Domain Estimation and Successive Difference Replication

Tim Trudell*† Khoa Dong† Yang Cheng† Eric Slud†‡

Abstract

Successive Difference Replication (SDR) is a variance estimation method originally used for systematic samples, which is applied to complex surveys at the U.S. Census Bureau, including the Current Population Survey. Often we are interested in estimates for domains or subpopulations. In practice, replicate factors are assigned to the full sample. For SDR variance estimation of an estimated domain total, we use a subset of the full replicate factors. The subsetting of the domain on the sorted full data translates to a skip pattern. The effects of different skip patterns on the SDR variance error are examined in this paper via ideal superpopulation models and a simulation study.

Key Words: Variance Estimation, Successive Difference Estimation, Systematic Sampling, Domain Estimation, Weight Replication, Current Population Survey

1. Introduction

Successive Difference Replication (SDR) is used for variance estimation at the U.S. Census Bureau for many surveys, including the Current Population Survey (CPS). SDR is a replication method that assigns "replicate factors" (i.e., weight multipliers) to sorted sample respondent data. In CPS practice for calculating total variance, SDR factors are assigned to groups of four households in self-representing primary sampling units (PSU), sorted by geographic information, prior to knowing household eligibility and response status. These factors lead to replicate estimates, which are in turn used to calculate the variance estimator.

SDR was inspired as a replicate version of a modified form of successive difference estimator described in Wolter [1984]. For a full description of the SDR factor assignment method, see Fay and Train [1995].

This paper examines a common situation where estimation is done on a large domain[1], U_D, $U_D \subset U$. Let S be the sample from U and denote the sampled units from the domain, S_D, $S_D = U_D \cap S$. Commonly, SDR replicate weights are released for a whole sample for use in analysis. We examine the effect of using these replicate weights when analyzing totals on U_D.

When we estimated March 2018 variances for estimated total counts of responding and nonresponding households, we unexpectedly found them to be very different. If X is our response total, N our total population of households, then $N - X$ would be the nonresponse total. Calculating variance of X and $N - X$ should give a similar result if the variance of N is small. In this case, the estimated variances should

*Disclaimer: Any views expressed here are those of the authors and not necessarily those of the U.S. Census Bureau.

†U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

‡University of Maryland, Mathematics Department, College Park, MD 20742

[1]Our interest in domain estimation arose from the need to produce household response and nonresponse total variance estimates. The domain we are considering is a time "slice" of one month from the larger annual sample which covers many months. We can view this annual sample as sampling from a product space of time-specific universes, U_l at month l, so that the notational yearly-sample universe is $U = \prod_{l=1}^{l'} U_l$, where l' is the last month considered by the sample.

be reasonably similar. This prompted investigation into the effects on variance estimation of domain subsetting patterns within the sorted sample.

1.1 Simulation Scenarios and Actions

In this paper we focus on two scenarios for the relation between domain and whole-sample indexing and two actions on replicate factors. Here "action" refers to a strategy of SDR replicate factor assignment.

Our goals are to provide evidence that the mechanism of replicate factor assignment explored in these scenarios explains the observed discrepancy in the estimated variances of the estimated counts of household responders and nonresponders, detailed in section 2, as well to draw attention to the broader issue of biases in SDR variance estimators for subdomains.

These two scenarios and two actions form the basis for the simulations. Let $i = 1, \ldots, n$ index the sample, S. Let $|S_D| = k$. Let a be the starting index for S_D. In the first scenario, S_D is of the form $\{a + l : l = 0, 1, 2, \ldots, k - 1\}$. In the second scenario, S_D is of the form $\{a + 2l : l = 0, 1, \ldots, k - 1\}$. In both scenarios, a is restricted so that S_D will always be of size k. For each scenario, we can choose from two actions: retain the original factors or reassign the SDR factors separately within the domain and its complement.

This paper is organized around some basic mathematical observations, using the mathematical equivalence demonstrated in Fay and Train [1995] then following up with simulations on internal CPS replicates and pure data simulations to provide a mechanism to explain the discrepancy in variance estimates for estimated responders and nonresponders.

