Bayesian Estimation Under Informative Sampling with Unattenuated Dependence

Matt Williams[1] Terrance Savitsky[2]

[1]Substance Abuse and Mental Health Services Administration
Matthew.Williams@samhsa.hhs.gov

[2]Bureau of Labor Statistics
Savitsky.Terrance@bls.gov

Federal Committee on Statistical Methodology
Research and Policy Conference
March 9, 2018

Inference from Survey Samples

- Goal of Analyst: perform inference about a finite population generated from an unknown model, P_0.
- Data Collected: from under a complex sampling design distribution, P_ν
  - Probabilities of inclusion 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 P_0 without accounting for P_ν.

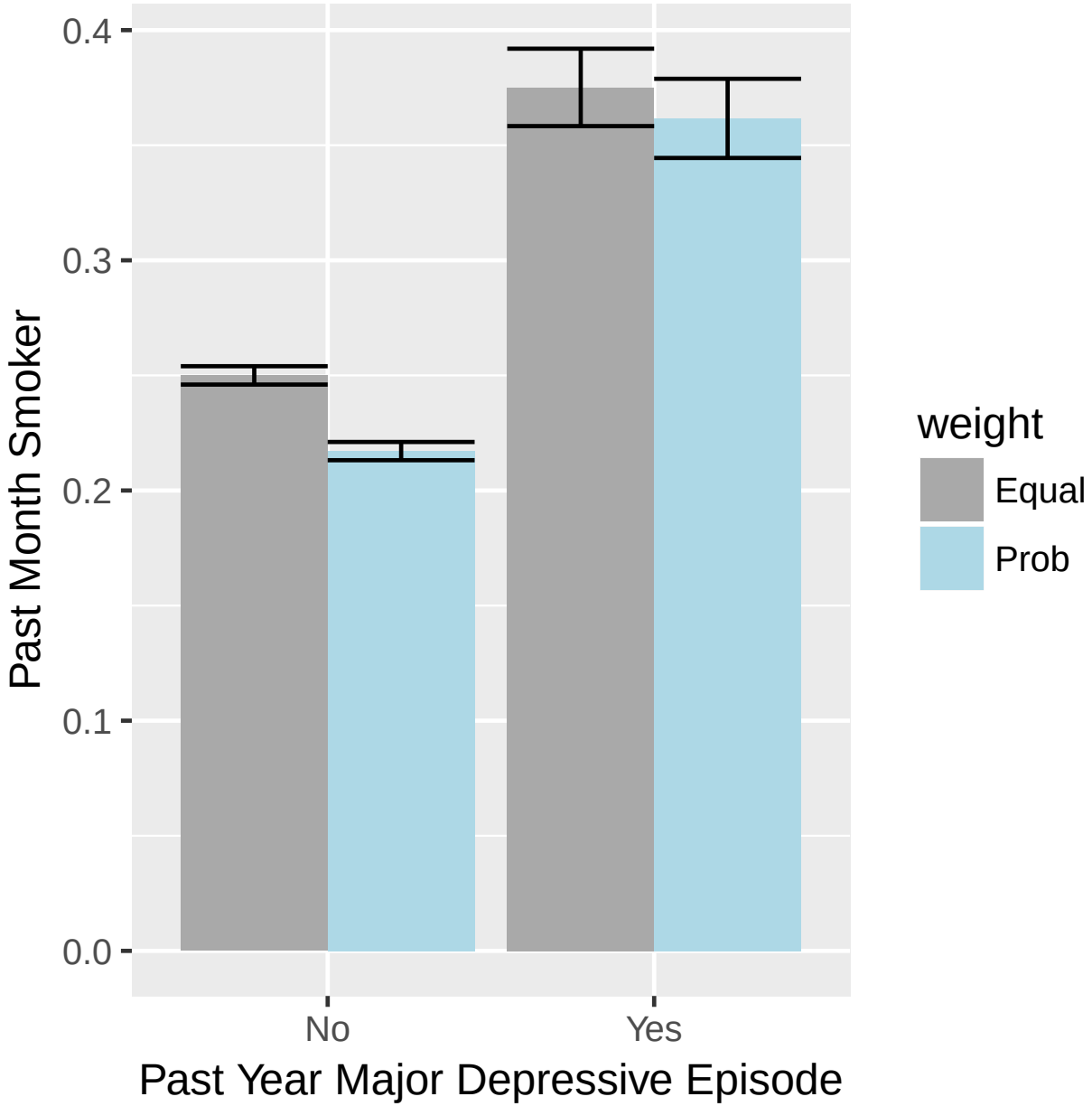
What does it mean for a Design to be Informative?

