CDC's National Center for Health Statistics

Design-Based Methods for State-Level Estimation under Three-Year Data Pools

William Waldron, PhD
Mathematical Statistician
Division of Research and Methodology
National Center for Health Statistics

August 8, 2024
Survey Data Collection, Estimation, and Disclosure Limitation Methods
Joint Statistical Meetings

Objective

Many federal nation-wide household surveys are increasingly being used for state-level estimation.

The reliability of state-level estimates differs from national and domain estimation due to lower degrees of freedom.

We explore a series of cost-effective design choices that can collectively impact the reliability of both larger and smaller states.

Motivation 1

Pooling years from cross-sectional surveys can have diminished impact for two-stage sample designs if the PSUs are static across years.

Many designs will fix PSU geographies for a period of ten years, followed by a dependent sample selection that promotes selecting the same PSUs (Method of Maximum Overlap).

When there are cost savings in moving year-to-year PSUs to adjacent areas, the *PSU Random Walk* can balance independent PSUs and recruitment costs.

Motivation 2

Small Area Estimation under hierarchical Bayesian models can improve reliability at the cost of some added model bias.

$$\hat{y}_i = \lambda y_i + (1 - \lambda) X_i{}'\beta$$

Improved design variance and variance estimators can…

- reduce reliance on the synthetic SAE estimator $X_i{}'\beta$ (or other predictors).
- better satisfy the SAE assumption of known sampling variances used in $\lambda(.)$
- lower the overall model mean-square error of $\hat{y}_i$

Contents

- **Single-Year Design Changes**
  - Systematic Sampling
  - Weighting Issues
- **PSU Cycling under three-year pooled data**
  - Taylor Series Variance Estimation for Survey Data
  - Two examples with double samples
- **PSU Random Walk**
  - Markov Chains for Adjacent PSU Sampling
  - Building an appropriate probability transition matrix

Single-Year Sample Design Changes

Systematic Sampling and Weighting

Single-Year Design Changes

- **Systematic Sampling**
  - Use a *continuous* variable.
  - Avoid “implicit stratification”.
  - Consider for *all* stages.

- **Weighting (at the State Level)**
  - Balance standard errors with # controls.
  - Consider weight trimming with caution.
  - Reweight under three-year pooled data.

- **Other Issues**
  - Keep # Strata low.
  - Keep # PSUs high.
  - Proper Estimation of # Degrees of Freedom when PSUs sizes differ.
  - Consider using FPC under higher first-stage sampling rates.

Systematic Sampling of ABS Frames

While SAE can use *auxiliary information* on the back end for post data collection modeling, it can also be used on the front end to improve the **direct estimate**.

The best sorting variable(s) will be correlated with the main survey response of interest.

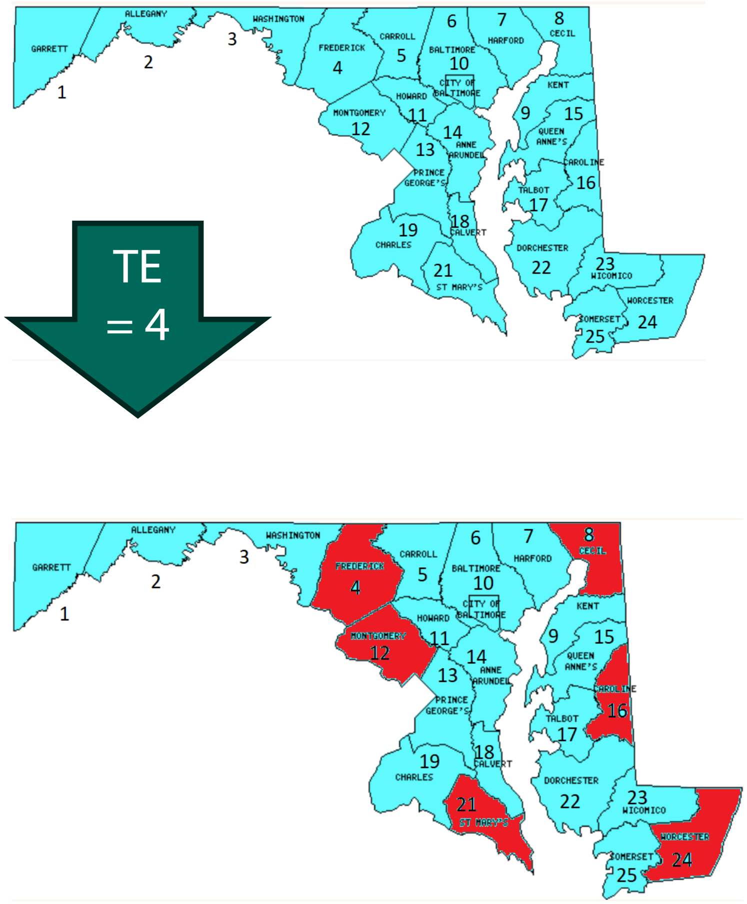
- By acquired frame information (can be *inaccurate).*
- Sorting for *spatial* dispersion using serpentine labeling.
- Sorting geographies by survey information (e.g., ACS):
  - Health Insurance Access
  - Poverty Status
  - CDC’s Social Vulnerability Index

Serpentine Labeling

Serpentine labeling maps the plane onto the real line.

Used for labeling Census Tracts and Blocks by Census Bureau.

Imperfect because the plane *cannot* be continuously mapped onto the real line (see Netto’s Theorem).


TE = 4

Sorting by County

State:

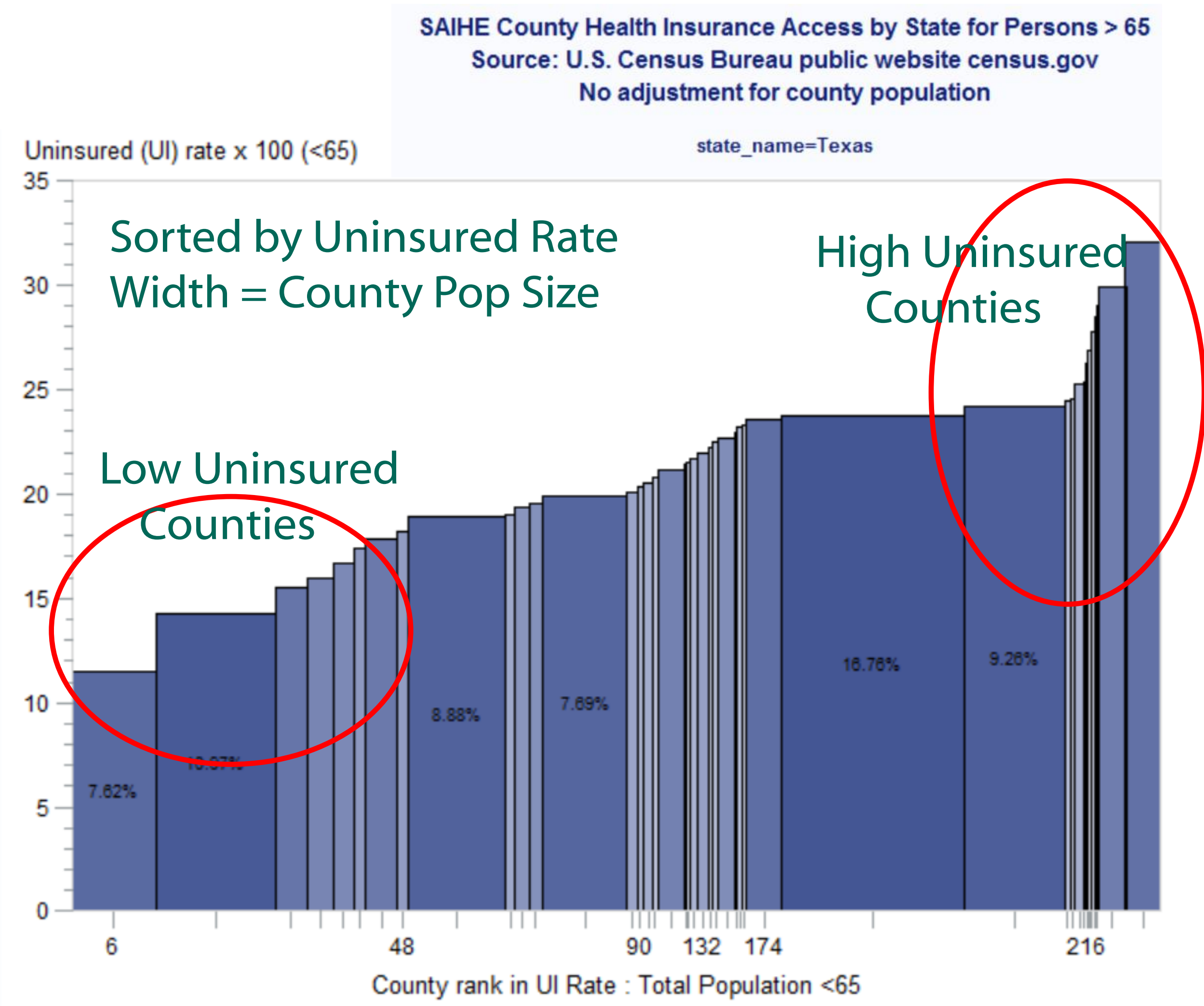
Texas, w/ 254 counties!

Variable of interest:

Uninsured Rate <65

Bar height: Uninsured Rate

Bar width: Population < 65



SAIHE County Health Insurance Access by State for Persons > 65
Source: U.S. Census Bureau public website census.gov
No adjustment for county population

state_name=Texas

Sorted by Uninsured Rate
Width = County Pop Size

Low Uninsured Counties

High Uninsured Counties

Sorting by County

State:

Texas, w/ 254 counties!

