NATIONAL CENTER FOR HEALTH STATISTICS

Two-Stage Systematic Cluster Sampling in the NHIS 2020 Design

William Waldron, Division of Research and Methodology
Matthew D. Bramlett, Jonaki Bose, Division of Health Interview Surveys

BACKGROUND

The National Health Interview Survey (NHIS) is an annual cross-sectional survey that obtains health information for approximately 30K civilian non-institutionalized adults. The NHIS asks participants over *600* questions on physical and mental health.

SAMPLE DESIGN

A stratified **two-stage design** is used with initial sampling of PSUs (county clusters). This is followed by within-PSU systematic sample of clusters dispersed in a sorted address-based commercial frame. Field representatives will visit each housing unit.

- **PSU Sample**: Sorted based on aggregated population sizes of each county cluster.
- **Within-PSU Sample:** *Clusters* are spaced apart based on a *Take-Every 1 Parameter*.
- **Within-Cluster Sample**: *Units* are spaced apart based on a *Take-Every 2 Parameter.*
- **Self-Reported PSUs**: For variance estimation purposes, these PSUs are reclassified as strata. *Pseudo-PSUs* are constructed by grouping the secondary-stage clusters.
- **Variance Estimation:** Taylor Series Linearization w/ Strata and PSUs specified.

OBJECTIVE

A systematic cluster design that balances in-person recruitment costs with the sampling variance is presented. The complexity of the design results in ambiguities when considering reliability. We investigated ways to enhance the underlying variance and improve the standard errors with emphasis on state-level estimation.

METHODS

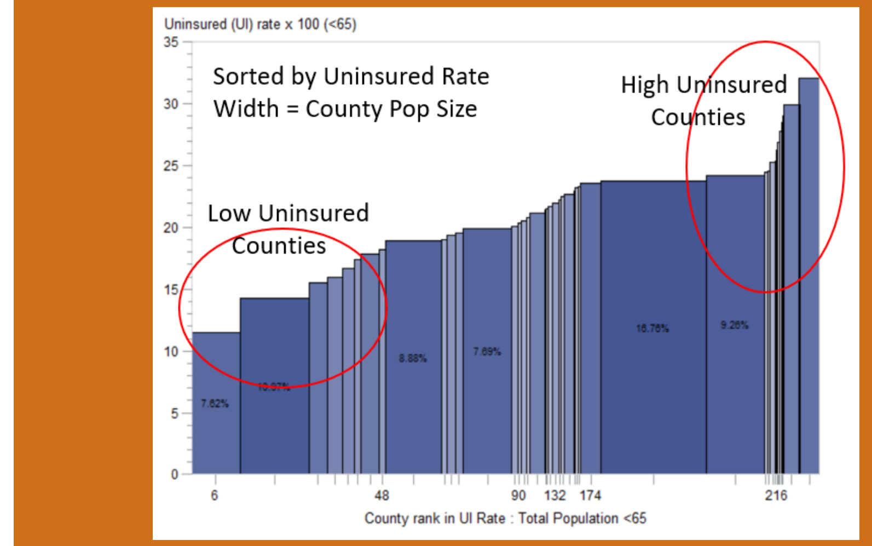
- **1st Stage Sorting.** We considered *four* different PSU sorting variables: Random, Geographical, Population Size, and Health Insurance access. These were compared in the context of measuring state-level Health Insurance estimates from SAHIE.
- **DF.** The naïve degrees of freedom: # PSUs - # Strata is modified for differing PSU sizes and numbers. It's compared to a Satterthwaite approximation for states.
- **Linearization.** A new Taylor Series linear approximation was explored that better reflects the systematic sampling. The technique was assessed using simulation.
- **FPC.** Under a πps (i.e., wor pps) sampling of the PSUs, the finite population correction factor is generally recommended at the state-level. Computation of the fpc is non-trivial due to binary (0,1) joint PSU probabilities of selection.

