Designing Efficient Samples While Accounting for Anticipated Response Rates

Jonathan Mendelson[1] and Michael Elliott[2]

[1] U.S. Bureau of Labor Statistics

[2] University of Michigan

COPAFS Quarterly Meeting, March 7, 2025

Any opinions expressed in this paper are those of the author(s) and do not constitute policy of the Bureau of Labor Statistics.

Bottom line up front

- Efficient samples maximize precision for fixed data collection costs or minimize costs for fixed precision
- Survey designers have long used Neyman allocation (1934) and its extension, optimal allocation, to design efficient samples
  - Existing theory assumes complete response
  - Very little work considers efficient allocation under nonresponse
- We derive optimal allocations under nonresponse, observing efficiency gains of 25% when response rates vary highly by strata

Background concepts: stratified sample allocation

- Stratified random sampling (STSRS) designs are as follows:
  - Step 1. Divide population of N units into H *strata.*
  - Step 2. Within stratum h (for $h = 1,2, \ldots, H$), draw a simple random sample of n_h units from the N_h population units.
- *Sample allocation* here refers to choice of sample sizes, $\{n_h\}$.
- Under 100% response, the *optimal design* minimizes cost or (design-based) variance, holding the other constant.

Background concepts: Neyman allocation

- Neyman (1934) showed that the optimal STSRS allocation for estimating population means or totals is $n_h \propto N_h S_h$.
  - S_h is the stratum h standard deviation.
  - Assumes unit costs, c_h, are equivalent across strata, i.e., $c_h = c$.
- The optimal design under unequal costs is $n_h \propto N_h S_h / \sqrt{c_h}$.
- These results are hugely useful, and underlie many of today's probability samples, but assume 100% response rates.

Gap: how to handle nonresponse in STSRS allocation

- Existing optimal allocation theory (e.g., Stuart 1954; Cochran 1977) generally assumes complete response
  - Exception for dual-frame telephone surveys (Lohr & Brick, 2014)
  - Gap is evidenced by key sampling textbooks' lack of theoretical treatment for how to efficiently allocate samples while accounting for nonresponse
- Gap seems especially problematic given recent transitions toward self-administration and mixed-mode surveys (Olson et al., 2021)
  - Nonresponse can vary by subgroup and meaningfully affect costs

Our setup and notation

Category	Set of units (stratum h)	Number (stratum h)
Population units	U_h	N_h
Original sample (via STSRS)	$s_h \subset U_h$	n_h
Responding sample	$s_{Rh} \subset s_h$	r_h

- Estimate $\bar{Y}$ via $\hat{\bar{Y}} = \sum_{h=1}^{H} \frac{N_h}{N} \sum_{i \in s_{Rh}} \frac{y_{hi}}{r_h}$
  - $\hat{\bar{Y}}$ is a poststratified estimator under nonresponse
  - $\hat{\bar{Y}}$ arises by adjusting the (complete response) design-based estimator by the inverse of strata response rates

(Unconditional) variance of $\hat{\bar{Y}}$

- Assume stratum h units have the same response propensity, $\bar{\phi}_h$
  - Implies responding sample is conditionally STSRS
- Assume at least one respondent per stratum
  - Model as binomial with support for 0 removed
- Then $\mathrm{Var}\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \frac{N_h^2 S_h^2 \zeta_h(n_h, \bar{\phi}_h)}{N^2 n_h \bar{\phi}_h} - \frac{N_h S_h^2}{N^2}$

  where $\zeta_h(n_h, \bar{\phi}_h) := \mathrm{E}\left(\frac{1}{r_h}\right)\mathrm{E}(r_h)$ is a variance inflation factor that reflects the effect of the uncertainty in the responding sample sizes

We assume (variable) strata costs can be decomposed based on response status

- Ignoring fixed costs, we assume that total costs in stratum h are

$$C_h = r_h c_{R_h} + (n_h - r_h) c_{NR_h}, \text{ where}$$

c_{R_h} and c_{NR_h} denote unit costs for a single respondent or nonrespondent.