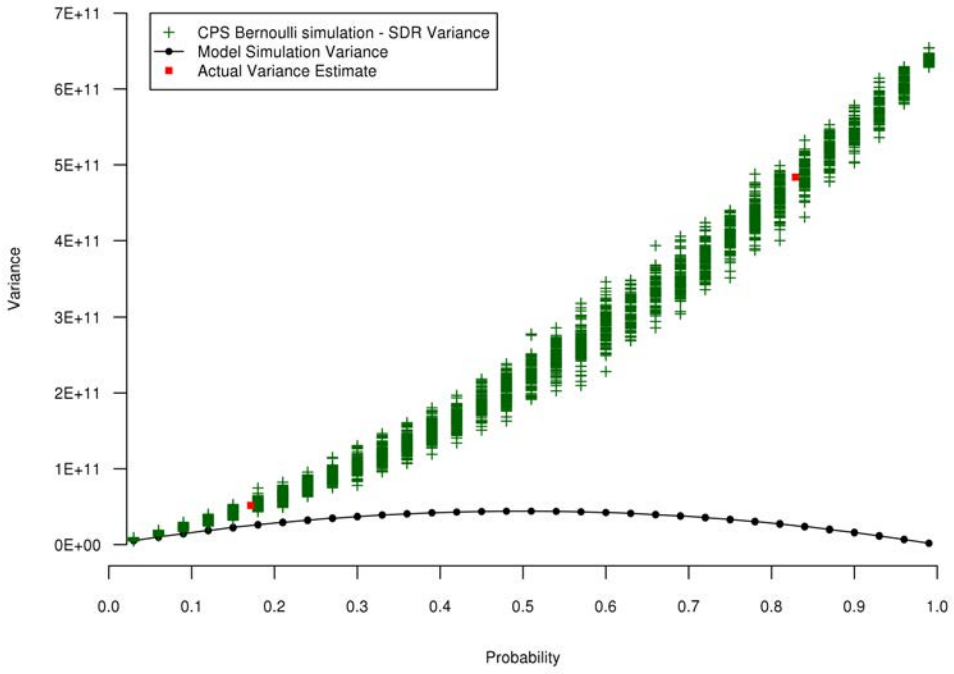
2. Background

The motivation for this work was an investigation into the variance of the CPS household response total. Our work was specifically restricted to self-representing PSUs where systematic sampling was the only stage of sampling. For a random variable X and constants a and b, we know that the variance of X, $V(a+bX) = b^2V(X)$. Letting $b = -1$ and $a = N$, we expect that $V(X) = V(N - X)$, where N is the total households, and the corresponding variance estimates should be approximately equal.

We ran a Bernoulli simulation on the March 2018 data, varying p across the interval $(0, 1)$. Figure 1 shows the variance estimates for March 2018 household response and nonresponse total, variance estimates from Bernoulli simulation data using March 2018 SDR replication factors (CPS Bernoulli simulation), and the model variance from the Bernoulli simulation. The red squares are the actual household response and nonresponse total variance estimates for March 2018. The green plus signs are individual simulation runs for a given probability, p. For each value of p, the simulation was run 100 times.

The CPS Bernoulli simulation shows that there is something unexpected in the CPS SDR replicate factors, hereafter referred to as the *variance anomaly*. The March 2018 SDR variance estimates for household response and nonresponse total lie within the range of Bernoulli simulation values. We note the lack of symmetry expected from the SDR variance estimator as p varies from 0 to 1 with the CPS Bernoulli simulation diverging from the model variance.

Figure 1: March 2018 CPS Bernoulli Simulation SDR Variance, Actual Household Response and Nonresponse Estimates, and Model Variance



Actual Variance Estimates (red squares): (left square) estimated variance of household nonresponse total (right square) estimated variance of household response total.
CPS Bernoulli simulation (Green Pluses): single Bernoulli simulations from CPS March 2018 replicate factors, 100 for each p.
Model variances (black circles): Bernoulli variances from the simulation.

3. Methods

3.1 Definitions

Let n be our sample size. Let i index our sample, $i = 1, \ldots, n$. Let y_i be the value of the ith observation. Following similar population assumptions in Huang and Bell [2009], we assume that y_i values are generated independently and identically distributed from density $f(y)$, with common mean, μ, and common variance, σ^2. Let w_i be a final survey weight for observation i. In this paper, we assume $w_i = w$ for all i, assuming our sample size, n, is fixed. Let our estimate be $\hat{\theta}_0 = \sum_{i=1}^{n} w_i y_i = w \sum_{i=1}^{n} y_i$. Let f_{ir} be the rth replicate factor for the ith observation, to be defined later. Then our replicate estimate $\hat{\theta}_r$ is defined as $\sum_{i=1}^{n} w_i f_{ir} y_i = w \sum_{i=1}^{n} f_{ir} y_i$. Finally, let our replicate variance estimate, $\hat{V}$, be

$$\hat{V} = \frac{1}{2\gamma^2 R} \sum_{r=1}^{R} \left[\hat{\theta}_r - \hat{\theta}_0\right]^2,$$

where γ is a parameter such that $0 < \gamma < \infty$. In Fay and Train [1995], γ was chosen to be $\frac{1}{2\sqrt{2}}$ which is the value we adopt in this paper.

As above, let S_D denote the set of indices of sampled units within our domain of interest. Let $z_i = y_i I_{\{i \in U_D\}}$.

We will examine the effect of different index patterns for the set S_D and the effect on the estimator, $\hat{V}^D$,

$$\hat{V}^D \coloneqq \frac{1}{2\gamma^2 R} \sum_{r=1}^{R} \left[\hat{\theta}_r^D - \hat{\theta}_0^D\right]^2,$$

where $\hat{\theta}_0^D = \sum_{i \in S_D} w_i y_i = \sum_{i=1}^{n} w_i z_i$ and $\hat{\theta}_r^D = \sum_{i \in S_D} w_i f_{ir} y_i = \sum_{i=1}^{n} w_i f_{ir} z_i$.

