(Pseudo-Bayesian) Inference for Complex Survey Data

Matt Williams[1]

[1]Center for Official Statistics
RTI International
mrwilliams@rti.org

Collaborative Research Cycle Explanatory Workshop
Dec 17, 2023

Thank you!

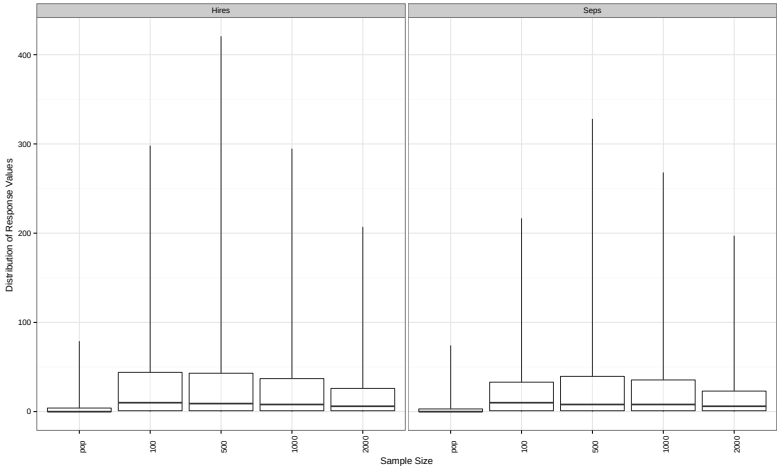
- Terrance Savitsky (BLS) for being a great collaborator and mentor.
- Christine Task (Knexus) and Gary Howarth (NIST) for keeping me in the loop over the years!
- To you all for your time and energy!

Bio

1. Work
   - 2 years as senior research statistician at RTI: National Survey on Drug Use and Health (SAMHSA) and Model-based early estimates (NCHS)
   - 10 years as mathematical statistical for federal government: USDA, HHS, NSF
   - Sample design, weighting, imputation, estimation, disclosure limitation (production and methods development)
2. Consulting
   - International surveys for agricultural production (USAID) and vaccination knowledge, attitudes, and behaviors (UNICEF)
3. Research (ORCID: 0000-0001-8894-1240)
   - Constrained Optimization for Survey Applications (weight adjustment, benchmarking model estimates)
   - Applying Bayesian inference methods to data from complex surveys.
   - Bayesian synthetic data for privacy protection.

Outline

Distributions of **y** in Informative Samples



Population Inference from Informative Samples

- Goal: perform inference about a finite population generated from an unknown model, $\mathbb{P}_{\theta_0}(\mathbf{y})$.
- Data: from under a complex sampling design distribution, $\mathbb{P}_{\nu}(\boldsymbol{\delta})$
  - Probabilities of inclusion $\pi_i = Pr(\delta_i = 1|\mathbf{y})$ are often associated with the variable of interest (purposefully)
  - Sampling designs are "informative": the balance of information in the sample $\neq$ balance in the population.
- Biased Estimation: estimate $\mathbb{P}_{\theta_0}(\mathbf{y})$ without accounting for $\mathbb{P}_{\nu}(\boldsymbol{\delta})$.
  - Use inverse probability weights $w_i = 1/\pi_i$ to mitigate bias.
- Incorrect Uncertainty Quantification:
  - Failure to account for dependence induced by $\mathbb{P}_{\nu}(\boldsymbol{\delta})$ leads to standard errors and confidence intervals that are the wrong size.

Pseudo Posterior

- Pseudo posterior $\propto$ Pseudo Likelihood $\times$ Prior

$$p^{\pi}(\boldsymbol{\theta}|\mathbf{y}, \tilde{\mathbf{w}}) \propto \left[\prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})^{\tilde{w}_i}\right] p(\boldsymbol{\theta})$$

$$w_i := \frac{1}{\pi_i}$$

$$\tilde{w}_i = \frac{w_i}{\frac{\sum w_i}{n}}, \; i = 1, \ldots, n$$

Outline

Frequentist Consistency of a (Pseudo) Posterior

- Estimated distribution $p^{\pi}(\boldsymbol{\theta}|\mathbf{y}, \tilde{\mathbf{w}})$ collapses around generating parameter θ_0 with increasing population N_ν and sample n_ν sizes.
  - Evaluated with respect to joint distribution of population generation $\mathbb{P}_{\theta_0}(\mathbf{y})$ and the sample inclusion indicators $\mathbb{P}_\nu(\boldsymbol{\delta})$.
- Conditions on the model $\mathbb{P}_{\theta_0}(\mathbf{y})$ (standard)
  - Complexity of the model limited by sample size
  - Prior distribution not too restrictive (e.g. point mass)
- Conditions on the sampling design $\mathbb{P}_\nu(\boldsymbol{\delta})$ (new-ish)
  - Every unit in population has non-zero probability of inclusion $\implies$ finite weights
  - Dependence restricted to countable blocks of bounded size $\implies$ arbitrary dependence within clusters, but approximate independence between clusters.