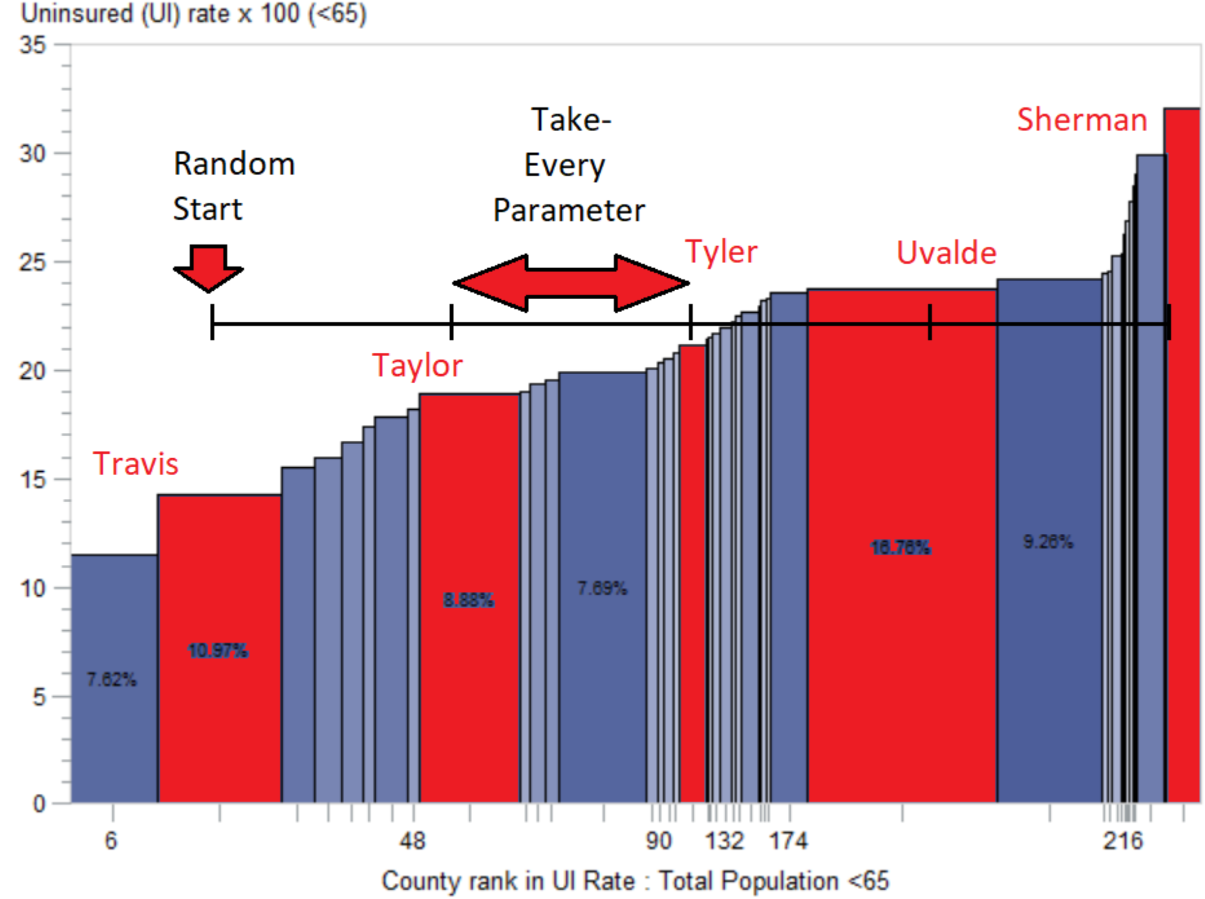
Variable of interest:

Uninsured Rate <65

Bar height: Uninsured Rate

Bar width: Population < 65

SAIHE County Health Insurance Access by State for Persons > 65
Source: U.S. Census Bureau public website census.gov
No adjustment for county population

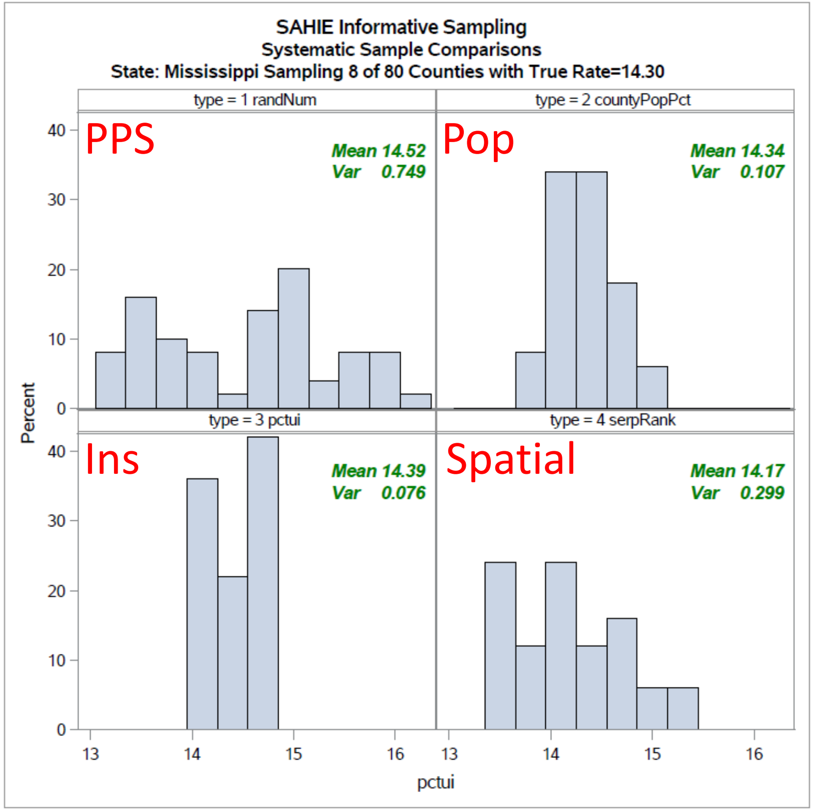


Systematic Sampling Example Using SAHIE

Small Area Health Insurance Estimates for each county were used in a simulation (n=50 for each state).

Different sort orders for selecting counties were compared.

While all sample approaches were unbiased, some samples were higher/lower than the state average.



Weighting Procedures

1. Select the **best set** of weighting variables for raking controls (use ACS or Census Pop.)

- More variables reduces bias. Less variables for lower variation.
- Optimize MSE = Bias^2 + Variance across multiple survey outcomes.
- Estimate Bias using Bootstrap or Jackknife or consider Full Model to be Unbiased.

Weighting Procedures

1. Select the **best set** of weighting variables for raking controls (use ACS or Census Pop.)

- More variables reduces bias. Less variables for lower variation.
- Optimize MSE = Bias^2 + Variance across multiple survey outcomes.
- Estimate Bias using Bootstrap or Jackknife or consider Full Model to be Unbiased.

2. Try to not to constrict weighting to controls (e.g., Quarterly Weights).

- When combining three years of data, it is best to *rerake*.

Weighting Procedures

1. Select the **best set** of weighting variables for raking controls (use ACS or Census Pop.)

- More variables reduces bias. Less variables for lower variation.
- Optimize MSE = Bias^2 + Variance across multiple survey outcomes.
- Estimate Bias using Bootstrap or Jackknife or consider Full Model to be Unbiased.

2. Try to not to constrict weighting to controls (e.g., Quarterly Weights).

- When combining three years of data, it is best to *rerake*.

3. Be conscious of weight sizes, perform weight trimming when possible.

- Some recommend trimming any weights larger than 5 × mean weight.
- Some recommend *against* weight trimming.

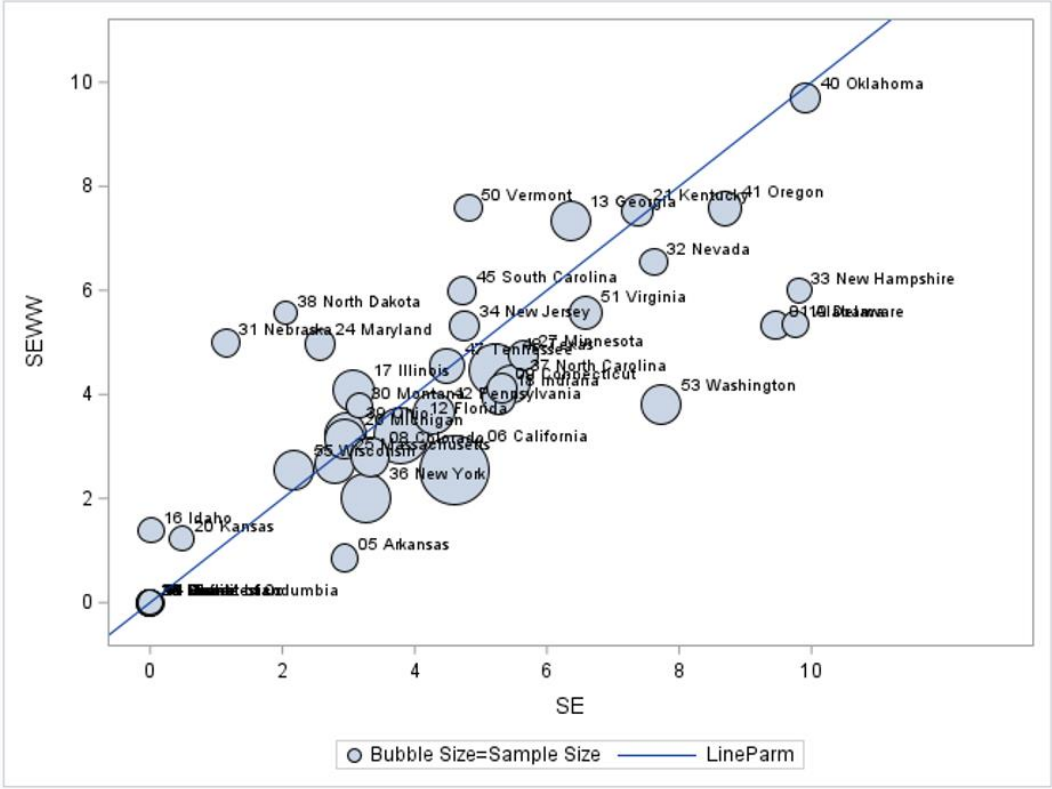
NHIS 2021: Standard Errors with and w/o Weights

Variable: Health Insurance for Persons 18-64 years old

Takeaways:

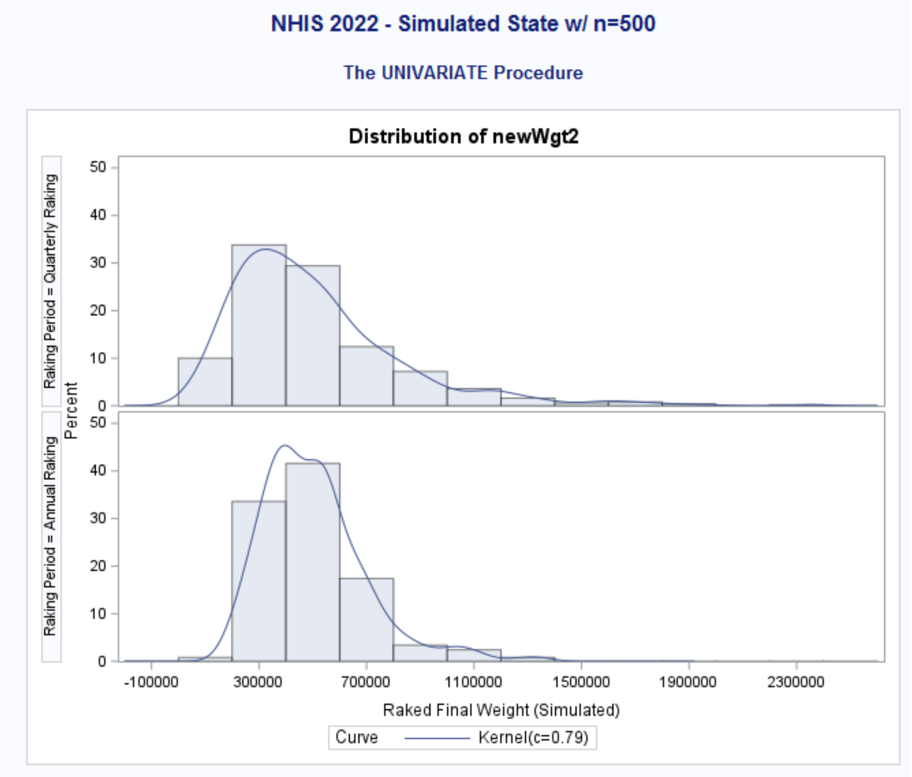
Adding weights increased standard errors for most states, but reduced bias.

Some states with small sample sizes had lower standard errors w/ weighting.



Raking Quarterly (4 × n=125) vs. Annually (n=500)

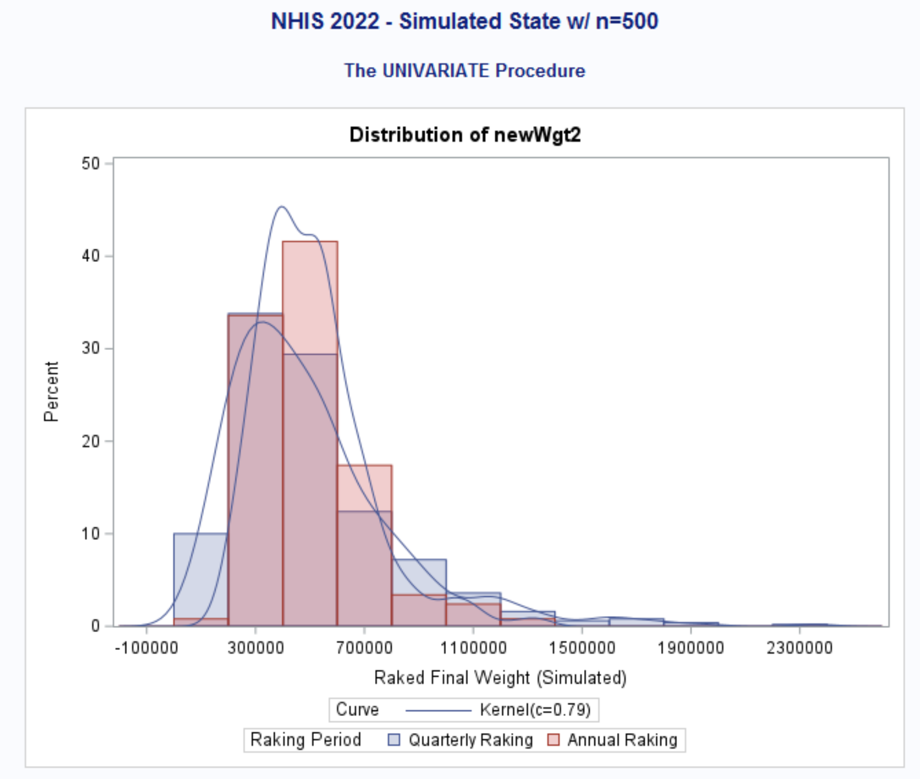
Raking Variables: Age Group, Gender, Race/Ethnicity, Poverty Status, Health Insurance



NHIS 2022 - Simulated State w/ n=500

The UNIVARIATE Procedure

Distribution of newWgt2



NHIS 2022 - Simulated State w/ n=500

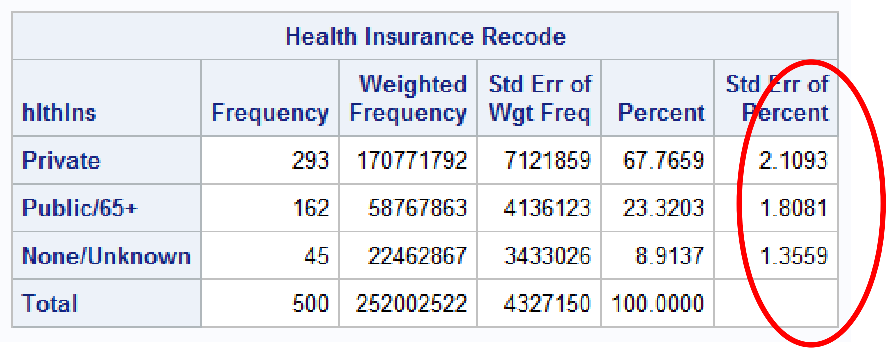
The UNIVARIATE Procedure

Distribution of newWgt2

Raking Simulation Results

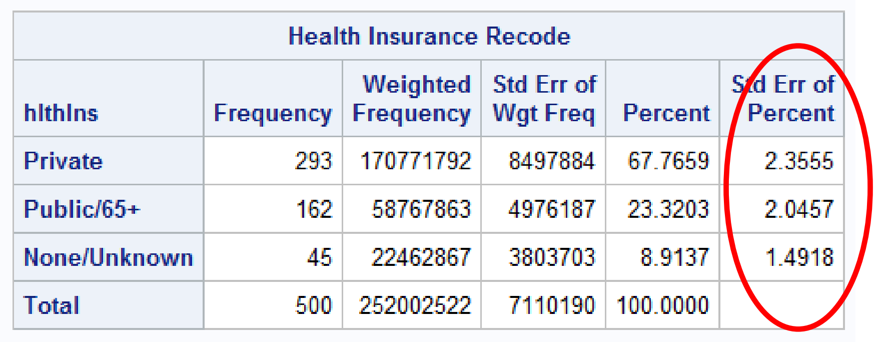
Raking Type	Average Weight	Std of Weights	# Raking Iterations
Annual	504,005	193,516	18
Quarterly	504,005	317,977	27.5

Annual Weighting Controls



Health Insurance Recode					
hlthIns	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
Private	293	170771792	7121859	67.7659	2.1093
Public/65+	162	58767863	4136123	23.3203	1.8081
None/Unknown	45	22462867	3433026	8.9137	1.3559
Total	500	252002522	4327150	100.0000	