  - Let $\tau_h = c_{R_h} / c_{NR_h}$ denote the ratio of unit costs for resps. relative to nonresps.

- We consider a few scenarios:

Cost structure scenario	Assumptions	Expected cost per invitee
General (strata-specific)	$\{\tau_h\}$, $\{c_{NR_h}\}$ known	$c_h = c_{NR_h}(\bar{\phi}_h(\tau_h - 1) + 1)$
Common cost structure	$\tau_h = \tau;\ c_{NR_h} = c_{NR}$	$c_h = c_{NR}(\bar{\phi}_h(\tau - 1) + 1)$
Constant cost per invitee	$\tau_h = 1;\ c_{NR_h} = c_{NR}$	$c_h = c_{NR}$

We find the optimal allocation for minimizing the (unconditional) variance or expected costs

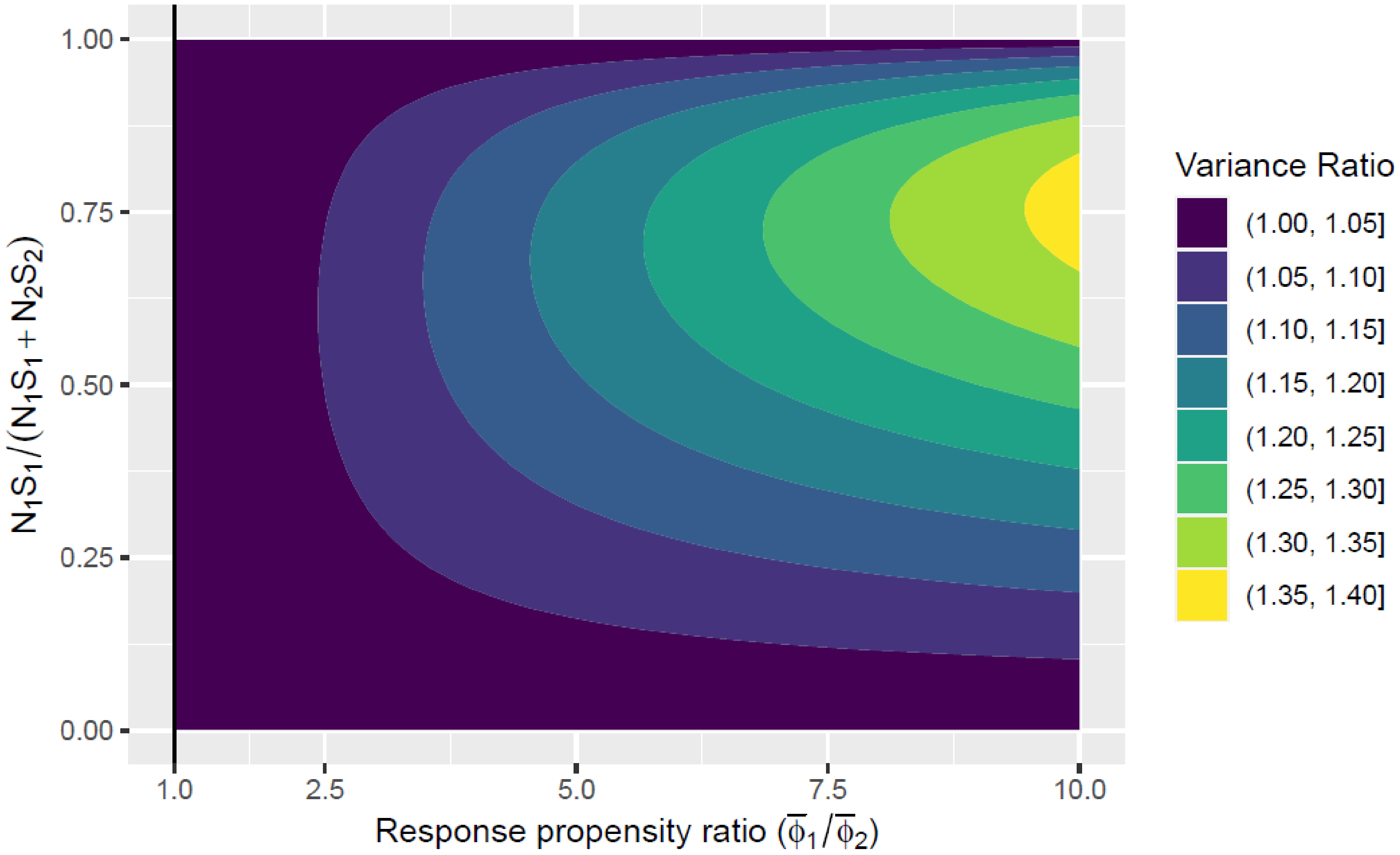
- $n_h \propto \frac{N_h S_h \sqrt{\zeta_h(n_h, \bar{\phi}_h)}}{\sqrt{\bar{\phi}_h c_h}}$
  - Note: n_h and c_h are defined re: invited sample
- Can compute $\{n_h\}$ iteratively by alternating between computations for $\{n_h\}$ and $\{\zeta_h(.)\}$
- We provide an R software implementation of the allocation with our JSSAM paper and via GitHub

