STA322 Project 1

Alicia Gong, Annie Lin

23 October, 2023

Introduction

Sample Frame

The sampling frame includes 3142 U.S. counties and equivalents plus District of Columbia, giving 3143 in total. There are two important notes: First, we excluded Guam, Virgin Islands, Puerto Rico, American Samoa, Northern Mariana Islands, and U.S. Minor Outlying Islands from the sampling frame. Second, Connecticut petitioned to replace 8 counties with 9 planning regions as county-equivalent geographic units in 2019, and the U.S. Census Bureau adapted this change in 2023 (see announcement). Most Connecticut data are still in county level instead of region level. Therefore, we chose to use county level data for Connecticut in this analysis, although it is not align with the latest U.S. Census Bureau standard.

Data

Population

All the county level population data, except Connecticut's, are from the 2020 decennial population census. Connecticut's county level population data are collected from Connecticut Demographics.

Hispanic Population

The county level Hispanic population data in 2020 and 2010 are downloaded from U.S. Census Bureau's website.

2020 Presidential Election

The county level voting data, except Alaska's, are collected from Politico. Alaska's county level data are collected from each county's Wikipedia page. The total voting population is calculated by using number of votes divided by percentage of votes.

Area

The data on county area in mi^2 are collected from the 2020 decennial population census. Connecticut's county level area data are collected separately from National Association of Counties

Survey Design

First, we used state as natural stratification. We used proportional allocation to decide the sample sizes within each strata. Then, we applied PPS (probability proportional to size) Systematic Sampling within each strata.

We decided to sample 300 counties, roughly 1/10 of all counties. We calculated sample size of each strata using $N_h/N = n_h/n$, where $N = 3143$ is the total number of U.S. counties, N_h is total number of counties in each strata, and $n = 300$ is the total number of sample size. We rounded all sample size to integer.

Then, within each state, we calculated each county's weight using $weights = t_x/(n * x_i)$, where t_x is the total population of each state (strata), n is the sample size of each state (i.e., number of sampled counties calculated using proportional allocation), and x_i is the population of each county.

We used the 'ppssstrat' command in pps package to help with the sampling process. 'ppssstrat' is a command for Stratified PPS Systematic Sampling, and it does sample by using systematic random sampling with probability proportional to size. Stratified PPS Systematic Sampling selects units at a fixed interval throughout the stratum or sampling frame after a random start, then it chooses the first unit randomly from the entire stratum with probability proportional to size and then treats the stratum observations as a closed loop.

In systematic PPS sampling, the list of units is first ordered randomly and the cumulative total of the auxiliary variable x_i is calculated. Selection of units then takes place using interval sampling with the interval value calculated by dividing the cumulative total $\sum_{i \in N} x_i$ by the sample size n. This is done by generating a random starting point between zero and the interval value in order to select the first unit. The second random number is generated by adding the interval value to the starting point. This is then used to select the second unit. This process of adding the interval value to the previous random number, and selecting the corresponding units, is repeated until the requisite number of units has been sampled.

Questions

1. What is an estimate of the average population density per county in the U.S. in 2020?

```
## Ratio estimator: svyratio.survey.design2(~area, ~pop_county, design = svydes)
## Ratios=
##      pop_county
## area   56.72258
## SEs=
##      pop_county
## area   8.203689
```

The average population density per county in the U.S. in 2020 is 52 per mi^2.

2. What is an estimate of the total number of people in the U.S. in 2020 who identify as Hispanic or Latino, any race?

```
##                         total      SE
## total_hispanic_2020 62740192 5659594
```

```
##                         2.5 %   97.5 %
## total_hispanic_2020 51647592 73832791
```

An estimate of 63435246 people in the U.S. in 2020 who identify as Hispanic or Latino.

3. What is an estimate of the total change in the number of people in the U.S. who identify as Hispanic or Latino, any race, between the 2010 and 2020 censuses?

```
##                   total     SE
## hispanic_diff 11380114 841022
```

```
##                  2.5 %   97.5 %
## hispanic_diff 9731742 13028487
```

The estimated difference of people in the U.S. who identify as Hispanic or Latino between 2010 and 2020 is 11217022.

4. What are estimates of the percentages of people in 2020 in the U.S. who voted Republican? How about Democrat? How about a third party?

To answer this question, we used ratio estimator to estimate the percentage.

```
## Ratio estimator: svyratio.survey.design2(~rep_vote, ~total_vote_pop, svydes)
## Ratios=
##          total_vote_pop
## rep_vote      0.5728288
## SEs=
##          total_vote_pop
## rep_vote      0.0115836
```

```
##                              2.5 %    97.5 %
## rep_vote/total_vote_pop 0.5501254 0.5955322
```

```
## Ratio estimator: svyratio.survey.design2(~dem_vote, ~total_vote_pop, svydes)
## Ratios=
##          total_vote_pop
## dem_vote      0.4099367
## SEs=
##          total_vote_pop
## dem_vote     0.01149594
```

```
##                             2.5 %    97.5 %
## dem_vote/total_vote_pop 0.387405 0.4324683
```

```
## Ratio estimator: svyratio.survey.design2(~third_vote, ~total_vote_pop, svydes)
## Ratios=
##            total_vote_pop
## third_vote     0.01723452
## SEs=
##            total_vote_pop
## third_vote   0.0002918687
```

An estimate of 56.3% of the votes in 2020 Presidential Election were for Republican, 42% were for Democratic, and 1.7% were for third party.

5. Answer one other question that interests you from the data that you could collect from the website.

We estimate percentage of population who did not vote in 2020 election.

```
##             total      SE
## no_vote 168948014 7747567
```

```
##             2.5 %    97.5 %
## no_vote 153763062 184132967
```

```
## [1] 0.524215
```

About 170903041 people didn't vote in 2020, and that's about 52% of the whole population.

We area also interested in the total population per county.

```
##              mean    SE
## pop_county 107028 11070
```

```
##               2.5 %   97.5 %
## pop_county 85332.12 128724.2
```

An estimated total population per county is 104980.