Quarterly Weighting Controls

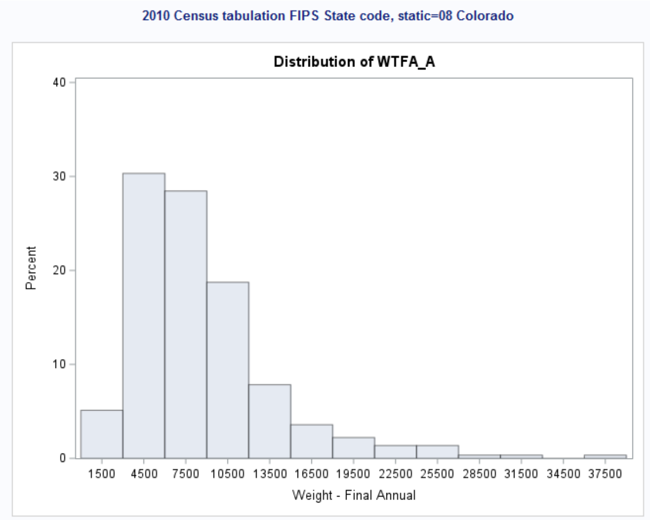


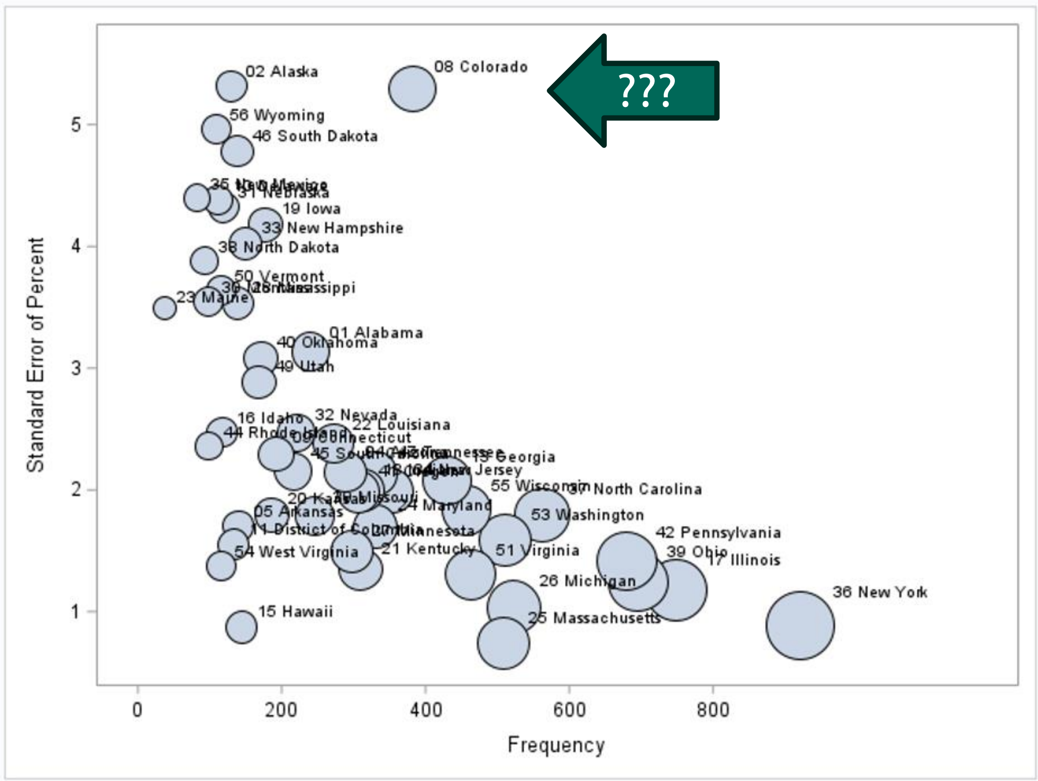
Health Insurance Recode					
hlthIns	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
Private	293	170771792	8497884	67.7659	2.3555
Public/65+	162	58767863	4976187	23.3203	2.0457
None/Unknown	45	22462867	3803703	8.9137	1.4918
Total	500	252002522	7110190	100.0000	

NHIS 2021: Weight Trimming

Variable: Health Insurance for Persons 18-64 years old

Colorado had one of the largest standard errors, despite having nearly 400 sample respondents.

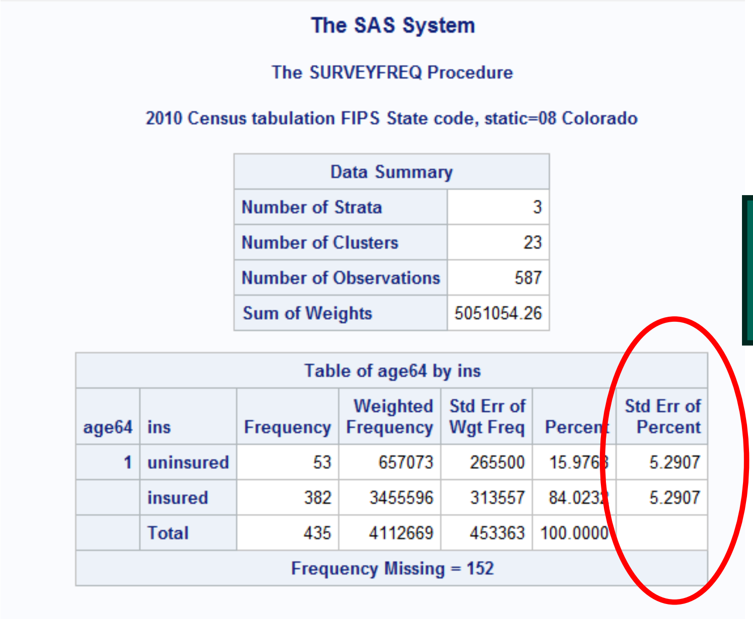




NHIS 2021: “Weight Trimming” in Colorado

Variable: Health Insurance for Persons 18-64 years old

Takeaway: Removing 5 respondents with the largest weights reduced the SE by 20% and moved the health insurance rate closer to the ACS standard (89%) for CO.



The SAS System

The SURVEYFREQ Procedure

2010 Census tabulation FIPS State code, static=08 Colorado

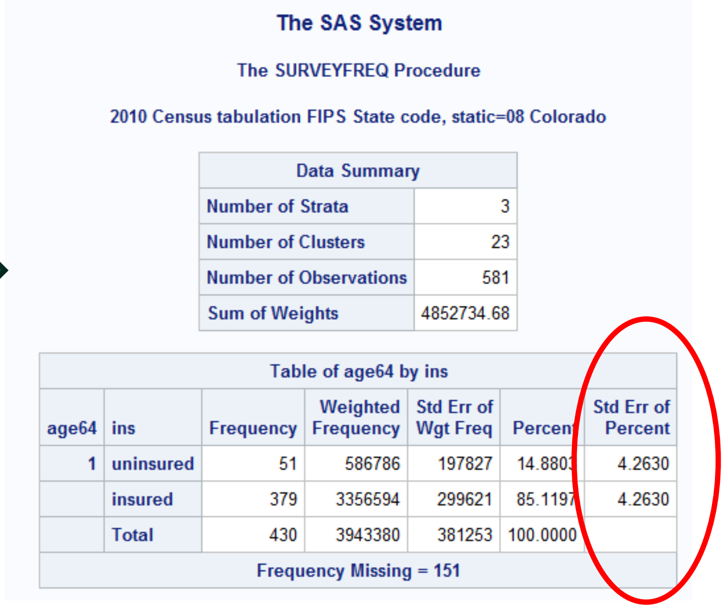
Data Summary	
Number of Strata	3
Number of Clusters	23
Number of Observations	587
Sum of Weights	5051054.26

Table of age64 by ins

age64	ins	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
1	uninsured	53	657073	265500	15.9768	5.2907
	insured	382	3455596	313557	84.0232	5.2907
	Total	435	4112669	453363	100.0000	

Frequency Missing = 152

Minus 5 Responses



The SAS System

The SURVEYFREQ Procedure

2010 Census tabulation FIPS State code, static=08 Colorado

Data Summary	
Number of Strata	3
Number of Clusters	23
Number of Observations	581
Sum of Weights	4852734.68

Table of age64 by ins

age64	ins	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
1	uninsured	51	586786	197827	14.8803	4.2630
	insured	379	3356594	299621	85.1197	4.2630
	Total	430	3943380	381253	100.0000	

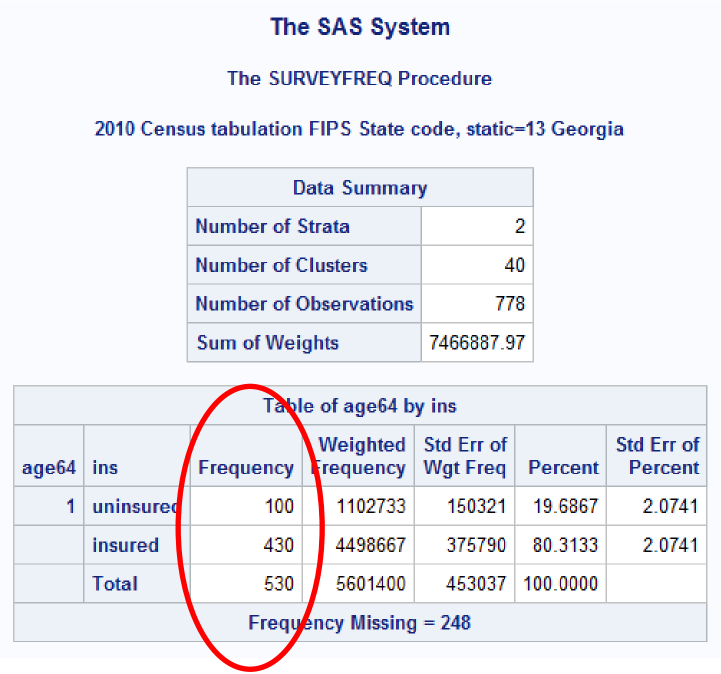
Frequency Missing = 151

PSU Cycling

Rotating PSUs Every Two Years

Sample Size Paradox

In a two-stage design, doubling the sample size may not improve reliability if the PSU sample sizes are already adequate.



The SAS System

The SURVEYFREQ Procedure

2010 Census tabulation FIPS State code, static=13 Georgia

Data Summary	
Number of Strata	2
Number of Clusters	40
Number of Observations	778
Sum of Weights	7466887.97

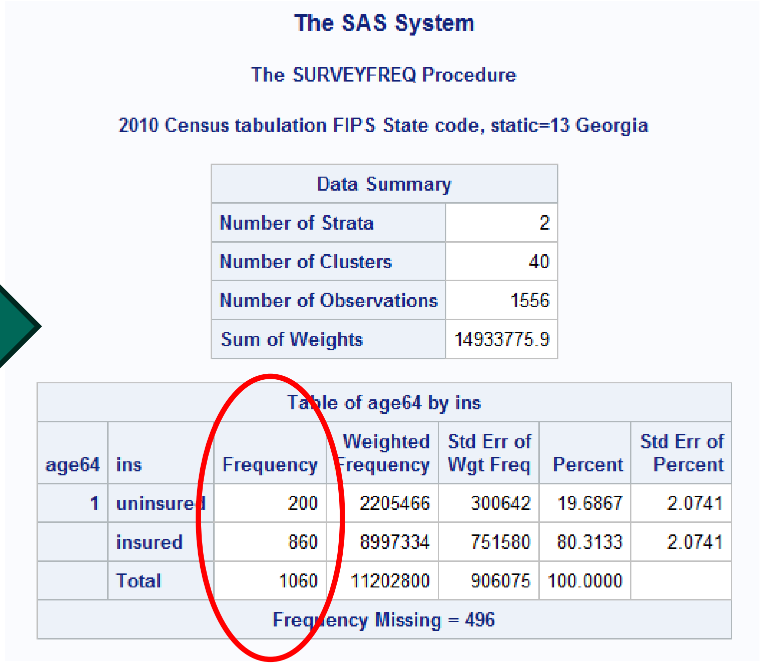
Table of age64 by ins

age64	ins	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
1	uninsured	100	1102733	150321	19.6867	2.0741
	insured	430	4498667	375790	80.3133	2.0741
	Total	530	5601400	453037	100.0000	

Frequency Missing = 248

NHIS 2021
State: GA

× 2
(Same PSUs)