RESULTS

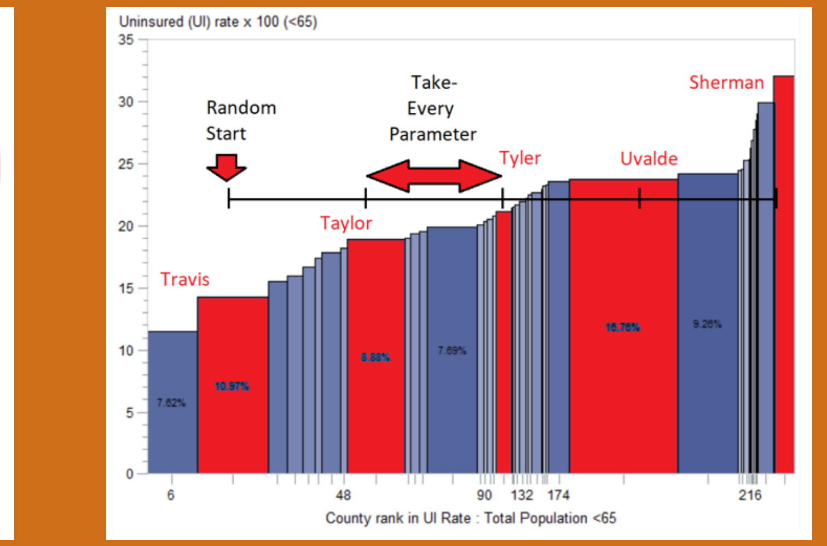
- While Health Ins. sorting fared the best, population-based sorting also performed well in many states. Geo-based sorting may be hampered by Euclidean labeling.
- Satterthwaite-based degrees of freedom were significantly lower in many states, although only a handful of state Health Ins. MOE's were significantly reduced.
- The systematic Taylor Series generated significantly lower standard errors for the linear simulations but were higher under the zero-slope high-variance simulation.
- The Bayesian ranking model produced viable estimates for the joint probabilities, however, the resulting fpc factors were still negative or close to zero.

First-Stage Systematic Sampling of PSUs

Texas Counties Sorted by Health Insurance with Population Size



Texas Counties Selected based on RS and TE Parameters



Population: Persons 18-64
Source: 2021 Small Area Health Insurance Estimates (SAHIE)

Updating the Degrees of Freedom for State Estimation

Two or more Pseudo-PSUs are required for any Self-Representing PSU. These Pseudo-PSUs differ in number and size compared to other PSUs. The DOF may need adjustment.

Under normality and PSU homoscedasticity, $\hat{V}_{TS}(\bar{y}) \sim \left(\frac{\sigma^2}{n}\right)\chi_d/d$ for degrees of freedom d and # PSUs n and chi-squared distribution χ_d with parameter d.

The DOF is d = # PSUs – # Strata when each stratum has exactly k PSUs each with uniform sample size. A better Satterthwaite approximation for the State DOF is…

$$DOF_{State} = \frac{1}{\frac{p_{SR}^2}{DOF_{SR}} + \frac{p_{SR}^2}{DOF_{NSR}}}$$

for population proportions $p_{SR} + p_{NSR} = 1,$

Bayesian fpc under Systematic Sampling

Wolter (2007) lists the first stage variance estimator of a πps sample as:

$$\hat{V}_{1^{st}}(\bar{y}) = \sum_{i=1}^{n}\sum_{j<i}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{M_i\bar{y}_i}{\pi_i} - \frac{M_j\bar{y}_j}{\pi_j}\right)^2 \qquad with\ fpc = \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}$$

In a systematic 1st stage selection, we tend to have fpc < 0 since for a given sorting of the data:

$$\pi_{ij}|sorting = \begin{cases} 1 & when\ i = RS + mod(j,TE) \\ 0 & otherwise \end{cases}.$$