We compare the (approximate) proposed allocation to two standard approaches:

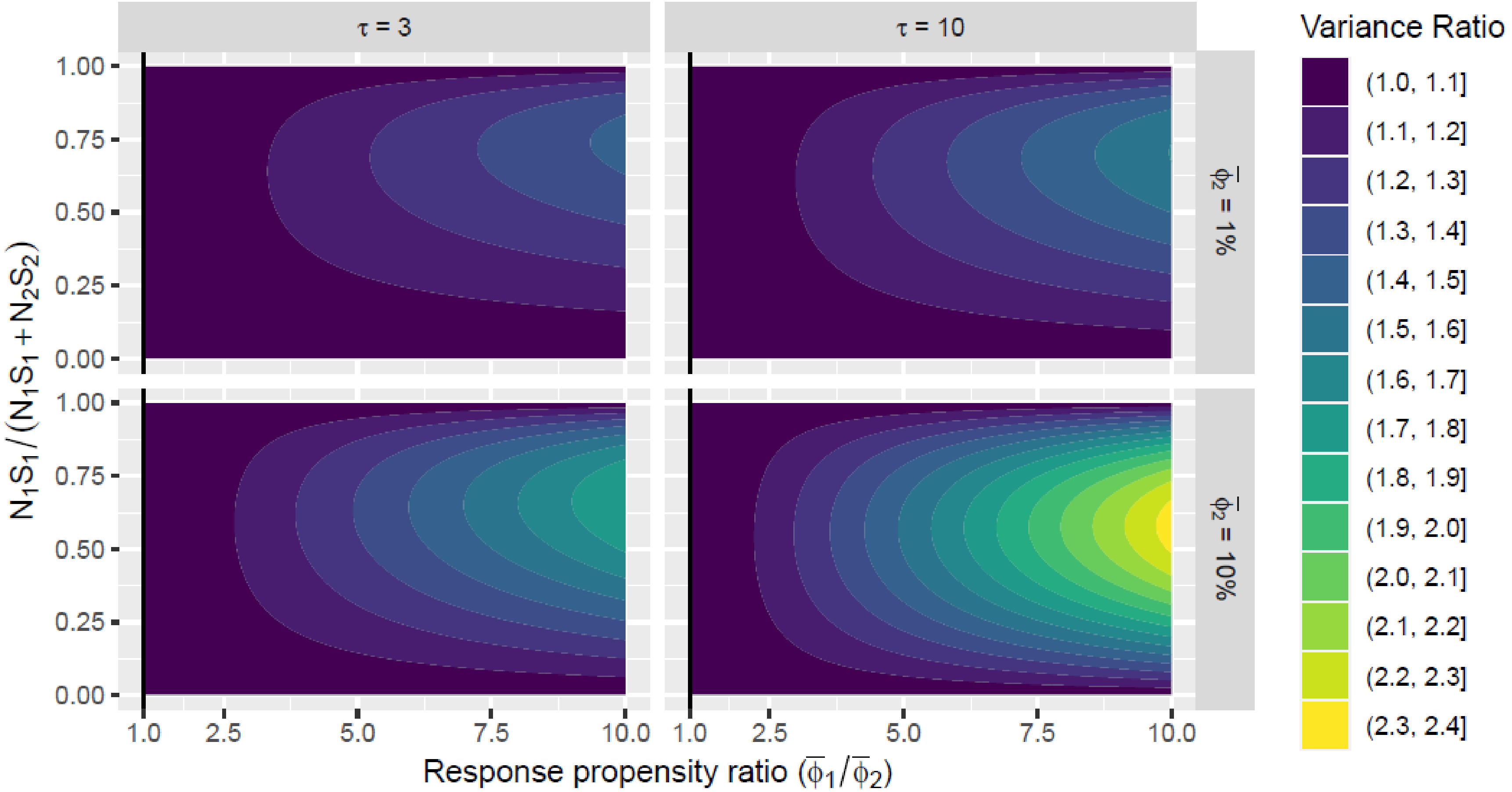
Allocation description	**Notation**	**Source of inefficiency**
Neyman/invitees	$n_h^{Ninv} \propto N_h S_h$	Excessive design effects
Neyman/respondents	$n_h^{Nresp} \propto N_h S_h / \bar{\phi}_h$	Excessive interview costs