The SAS System

The SURVEYFREQ Procedure

2010 Census tabulation FIPS State code, static=13 Georgia

Data Summary	
Number of Strata	2
Number of Clusters	40
Number of Observations	1556
Sum of Weights	14933775.9

Table of age64 by ins

age64	ins	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
1	uninsured	200	2205466	300642	19.6867	2.0741
	insured	860	8997334	751580	80.3133	2.0741
	Total	1060	11202800	906075	100.0000	

Frequency Missing = 496

Taylor Series Variance Estimation

All reported standard errors are based on the following formula for n *PSUs*:

$$\hat{V}_{TS}(\bar{y}) = \frac{n}{n-1}\frac{1}{W^2}\sum_{i=1}^{n}\left(W_i(\bar{y}_i - \bar{y}) - \sum_{s=1}^{n} W_s(\bar{y}_s - \bar{y})\right)^2$$

$$\bar{y} = \frac{1}{W}\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}y_{ij}\,, \bar{y}\ _i = \frac{1}{W_i}\sum_{j=1}^{n_i} w_{ij}y_{ij,}\ \ W_i = \sum_{j=1}^{n_i} w_{ij}\,, W = \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}$$

Taylor Series Variance Estimation

All reported standard errors are based on the following formula for n *PSUs*:

$$\hat{V}_{TS}(\bar{y}) = \frac{n}{n-1}\frac{1}{W^2}\sum_{i=1}^{n}\left(W_i(\bar{y}_i - \bar{y}) - \sum_{s=1}^{n} W_s(\bar{y}_s - \bar{y})\right)^2$$

$$\bar{y} = \frac{1}{W}\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}y_{ij}\,, \bar{y}_{\,i} = \frac{1}{W_i}\sum_{j=1}^{n_i} w_{ij}y_{ij,}\;\; W_i = \sum_{j=1}^{n_i} w_{ij}\,, W = \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}$$

Under normality of y_i and equal-sized weights, $\hat{V}_{TS}(\bar{y}) \sim \frac{\left(\frac{\sigma^2}{n}\right)\chi^2_{n-1}}{\mathrm{n-1}}$.

When more than one stratum is involved, d.f. = # PSUs - # Strata (software default).

Rules for Pooling Years with SR Strata

It is common to treat Self-Reporting PSUs as Strata and create Pseudo-PSUs from the secondary stage sampling units.

NSR PSUs: When PSUs are fixed across multiple years, they should not be considered independent. No increase in degrees of freedom.

SR PSUs: The generated pseudo-PSUs should be treated as *independent* PSUs in every year in the same separate stratum.

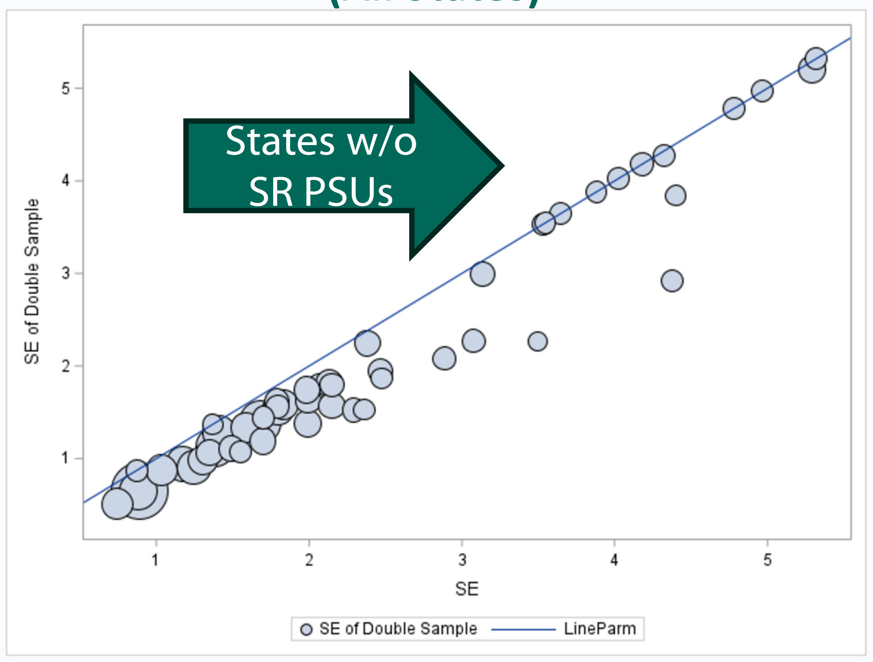
Rules for Pooling Years with SR Strata

It is common to treat Self-Reporting PSUs as Strata and create Pseudo-PSUs from the secondary stage sampling units.

NSR PSUs: When PSUs are fixed across multiple years, they should *not* be considered independent. No increase in degrees of freedom.

SR PSUs: The generated pseudo-PSUs should be treated as *independent* PSUs in every year in the same separate stratum.

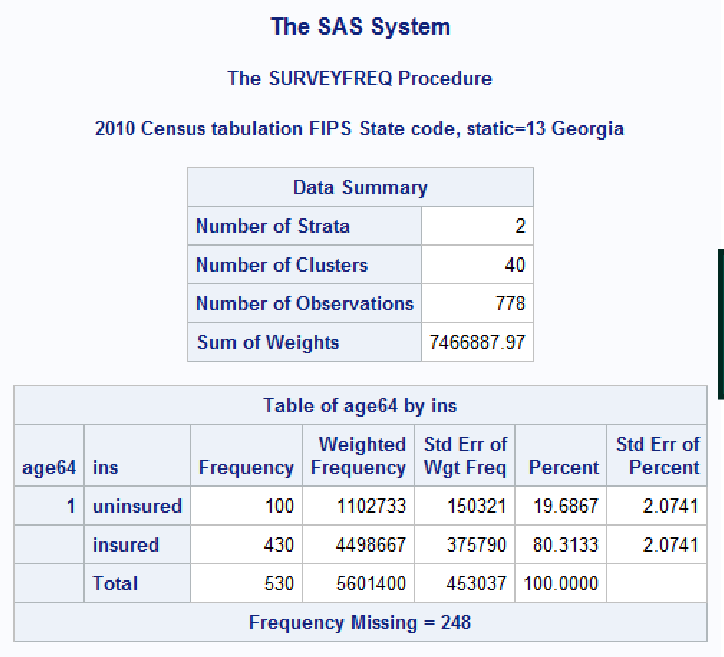
NHIS 2021 Double Sample SE's (All States)



States w/o SR PSUs

Increased # PSUs

In a two-stage design, doubling the **number of PSUs** will improve reliability if the PSU sample sizes are all adequate.



The SAS System

The SURVEYFREQ Procedure

2010 Census tabulation FIPS State code, static=13 Georgia

Data Summary	
Number of Strata	2
Number of Clusters	40
Number of Observations	778
Sum of Weights	7466887.97

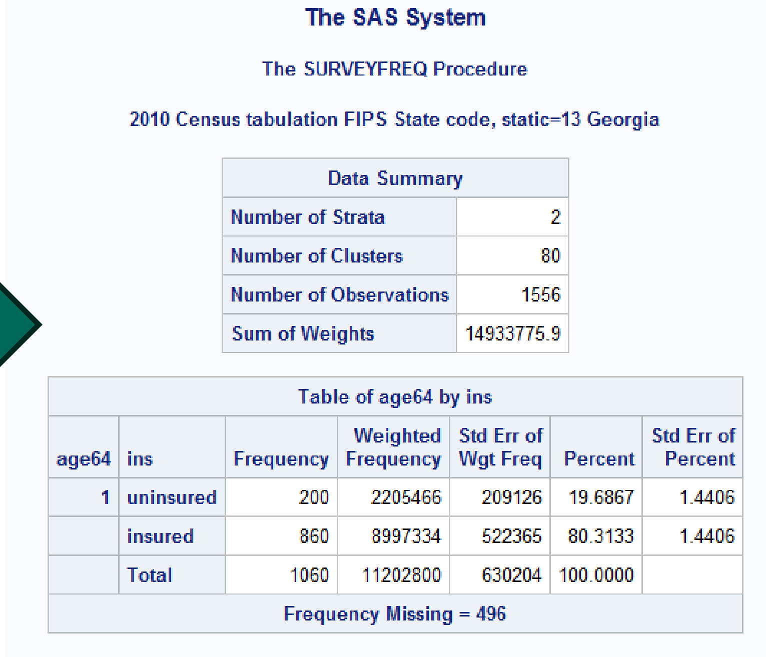
Table of age64 by ins

age64	ins	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
1	uninsured	100	1102733	150321	19.6867	2.0741
	insured	430	4498667	375790	80.3133	2.0741
	Total	530	5601400	453037	100.0000	

Frequency Missing = 248

NHIS 2021
State: GA

× 2
(w/ new PSUs)



The SAS System

The SURVEYFREQ Procedure

2010 Census tabulation FIPS State code, static=13 Georgia

Data Summary	
Number of Strata	2
Number of Clusters	80
Number of Observations	1556
Sum of Weights	14933775.9

Table of age64 by ins

age64	ins	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
1	uninsured	200	2205466	209126	19.6867	1.4406
	insured	860	8997334	522365	80.3133	1.4406
	Total	1060	11202800	630204	100.0000	

Frequency Missing = 496

PSU Cycling

To save interview training and recruitment costs, some nationwide federal household surveys will fix PSUs for a ten-year period.

At the same time, pooling three years’ worth of data is less effective for state-level estimates in some states due to the static non-representing PSUs.

In practice, many survey interviewers work part-time and do not remain employed in data collection services over a full ten-year period.

Changing PSUs every year is prohibitive. Yet large inferential gains can still be made if a PSU tenure is set to two years and if half the sampled PSUs are rotated every year.

PSU Cycling: Computing the Degrees of Freedom

In the NHIS 2020 sample design, NSR PSUs will be rotated every two years.

After a burn-in period, only half the PSUs will be rotated every year to maximize the number of PSUs in three-year pooled datasets.

Coupled with the Pseudo-PSUs, this causes even more discrepancies in the PSU sample sizes within states, requiring a Satterthwaite approximation to the *effective* df.

