the foreach package
library(foreach)
# To run this code, first download and place these files in your working directory
cps <- read_dta("PNAS_CPS_Complete_Labeled.dta")
nces <- read_csv("PNAS_Degree_Ratios.csv")
```

Part 1. Sex gap in employment (15 points)

In this part, you will reproduce Fig 2 from the assigned paper.

- Begin with the CPS data: `PNAS_CPS_Complete_Labeled.dta`
- Filter to the ages of interest: `age >= 25 & age <= 54`
- Filter to `asecwt > 0` (see paper footnote on p. 6995 about negative weights)
- Mutate to create an `employed` variable indicating that `empstat == 10 | empstat == 12`
- Mutate to convert `sex` to a factor variable using `as_factor`
- Group by `sex` and `year`
- Summarize the proportion employed: use `weighted.mean` to take the mean of `employed` using the weight `asecwt`
- Pivot wider to have a column `Male` and a column `Female` with your summary from above in each column. Each row will be a year
- Mutate to create the Female / Male ratio in employment
- Use `ggplot()` to produce a line graph as in the paper, Figure 2

Based on the graph and your reading of the paper, interpret in a few sentences. How did the pattern from 1970 to about 1990 differ from the pattern after 1990?

Answer.

```
# Your code here
```

Part 2: Sex gap in higher education (10 points)

In this part, you will reproduce Fig 5 from the paper.

- Use the NCES data: `PNAS_Degree_Ratios.csv`
- Create a `ggplot()` as in Figure 5 of the paper
- Make one additional tweak: use `geom_hline()` to add a dashed horizontal line at $y = 1$

Based on the graph and your reading of the paper, interpret in a few sentences. How did the Female / Male ratio in BA and doctoral degrees granted change from 1970 to 2015?

Answer.

```
# Your code here
```

Part 3: Integrate Parts 1 and 2 (5 points)

Look at the figures from Parts 1 and 2. What strikes you about the difference? Answer in 1–3 sentences.

Answer.

Part 4: Other reading questions (20 points)

4.1 (4 points) In the discussion, the authors write that "change in the gender system has been deeply asymmetric." Explain this in a sentence or two to someone who hasn't read the article.

Answer.

4.2 (4 points) The authors discuss cultural changes that could lead to greater equality in the future. Give one concrete example of a possible cultural change.

Answer.

4.3 (4 points) The authors discuss institutional changes that could lead to greater equality in the future. Give one concrete example of a possible institutional change.

Answer.

4.4 (4 points) What was one fact presented in this paper that most surprised you?

Answer.

4.5 (4 points) What do you think would be a good next question to ask with these data?

Answer.

Bonus: Sampling uncertainty (20 points for graduate students)

Overview of this question

This question is required for graduate students. It is optional for undergraduates with no extra credit. It will introduce a useful method for statistical inference in complex surveys, as well as a few coding skills: writing your estimator in a function and creating a `foreach` loop. The instructions are wordy to help support you in these skills!

Tutorial: Why we are doing this

Population inference is challenging. We would like to know the female-male employment ratio in the full population of U.S. adults ages 25–54. Due to cost limitations, the Bureau of Labor Statistics does not conduct a census but instead samples from the population. Who is included in the sample is random, and by extension any estimate from that sample is a random variable. We would like to better understand the degree to which this randomness implies uncertainty about the unknown population-based quantities when we conduct estimation in a sample.

We would like to produce a **95% confidence interval**. From past courses, you may recall the motivation for a confidence interval. Suppose we hypothetically re-drew the sample from the population many times. Suppose each time we followed a procedure to produce an interval. If those hypothetical intervals would contain the true population value 95% of the time, then the procedure yields a valid 95% confidence interval.

You might also recall mathematical formulas for a 95% confidence interval, which hold in settings such as a simple random sample. But the CPS-ASEC is not a simple random sample! Units are not selected with equal probabilities—some (e.g., those living in Wyoming) are selected with much higher probabilities than others (e.g., those living in California). The inclusion of one unit is also not independent of the inclusion of another unit—for example, because the sample is geographically clustered you are more likely to be included if your neighbor is included. Statistical inference is very hard! Mathematical formulas are not available.

Thankfully, a computational method for statistical inference is readily available. **Replicate weights** are a series of variables provided by the data collector. Each replicate weight up- and down-weights various observations so that sample-based estimates mimic the variability that would be seen if an entirely new sample were re-drawn from the population. To use the weights, one repeats the analysis many times, replacing the weight `asecwt` with one of the replicate weight columns `repwtp*`. The standard deviation of the estimate across these replicates is an estimate of the standard error of the point estimate that you created using `asecwt`.

Concrete instructions for the bonus question

In this bonus question, you will use the replicate weights.

1. Go to cps.ipums.org. Select the following variables: `age`, `year`, `empstat`, `sex`, `asecwt` and `repwtp*` where the latter will yield 160 columns of replicate weights when you download the data.
2. Select the ASEC samples for 2018–2022. The reason we are not selecting all years is so that the data file does not become huge.
3. Using R, create a prepared data frame as in the beginning of Part 1
   - Filter to the ages of interest: `age >= 25 & age <= 54`
   - Filter to `asecwt > 0` (see paper footnote on p. 6995 about negative weights)
   - Mutate to create an `employed` variable indicating that `empstat == 10 | empstat == 12`
   - Mutate to convert `sex` to a factor variable using `as_factor`

4. Following the steps from Part (1), write a function that
   - accepts `data` (which will be your prepared object) and a string `weight_name` as the argument
   - calculates the female/male employment ratio using the weight `weight_name`, as in Part 1
   - returns a data frame where each row is a year and the columns are `year`, `estimate`, and `weight_name`
5. Call your function with the weight name `asecwt`. The result is your point estimate.
6. Call your function for each replicate weight `repwtp1` to `repwtp160`. We suggest a `foreach` loop as in the pseudocode below, which we have structured to combine the results in one data frame with many rows.
7. Within each year, take the standard deviation of replicate weight estimates. This is the estimated standard error for your point estimate.
8. Join your point estimate and standard error into one data frame.
9. Mutate to create new variables `ci.min` and `ci.max` for the lower and upper bounds of a confidence interval. We will assume a normal sampling distribution (drawing on the Central Limit Theorem) and place the bounds of this interval at the point estimate plus or minus `qnorm(.975)` times your estimated standard error.
10. Visualize the result using `ggplot`. Use `geom_point()` for the point estimate, and use `geom_errorbar()` for the 95% confidence interval estimate.

Then interpret. Given your confidence intervals, does the jump in Female / Male employment in 2020 seem likely attributable to a population-level change, or to randomness in sampling? If a population-level change, propose a possible substantive explanation for this jump.

Answer.