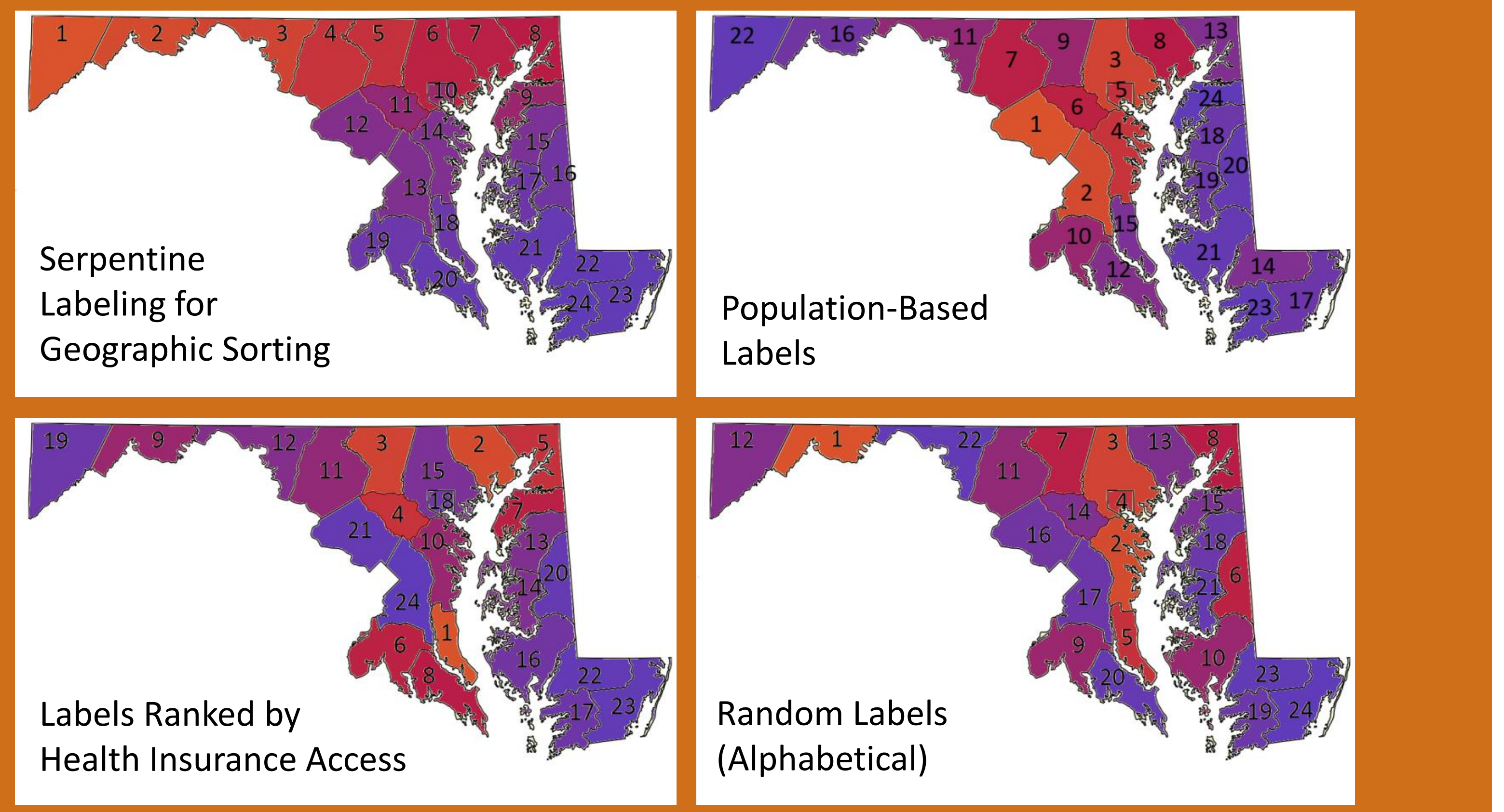
We therefore consider π_{ij} to be random variable, under the which above now is *conditional* on a particular data sorting. A suitable sorting prior leads to the following Bayesian estimator:

$$\hat{\pi}_{ij} = \int_{\Omega}\mathbb{E}\big(1_{ij}|s(\omega)\big)p(s(\omega)) = \int_{P}\mathbb{E}\big(1_{ij}|s\big)p(s)$$

Let $\xi_1, \xi_2, \dots, \xi_N$ denote the 1st stage sorting variable estimates, with $\xi_i \sim \mathbb{N}(\hat{\xi}_i, \sigma_i^2)$. Using the design-based estimator $\hat{\sigma}_i^2$ as a plug-in, we can update the fpc using Monte Carlo methods.

Research Question: How to select the *best* sorting variable?