PSU Cycling: Computing the Degrees of Freedom

In the NHIS 2020 sample design, NSR PSUs will be rotated every two years.

After a burn-in period, only half the PSUs will be rotated every year to maximize the number of PSUs in three-year pooled datasets.

Coupled with the Pseudo-PSUs, this causes even more discrepancies in the PSU sample sizes within states, requiring a Satterthwaite approximation to the *effective* df.

$$df_{eff} = \frac{\left(\sum_{h=1}^{L} n_h \sigma_h^2\right)^2}{\sum_{h=1}^{L}\left(\frac{n_h^2}{n_h-1}\right)\sigma_h^4}$$

For L strata with n_h PSUs in the h^{th} stratum and stratum variance σ_h^2.

PSU Random Walk

Moving PSUs to Adjacent Areas

PSU Random Walk

The ideal situation is to always sample PSUs *independently* as in PSU Cycling.

However, there can be cases where extra clustering can be used to keep sampled PSUs *nearby*. Some practitioners are willing to keep the same PSU in place (overlap).

Traditional stratification based on geography can reduce the degrees of freedom. The reduction is lower under the PSU Random Walk.

We develop a sampling methodology that defines advantageous *conditional* probabilities of selection while preserving the appropriate *unconditional* probabilities of selection.

PSU Random Walk Implementation

Step 1. Initially select each PSU independently and without replacement based on the relative measure of size as the probability of selection (POS).

Step 2. Determine the *transition*, or next step, probabilities of selection that move the active PSUs to an adjacent PSU. This is done in such a way as to preserve the original, or unconditional, probabilities of selection.

Step 3. Iterate each PSU according to the adjacent conditional probabilities of selection. If any two PSUs overlap, continue the iterative sequence on as many PSUs as needed until a without replacement sample is achieved.

Research Problem

Suppose there are n PSUs. Let t $= \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix}$ denote the vector of probabilities of selection.

We must find a nontrivial $n \times n$ matrix $A \neq I$ such that the following hold:

$$t'A = t'$$

$$0 \leq A_{ij} \leq 1 \; \forall \; 1 \leq i, j \leq n$$

$$\sum_{j=1}^{n} A_{ij} = 1 \; \forall \; i$$

The Probability Transition Matrix

Any matrix satisfying those conditions is called a *probability transition matrix*. We also add the requirement that $A_{ii} = 0$.

- This requirement states that the random walk cannot stay in the same PSU and must transition to another geographical area with each iteration.
- It also means that $\det(A) = 0$ and that A is a singular matrix.

The Probability Transition Matrix

Any matrix satisfying those conditions is called a *probability transition matrix*. We also add the requirement that $A_{ii} = 0$.

- This requirement states that the random walk cannot stay in the same PSU and must transition to another geographical area with each iteration.
- It also means that $\det(A) = 0$ and that A is a singular matrix.

It is not clear how to find such a matrix, in general, if one even exists for a given topology and vector t. Matrix problems usually have the form $Ax = b$, solve for x.

Since the PSU Random Walk will have a finite state space and because every state communicates, we expect to see $\lim_{n \to \infty} A^n = [t', \quad t', \quad t', \ldots, t']'$.

Finding the Probability Transition Matrix

Numerical methods may be used to find suitable matrices that satisfy the conditions.

An objective function is required. The objective function should attempt to minimize the design-based variance of the survey estimator.

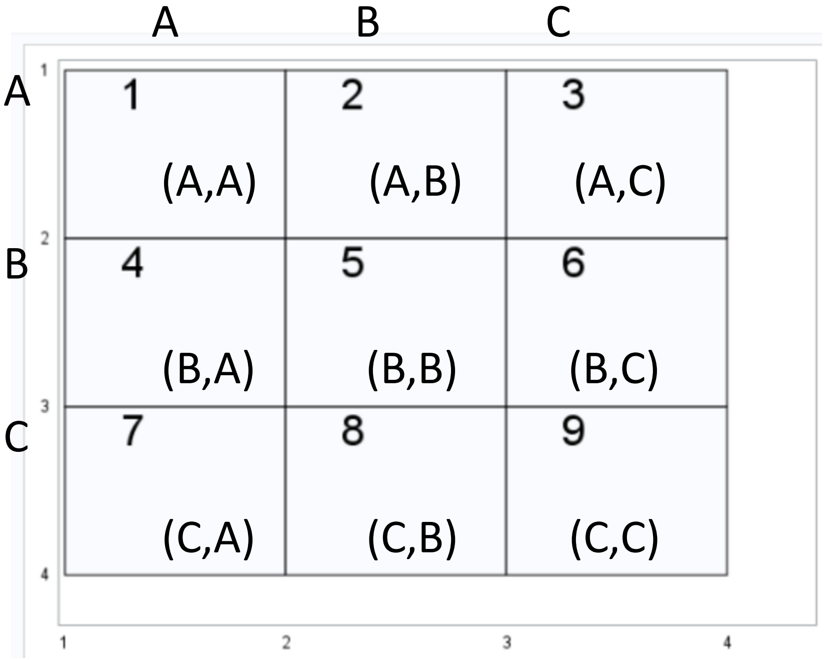
Minimize $Cov(1_i, 1_j)$, where 1_i is the event that the i^{th} PSU is selected.

Suitable bounds should be placed on the matrix entries to reduce variance.

Example: 3 × 3 Square

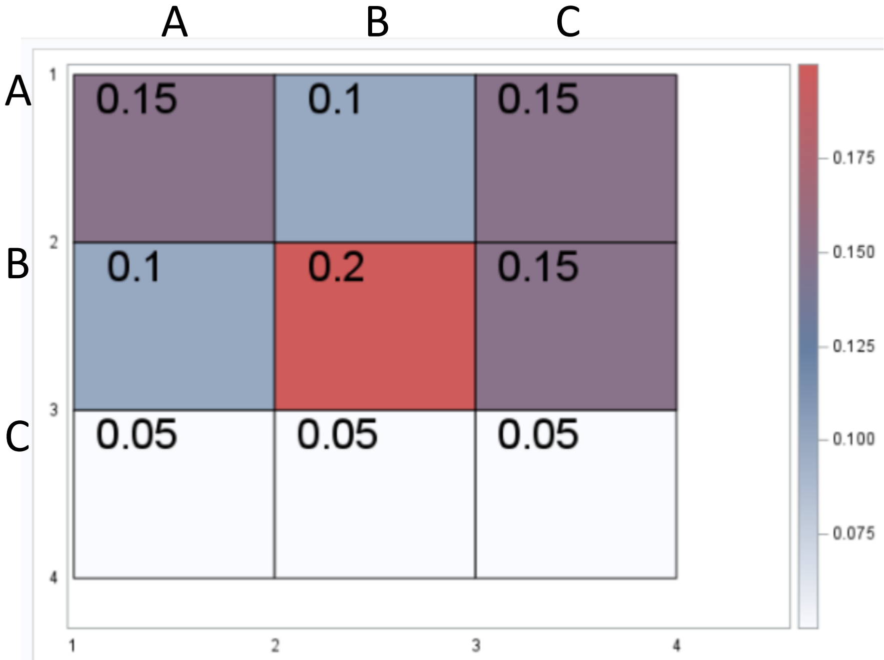
Suppose there are nine PSUs arranged in a 3 × 3 grid with varying probabilities of selection. How to find a probability transition matrix?

PSU Topology



	A	B	C
A	1 (A,A)	2 (A,B)	3 (A,C)
B	4 (B,A)	5 (B,B)	6 (B,C)
C	7 (C,A)	8 (C,B)	9 (C,C)

PSU MOS

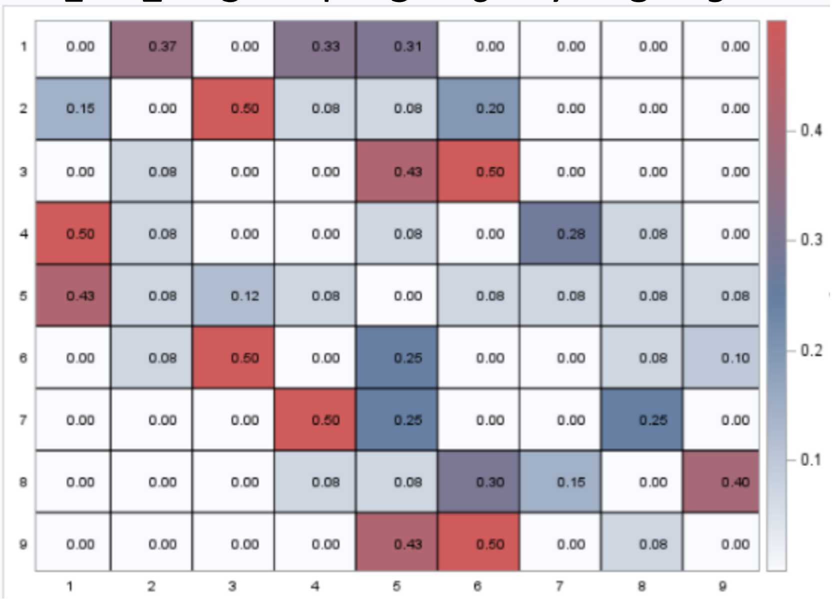


	A	B	C
A	0.15	0.1	0.15
B	0.1	0.2	0.15
C	0.05	0.05	0.05

Example: 3 × 3 Square Solution

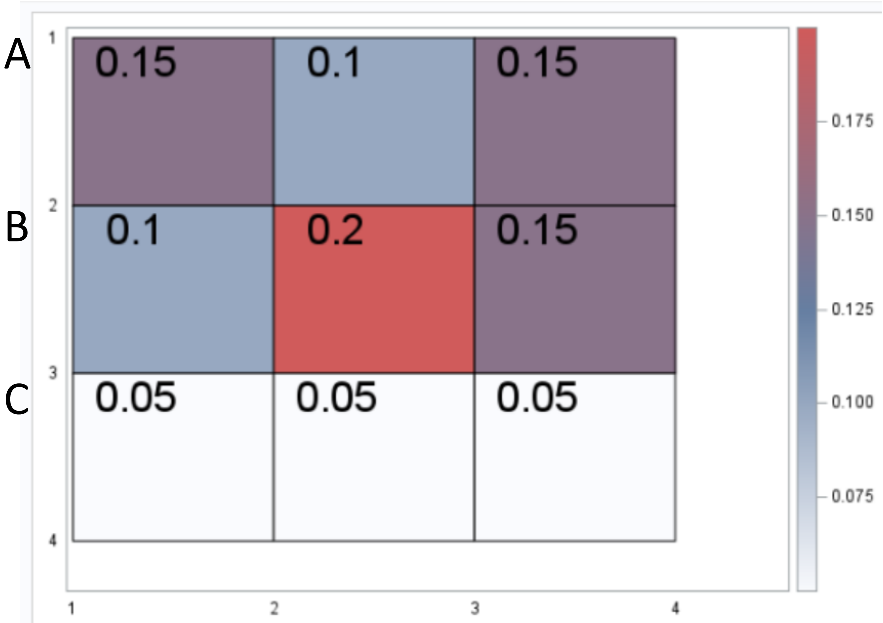
A linear programming algorithm was used to solve for the matrix with constraints: $0.075 \leq A_{ij} \leq 0.5$, $t'A = t'$, and $\sum_{j=1}^{n} A_{ij} = 1 \; \forall \, i$.