- We consider the approximate variances of the standard approaches relative to that of the proposed design
  - We assume small r_h / N_h and large $n_h \bar{\phi}_h$
- We prove that under the *constant cost per invitee* scenario, Neyman allocations of invitees and respondents are equally inefficient!

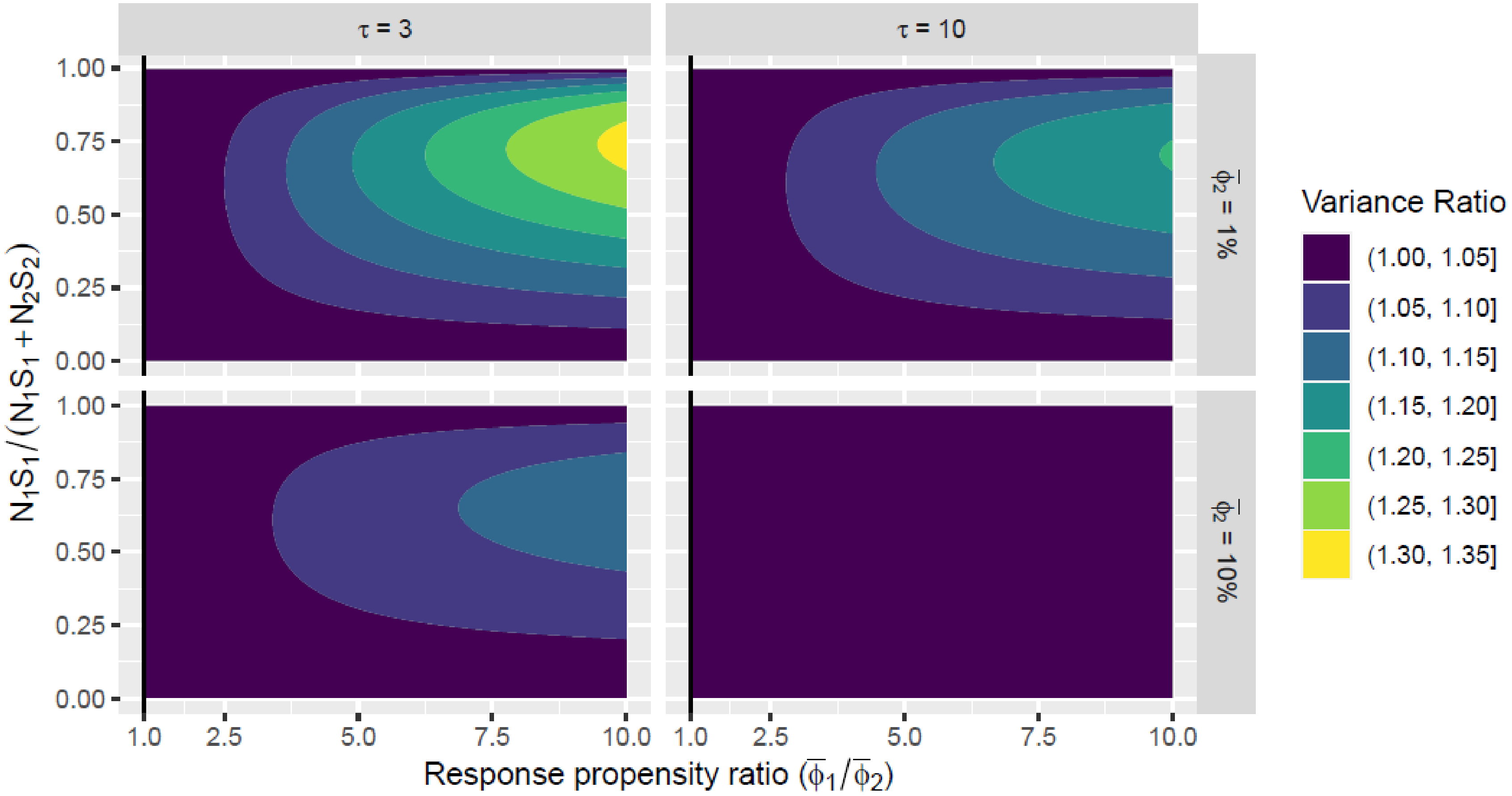
Variance ratio (VR) of standard approach (N/inv or N/resp) to proposed method under *common cost per invitee* ($H = 2$ example)