Maryland Counties Labeled by Different Factors


Serpentine Labeling for Geographic Sorting
Population-Based Labels
Labels Ranked by Health Insurance Access
Random Labels (Alphabetical)

Taylor Series Variance Formula for Systematic Samples

Ignoring the fpc, the Taylor Series variance estimator for a given stratum is given by:

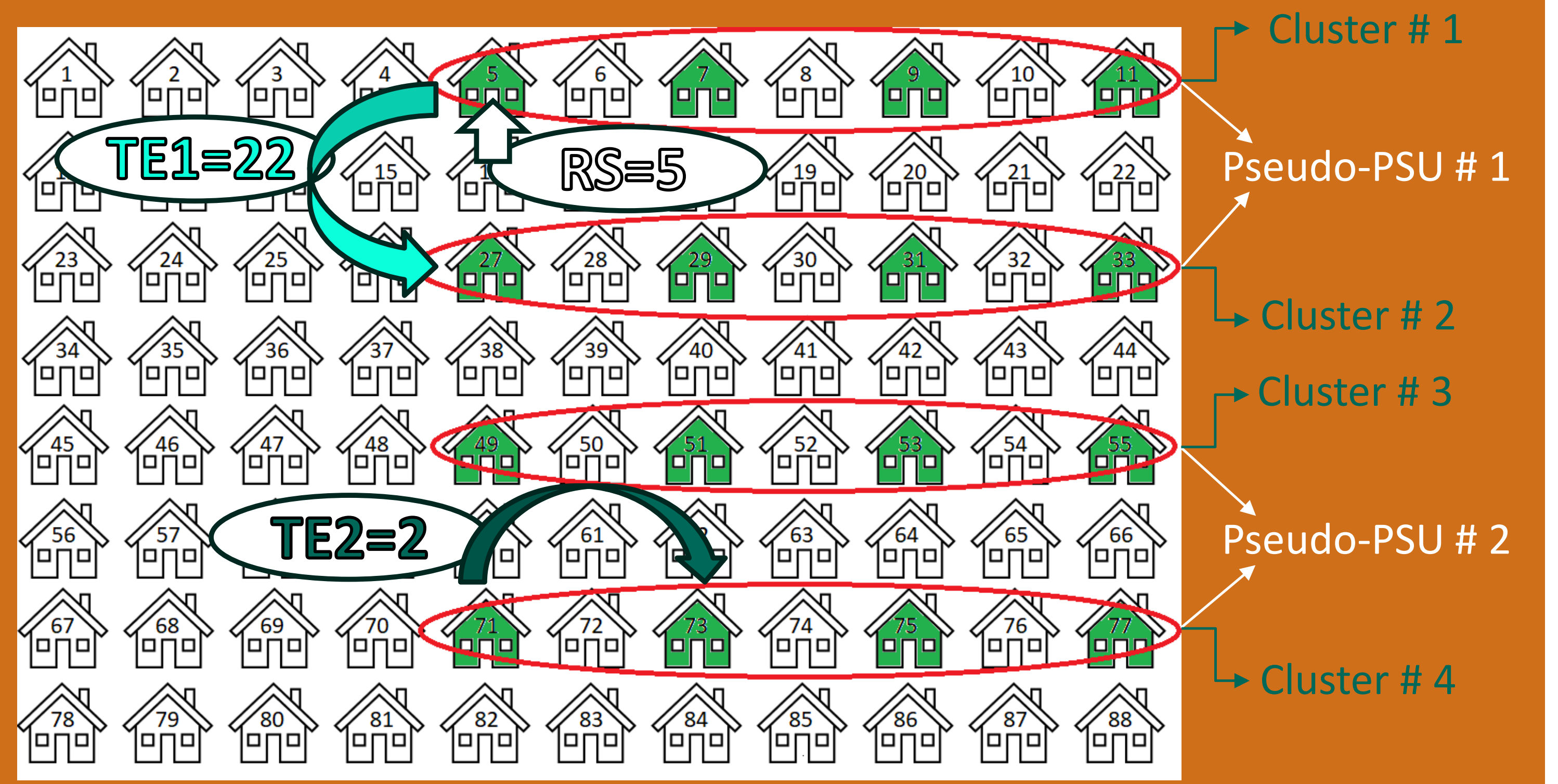
$$\hat{V}_{TS}(\bar{y}) = \frac{n}{n-1}\frac{1}{W^2}\sum_{i=1}^{n}\left(W_i(\bar{y}_i - \bar{y}) - \frac{1}{n}\sum_{t=1}^{n}W_t(\bar{y}_t - \bar{y})\right)^2$$