Transition Matrix



	1	2	3	4	5	6	7	8	9
1	0.00	0.37	0.00	0.33	0.31	0.00	0.00	0.00	0.00
2	0.15	0.00	0.50	0.08	0.08	0.20	0.00	0.00	0.00
3	0.00	0.08	0.00	0.00	0.43	0.50	0.00	0.00	0.00
4	0.50	0.08	0.00	0.00	0.08	0.00	0.28	0.08	0.00
5	0.43	0.08	0.12	0.08	0.00	0.08	0.08	0.08	0.08
6	0.00	0.08	0.50	0.00	0.25	0.00	0.00	0.08	0.10
7	0.00	0.00	0.00	0.50	0.25	0.00	0.00	0.25	0.00
8	0.00	0.00	0.00	0.08	0.08	0.30	0.15	0.00	0.40
9	0.00	0.00	0.00	0.00	0.43	0.50	0.00	0.08	0.00

PSU MOS

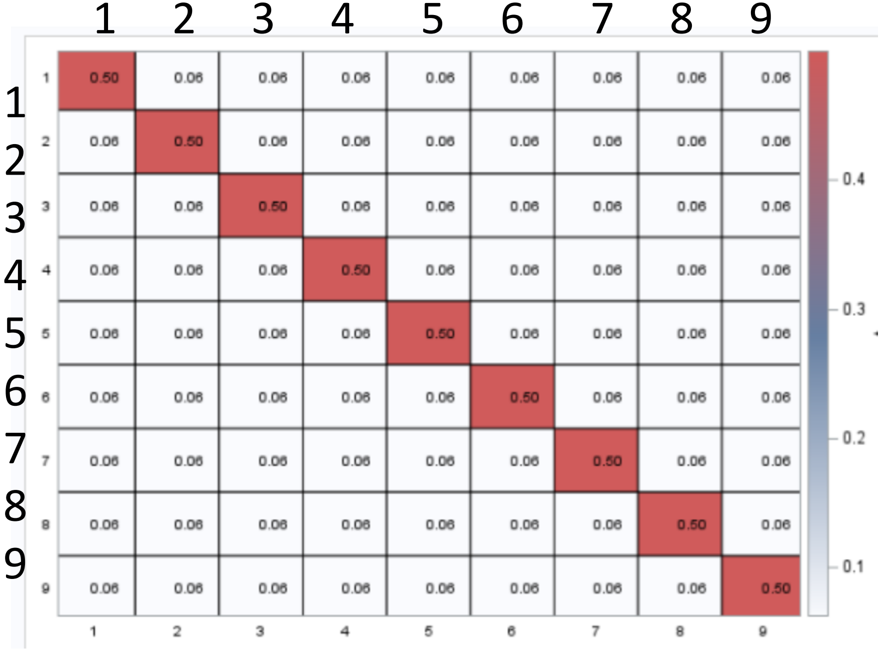


	A	B	C
A	0.15	0.1	0.15
B	0.1	0.2	0.15
C	0.05	0.05	0.05

Example: 3 × 3 Maximum Overlap Solution

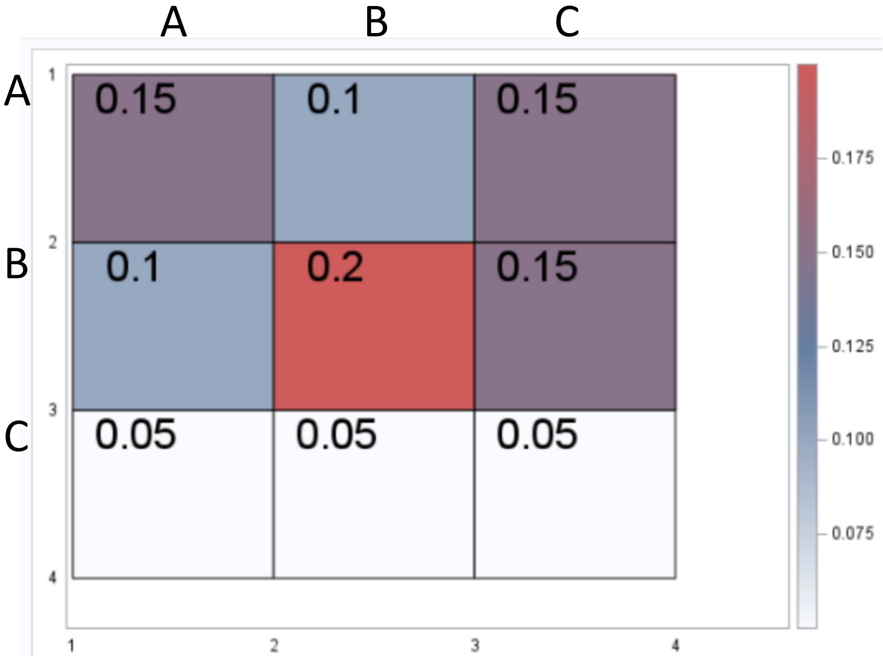
Some practitioners advocate for conditional probabilities that promote staying in the same PSU no matter the POS. Permanent random numbers are used to preserve POS.

Transition Matrix



	1	2	3	4	5	6	7	8	9
1	0.50	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
2	0.06	0.50	0.06	0.06	0.06	0.06	0.06	0.06	0.06
3	0.06	0.06	0.50	0.06	0.06	0.06	0.06	0.06	0.06
4	0.06	0.06	0.06	0.50	0.06	0.06	0.06	0.06	0.06
5	0.06	0.06	0.06	0.06	0.50	0.06	0.06	0.06	0.06
6	0.06	0.06	0.06	0.06	0.06	0.50	0.06	0.06	0.06
7	0.06	0.06	0.06	0.06	0.06	0.06	0.50	0.06	0.06
8	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.50	0.06
9	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.50

PSU MOS



	A	B	C
A	0.15	0.1	0.15
B	0.1	0.2	0.15
C	0.05	0.05	0.05

Example: Transition Matrix A

	1	2	3	4	5	6	7	8	9
1	0	0.366667	0	0.325	0.308333	0	0	0	0
2	0.15	0	0.5	0.075	0.075	0.2	0	0	0
3	0	0.075	0	0	0.425	0.5	0	0	0
4	0.5	0.075	0	0	0.075	0	0.275	0.075	0
5	0.425	0.075	0.125	0.075	0	0.075	0.075	0.075	0.075
6	0	0.075	0.5	0	0.25	0	0	0.075	0.1
7	0	0	0	0.5	0.25	0	0	0.25	0
8	0	0	0	0.075	0.075	0.3	0.15	0	0.4
9	0	0	0	0	0.425	0.5	0	0.075	0

Example: 3 × 3 Square POS Computation

POS Computation for Cell 1

$$0.15 = p_1 =$$

$$\sum P(X_{k+1} = 1 | X_k = i) P(X_k = i) = \sum A_{i1} p_i = A_{21} p_2 + A_{41} p_4 + A_{51} p_5$$

$$= 0.15 * 0.1 + 0.5 * 0.1 + 0.425 * 0.2$$

$$= 0.15$$

Example: A^2

	1	2	3	4	5	6	7	8	9
1	0.348542	0.047500	0.221875	0.050625	0.051875	0.096458	0.112500	0.047500	0.023125
2	0.069375	0.118750	0.109375	0.054375	0.314375	0.255625	0.026250	0.026250	0.025625
3	0.191875	0.069375	0.340625	0.037500	0.130625	0.046875	0.031875	0.069375	0.081875
4	0.043125	0.188958	0.046875	0.316875	0.234167	0.043125	0.016875	0.074375	0.035625
5	0.048750	0.176458	0.075000	0.186875	0.270417	0.137500	0.031875	0.035625	0.037500
6	0.117500	0.056250	0.068750	0.030000	0.266250	0.356250	0.030000	0.026250	0.048750
7	0.356250	0.056250	0.031250	0.037500	0.056250	0.093750	0.193750	0.056250	0.118750
8	0.069375	0.033750	0.159375	0.080625	0.288125	0.205625	0.026250	0.101250	0.035625
9	0.180625	0.069375	0.303125	0.037500	0.130625	0.054375	0.043125	0.069375	0.111875

Example: A^4

	1	2	3	4	5	6	7	8	9
1	0.230945	0.071270	0.189095	0.066093	0.136595	0.121413	0.076973	0.054082	0.053533
2	0.116912	0.109253	0.122109	0.101780	0.233410	0.188064	0.039895	0.041866	0.046709
3	0.181497	0.083570	0.217834	0.073818	0.165319	0.110192	0.051995	0.057736	0.058039
4	0.084888	0.137293	0.104763	0.167644	0.243802	0.134698	0.032169	0.054254	0.040489
5	0.101622	0.124833	0.112199	0.134224	0.238250	0.159904	0.037133	0.047294	0.044540
6	0.135494	0.095676	0.121406	0.086604	0.222253	0.204614	0.045171	0.040623	0.048159
7	0.243816	0.069090	0.159277	0.063727	0.131816	0.129145	0.091926	0.052652	0.058551
8	0.121596	0.103368	0.140652	0.107368	0.221871	0.164355	0.039764	0.051096	0.049929
9	0.180688	0.083570	0.212525	0.073614	0.166385	0.112706	0.053232	0.057511	0.059767

