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, the percentage of Hispanic and Latinx residents, and partisanship. 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 Wikipedia article link. More detailed data for each county was extracted from their respective wikipedia page (i.e., data for DeKalb County, AL: Wikipedia article link). The pages for each county usually contain census data from the past few decades.

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.

```
# 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. This decision was mainly due to the lack of data for our other variables of interest for each county.

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

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 sampled from its respective state divided by the total number of counties in that state. For instance, Alabama has 67 counties and 3 of them were sampled. Therefore, its weight is 0.045. 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()
```

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

```r
# 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
weights_df <- data.frame(state = state_nh$state, weight = state_nh$nh/state_nh$Nh)
sample <- merge(sample,weights_df, by = "state")
```

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
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

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

Partisan Breakdown in the U.S. in 2020