This estimator ignores the systematic selection of the PSUs leading to $\bar{y}_1 \leq \bar{y}_2 \leq \ \dots \leq \bar{y}_n$. We explored the following estimator and drew comparisons through simulation:

$$\tilde{V}_{TS}(\bar{y}) = \frac{n-1}{n-2} \times \frac{1}{(W-W_1)^2} \times \sum_{i=2}^{n}\left(W_i(\bar{y}_i - \bar{y}_{i-1}) - \frac{1}{n-1}\sum_{t=2}^{n}W_t(\bar{y}_t - \bar{y}_{t-1})\right)^2$$

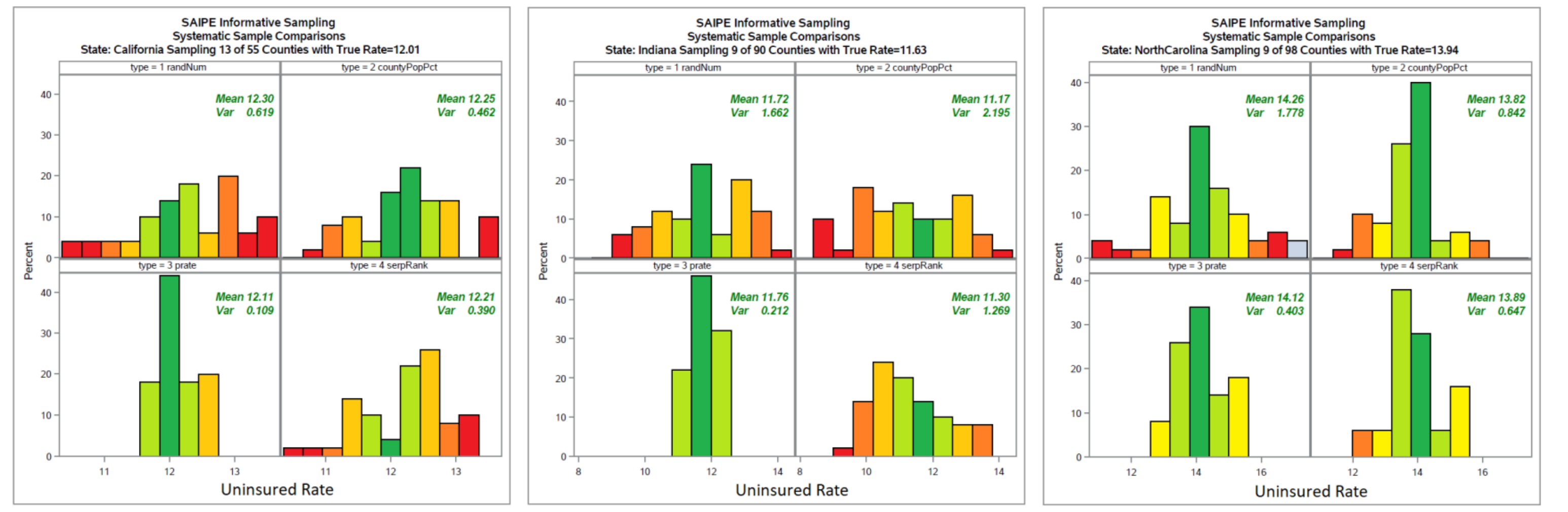
Second-Stage Systematic Cluster Sampling of Households