VR: Neyman/invitees to proposed method under *common cost structure* $(H = 2)$



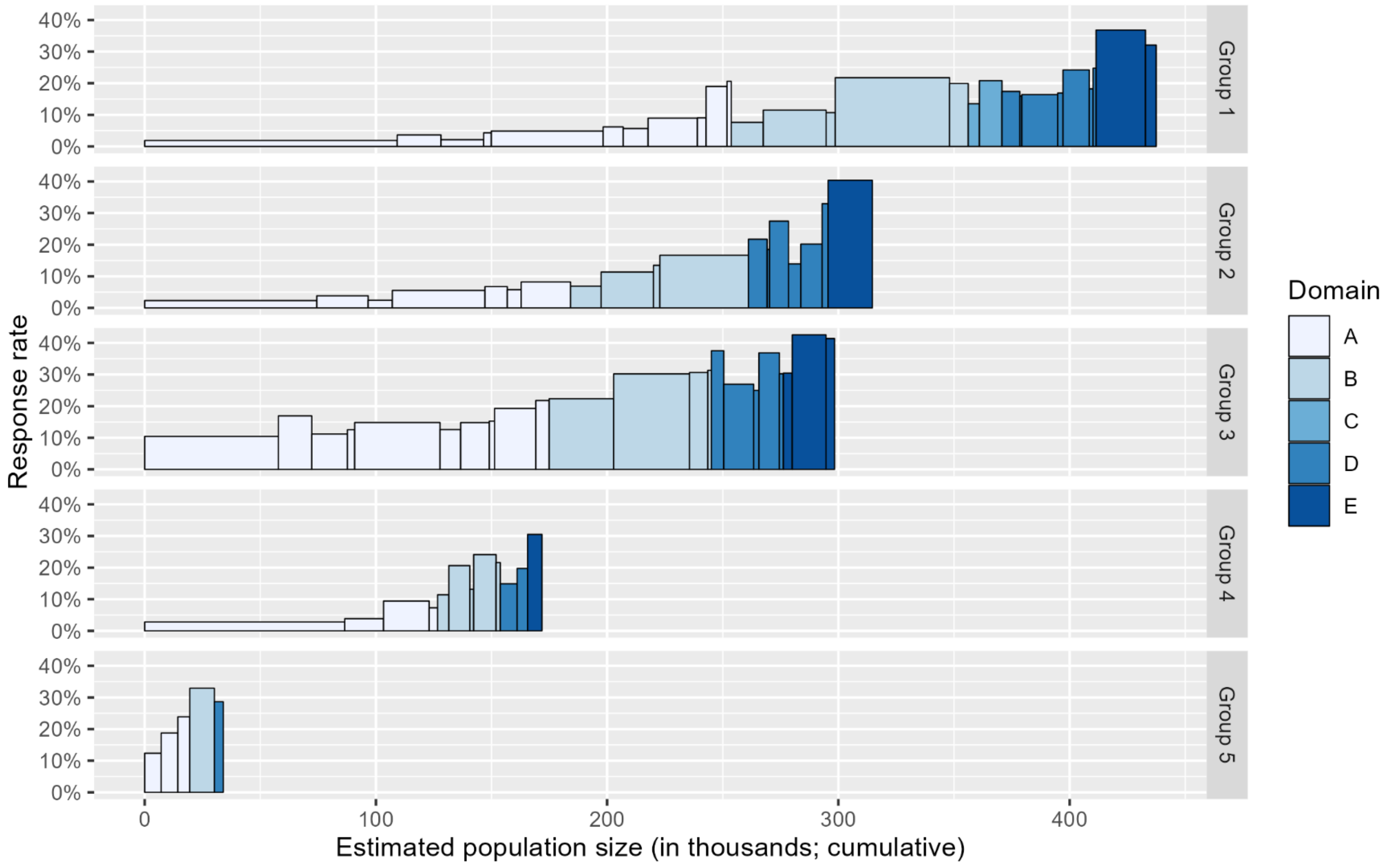
VR: Neyman/<u>respondents</u> to proposed method under *common cost structure* $(H = 2)$



Example. Application to Self-Administered Survey

- We compared allocations using a relevant public use dataset that was available at the time of conducting this research
- Survey characteristics
  - Sponsor type: federal agency (not a PSA)
  - Frame type: population list
  - Contact modes: mail and email (up to 4 mail contacts and 8 email contacts, for a total of 12)
  - Response mode: web
  - Sample analyzed: *n* = 75,548 invitees
    - Public data have base weights, poststrata, disposition codes
    - 14% response rate (AAPOR RR2)

Response rates varied substantially across subgroups



We allocated *n* = 50k under *constant cost per invitee*

Allocation	Notation	Response rate (%)	Respondents	$deff_w$	Effective respondents
Neyman/invitees	$n_h \propto N_h S_h$	13.6	6,794	2.37	2,865

Note. Assumes $r_h = n_h \bar{\phi}_h$ *and* S_h *constant across strata.*

We allocated *n* = 50k under *constant cost per invitee*

Allocation	Notation	Response rate (%)	Respondents	$deff_w$	Effective respondents
Neyman/invitees	$n_h \propto N_h S_h$	13.6	6,794	2.37	2,865
Neyman/resp.	$n_h \propto N_h S_h / \bar{\phi}_h$	5.7	2,865	1.00	2,865

Note. Assumes $r_h = n_h \bar{\phi}_h$ *and* S_h *constant across strata.*

We allocated *n* = 50k under *constant cost per invitee*

Allocation	Notation	Response rate (%)	Respondents	$deff_w$	Effective respondents
Neyman/invitees	$n_h \propto N_h S_h$	13.6	6,794	2.37	2,865
Neyman/resp.	$n_h \propto N_h S_h / \bar{\phi}_h$	5.7	2,865	1.00	2,865
Proposed/approx.	$n_h \propto N_h S_h / \sqrt{c_h \bar{\phi}_h}$	9.0	4,494	1.26	3,570
Proposed/exact	$n_h \propto N_h S_h \sqrt{\zeta_h(.)} / \sqrt{c_h \bar{\phi}_h}$	9.0	4,496	1.26	3,570

