A Comparison of Alternative Optimization Techniques for Sample Allocation in Surveys with National and Sub-National Precision Requirements*

John Chesnut and Shawn Baker

Demographic Statistical Methods Division, U.S. Census Bureau

2019 Joint Statistical Meetings
Denver, Colorado

*Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau, DAO approval # CBDRB-FY19-ROSS-B0165

Sample Redesign Research Program

- Research and develop innovative cross-cutting improvements to the sample designs used by the major household surveys at Census
  - Current Population Survey - labor force characteristics of the U.S. population - (BLS)
  - Survey of Income and Program Participation - income and govt. program participation of individuals and households in the U.S. (Census)
  - National Crime Victimization Survey - characteristics and consequences of criminal victimization in the U.S. (BJS)
  - Consumer Expenditure Surveys - two surveys that characterize the buying habits of American consumers (BLS)
  - American Housing Survey - collects data on the Nation's housing stock, household characteristics, housing and neighborhood quality, housing costs, and recent movers (HUD)

Research Focus

- Are the household surveys with multiple objectives meeting their objectives in an optimal manner? (e.g., multiple key estimates and domains of interest)
- As a first step, consider the sample allocation problem for the CPS with precision standards at the national and state-level, explore alternative allocation methods

Previous Work on Sample Allocation for Multipurpose Surveys

- Neyman allocation (1934) - univariate, insufficient precision for small strata
- Compromise solution - for P variables, average the Neyman allocation solutions for each variable (Huddleston, Claypool, and Hocking, 1970)
- Define an objective function that is a weighted average of the variances $\bar{V} = \sum_P H_p V_p$, (e.g., Valliant and Gentle, 1996)
  - Choice of the importance weights $\{H_p\}$ is arbitrary - optimality not clear
- Minimization of a convex objective function while satisfying inequality constraints for the variances, $V_p \leq V_{p0}$ (e.g., Bethel, 1989)
  - Gives the optimal solution
  - Complex analytical solutions, but can use numerical methods

Current Population Survey - Sample Design Requirements

- Designed primarily to produce national and state estimates of labor force characteristics
- Official design requirements (CPS Technical Paper 66, 2006)
  - A 0.2% change in the unemployment rate from month-to-month is statistically significant at the 10% level assuming a 6% unemployment rate
  - A maximum coefficient of variation (*cv*) of the annual average unemployment level for each state, the District of Columbia, and the metropolitan areas of New York and Los Angeles is 8%
- Unofficial design requirements
  - Approximate sample size of 60,000 housing units
  - Reliability for other labor force characteristics
  - Approximately self weighting national sample

Define the Sample Allocation Problem - Objective Function

- Rottach and Erkens (2012) developed a mathematical model that relates the national and state-level design requirements
  - Both requirements are converted into *cv* requirements for national and state-level monthly unemployment totals
  - At the national-level for a given month t, $cv(\hat{\hat{Y}}_t) = cv(\hat{Y}_t)$ based on the linearization $cv^2(A/B) \cong cv^2(A) - cv^2(B)$ and a negligible $cv^2(B)$
  - Assume $\{\hat{Y}_{t,s}\}_{s \in States}$ are independent and assume the national and state-level unemployment rates are approximately equal
  - Therefore, $cv^2(\sum_s \hat{Y}_{t,s}) = \sum_s p_s^2 cv^2(\hat{Y}_{t,s})$ where $p_s = CLF_s/CLF$, (Civilian Labor Force)

Defining the Sample Allocation Problem - Objective Function

- Given direct estimates of the current state cv values and assuming that $cv^2 \propto \frac{1}{n}$,
  $cv^2_{new,s}(\hat{Y}_{t,s}) = \frac{SI_{new,s}}{SI_{current,s}} cv^2(\hat{Y}_{t,s})$
- Therefore, $cv^2(\sum_s \hat{Y}_{t,s}) = \sum_s \left(\frac{CLF_{new,s}}{n_{new,s}}\right)\left(\frac{1}{SI_{current,s}}\right) p_s^2 cv^2(\hat{Y}_{t,s})$
- The decision variables are the set of new state sample sizes $\{n_{new,s}\}$