Simulation Example: Three-Stage Sample

Area (PPS), Household (Systematic, sorting by Size), Individual (PPS)

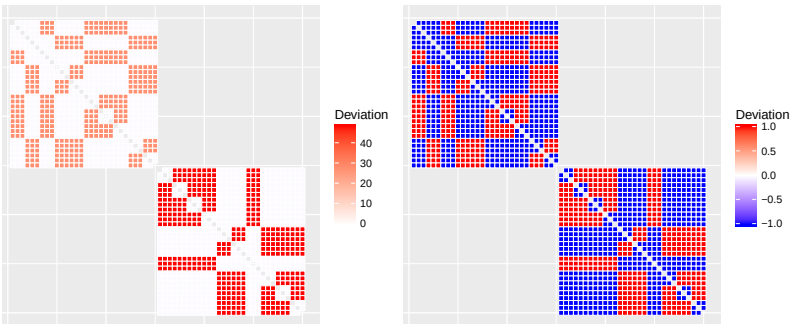


Figure: Factorization matrix $(\pi_{ij}/(\pi_i\pi_j) - 1)$ for two PSU's. Magnitude (left) and Sign (right). Systematic Sampling ($\pi_{ij} = 0$). Clustering and PPS sampling ($\pi_{ij} > \pi_i\pi_j$). Independent first stage sample ($\pi_{ij} = \pi_i\pi_j$)

Simulation Example: Three-Stage Sample (Cont)

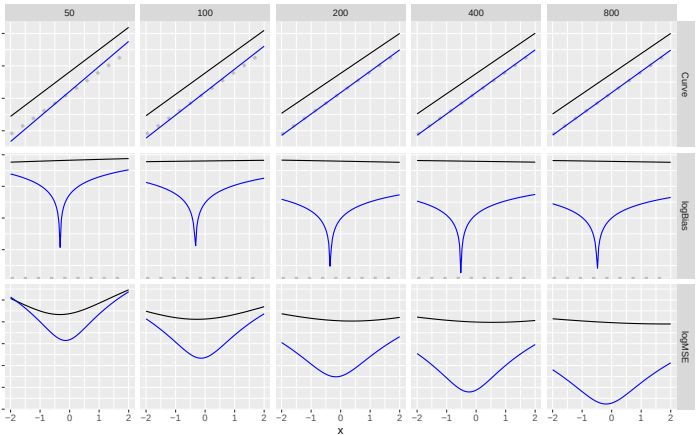


Figure: The marginal estimate of $\mu = f(x_1)$. **population curve**, sample with **equal weights**, and **inverse probability weights**. Top to bottom: estimated curve, log of BIAS, log MSE. Left to right: sample size (50 to 800).

Outline

Variance Estimation

- The de-facto approach:
  - approximate sampling independence of the primary sampling units (Heeringa et al., 2010).
  - within-cluster dependence treated as nuisance
- Two common methods:
  - Taylor linearization and replication based methods.
  - A variety of implementations are available (Binder, 1996; Rao et al., 1992).

Taylor Linearization

Let y_{ij} and w_{ij} be the observed data for individual i in cluster j of the sample. Assume the parameter θ is a vector of dimension d with population model value θ_0.

1. Approximate an estimate $\hat{\theta}$, or a 'residual' $(\hat{\theta} - \theta_0)$, as a weighted sum: $\hat{\theta} \approx \sum_{i,j} w_{ij} z_{ij}(\theta)$ where z_{ij} is a function evaluated at the current values of y_{ij}, and $\hat{\theta}$ (e.g. $z_i(\hat{\theta}) = H_{\theta_0}^{-1} \dot{\ell}_{\hat{\theta}}(\mathbf{y}_i)$).
2. Compute the weighted components for each cluster (e.g., primary sampling units (PSUs)): $\hat{\theta}_j = \sum_i w_{ij} z_{ij}(\theta)$.
3. Compute the variance between clusters: $\widehat{Var(\hat{\theta})} = \frac{1}{J-d} \sum_{j=1}^{J} (\hat{\theta} - \hat{\theta}_j)(\hat{\theta} - \hat{\theta}_j)^T$
4. For stratified designs, compute $\hat{\theta}_s$ and $\widehat{Var(\hat{\theta}_s)}$ within strata and sum $\widehat{Var(\hat{\theta})} = \sum_s \widehat{Var(\hat{\theta}_s)}$.

Replication

Let y_{ij} and w_{ij} be the observed data for individual i in cluster j of the sample. Assume the parameter θ is a vector of dimension d with population model value θ_0.