Let h_{ir} be the ith row and rth column element of a Hadamard matrix of order R, $R > n + 2$. A Hadamard matrix, H, is a square matrix of order $4l$ for some integer l, with entries either -1 or 1. The defining property is that $HH^T = 4lI_{4l}$, where I_{4l} is the identity matrix of order $4l$ and H^T is the transpose of H. Finally, define f_{ir} as

$$f_{ir} \coloneqq 1 + \gamma(h_{i+1,r} - h_{i+2,r})$$

and for $i = n$,

$$f_{nr} \coloneqq 1 + \gamma(h_{n+1,r} - h_{2,r}),$$

where γ is a constant such that $0 < \gamma < \infty$. This definition is the same as in Fay and Train [1995], but in practice $n > R$. In surveys such as CPS, we recycle rows from a smaller Hadamard matrix to assign factors to the whole sample. One reference for insight into this row recycling is Ash [2014]. Unless otherwise specified, we will use a Hadamard matrix of dimension $R > n + 2$.

3.2 Some Algebraic Results

With the setup described in section 3.1, we now define two different structures for S_D and then evaluate their expected values under the described superpopulation model.

Block Scenario: $S_D = \{a + l \mid l = 0, 1, \ldots, k-1\}$ for some $a = 1, 2, \ldots, n - k + 1$.

Alternating Scenario: $S_D = \{a + 2 \cdot l \mid l = 0, 1, \ldots, k-1\}$ for some $a = 1, 2, \ldots, n - 2k + 2$.

For these two scenarios, we can calculate the expected relative error of the SDR variance estimator versus the superpopulation variance. Expected relative error in this paper is defined as follows[2]:

$$\frac{E(\hat{V}_{\text{SDR}}) - V(\hat{\theta}^D)}{V(\hat{\theta}^D)}$$

where expectations are done under the superpopulation model. Let $\hat{V}_{\text{SDR}}^D$ denote the SDR variance estimate for the sample domain conditioned on a finite population. The variance of the estimator is the expected superpopulation variance, $V(\hat{\theta}^D) = kw^2\sigma^2$. The expectation is taken with respect to the superpopulation density. For details, see Appendix A.

[2]The expressions $E(\hat{V}_{\text{SDR}})$ and $V(\hat{\theta}^D)$ are with respect to the superpopulation model, not design. The expectation is approximately equal to the anticipated variance of Isaki and Fuller [1982] where we assume that our estimator $\hat{\theta}$ is design unbiased and that $\hat{V}_{\text{SDR}}$ is approximately equal to the design variance.

The expected relative error for the *block scenario* is (also found in Huang and Bell [2009]):

$$E\left(\frac{\hat{V}^D_{\text{SDR}} - kw^2\sigma^2}{kw^2\sigma^2}\right) = \frac{\mu^2}{k\sigma^2} \tag{1}$$

For the *alternating scenario* it is:

$$E\left(\frac{\hat{V}^D_{\text{SDR}} - kw^2\sigma^2}{kw^2\sigma^2}\right) = \frac{\mu^2}{\sigma^2} \tag{2}$$

The major difference is that in the block scenario we have the expected relative error, for fixed superpopulation parameters, of the order $\frac{1}{k}$ while for the alternating scenario we have no such improvement with increasing size, as the error is constant.

3.3 Simulation

In CPS, SDR factors are assigned to a whole annual sample. When a particular month's data is estimated, the appropriate factors are subsetted from the annual sample, then run through the weighting process and variance estimation. We can think of one month's data as approximately 1/12 of the annual sample. Simulation 1 explores the skipping mechanism and influence of our actions on the SDR replicates. Simulation 2 seeks to explain the CPS Bernoulli simulation as function of the skip pattern. Both simulations do not involve CPS data.

For the action of retaining the original SDR factors, we can use equations 1 and 2 with Binomial(m,p) parameters to get for the block scenario:

$$E\left(\frac{\hat{V}^D_{\text{SDR}} - kw^2\sigma^2}{kw^2\sigma^2}\right) = \frac{mp}{k(1-p)}$$

and

$$E\left(\frac{\hat{V}^D_{\text{SDR}} - kw^2\sigma^2}{kw^2\sigma^2}\right) = \frac{mp}{1-p}$$

for the alternating scenario. Thus we expect our relative errors for both scenarios to grow to infinity as p approaches 1.

3.3.1 Simulation 1

The simulation explores the combination of two scenarios and two actions described in section 1.1 for a total of four situations. We will compare and contrast the four combinations of scenarios and actions with respect to the expected relative error defined earlier.

For each set of superpopulation parameters, we simulate a sample of size 12,000 with a Bernoulli response. The SDR factors are then assigned to the sample of 12,000. Then we subset 1,000 indices as our domain for the *block* and *alternating* scenarios with independent random start points for each. Each set of 1,000 rows for the *block* and *alternating* scenarios are independent. After subsetting, S_D, we choose our action: either retain the original SDR factors or create new SDR factors within the ordered sample of domain units. We then take averages across the simulations for each of the scenarios and then compute the relative error of the SDR variance estimate.