Note. Assumes $r_h = n_h \bar{\phi}_h$ *and* S_h *constant across strata.*

- The proposed allocation increased the effective (responding) sample size by 25%, under equivalent total costs

We saw similar gains when allocating for specific Y's

- The Neyman-type allocations had variances 21%–26% higher than the proposed method

$100{,}000 * \mathrm{Var}\left(\hat{\bar{Y}}\right)$ for specific Y's, by allocation				
Allocation	$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_3$	$\bar{Y}_4$
Neyman/invitees	1.60	1.50	1.61	1.41
Neyman/respondents	1.63	1.51	1.62	1.42
Proposed/approx.	1.32	1.20	1.29	1.12
Proposed/exact	1.32	1.20	1.29	1.12

Note. Assumes constant cost per invitee, $S_h = \hat{S}_h$, and $n = 50{,}000$.

Summary of results

- We extended classic theory for STSRS optimal allocation to allow for nonresponse
  - Our allocation strikes a better balance between design effects and cost-per-complete than existing practices
- We see the best gains when response rates vary greatly by strata
- N/inv and N/resp are equally inefficient if $\tau = 1$
  - Larger τ mitigates inefficiencies of N/resp but exacerbates that of N/inv
- We show the importance of incorporating anticipated nonresponse into the cost model assumed for design