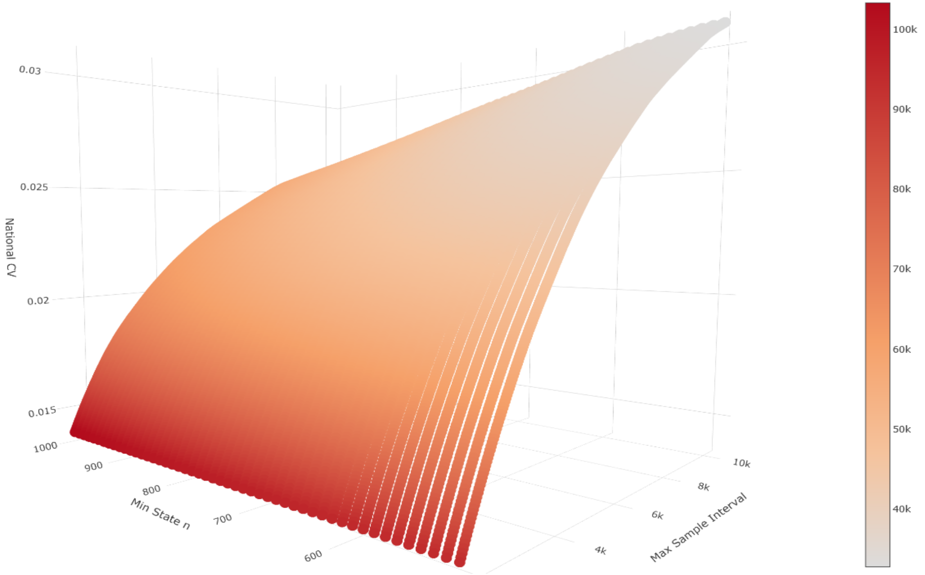
Defining the Sample Allocation Problem - Constraints

- Translate the national-level minimum detectable difference requirement into a *cv* requirement for the national unemployment total
  - Rottach and Erkens (2012) determine a modeled correlation between subsequent monthly unemployment rates - predict $corr(\bar{Y}_t, \bar{Y}_{t+1}) = 0.41$
  - Assuming a 6% unemployment rate, to detect a $\hat{\hat{Y}}_t - \hat{\hat{Y}}_{t+1} = 0.002$ for an $\alpha = 0.10$ requires a $cv(\hat{\hat{Y}}_t) = 0.0187$
  - Nation-level constraint, $cv^2(\sum_s \hat{Y}_{t,s}) \leq 0.0187^2$

Defining the Sample Allocation Problem - Constraints

- Translate the state-level *cv* requirement for the annual average monthly unemployment level into a *cv* requirement for the state monthly unemployment total
  - $cv_s^2(\hat{Y}_s) = \dfrac{.08^2}{0.71\alpha_s + 0.20(1-\alpha_s)}$ where $\alpha_s = \dfrac{V_{b,s}(\hat{Y})}{V_s(\hat{Y})}$
  - Note that Rottach and Erkins derive the 0.71 and 0.20 factors using a model-based approach to predict the between and within correlation component values for all pair-wise combinations of months within a 12 month period
- State-level constraints, $\{cv_s \leq cv_{s0}\}_{States}$
- Additional soft constraints
  - $n_s \geq n_0$
  - $SI_s \leq SI_0$
  - $n \approx 60,000$

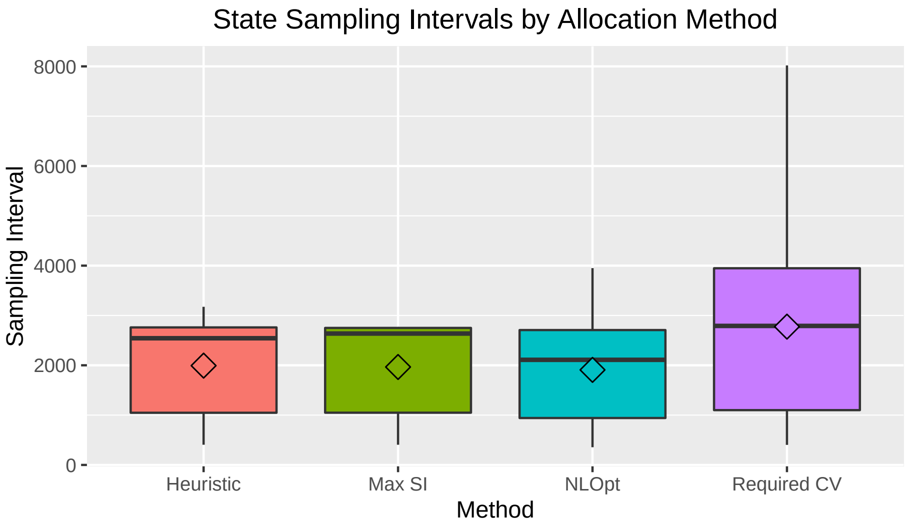
Defining the Sample Allocation Problem - Selecting the Soft Constraints