3.3.2 Simulation 2

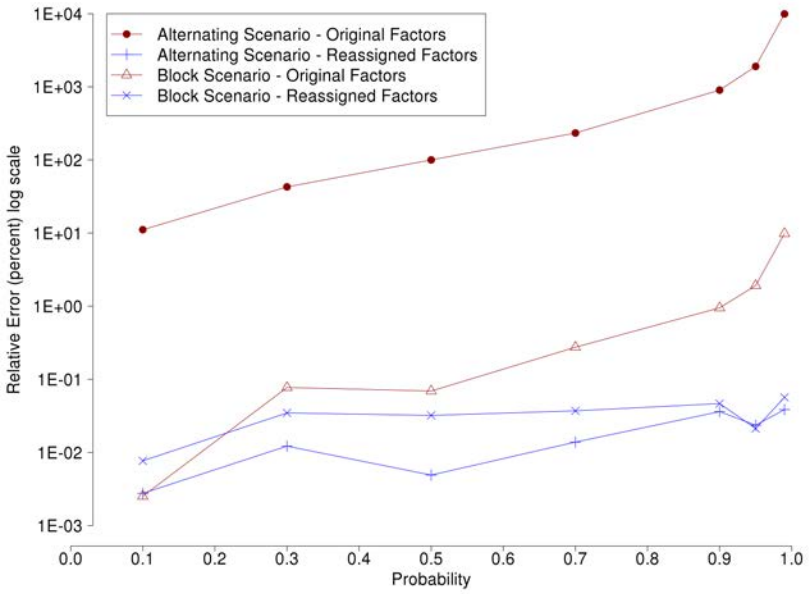
The second simulation refines simulation 1 in order to better mimic CPS SDR factor assignment in order to explain the behavior of the CPS Bernoulli simulation, observed in Section 2. We first group households into groups of four and then assign factors to groups, rather than to each individual household. The other potential change is in the Hadamard matrix dimension and algorithm for row assignment. For one version of simulation 2, we retain the original Fay-Train algorithm that requires a Hadamard matrix of order $R > n + 2$, where $n = 12,000$. In the second version, we use a Hadamard matrix of order 160 (following CPS practice), and recycle the rows mimicking the CPS production algorithm for row assignment. We then compare these simulations with the CPS Bernoulli simulation to see how well this mechanism can explain the observed behavior.

4. Results

4.1 Simulation 1

Figure 2 shows the results of simulation 1, plotting relative error (in percent) on a log scale versus our superpopulation parameter, p. The Monte Carlo simulation error was on the order of 10^{-2} so some of the "signal" in the two reassigned factors plots is merely simulation noise. The scenario that performs the worst is the alternating scenario with the original factors. Here, even when we have p around 50%, we have an expected relative error of around 100% which continues to grow as p approaches 1, matching results from section 3.2 in equation 2. For the block scenario with original factors, the situation is better, with 1% relative error around $p = 0.9$. This is because the size of the domain improves the error. If you have this skip pattern and a domain sample size of roughly 1,000, the relative error is not bad unless p is close to 1. That is expected due to the result in equation 1 in section 3.2. Finally, if we choose to reassign SDR factors, the relative error becomes very small for each scenario, across all values of p.

Figure 2: Comparison of two scenarios and two actions: Relative Error of Variance Estimate (log scale) versus Probability P



Simulation Setup Each point represents the average relative error across 1,000 simulations from a full sample of size 12,000 and a domain sample size of 1,000, for the given value p. Lines simply connect simulation point estimates to aid the viewer.
Alternating Scenario - Original Factors High relative error that grows quickly as p increases.
Alternating Scenario - Reassigned Factors Low relative error (within simulation error) for all p.
Block Scenario - Original Factors Low relative error when p is small but grows larger as p, accelerating towards infinity. Sample domain size = 1,000 reduces relative error.
Block Scenario - Reassigned Factors Low relative error (within simulation error) for all p.

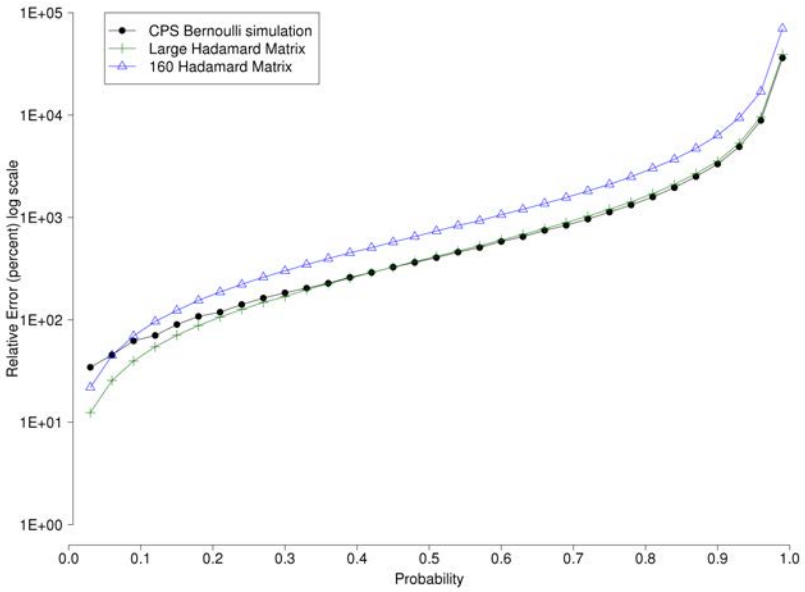
The most conservative practice would be to always reassign SDR factors to whatever analytical subset you considered. If you knew your data had the block skip pattern and had some tolerance for variance estimation error and a reasonable sample size, you could accept the original factors. A typical domain sample will not generate a skip pattern resembling blocks, but will be more similar to the alternating pattern. Here our recommendation is to reassign SDR factors rather than use the original ones. This is really only feasible if the number of domains for analysis is relatively small or the computational cost for each domain is small.

