Optimal Stratified Sampling for Probability-Based Online Panels

DC AAPOR 2025

Jonathan Eggleston
Senior Economist
U.S. Census Bureau

September 16th, 2025

Any opinions and conclusions expressed herein are those of the author and do not represent the views of the U.S. Census Bureau. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection (Project 7532352: CBDRB-FY25-CES005-009).

United States® Census Bureau

Introduction + Literature Review

- Recently in survey methodology, there has been advancements in augmenting sampling frames with auxiliary microdata sources to improve sampling efficiency
  - Household Trends and Outlook Pulse Survey (HTOPS) from the U.S. Census Bureau: Using the Demographic Frame (Ratcliffe 2021) for sampling
  - NORC uses various commercial data for sampling in the AmericSpeak Panel
- These innovations has been made possible by recent advances in
  - Acquiring and linking auxiliary microdata sources
  - Using machine-learning techniques to create sampling strata

United States® Census Bureau

Introduction + Literature Review

- However, there been much less research on how to construct *sampling rates* for stratified sampling
  - The assignment of sampling rates has important implications for cost-savings and variances of estimates
- Much of current survey practice is based on formulas in papers and textbooks from the 20th century (e.g. Cochran 1977). These formulas typically assume 100% response rates
- Mendelson and Elliott (2024) derive sampling rate formulas under anticipated nonresponse, but their results are based on
  1. A cross-sectional survey and
  2. Only one survey-outcome of interest

United States® Census Bureau

Contribution

- In this work, I will investigate how to construct optimal sampling rates for online probability-based panels
  - Online panels: A popular means for conducting surveys in a cost-efficient manner in the internet age
- Model gives a framework to set sampling rates for online panels in a principled manner
- Implementation of model's sampling rates can propose a path forward for maintaining the precision of estimates when reducing a survey's fielding budget

United States® Census Bureau

Outline

1. Model Overview
2. Data
3. Results

United States® Census Bureau

External

Outline

1. Model Overview
2. Data
3. Results

Setup of Sampling Problem

- Stratified sampling problem
  - Goal is to minimize the standard errors of estimates subject to the fielding budget equaling some pre-specified amount
  - The only variables that can be change is the number of cases drawn in each strata. Other features of the survey are held constant
- Mendelson and Elliott (2024): For a cross-sectional survey, a stratum should have higher sampling rates if
  1. Its expected response rates are lower: Need to sample more in order to have an adequate respondent sample size
  2. Its expected cost per case is lower: Could sample more from this strata to boost the overall respondent sample size

United States® Census Bureau

Setup of Sampling Problem

- For online panels: develop a model that's similar to a cross-sectional sampling model. A key difference is to incorporate baseline and topical surveys, which can have their own response rates and cost structures
  - Baseline survey: The initial survey that recruits panelists
  - Topical survey: Individual surveys panelists are asked to participate in
- Goal is to try to incorporate many features of online panels, but make some assumptions to help simplify the sampling problem

United States® Census Bureau

Simplifying Assumptions

1. Assume the panel exists for a fixed period of time (5-years in this analysis)
2. Content of topical surveys is constant over time
3. No attrition: topical response rates are constant over time
4. Within a strata, response propensities and survey outcomes are independent over time
5. Budgets surpluses or shortfalls from the initial baseline survey can be carried over to the budget for the topical surveys at a constant real interest rate
    - This assumption is unrealistic in many settings (e.g. government agencies often have to spend funds within a fiscal year). But this help to greatly simplify the budget constraint and the optimization problem

United States® Census Bureau

Model

$$\min_{\{n_1^s,\ldots,n_H^s\}} \frac{\beta(1-\beta^T)}{1-\beta}\left(\sum_{s=1}^{S}\sum_{k=1}^{K^s}\frac{\omega_{s,k}E\left[\left(\widehat{\theta_t^{s,k}}-\theta^{s,k}\right)^2\right]}{(\theta^{s,k})^2}\right)$$

Subject to

$$C_1 + \delta C_o = \sum_{h=1}^{H} n_h^I \alpha_h$$

United States® Census Bureau

Model

$$\min_{\{n_1^s,\ldots,n_H^s\}} \frac{\beta(1-\beta^T)}{1-\beta}\left(\sum_{s=1}^{S}\sum_{k=1}^{K^s}\frac{\omega_{s,k}E\left[\left(\widehat{\theta_t^{s,k}}-\theta^{s,k}\right)^2\right]}{(\theta^{s,k})^2}\right)$$

Subject to

$$C_1 + \delta C_o = \sum_{h=1}^{H} n_h^I \alpha_h$$

Mean Square Error (Normalized)

