Project 1

Elyse McFalls and Holly Cui

2023-10-20

Data and Sampling Frame

```
# loading county data
county <- read_excel("county.xlsx", col_names = TRUE)
```

This project seeks to understand facets the United States using county-level data. Specifically, we are interested in examining the population density and demographic information Data on the states and counties in the United States were collected from the wikipedia page "List of states and territories of the United States," which contains the 2020 census estimates for each county's population and area (source). More detailed data for each county was extracted from the Unites States Census Bureau and MIT's Election Data and Science Labs (MEDSL). The data from the Census contained demographic estimates for each county from 2010 to 2020 (source). MEDSL, on the other hand, had data on political participation by county from 2000 to 2020 (source).

All counties from the 50 U.S. states and the District of Columbia were included in the sampling frame. We did not include counties from U.S. territories because we are primarily interested in regions that are enfranchised. We also focused in data from the years 2010 and 2020 for the analyses.

```
# removing US territories
US_territories <- c("American Samoa", "Guam", "Northern Mariana Islands",
                    "Puerto Rico", "U.S. Minor Outlying Islands",
                    "Virgin Islands (U.S.)")

clean_county <- county %>%
  filter(!state %in% US_territories) %>%
  group_by(state) %>%
  mutate(Nh = n(), Sh = sd(pop)) %>%
  ungroup() %>%
  mutate(id = row_number())
```

Sampling Procedure

Sampling Design

In deciding on the sampling design, we wanted to chose a framework that best fit the goals of this project. This framework would ideally give us a representative sample of the counties in the U.S. that bears in mind the variedness of each state. Therefore, we decided on a simple stratified sample with each of the 50 states serving as individual stratum. Moreover, we are using optimally allocation to determine how many counties

to sample from each state. The District of Columbia, which operates like a city, state, and county, was treated as a state and a county. It was sampled with 100% certainty (source).

By taking a stratified sample, we are ensuring that every state in the U.S. is being accounted for. Stratified samples also provide estimates with lower standard errors when there is more between variance than within variance. We expect this to be the case with counties in each state. For instance, counties in Virginia are expected to exhibit more similarities among themselves than when compared to counties in Wyoming. However, we also believed that the variance of our variables of interest within each state may differ by state. Consider a state like North Dakota where most of the counties are rural compared to a state like North Carolina that has a mixture of rural and urban areas. We'd anticipate that county-level population in North Dakota would be more uniform and the same statistics in North Carolina would be more varied. source. Therefore, instead of sampling proportional to size, we opted to for an optimal allocation framework (equation 1). This way, we would sample more from states that have high variances for our variables of interest. We decided to focus on the populations of each county to optimally allocate the samples of each stratum. We expect population size to be related to demographics. Therefore, this framework will reduce the variance in our population estimates and likely the variance in our demographic related estimates as well.

$$\text{Eq. } 1 : n_h = n\frac{N_h S_h / N}{\sum_{h=1}^{H} N_h S_h / N}$$

where N is the total number of counties in the population, n is the sample size,

H is the total number of states, N_h is the number of counties in state h, and

S_h is the variance of the county populations in state h

Moreover, taking a simple random sample of counties allows us to get a representative sample from each state. We considered taking a sample proportional to size, but we felt it was unnecessary to prioritize more populated ares. We also decided to sample the District of Columbia with 100% certainty because, as the nation's capital, it is an important region and it functions like a state and a county. However, if we treated it like a state with one county, our optimal allocation formula would assign no samples from it due to there being no variance. Therefore, in order to have D.C. in our sample, we purposefully included it.

Sample Sizes and Weights

We used an initial sampled size of 314, which is 10% of our population. However, the output of the optimal allocation formula gave us some states with no counties sampled. To have all states accounted for, we rounded any calculated nh values less than 0.5 to 1. This left us with a sample of 317.

The weight for each county is the total number of counties in its respective state divided by the number of counties state. For instance, Alabama has 67 counties and 3 of them were sampled. Therefore, its weight was 22.33. The weight for D.C. is 1 since it is a certainty primary sampling unit.

```
# state-level strata statistics (without DC - certainty PSU)
clean_state <- county %>%
  filter(!state %in% US_territories) %>%
  filter(state != "District of Columbia") %>%
  group_by(state) %>%
  summarise(Nh = n(), Sh = sd(pop)) %>%
  ungroup()
```

```r
# calculate denominator for optimal allocation
n = 314
denominator = sum(clean_state$Nh * clean_state$Sh)

# summarize nh for each state strata
state_strata <- clean_state %>%
  mutate(nh_round = round((n-1)*Nh*Sh / denominator)) %>%
  mutate(nh = ifelse(nh_round == 0, nh_round+1, nh_round))

# get final sampling schema
state_nh <- state_strata %>%
  select(state, Nh, Sh, nh)
```

```r
set.seed(123)

sample_county = c()
for (i in 1:nrow(state_nh)) {
  sample_by_state = sample(clean_county$id[clean_county$state == state_nh$state[i]],
                           state_nh$nh[i])
  sample_county = c(sample_county, sample_by_state)
}

# add in DC (certainty PSU)
final_sample = c(sample_county, 317)
```

```r
sample = clean_county %>%
  filter(id %in% final_sample)
```

```r
# excluding D.C. for now, its weight is 1
sample_final <- sample %>%
  left_join(state_nh, by = c("state", "Nh", "Sh")) %>%
  select(-id) %>%
  mutate(weights = Nh/nh,
         pop_density = pop/area,
         county = ifelse(grepl("city", county, ignore.case = TRUE),
                         sub(",.*", "", county),
                         county))
```