Within-PSU Sampling Example w/ Four Clusters of Size Four



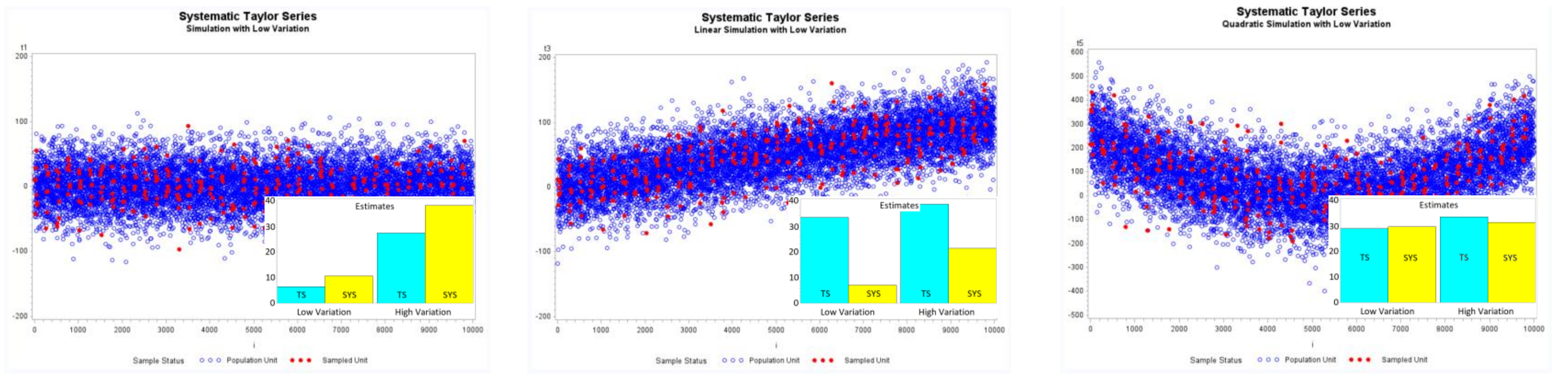
Systematic Sampling Simulations with Estimated State Health Insurance Rates

100 Repeated 1st-Stage Samples per State w/ Histograms (by Sorting Variable)

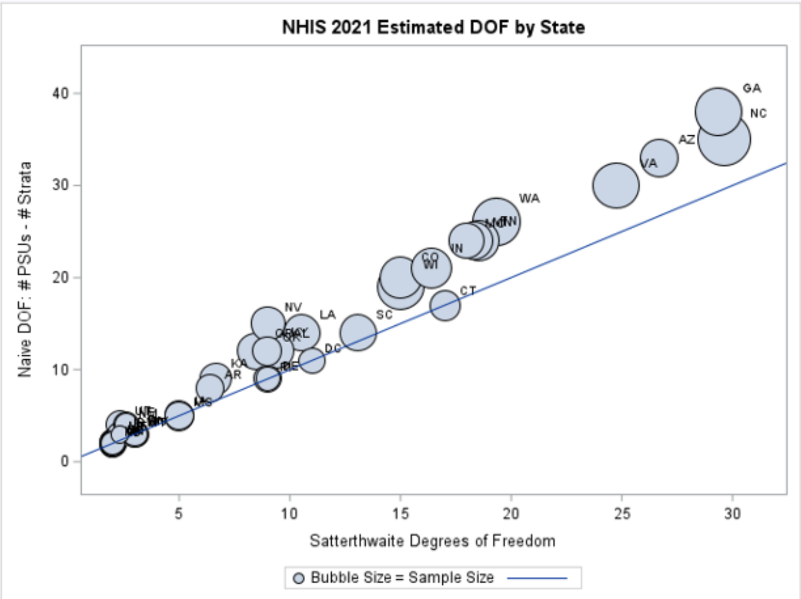


Simulated Populations for Systematic Taylor Series

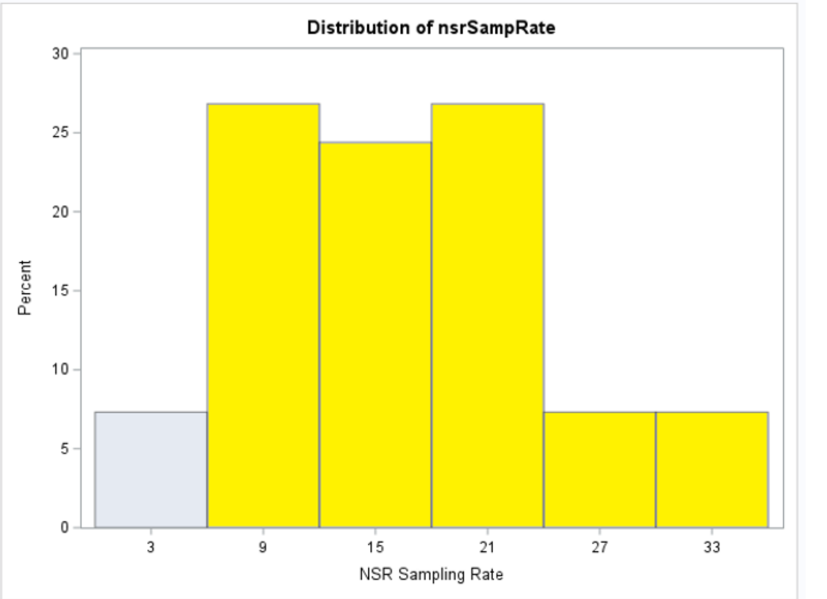
Populations of 10K Units w/ Zero-Slope/Linear/Quadratic Base; Low/High Variation, and 5% Sampling Rate.