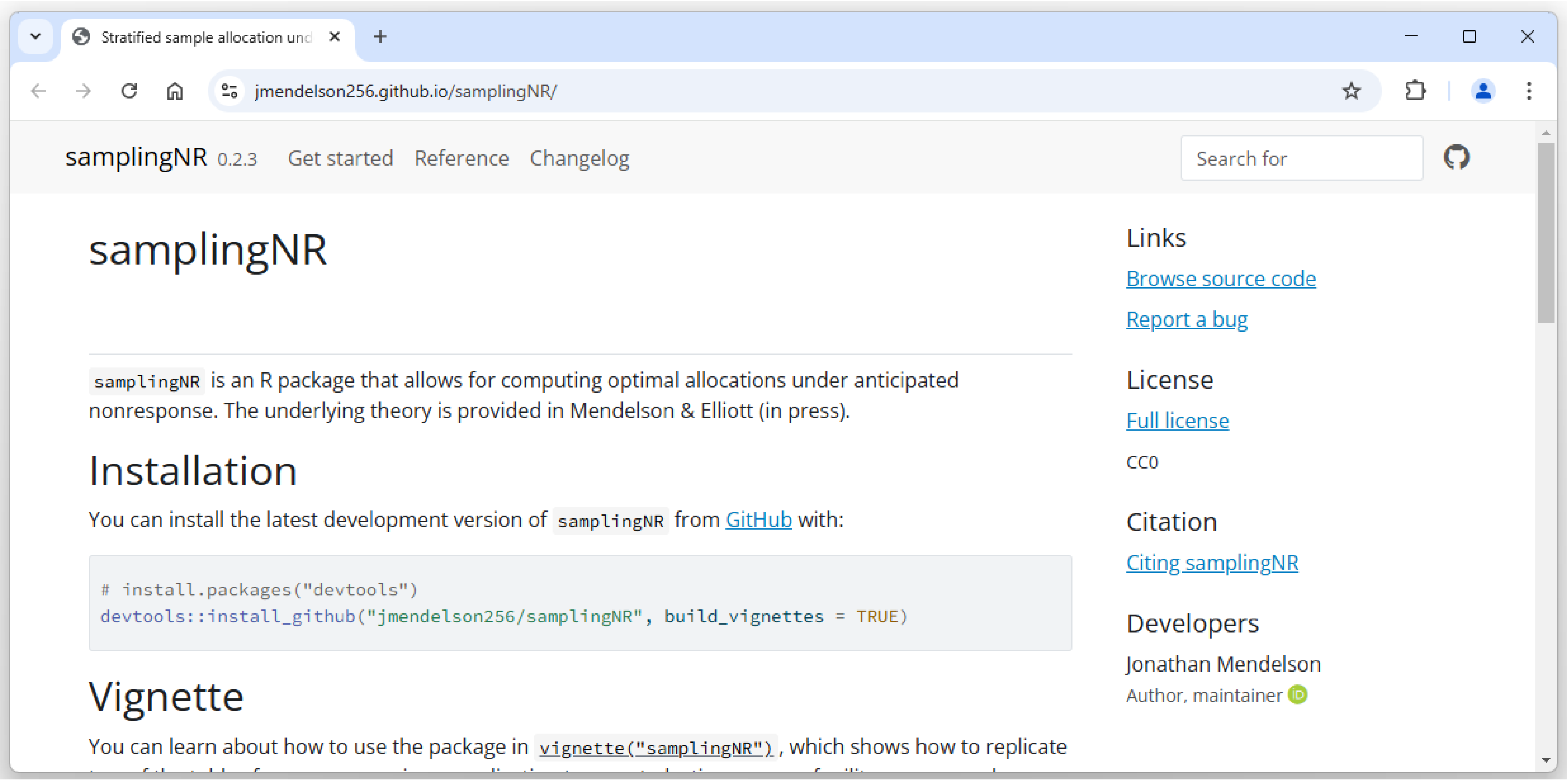
Limitations and future directions

- We assumed constant response propensities within strata, use of PS estimator under nonresponse, and known $\{\bar{\phi}_h\}$ and $\{S_h\}$
  - However, existing theory has similar assumptions but under 100% response!
  - Our allocation might be less sensitive to misspecified response rates than N/resp (due to milder oversampling of low response rate strata)
- Future work could consider different variance and/or cost structures
  - Could treat unknown eligibility via analogy to domain estimation
  - Likewise, more work needed on costs for other contexts (e.g., multi-stage, sequential mixed-mode)

Potential implications for statistical agencies

- Applicability is clearest for cross-sectional surveys with response rates that vary greatly by groups and in a manner driving costs
  - Provides another tool for dealing with response rate challenges, although potential gains will vary by survey
  - Note the potential trade-offs with domain (subgroup) precision
- Research illustrates utility of examining sampling assumptions, especially regarding costs and response rates

Software implementation in R is freely available at https://github.com/jmendelson256/samplingNR/



