Project 1

Elyse McFalls and Holly Cui

2023-10-20

Data and Sampling Frame

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.

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 S_h is the variance of the county populations for each 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

We calculated a sample size of 342 counties using the sample size for a known population formula with a 95% confidence interval (equation 2). We also set the error to be 5% to coincide with our confidence interval and the population proportion to 0.5 for the worst case-scenario estimate.

$$\text{Eq. 2: } n = \frac{\frac{Z^2 \times p(1-p)}{e^2}}{1 + \frac{Z^2 \times p(1-p)}{e^2 N}}$$

where

- Z: Z-score for the 95% confidence intervals
- p: Population proportion of ?
- e: Designated margin of error
- N: Population size (3,113 counties in the States + D.C.)

The breakdown of the number of counties sampled by state is as follows:

State	nh	State	nh	State	nh
Alabama		Louisiana		Ohio	
Alaska		Maine		Oklahoma	
Arizona		Maryland		Oregon	
Arkansas		Massachusetts		Pennsylvania	
California		Michigan		Rhode Island	
Colorado		Minnesota		South Carolina	
Connecticut		Mississippi		South Dakota	
Delaware		Missouri		Tennessee	
Florida		Montana		Texas	
Georgia		Nebraska		Utah	
Hawaii		Nevada		Vermont	
Idaho		New Hampshire		Virginia	
Illinois		New Jersey		Washington	
Indiana		New Mexico		West Virginia	
Iowa		New York		Wisconsin	
Kansas		North Carolina		Wyoming	
Kentucky		North Dakota		District of Columbia*	1

*certainty primary sampling unit

Weights

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