4.2 Simulation 2

We have explored the consequences of our actions under the two skip pattern scenarios. We need to examine how well this mechanism explains the *variance anomaly* found in Figure 1. Figure 3 plots the relative error in variance estimates for the CPS Bernoulli simulation with the two versions of Simulation 2. Both versions of simulation 2 assume the alternating skip pattern scenario and the retain SDR factor action. We denote the version using the Fay-Train assignment with the Hadamard matrix with order $R > n + 2$ as the Large Hadamard Matrix version. The version that uses the 160 order Hadamard matrix mimicking CPS SDR factor assignment is denoted as 160 Hadamard Matrix.

With the two versions of Simulation 2, because their plotted relative errors in Figure 3 are very close, we come close to predicting the *variance anomaly*, except when $p < 0.2$, when there is an effect not yet explained. The alternating scenario skip pattern accounts for much of the variation in the data. The effect of using the 160 order Hadamard matrix with row recycling versus the full Hadamard matrix is a much smaller effect, shown by the differences between the two Simulation 2 versions.

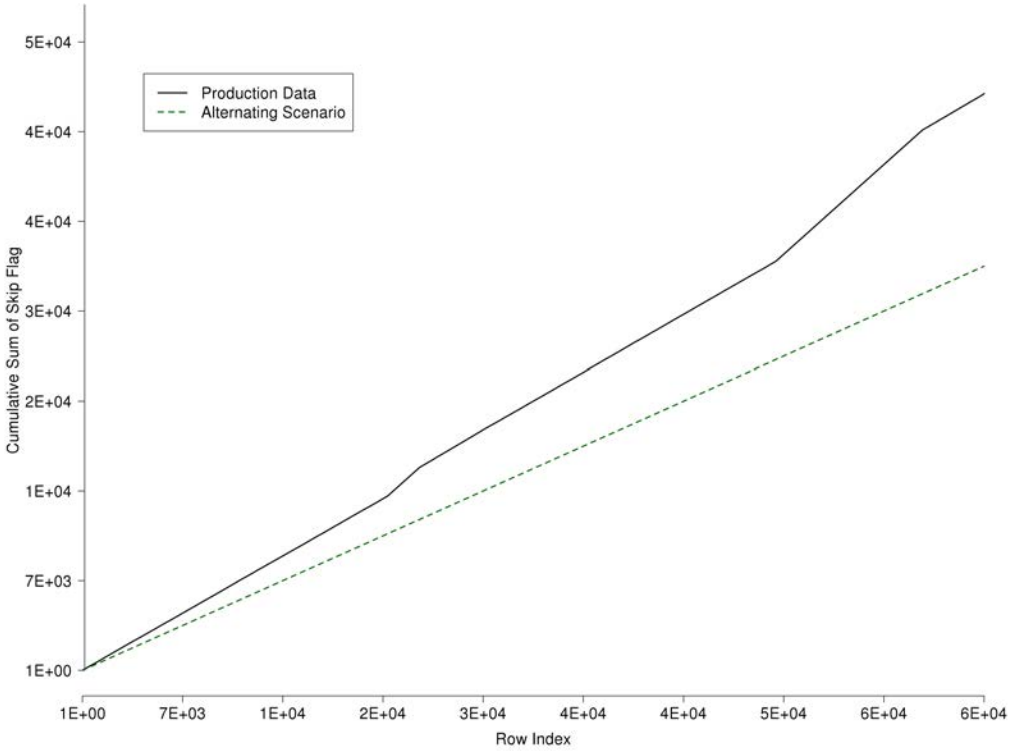
Figure 3: CPS Bernoulli Simulation versus Alternating Skip Pattern - Relative Error (log scale) versus p



Simulation Setup Each point represents the average relative error across 1,000 simulations from a full sample of size 12,000 and a domain sample size of 1,000, for the given value p. Lines simply connect simulation point estimates.
CPS Bernoulli Simulation The variance anomaly we wish to explain with the SDR factor assignment mechanism via simulation 2.
Large Hadamard Matrix version Using Alternating skip pattern scenario, original factors. Simulation 2 using a large (order greater than 12,000) Hadamard matrix, each SDR factor is assigned to groups of four households to mimic CPS.
160 Hadamard Matrix version Using Alternating skip pattern scenario, original factors. Simulation 2 using a small (160 by 160 Hadamard matrix, using row assignment algorithm similar to CPS production row assignment algorithm.

We see that the alternating skip pattern scenario with the superpopulation model gives a reasonable approximation of the variance anomaly seen with the skip pattern observed in the March 2018 CPS data. To measure the similarity between the actual CPS skip pattern and the one in the simulated alternating scenario, we created a flag variable for every row skipped in the CPS SDR factor file. Figure 4 shows a cumulative sum of these flag variables for the actual CPS SDR factor file and the alternating scenario. The alternating scenario is the straight line, with the cumulative sum increasing by 1 every two rows. We can see that in the actual data, we are skipping more frequently than the alternating scenario with some variation in the number of successive rows skipped.