1. Through randomization (bootstrap), leave-one-out (jackknife), or orthogonal contrasts (balanced repeated replicates), create a set of K replicate weights $(w_i)_k$ for all $i \in S$ and for every $k = 1, \ldots, K$.
2. Each set of weights has a modified value (usually 0) for a subset of clusters, and typically has a weight adjustment to the other clusters to compensate: $\sum_{i \in S} (w_i)_k = \sum_{i \in S} w_i$ for every k.
3. Estimate $\hat{\theta}_k$ for each replicate $k \in 1, \ldots, K$.
4. Compute the variance between replicates: $\widehat{Var(\hat{\theta})} = \frac{C}{K-d} \sum_{k=1}^{K} (\hat{\theta} - \hat{\theta}_k)(\hat{\theta} - \hat{\theta}_k)^T$.
5. For stratified designs, generate replicates such that each strata is represented in every replicate.

Challenges

There are two notable trade-offs associated with these methods:

- Taylor linearization: value $\hat{\theta}$ computed once then used in a plug in for $z_i(\theta)$.
  - Replication methods: estimate $\hat{\theta}_k$ computed K times.
  - Sizable differences in computational effort
- Replication methods: no derivatives are needed.
  - Taylor linearization: requires the calculation of a gradient to derive the analytical form of the first order approximation $z_i(\theta)$.
  - This poses significant analytical challenges for all but the simplest models.

Some Improvements

- Abstraction of Derivatives (less analytic work for Taylor Linearization)
  - Recent advances in algorithmic differentiation (Margossian, 2018), allows us to specify the model as a log density but only treat the gradient in the abstract without specifying it analytically.
  - Implemented in Stan and Rstan (Carpenter, 2015; Stan Development Team, 2016)
- Hybrid Approach or Taylor Linearization for replicate designs (less computation for Replication approaches)
  - Survey package (Lumley, 2016) to calculate replication variance of gradient $\hat{\ell}_\theta$
  - Use plug in for θ, only estimate once

$$(\hat{\psi} - \psi_0) = H_{\theta_0}(\hat{\theta} - \theta_0) \approx \sum_{i \in S} w_i \dot{\ell}_{\hat{\theta}}(\mathbf{y}_i) = \sum_{i \in S} w_i z_i(\hat{\theta}),$$

with $\text{Var}_{P_{\theta_0}, P_\nu}(\hat{\psi} - \psi_0) = J^{\pi}_{\theta_0}$.

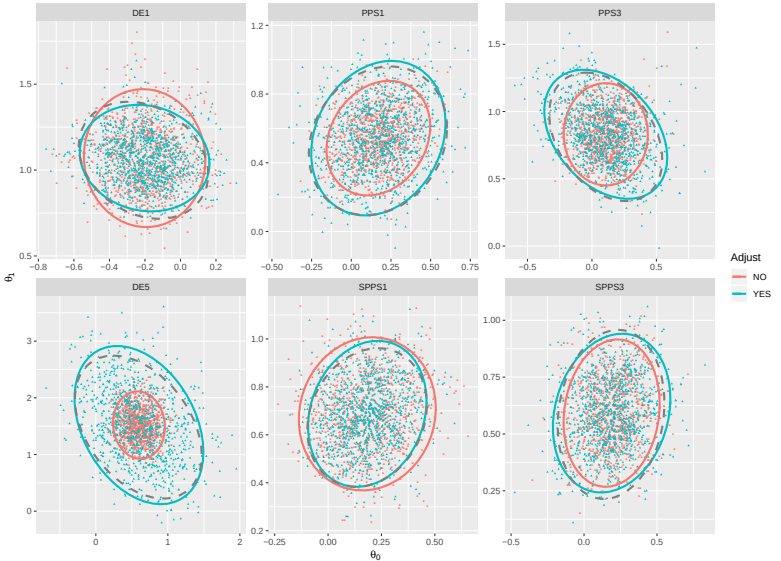
Example: Design Effect for Survey-Weighted Bayes

- Pseudo posterior $\propto$ Pseudo Likelihood $\times$ Prior

$$p^{\pi}\left(\boldsymbol{\theta}|\mathbf{y},\tilde{\mathbf{w}}\right) \propto \left[\prod_{i=1}^{n} p\left(y_i|\boldsymbol{\theta}\right)^{\tilde{w}_i}\right] p\left(\boldsymbol{\theta}\right)$$