The breakdown of the number of counties sampled by state using the optimal allocation formula is as follows:

State	nh	State	nh	State	nh
Alabama	3	Louisiana	3	Ohio	9
Alaska	1	Maine	1	Oklahoma	4
Arizona	8	Maryland	3	Oregon	3
Arkansas	2	Massachusetts	4	Pennsylvania	9
California	39	Michigan	10	Rhode Island	1
Colorado	5	Minnesota	6	South Carolina	3
Connecticut	1	Mississippi	2	South Dakota	1
Delaware	1	Missouri	7	Tennessee	6
Florida	16	Montana	1	Texas	48
Georgia	11	Nebraska	3	Utah	3

State	nh	State	nh	State	nh
Hawaii	1	Nevada	4	Vermont	1
Idaho	2	New Hampshire	1	Virginia	8
Illinois	25	New Jersey	3	Washington	7
Indiana	5	New Mexico	2	West Virginia	1
Iowa	3	New York	16	Wisconsin	5
Kansas	4	North Carolina	8	Wyoming	1
Kentucky	4	North Dakota	1	District of Columbia*	1

*certainty primary sampling unit

```
sample_complete <- read.csv('sample_complete.csv')
```

```
sample_complete[61, 9] = 1
```

```
sample_complete <- sample_complete %>% mutate(hisp_change = hisp_2020-hisp_2010)
```

```
# survey design
options(survey.lonely.psu="certainty")
#sample_nodc <- sample_complete[-c(61),]
des <- svydesign(~1, strata = sample_complete$state,
                 weights = sample_complete$weights,
                 fpc = sample_complete$Nh,
                 data=sample_complete)
```

Results

Average Population Density per County in the U.S. in 2020

Amount of Hispanics and Latinos in the U.S. in 2020

Change in the Amount of Hispanics and Latinos in the U.S. between 2010 and 2020

```
svytotal(~hisp_change, des)
```

```
##                total      SE
## hisp_change 10391602 1603186
```

```
confint(svytotal(~hisp_change, des))
```

```
##                2.5 %   97.5 %
## hisp_change 7249416 13533788
```

We estimate that a total of 10,391,602 more U.S. residents identified as Hispanic or Latino in 2020 compared to 2010. We have a 95% confidence interval of (7,249,416, 13,533,788) for this estimate.

It's clear that a lot more Hispanic/Latino identifying Americans now compared to a decade ago. Out of curiosity, we also wanted to see if the proportion of Hispanic and Latinx residents rose over time. Since the total population data and Hispanic/Latino data came from two different sources, we will use a ratio estimate to estimate both quantities then divide.

```
# hipanic pop in 2010
svyratio(~hisp_2010, ~pop.2010, des)
```

```
## Ratio estimator: svyratio.survey.design2(~hisp_2010, ~pop.2010, des)
## Ratios=
##            pop.2010
## hisp_2010 0.1816844
## SEs=
##             pop.2010
## hisp_2010 0.02208595
```

```
confint(svyratio(~hisp_2010, ~pop.2010, des))
```

```
##                         2.5 %    97.5 %
## hisp_2010/pop.2010 0.1383968 0.2249721
```

```
# hipanic pop in 2020
svyratio(~hisp_2020, ~pop.2020, des)
```

```
## Ratio estimator: svyratio.survey.design2(~hisp_2020, ~pop.2020, des)
## Ratios=
##            pop.2020
## hisp_2020 0.2062212
## SEs=
##             pop.2020
## hisp_2020 0.02309742
```

```
confint(svyratio(~hisp_2020, ~pop.2020, des))
```

```
##                         2.5 %    97.5 %
## hisp_2020/pop.2020 0.1609511 0.2514913
```

We estimate that the percentage of Hispanic and Latino residents in 2010 was 18.17% (95% CI: (13.84%, 22.5%)) while the same statistic in 2020 was estimated to be 20.62% (95% CI: (16.1%, 25.15%)). These results show that we estimate the proportion of Hispanic and Latino residents to be higher in 2020 compared to 2010, but the overlapping confidence intervals suggest this is not a significant difference.