Figure 4: Cumulative Sum of Skip Flag for March 2018 SDR Replicate Factors Versus Idealized Alternating Scenario



Because the relative error of pure simulation models from Simulation 2 closely tracks the CPS Bernoulli simulation from Figure 1, the effect of the alternating scenario plausibly explains the variance anomaly seen in Figure **??**, despite the many simplifying assumptions made in the simulations versus the actual data.

5. Conclusions and Future Work

Based on the expected relative error metric, reassigning SDR factors to the domain sample is preferred regardless of skip pattern. When we assign new SDR factors for our sample domain, S_D, we remove any artificial error caused by subsetting from preassigned SDR factors.

Subsetting domains for variance estimation after SDR factors have been assigned can lead to large error in variance estimation, depending on the pattern of subset-

ting. In general, the skip pattern will be closer to the alternating row scenario than to the block scenario. For this reason, our simulations suggest potential errors in variance estimation if SDR factors are assigned before domain subsetting.

Our recommendation would be to apply SDR factors only within the domain of interest. This factor reassignment could be implemented in a software package. Unless computational power is very cheap, this solution may not be feasible if a large number of domains need to be analyzed quickly. The additional burden of incorporating this approach could add maintenance overhead for the survey analysis infrastructure.

For CPS[3], specifically, we can reduce variance errors by altering the SDR row assignment sort to ensure that a single month's data follows the *block* scenario rather than the *alternating* scenario. This modification still would not prevent variance estimation biases from arising from analysis done with different types of domains.

The simulation differs from reality in a few ways. In CPS practice, SDR factors are assigned to groups of four households prior to knowing eligibility and response status. Both of these would result in the inclusion of extra nonrespondent cases before we even consider our domain, U_D. This simulation also assumes no calibration or non-response adjustments to weights and assumes a self weighting scheme like CPS. Additionally, our data analysis only focuses on self-representing (SR) primary sampling units (PSU). These comprise approximately three quarters of all households. Estimating the effect of this bias on the total CPS variance would be useful. For our preliminary variance results using CPS base-weights which have small variation between states and zero within a state, this was a reasonable assumption. We did some simulations for final weight replicates and still saw the effect described in this paper, but further work is necessary to assess it properly.

We did explore other superpopulation models (normal, AR1, etc) but the conclusions there were similar to the Binomial case. In all cases, we assumed independence in our y_i's in the superpopulation model. Relaxing that assumption would be another avenue for future exploration.

References

Stephen Ash. Using successive difference replication for estimating variances. *Survey Methodology*, 40(1):47–59, 2014.

Robert E. Fay and George Train. Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. In *Proceedings of the Section on Government Statistics, American Statistical Association, Alexandria, VA*, pages 154–159, 1995.

Elizabeth T. Huang and William R. Bell. A simulation study of the distribution of fay's successive difference replication variance estimator. In *JSM Proceedings, Survey Research Methods Section*, 2009.

Cary T. Isaki and Wayne A. Fuller. Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96, 1982.

Kirk M. Wolter. An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, 79(388):781–790, 1984.

[3]Here we again are considering our universe, U, as a product of monthly universes which we draw sample for each individual monthly universe and our domain is just one monthly universe.

A. Derivation of Expected Relative Error for Block and Alternating Scenarios

A.1 Block Scenario

Lemma A.1 *Let i index the sample S, $i = 1, \ldots, n$. Let y_i be distributed iid for some distribution function, f, and assume $E(y_i) = \mu$ and $Var(y_i) = \sigma^2$. Let $S_D \subset S$, where $S_D = \{a + l : l = 0, 1, \ldots, k-1\}$, $a \in \{1, \ldots, n-k+1\}$. Then the expected relative error is*

$$E\left(\frac{\hat{V}_{SDR} - kw^2\sigma^2}{kw^2\sigma^2}\right) = \frac{\mu^2}{k\sigma^2}.$$

In the first scenario, all indices are contiguous while in the second scenario, indices skip every other row. For the *block* scenario, after plugging in for $\hat{\theta}_r^D$, etc and expanding the square we get:

$$\hat{V}^D = \frac{w^2}{2\gamma^2 R}\sum_{r=1}^{R}\left[\sum_{i=1}^{n}(f_{ir}-1)^2 z_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} z_i z_j (f_{ir}-1)(f_{jr}-1)\right]$$

Since $z_j = 0$ when $j \notin S_D$, we can simplify this expression:

$$\hat{V}^D = \frac{w^2}{2\gamma^2 R}\sum_{r=1}^{R}\left[\sum_{i=a}^{a+k-1}(f_{ir}-1)^2 y_i^2 + 2\sum_{i=a}^{a+k-2}\sum_{j=i+1}^{a+k-1} y_i y_j (f_{ir}-1)(f_{jr}-1)\right]$$