- Variances Differ:
  - Weighted MLE: $H_{\theta_0}^{-1} J_{\theta_0}^{\pi} H_{\theta_0}^{-1}$ (Robust)
  - Weighted Posterior: $H_{\theta_0}^{-1}$ (Model-Based)
- Adjust for Design Effect: $R_2^{-1} R_1$
  - $\hat{\theta}_m \equiv$ sample pseudo posterior for $m = 1, \ldots, M$ draws with mean $\bar{\theta}$
  - $\hat{\theta}_m^a = \left(\hat{\theta}_m - \bar{\theta}\right) R_2^{-1} R_1 + \bar{\theta}$
  - where $R_1' R_1 = H_{\theta_0}^{-1} J_{\theta_0}^{\pi} H_{\theta_0}^{-1}$
  - $R_2' R_2 = H_{\theta_0}^{-1}$

Joint Distribution



Related Papers

- Consistency of the Pseudo-Posterior (Savitsky and Toth, 2016)
- Uncertainty Quantification (Williams and Savitsky, 2021)
- Extension to multistage surveys (Williams and Savitsky, 2020; Han and Wellner, 2021)
- Extension to pairwise weights and outcomes (Williams and Savitsky, 2018)
- Extension to Divide and Conquer computational methods (Savitsky and Srivastava, 2018)
- Correction of asymptotic coverage (Williams and Savitsky, 2021)
- Joint modeling of Outcome and Weights (León-Novelo and Savitsky, 2019; Leon-Novelo and Savitsky, 2021)

References I

Binder, D. A. (1996), 'Linearization methods for single phase and two-phase samples: a cookbook approach', *Survey Methodology* **22**, 17–22.

Carpenter, B. (2015), 'Stan: A probabilistic programming language', *Journal of Statistical Software* .

Han, Q. and Wellner, J. A. (2021), 'Complex sampling designs: Uniform limit theorems and applications', *The Annals of Statistics* **49**(1), 459–485.

Heeringa, S. G., West, B. T. and Berglund, P. A. (2010), *Applied Survey Data Analysis*, Chapman and Hall/CRC.

Leon-Novelo, L. G. and Savitsky, T. D. (2021), 'Fully bayesian estimation under dependent and informative cluster sampling', *arXiv preprint arXiv:2101.06237* .

León-Novelo, L. G. and Savitsky, T. D. (2019), 'Fully bayesian estimation under informative sampling', *Electron. J. Statist.* **13**(1), 1608–1645.
URL: *https://doi.org/10.1214/19-EJS1538*

Lumley, T. (2016), 'survey: analysis of complex survey samples'. R package version 3.32.

Margossian, C. C. (2018), 'A review of automatic differentiation and its efficient implementation', *CoRR* **abs/1811.05031**.
URL: *http://arxiv.org/abs/1811.05031*

References II

McGuire, F. H., Beccia, A. L., Peoples, J., Williams, M. R., Schuler, M. S. and Duncan, A. E. (2023), 'Depression at the intersection of race/ethnicity, sex/gender, and sexual orientation in a nationally representative sample of us adults: A design-weighted maihda', *medRxiv* .
URL: *https://www.medrxiv.org/content/early/2023/04/17/2023.04.13.23288529*

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992), 'Some recent work on resampling methods for complex surveys', *Survey Methodology* **18**, 209–217.

Savitsky, T. D. and Srivastava, S. (2018), 'Scalable bayes under informative sampling', *Scandinavian Journal of Statistics* **45**(3), 534–556. 10.1111/sjos.12312.
URL: *http://dx.doi.org/10.1111/sjos.12312*

Savitsky, T. D. and Toth, D. (2016), 'Bayesian Estimation Under Informative Sampling', *Electronic Journal of Statistics* **10**(1), 1677–1708.

Savitsky, T. D. and Williams, M. R. (2022), 'Pseudo bayesian mixed models under informative sampling', *Journal of Official Statistics* **38**(3), 901–928.

Savitsky, T. D., Williams, M. R., Gershunskaya, J., Beresovsky, V. and Johnson, N. G. (2023), 'Methods for combining probability and nonprobability samples under unknown overlaps'.

Savitsky, T. D., Williams, M. R. and Srivastava, S. (2020), 'Pseudo bayesian estimation of one-way anova model in complex surveys', *arXiv preprint arXiv:2004.06191* .

References III

Stan Development Team (2016), 'RStan: the R interface to Stan'. R package version 2.14.1. **URL:** *http://mc-stan.org/*

Williams, M. R. and Savitsky, T. D. (2018), 'Bayesian pairwise estimation under dependent informative sampling', *Electron. J. Statist.* **12**(1), 1631–1661.

Williams, M. R. and Savitsky, T. D. (2020), 'Bayesian estimation under informative sampling with unattenuated dependence', *Bayesian Anal.* **15**(1), 57–77. **URL:** *https://doi.org/10.1214/18-BA1143*

Williams, M. R. and Savitsky, T. D. (2021), 'Uncertainty estimation for pseudo-bayesian inference under complex sampling', *International Statistical Review* **89**(1), 72–107. **URL:** *https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12376*