Partisan Breakdown in the U.S. in 2020

```
# percentage of people who voted republican
svyratio(~REPUBLICAN, ~totalvotes, des)
```

```
## Ratio estimator: svyratio.survey.design2(~REPUBLICAN, ~totalvotes, des)
## Ratios=
##            totalvotes
## REPUBLICAN  0.4839342
## SEs=
##            totalvotes
## REPUBLICAN 0.01721629
```

```
confint(svyratio(~REPUBLICAN, ~totalvotes, des))
```

```
##                             2.5 %    97.5 %
## REPUBLICAN/totalvotes 0.4501909 0.5176775
```

We estimate that 48.39% of voters in the 2020 election voted for Trump and the Republican party. We have a 95% confidence interval of (45.02%, 51.77%). Based on the 2020 Presidential Popular Vote Summary, we are 1.54% off from the actual total of 46.85% (source).

```
# percentage of people who voted democrat
svyratio(~DEMOCRAT, ~totalvotes, des)
```

```
## Ratio estimator: svyratio.survey.design2(~DEMOCRAT, ~totalvotes, des)
## Ratios=
##          totalvotes
## DEMOCRAT  0.4991369
## SEs=
##          totalvotes
## DEMOCRAT 0.01747804
```

```
confint(svyratio(~DEMOCRAT, ~totalvotes, des))
```

```
##                           2.5 %    97.5 %
## DEMOCRAT/totalvotes 0.4648806 0.5333932
```

We also predict that 49.91% of voters in the 2020 election voted for Biden and the Democratic part. We have a 95% confidence interval of (46.49%, 53.34%). Using the same resource as before, are estimate is off by 1.4%. The actual percentage is 51.31% (source).

```
# percentage of people who voted third party
svyratio(~THIRD, ~totalvotes, des)
```

```
## Ratio estimator: svyratio.survey.design2(~THIRD, ~totalvotes, des)
## Ratios=
##       totalvotes
## THIRD 0.01692889
## SEs=
##         totalvotes
## THIRD 0.0006317843
```

```
confint(svyratio(~THIRD, ~totalvotes, des))
```

```
##                         2.5 %     97.5 %
## THIRD/totalvotes 0.01569061 0.01816716
```

Finally, the expected percentage of voters who voted for a third party in the 2020 election was 1.69%. This estimate had a 95% confidence interval of (1.57%, 1.82%). Given the actual value of 1.18%, we are off by 0.51% (source). This is a relatively large difference which is reflected by our confidence interval which does not include the true percentage value. Overall, it is clear that our survey design did not adequately estimate this quantity.

The partisanship results from the 2020 election indicate how close it was. The 95% confidence intervals for the Republican and the Democratic party overlap each other by a great deal.

Change in Proportion of 65+ Population Overtime

```
# proportion of residents 65 and over in 2010
svyratio(~pop_2010_65, ~pop.2010, des)
```

```
## Ratio estimator: svyratio.survey.design2(~pop_2010_65, ~pop.2010, des)
## Ratios=
##              pop.2010
## pop_2010_65 0.1313276
## SEs=
##                pop.2010
## pop_2010_65 0.002827304
```

```
# CI for proportion of residents 65 and over in 2010
confint(svyratio(~pop_2010_65, ~pop.2010, des))
```

```
##                           2.5 %   97.5 %
## pop_2010_65/pop.2010 0.1257862 0.136869
```

We expect that roughly 13.1% of residents in the U.S. were 65 and over in 2010. Our 95% confidence interval for this estimate is (12.6%, 13.7%).

```
# proportion of residents 65 and over in 2020
svyratio(~pop_2020_65, ~pop.2020, des)
```

```
## Ratio estimator: svyratio.survey.design2(~pop_2020_65, ~pop.2020, des)
## Ratios=
##              pop.2020
## pop_2020_65 0.1682854
## SEs=
##                pop.2020
## pop_2020_65 0.003539639
```

```
# CI for proportion of residents 65 and over in 2020
confint(svyratio(~pop_2020_65, ~pop.2020, des))
```

```
##                           2.5 %   97.5 %
## pop_2020_65/pop.2020 0.1613478 0.175223
```

Moreover, we estimate that the percentage of U.S. residents who are 65 and over in 2020 was 16.8%. We have a 95% confidence interval of (16.1%, 17.5%) for this estimate.

Given that our confidence intervals do not overlap, we have some evidence to suggest that proportion of residents in the U.S. who are 65 and over was higher in 2020 than in 2010. In other words, people in the U.S. have gotten older overtime (more evidence from the Census Bureau: source). Future studies may seek to model the trajection of the 65+ age group to determine if this trend is present throughout multiple decades. We may also want to run a regression to see what factors are significantly associated with this pattern while controlling for other variables. These results could aid in understanding the growth of the 65+ age group and what this growth may mean for future health care policies, social security benefits, and so on.