Rearranging the sums to bring the sum over r inside, we get

$$\hat{V}^D = \frac{w^2}{2\gamma^2 R}\left[\sum_{i=a}^{a+k-1} y_i^2 \sum_{r=1}^{R}(f_{ir}-1)^2 + 2\sum_{i=a}^{a+k-2}\sum_{j=i+1}^{a+k-1} y_i y_j \sum_{r=1}^{R}(f_{ir}-1)(f_{jr}-1)\right]$$

For the first sum, it reduces to $2\gamma^2 R$. For the second term,

$$\sum_{r=1}^{R}(f_{ir}-1)(f_{jr}-1)$$

it depends on i and j. When $j > i+1$, the sum is zero. When $j = i+1$, the sum becomes $-\gamma^2 R$. This leads to

$$\hat{V}^D = w^2\left[\sum_{i=a}^{a+k-1} y_i^2 - \sum_{i=a}^{a+k-2} y_i y_{i+1}\right]$$

Using the superpopulation assumptions, we can now compute an expected relative error for the variance estimator, since we know the variance under the superpopulation model (common mean μ, variance σ^2, and uncorrelated values). Taking expectations over the superpopulation we get

$$E\hat{V}^D = kw^2(\sigma^2 + \mu^2) - w^2(k-1)\mu^2 = kw^2\sigma^2 + w^2\mu^2.$$

We can also calculate the expected variance given the superpopulation, which becomes $V(\hat{\theta}) = kw^2\sigma^2$. Putting this together our expected relative error in the block scenario is:

$$E\left(\frac{\hat{V}_{\text{SDR}} - kw^2\sigma^2}{kw^2\sigma^2}\right) = \frac{\mu^2}{k\sigma^2}.$$

Thus as our domain size, k, grows larger, our expected relative error decreases on the order of k^{-1} for a fixed superpopulation.

A.2 Alternating Scenario

Lemma A.2 *Let i index the sample S, $i = 1, \ldots, n$. Let y_i be distributed iid for some distribution function, f, and assume $E(y_i) = \mu$ and $Var(y_i) = \sigma^2$. Let $S_D \subset S$, where $S_D = \{a + 2l : l = 0, 1, \ldots, k-1\}$, $a \in \{1, \ldots, n - 2k + 2\}$. Then the expected relative error is*

$$E\left(\frac{\hat{V}_{SDR} - kw^2\sigma^2}{kw^2\sigma^2}\right) = \frac{\mu^2}{\sigma^2}.$$

In this scenario, the indexing set, S_D, skips every other row after starting at a particular index, a. For clarity, we assume $a \neq 1$ and $a \neq n$. Since the successive difference estimator is invariant under a shift in labels, this can always be avoided.

$$S_D = \{a + 2 \cdot l \mid l = 0, 1, \ldots, k-1\}$$

for some $a = 1, 2, \ldots, n - 2k + 2$.

Borrowing notation from the previous section, we have

$$|\hat{V}^D| = \frac{w^2}{2\gamma^2 R}\left[\sum_{l=0}^{k-1} y_{a+2l}^2 \sum_{r=1}^{R}(f_{(a+2l)r} - 1)^2 + 2\sum_{l=0}^{k-2}\sum_{m=l+1}^{k-1} y_{a+2l}y_{a+2m}\sum_{r=1}^{R}(f_{(a+2l)r} - 1)(f_{(a+2m)r} - 1)\right].$$

The term $\sum_{r=1}^{R}(f_{(a+2l)r} - 1)^2 = 2\gamma^2 R$ as before. The cross term $\sum_{r=1}^{R}(f_{(a+2l)r} - 1)(f_{(a+2m)r} - 1)$, always vanishes due to the orthogonality of distinct rows of the Hadamard matrix. Thus this becomes

$$\hat{V}^D = w^2 \sum_{l=0}^{k-1} y_{a+2l}^2.$$

With the superpopulation assumptions, we can compute the expected value as:

$$E\hat{V}^D = w^2 k(\sigma^2 + \mu^2).$$

Again the variance is $kw^2\sigma^2$ and the relative variance is

$$\frac{\mu^2}{\sigma^2}.$$

Here we do not gain any reduction in error as our domain size, k, increases but are entirely dependent on the superpopulation parameters.

B. Derivation of Fay-Train Results

Assume we have sampled a systematic sample of size n. Let i, $i = 1, \ldots, n$, index the observations from the sample. Let y_i be the observed measurement. Let w_i be a survey weight and denote the sum $\sum_{i=1}^{n} w_i = W$. Let $\hat{\theta}_0 = \sum_{i=1}^{n} w_i y_i$. We wish to estimate the variance of $\hat{\theta}_0$.

Let H be a Hadamard matrix of order $R > n+2$, for some integer R and let h_{ij} represent the ith row and jth column from the matrix. Each $h_{ij} \in \{-1, 1\}$ and for $i \neq j$, $\sum_{k=1}^{R} h_{ik} h_{jk}$ equals zero. If $i = j$, $\sum_k h_{ik}^2 = R$. In other words, $HH^T = RI$ where I is the identity matrix.