Degrees of Freedom by State: Naïve vs. Satterthwaite

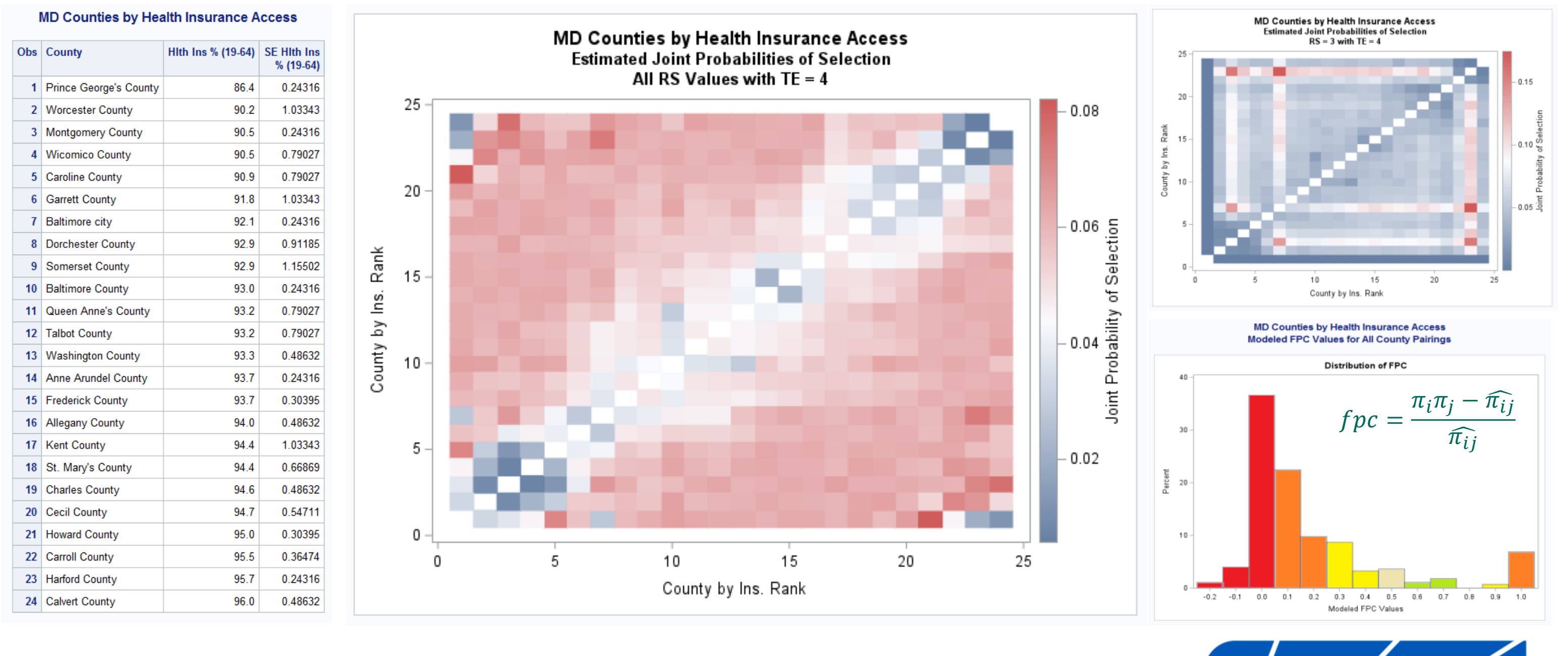


PSU Sampling Rates across States (NSR Only)



Estimating Joint PSU Probabilities of Systematic Selection of MD Counties (25% Sampling Rate)

Randomized Sorting of Counties by Health Insurance (Source: 2021 ACS)



CONTACT INFO

William Waldron, NCHS
nap2@cdc.gov
301-458-4079

CDC

www.cdc.gov | Contact CDC at: 1-800-CDC-INFO or www.cdc.gov/info 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.