Example: A^8

	1	2	3	4	5	6	7	8	9
1	0.166947	0.093173	0.161799	0.091059	0.186137	0.142730	0.055211	0.051265	0.051681
2	0.143035	0.102565	0.143901	0.103016	0.206199	0.155007	0.047962	0.049074	0.049241
3	0.159404	0.095995	0.160374	0.094475	0.191505	0.143427	0.052287	0.051217	0.051317
4	0.133684	0.107357	0.138268	0.110667	0.212870	0.154432	0.045259	0.049280	0.048184
5	0.138551	0.104902	0.141139	0.106757	0.209392	0.154662	0.046689	0.049179	0.048728
6	0.146829	0.100811	0.145745	0.100511	0.203433	0.154831	0.049172	0.049089	0.049580
7	0.168992	0.092483	0.160873	0.090286	0.184948	0.143358	0.056244	0.051110	0.051705
8	0.144555	0.102078	0.146481	0.102650	0.204441	0.152432	0.048258	0.049595	0.049509
9	0.159272	0.096046	0.160000	0.094538	0.191681	0.143719	0.052292	0.051161	0.051291

Example: A^{16}

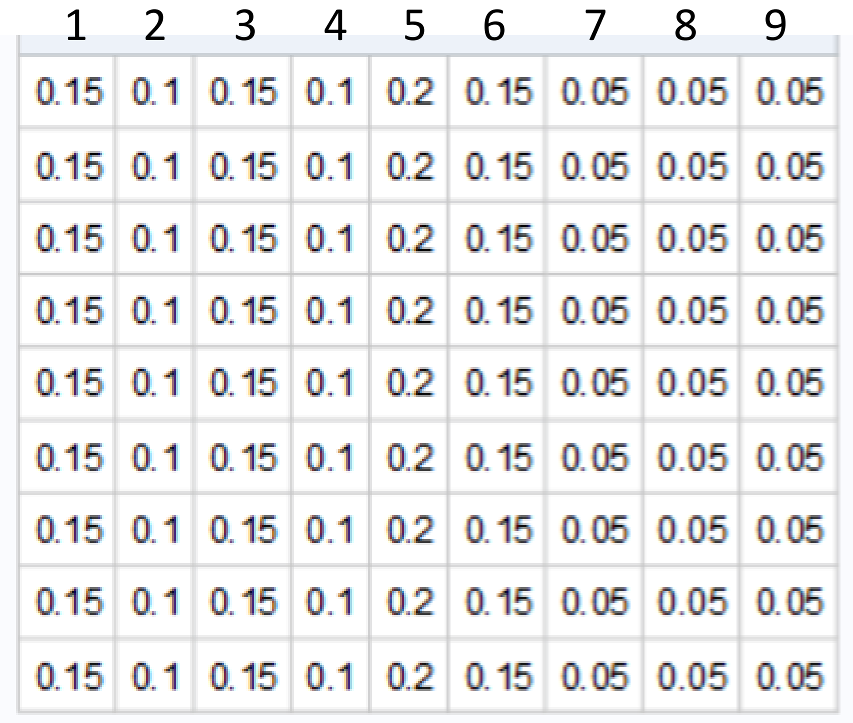
	1	2	3	4	5	6	7	8	9
1	0.150881	0.099637	0.150692	0.099513	0.199265	0.149584	0.050258	0.050074	0.050097
2	0.149630	0.100152	0.149701	0.100202	0.200311	0.150186	0.049893	0.049967	0.049959
3	0.150542	0.099776	0.150445	0.099701	0.199543	0.149727	0.050156	0.050049	0.050061
4	0.149123	0.100364	0.149316	0.100490	0.200728	0.150402	0.049744	0.049929	0.049904
5	0.149386	0.100254	0.149515	0.100341	0.200512	0.150290	0.049821	0.049949	0.049932
6	0.149823	0.100071	0.149847	0.100093	0.200152	0.150103	0.049950	0.049981	0.049980
7	0.150957	0.099605	0.150739	0.099472	0.199205	0.149558	0.050281	0.050078	0.050104
8	0.149723	0.100114	0.149783	0.100152	0.200231	0.150131	0.049919	0.049977	0.049970
9	0.150532	0.099781	0.150436	0.099706	0.199552	0.149733	0.050153	0.050048	0.050060

Example: A^{24}

	1	2	3	4	5	6	7	8	9
1	0.150047	0.099981	0.150037	0.099974	0.199961	0.149977	0.050014	0.050004	0.050005
2	0.149980	0.100008	0.149984	0.100011	0.200017	0.150010	0.049994	0.049998	0.049998
3	0.150029	0.099988	0.150023	0.099984	0.199976	0.149986	0.050008	0.050002	0.050003
4	0.149953	0.100019	0.149963	0.100026	0.200039	0.150022	0.049986	0.049996	0.049995
5	0.149967	0.100014	0.149974	0.100018	0.200027	0.150016	0.049990	0.049997	0.049996
6	0.149990	0.100004	0.149992	0.100005	0.200008	0.150005	0.049997	0.049999	0.049999
7	0.150051	0.099979	0.150040	0.099972	0.199957	0.149976	0.050015	0.050004	0.050006
8	0.149985	0.100006	0.149988	0.100008	0.200012	0.150007	0.049996	0.049999	0.049998
9	0.150029	0.099988	0.150023	0.099984	0.199976	0.149986	0.050008	0.050002	0.050003

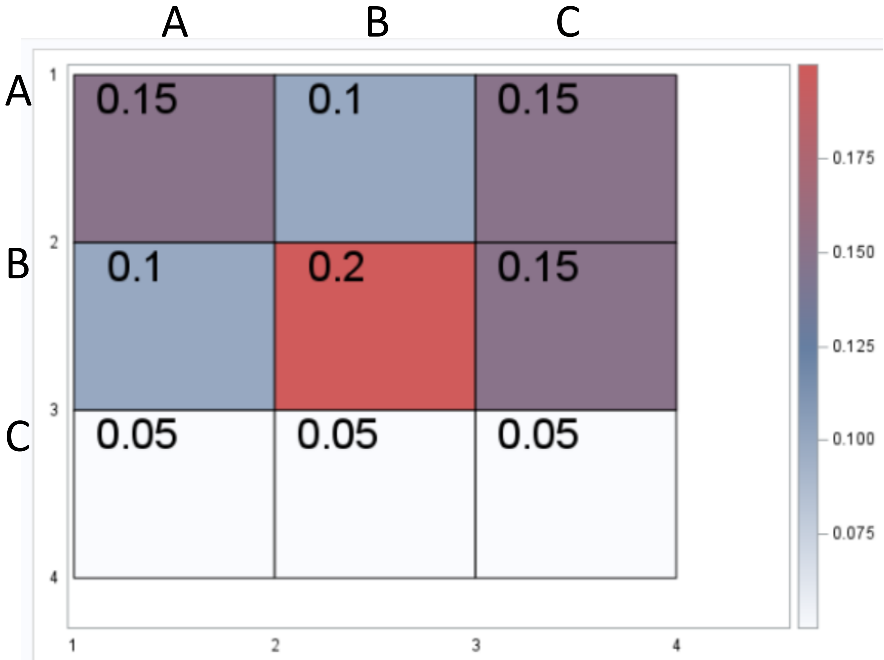
Example: 3 × 3 Square Limiting Distribution

A^{48}



1	2	3	4	5	6	7	8	9
0.15	0.1	0.15	0.1	0.2	0.15	0.05	0.05	0.05
0.15	0.1	0.15	0.1	0.2	0.15	0.05	0.05	0.05
0.15	0.1	0.15	0.1	0.2	0.15	0.05	0.05	0.05
0.15	0.1	0.15	0.1	0.2	0.15	0.05	0.05	0.05
0.15	0.1	0.15	0.1	0.2	0.15	0.05	0.05	0.05
0.15	0.1	0.15	0.1	0.2	0.15	0.05	0.05	0.05
0.15	0.1	0.15	0.1	0.2	0.15	0.05	0.05	0.05
0.15	0.1	0.15	0.1	0.2	0.15	0.05	0.05	0.05
0.15	0.1	0.15	0.1	0.2	0.15	0.05	0.05	0.05

PSU POS



	A	B	C
A	0.15	0.1	0.15
B	0.1	0.2	0.15
C	0.05	0.05	0.05

How to adjust the degrees of freedom?

Dependent PSUs. Since the PSUs selected within the same PSU Random Walk are correlated, we should not treat them as if they were sampled *independently*. The transition matrix should optimize the sample variance

We propose the following adjustment to the degrees of freedom to account for this relationship for each PSU pairing in a three-year pool.

$$\# \, PSUs = 2 \; - Corr(1_i, 1_j)$$

Recap

Systematic sampling can obtain a more representative sample w/o adding more degrees of freedom under multiple contexts: spatially or w.r.t. a highly correlated frame variable.

Survey practitioners should be cognizant of the overall bias-variance tradeoff when

- adding variables to weighting models
- correcting for non-response bias
- weighting trimming and other adjustments

Alternative first and second stage sampling methods can be used to improve reliability without adding prohibitive costs.

References

Durrett, R. (2016). *Essentials of Stochastic Process.* Springer.

Fay, R. (2014). The Sample Overlap Problem for Systematic Sampling. *Proceedings of the Joint Statistical Meetings (2014).* The American Statistical Association.

Korn, E., Graubard, B. (1999). *Analysis of Health Surveys.* John Wiley and Sons.

Valliant, R., Dever, J., Kreuter, F. (2013). *Practical Tools for Designing and Weighting Surveys.* Springer.

Thank you!

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.