```
# Your code here. We have included some helpers.

# Do steps 1-3 on your own.

# 4. Write a function to make an estimate
make_estimate <- function(data, weight_name) {
  # Use the data, with the weight weight_name
  # Return a data frame with year and the estimated sex ratio in employment rates
  # Below is a temporary return so the other pseudo-code can run
  return(data.frame(year = NA, estimate = NA, weight_name = weight_name))
}

# 5. Create a point estimate
point <- make_estimate(data = prepared, weight_name = "asecwt")

# 6. Create estimates with each replicate weight
estimate_star <- foreach(weight_name_case = paste0("repwtp",1:160), .combine = "rbind") %do% {
  make_estimate(data = prepared, weight_name = weight_name_case)
}

# 7. Aggregate the replicate estimates to a standard error estimate
se <- estimate_star %>%
  group_by(year) %>%
  summarize(se = sd(estimate), .groups = "drop")

# 8 and 9. Combine the point estimate and standard error with confidence interval
estimate <- point %>%
  left_join(se, by = "year") %>%
  # Construct a confidence interval
  mutate(ci.min = estimate - qnorm(.975) * se,
```

```
         ci.max = estimate + qnorm(.975) * se)

# 10. Create a ggplot
```

Computing environment

Leave this at the bottom of your file, and it will record information such as your operating system, R version, and package versions. This is helpful for resolving any differences in results across people.

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] foreach_1.5.2   haven_2.5.1     forcats_1.0.0   stringr_1.5.0
##  [5] dplyr_1.1.0     purrr_1.0.1     readr_2.1.3     tidyr_1.3.0
##  [9] tibble_3.1.8    ggplot2_3.4.0   tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.2.0    xfun_0.36           gargle_1.3.0
##  [4] colorspace_2.1-0    vctrs_0.5.2         generics_0.1.3
##  [7] htmltools_0.5.4     yaml_2.3.7          utf8_1.2.3
## [10] rlang_1.0.6         pillar_1.8.1        glue_1.6.2
## [13] withr_2.5.0         DBI_1.1.3           bit64_4.0.5
## [16] dbplyr_2.3.0        modelr_0.1.10       readxl_1.4.1
## [19] lifecycle_1.0.3     munsell_0.5.0       gtable_0.3.1
## [22] cellranger_1.1.0    rvest_1.0.3         codetools_0.2-18
## [25] evaluate_0.20       knitr_1.42          tzdb_0.3.0
## [28] fastmap_1.1.0       parallel_4.2.2      fansi_1.0.4
## [31] broom_1.0.3         scales_1.2.1        backports_1.4.1
## [34] googlesheets4_1.0.1 vroom_1.6.1         jsonlite_1.8.4
## [37] bit_4.0.5           fs_1.6.0            hms_1.1.2
## [40] digest_0.6.31       stringi_1.7.12      grid_4.2.2
## [43] cli_3.6.0           tools_4.2.2         magrittr_2.0.3
## [46] crayon_1.5.2        pkgconfig_2.0.3     ellipsis_0.3.2
## [49] xml2_1.3.3          reprex_2.0.2        googledrive_2.0.0
## [52] lubridate_1.9.1     timechange_0.2.0    iterators_1.0.14
## [55] assertthat_0.2.1    rmarkdown_2.20      httr_1.4.4
## [58] rstudioapi_0.14     R6_2.5.1            compiler_4.2.2
```