Denote the Successive Difference Replication weighting factor as f_{ir} for the ith row and rth replicate. For $i < n$, let

$$f_{ir} = 1 + \gamma(h_{i+1,r} - h_{i+2,r})$$

and for $i = n$,

$$f_{nr} = 1 + \gamma(h_{n+1,r} - h_{2,r}),$$

where γ is a constant such that $0 < \gamma < \infty$. In Fay and Train [1995], γ is set to $\frac{1}{2\sqrt{2}}$. Let $\hat{\theta}_r = \sum_{i=1}^{n} f_{ir} w_i y_i$.

Finally, let our replicate variance estimate be

$$\frac{1}{2\gamma^2 R} \sum_{r=1}^{R} \left[\hat{\theta}_r - \hat{\theta}_0\right]^2.$$

Plugging in for $\hat{\theta}_r$ and $\hat{\theta}_0$ and expanding the square, we get

$$\frac{1}{2\gamma^2 R} \sum_{r=1}^{R} \left[\sum_{i=1}^{n} w_i^2 y_i^2 (f_{ir} - 1)^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} y_i w_i y_j w_j (f_{ir} - 1)(f_{jr} - 1) \right]$$

Since all sums are finite, we can exchange summations to sum over r first.

$$\frac{1}{2\gamma^2 R} \left[\sum_{i=1}^{n} w_i^2 y_i^2 \sum_{r=1}^{R} (f_{ir} - 1)^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} y_i w_i y_j w_j \sum_{r=1}^{R} (f_{ir} - 1)(f_{jr} - 1) \right]$$

Substituting for f_{ir}, taking care about the special definition of f_{nr}, we get

$$\begin{aligned}
&\frac{1}{2\gamma^2 R} \Bigg[\sum_{i=1}^{n-1} w_i^2 y_i^2 \gamma^2 \sum_{r=1}^{R} (h_{i+1,r}^2 + h_{i+2,r}^2 - 2h_{i+1,r} h_{i+2,r}) \\
&+ w_n^2 y_n^2 \gamma^2 \sum_{r=1}^{R} \left(h_{n+1,r}^2 + h_{2,r}^2 - 2h_{n+1,r} h_{2,r}\right) \\
&+ 2 \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} y_i w_i y_j w_j \gamma^2 \sum_{r=1}^{R} (h_{i+1,r} h_{j+1,r} - h_{i+1,r} h_{j+2,r} - h_{i+2,r} h_{j+1,r} + h_{i+2,r} h_{j+2,r}) \\
&+ 2 \sum_{i=1}^{n-1} y_i w_i y_n w_n \gamma^2 \sum_{r=1}^{R} (h_{i+1,r} h_{n+1,r} - h_{i+1,r} h_{2,r} - h_{i+2,r} h_{n+1,r} + h_{i+2,r} h_{2,r}) \Bigg]
\end{aligned}$$

Many of these expanded terms can simplify, using properties of the Hadamard matrix. A general cross term, $\sum_{r=1}^{R} h_{m,r}h_{l,r}$, equals zero if $m \neq l$ and R if $m = l$. The first two terms in the sum simplify straightforwardly. For the third term,

$$2\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} y_i w_i y_j w_j \gamma^2 \sum_{r=1}^{R} (h_{i+1,r}h_{j+1,r} - h_{i+1,r}h_{j+2,r} - h_{i+2,r}h_{j+1,r} + h_{i+2,r}h_{j+2,r})$$

the four component sums vanish depending on the values of i and j. For $\sum_{r=1}^{R} h_{i+1,r}h_{j+1,r}$ and $\sum_{r=1}^{R} h_{i+2,r}h_{j+2,r}$, these sums are nonzero when $i = j$. This is impossible because we have $i + 1 \leq j \leq n - 1$. Thus these two components are always zero. For $\sum_{r=1}^{R} h_{i+1,r}h_{j+2,r}$, it is nonzero when $i + 1 = j + 2$. This implies $j = i - 1$ which is impossible since $j \geq i + 1$. For the third, $\sum_{r=1}^{R} h_{i+2,r}h_{j+1,r}$, this is nonzero only when $i + 2 = j + 1$ or $j = i + 1$. This three of the four components of the third sum are always zero, and one is only nonzero when $j = i+1$. The third term simplifies to

$$-2\sum_{i=1}^{n-2} y_i w_i y_{i+1} w_{i+1} \gamma^2 R.$$

The fourth component is similar to the third described above. The only nonzero component is $\sum_{r=1}^{R} h_{i+1,r}h_{2,r}$ when $i = 1$.

$$\frac{1}{2\gamma^2 R}\left[\sum_{i=1}^{n-1} w_i^2 y_i^2 \gamma^2 (R + R - 2\cdot 0) + w_n^2 y_n^2 \gamma^2 (R + R - 2\cdot 0) \right.$$
$$\left. -2\left(\sum_{i=1}^{n-2} y_i w_i y_{i+1} w_{i+1} \gamma^2 R\right) - 2y_1 w_1 y_n w_n \gamma^2 R\right]$$

Finally,

$$\sum_{i=1}^{n} w_i^2 y_i^2 - \left(\sum_{i=1}^{n-1} y_i w_i y_{i+1} w_{i+1}\right) - y_1 w_1 y_n w_n$$

equals

$$\frac{1}{2}\left[(w_1 y_1 - w_n y_n)^2 + \sum_{i=2}^{n} (y_i w_i - y_{i-1} w_{i-1})^2\right].$$