Sample Allocation Methods

- Nonlinear optimization algorithm (NLOpt - http://github.com/stevengj/nlopt)
  - Constrained optimization using the Augmented Lagrangian algorithm along with the Method of Moving Asymptotes (MMA)
- Maximum Sampling Interval (Max SI)
  - Iteratively decreases a "ceiling" for the sampling intervals to reduce the range of sample weights across the nation
- Greedy heuristic
  - Iteratively adds an additional sample to the stratum with the largest reduction in variance

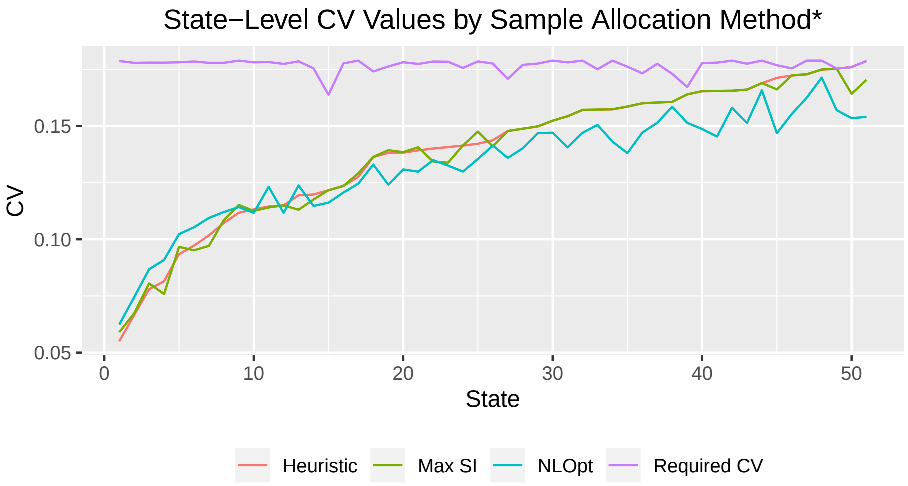
Results - Comparing Allocation Methods



Comparing Allocation Methods

	Sampling Int.			
Method	Max	Min	cv	$n_{cv=1.9}$
Required cv	402	8,019	2.532	–
NLOpt	355	3,951	1.937	62,500
Max SI	405	2,750	1.873	58,700
Heuristic	405	3,172	1.866	58,400

Comparing the State-Level Precision Results



*States are ordered by their CV values attained via the Heuristic method of allocation

Conclusions

- Model relating the state and national-level precision requirements works well...
  - Simultaneously optimizing on both requirements
  - Accounts for the correlational structure of the composite estimator
  - Useful tool for assessing current sample allocation
- Model limitations
  - Complex, may not generalize well to other surveys
    - Model or empirical-based correlation estimates required as inputs
    - Variance estimates rely on existing survey data
    - Assumes the global and domain estimates are equal
    - Univariate case with global and domain precision requirements
  - Assumes the first stage sample size is fixed

Conclusions{cont'd}

- Choice of allocation method depends on prioritization of a self-weighting design vs. desired precision levels
  - Maximum Sampling Interval method
    - ★ Good control over the self-weighting design properties
    - ★ Requires slightly more budget to meet the CV requirements
    - ★ Benefits of self-weighting - more relevant at the state level
  - Greedy Heuristic method
    - ★ Lower variance per unit cost - national level estimate
    - ★ Sacrifices some control of the self-weighting properties
  - Nonlinear Optimization Algorithm
    - ★ For the majority of states, effective at minimizing the state-level variances per unit cost
    - ★ More conservative allocation for states with smaller variances - penalized national-level precision result

Future Research

- Investigate whether we can apply the model and allocation to other surveys
- Refine or explore additional constraints, e.g., better control of interviewer workloads
- Can we generalize the approach to the multi-variate case?

John Chesnut
thomas.j.chesnut@census.gov