References

- Cochran, W. G. (1977), Sampling Techniques, New York, NY: John Wiley & Sons, Inc.
- Lohr, S. L., and Brick, J. M. (2014), "Allocation for Dual Frame Telephone Surveys with Nonresponse," *Journal of Survey Statistics and Methodology*, 2, 388–409.
- Mendelson, J., and Elliott, M. R. (2024), "Optimal Allocation Under Anticipated Nonresponse," *Journal of Survey Statistics and Methodology,* 12(5), 1405–1429*.* https://doi.org/10.1093/jssam/smae020
- Neyman, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, 97, 558–62
- Olson, K., Smyth, J. D., Horwitz, R., Keeter, S., Lesser, V., Marken, S., Mathiowetz, N. A., McCarthy, J. S., O'Brien, E., Opsomer, J. D., Steiger, D., Sterrett, D., Su, J., Suzer-Gurtekin, Z. T., Turakhia, C., and Wagner, J. (2021), "Transitions from Telephone Surveys to Self- Administered and Mixed-Mode Surveys: AAPOR Task Force Report," *Journal of Survey Statistics and Methodology*, 9, 381–411.
- Stuart, A. (1954), "A Simple Presentation of Optimum Sampling Results," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 16, 239–24

Contact Information

Jonathan Mendelson
Research Statistician,
Office of Survey Research Methods
Behavioral Science Research Center
202-691-7268
mendelson.jonathan@bls.gov