United States® Census Bureau

Model

$$\min_{\{n_1^s,\ldots,n_H^s\}} \frac{\beta(1-\beta^T)}{1-\beta}\left(\sum_{s=1}^{S}\sum_{k=1}^{K^s}\frac{\omega_{s,k}E\left[\left(\widehat{\theta_t^{s,k}}-\theta^{s,k}\right)^2\right]}{(\theta^{s,k})^2}\right)$$

Subject to

$$C_1 + \delta C_o = \sum_{h=1}^{H} n_h^I \alpha_h$$

Term with discount factor that ended up not affecting the optimal value. One benefit from assuming no attrition

United States® Census Bureau

Model

$$\min_{\{n_1^S,\ldots,n_H^S\}} \frac{\beta(1-\beta^T)}{1-\beta}\left(\sum_{s=1}^{S}\sum_{k=1}^{K^s}\frac{\omega_{s,k}E\left[\left(\widehat{\theta_t^{s,k}}-\theta^{s,k}\right)^2\right]}{(\theta^{s,k})^2}\right)$$

Subject to

$$C_1 + \delta C_o = \sum_{h=1}^{H} n_h^I \alpha_h$$

Budget constraint. More complicated than cross-sectional problem, but largely in a similar linear form

United States® Census Bureau

Optimization Techniques

- Have developed python code that uses numerical optimization techniques (e.g. Valliant et al. 2014) to find a solution
- Some details:
  - Use simulation to evaluate integrals in the problem
  - Use a basin-hopping global optimization algorithm in the SciPy library. This helps address some lack of smoothness in the objective function from the simulated integrals

United States® Census Bureau

Outline

1. Model Overview
2. Data
3. Results

Data: Census Household Panel (CHP)

- A survey from the U.S. Census Bureau
- One of the predecessors to the Household Trends and Outlook Pulse Survey
- CHP sampling used the Census Bureau's Demographic Frame
  - Frame harmonizes decennial census and administrative data to create a best guess of the address an individual resides at
- Sampling procedure
  - Flag households that likely have individuals who are Hispanic or non-White
  - The flagged addresses were sampled at twice the rate of addresses that primarily have White non-Hispanic individuals

United States® Census Bureau

Sampling Rate

- This oversampling rate was determined based on a rule of thumb using some rough sample size calculations and relative response rate information from past surveys
- Subsequent analyses: examine how the sampling rate would change under various model parameterizations

Outline

1. Model Overview
2. Data
3. Results

Response Rate and Mode Differences

Frame Strata	Response Rate Baseline	Response Rate Topical
1: White non-Hispanic	18.97%	59.72%
	(0.211)	(0.606)
2: Black	13.70%	46.56%
	(0.300)	(1.178)
3: Hispanic	13.48%	45.59%
	(0.275)	(1.092)
4: Other Race	18.72%	49.84%
	(0.390)	(1.154)

Other Race baseline response rate similar to first stratum's response rate

Model Parameterization

- Assume there is only one outcome variable for the survey, which has the same variance across the strata
  - Simplification helps focus on response rate and cost differences across strata for the model
- Importance weights for subgroup analyses:
  - Overall national estimate: 0.6
  - Mean for White non-Hispanics: 0.2
  - Mean for Hispanics and non-Whites: 0.2

United States® Census Bureau

Sampling Rate: Relative to Stratum 1

Frame Strata	Model's Optimal Sampling Rate	Sampling Rate Used in CHP
1: White non-Hispanic	1	1
2: Black	2.237	2
3: Hispanic	2.336	2
4: Other Race	1.578	2

United States® Census Bureau

Interpretation

- Can express change in sampling rates in terms of variation reduction or cost-savings
- If fielding budget is help constant, use of the optimal sampling rates would **reduce** the average variance of estimate **by 1.83%**
- Use of optimal sampling rates could allow one to **reduce** the fielding budget **by 1.80%** and still maintain the same average level of variance as before

United States® Census Bureau

Robustness Check

- In my paper, I examine changes in sampling rates by changing importance weights, and with equalizing certain parameters across strata
- Largest change in results come from changing the optimization problem to separate out the mean for Hispanics and non-Whites into 3 domain means (Black, Hispanic, Other Race)
  - Sampling rates for Strata 2-4 would increase, resulting in larger variance improvement (**about 7%)** if this is the “correct” specification

United States® Census Bureau

Contact Information

Jonathan Eggleston
Senior Economist
Survey Improvement Technical Lead

Survey and Economic Research Group
Center for Economic Studies
U.S. Census Bureau

Office: 301.763.2357
jonathan.s.eggleston@census.gov

United States® Census Bureau