- Consistency Theory (estimators collapsing to true P_0)
  - Empashize use of inverse probability weights, $w_i = 1/\pi_i$ using marginal (individual) probabilities of being selected
  - Weights are used as plug-in corrections (Savitsky and Toth, 2016) or co-modelled (Novelo and Savitsky, 2017)
  - Sampled elements assumed to become more and more independent with larger sample/population sizes ($\pi_{ij} \rightarrow \pi_i \pi_j$)
- Estimation in Practice
  - Weighting, stratification, and clustering are all incoporated.
  - Variances estimated by aggregating up to the (asymptotically) independent units (cluster ID's or primary sampling units (PSU's).
  - Individual units aren't assumed to be sampled independently in general ($\pi_{ij} \neq \pi_i \pi_j$).

Motivating Example: The National Survey on Drug Use and Health (NSDUH)

- Sponsored by Substance Abuse and Mental Health Services Administration (SAMHSA)
  - Primary source for: alcohol, illicit drug, substance use disorder, mental health issues and their co-occurrence
  - Civilian, non-institutionalized population for 12 and older in the US.
- Multistage, state-based sampling design (Morton et al., 2016)
  - Early stages defined by geography: Select dwelling units (DU's) nested within census block groups (PSU's)
  - Geographic units stratified 'implicitly' via sorting on socioeconomic indicators and selected proportional to compositive population size measure (Systematic PPS).
  - DU's selected within segment via random starting point and selecting every k^{th} unit (Systematic)
  - Individuals (0,1, or 2), selected jointly as pairs with different probabilities assigned to different age-pair combinations (Unequal joint selection of pairs)

NSDUH: Depression and Smoking (12 and older, 2014)



Motivating Example: NSDUH (continued)

- Smoking and depression have the potential to cluster geographically and within dwelling units, since related demographics (age, urbanicity, education, etc) may cluster.
- Current literature is silent on the issue of non-ignorable clustering that may be informative (i.e. related to the response of interest).
- We present conditions for asymptotic consistency for designs like the NSDUH with common features such as:
  - Cluster sampling, selecting only one unit per cluster, or selecting multiple individuals from a dwelling unit.
  - Population information used to sort sampling units along gradients.

Estimation Methods to Account for Informative Sampling

- First-order weights and likelihood
  - Plug-in to exponentiate likelihood (Savitsky and Toth, 2016)
  - Joint modeling of weights and outcome (Novelo and Savitsky, 2017)
- Second-order (pairwise weights) and composite likelihood
  - Clever way to perform hierarchical modelling when design and population model have same structure (Yi et al., 2016)
  - Only beats first-order weights under specific conditions for population model and sample (Williams and Savitsky, 2017)
  - E.g: conditional behavior of spouse-spouse pairs within households AND differential selection of pairs of individuals within a household related to outcome AND subsetting of sample based on information only available in the sample (pair relationship)
- For simplicity, we focus on the *first-order* weights using the *plug-in* method

The Pseudo-Posterior Estimator (Savitsky and Toth, 2016)

The plug-in estimator for posterior density under the analyst-specified model for $\boldsymbol{\lambda} \in \Lambda$ is

$$\hat{\pi}\left(\boldsymbol{\lambda}|\mathbf{y}_o, \tilde{\mathbf{w}}\right) \propto \left[\prod_{i=1}^{n} p\left(y_{o,i}|\boldsymbol{\lambda}\right)^{\tilde{w}_i}\right] \pi\left(\boldsymbol{\lambda}\right), \tag{1}$$

- pseudo-likelihood: $\prod_{i=1}^{n} p\left(y_{o,i}|\boldsymbol{\lambda}\right)^{\tilde{w}_i}$
- prior: $\pi\left(\boldsymbol{\lambda}\right)$
- values $\mathbf{y}_o$ and sampling weights $\{\tilde{\mathbf{w}}\}$ for individuals observed in sample

Consistency Conditions (Savitsky and Toth, 2016)

Based on Empirical Process functionals (Ghosal and van der Vaart, 2007). Population U_ν of size N_ν growing with index $\nu \uparrow \infty$.

(A1) (Local entropy condition - Size of model)

(A2) (Size of space)

(A3) (Prior mass covering the truth)

(A4) (Non-zero Inclusion Probabilities)

$$\sup_{\nu} \left[\frac{1}{\min_{i \in U_\nu} |\pi_{\nu i}|} \right] \leq \gamma, \text{ with } P_0 - \text{probability 1.}$$

(A6) (Constant Sampling fraction)
For constant $f \in (0, 1)$,

$$\limsup_{\nu} \left| \frac{n_\nu}{N_\nu} - f \right| = \mathcal{O}(1), \text{ with } P_0 - \text{probability 1.}$$

Consistency Conditions (NEW)

(A5.1) (Growth of dependence is restricted)
Binary partition $\{S_{\nu 1}, S_{\nu 2}\}$ of the set of all pairs $S_\nu = \{\{i,j\} : i \neq j \in U_\nu\}$ such that

$$\limsup_{\nu \uparrow \infty} |S_{\nu 1}| = \mathcal{O}(N_\nu),$$

and

$$\limsup_{\nu \uparrow \infty} \max_{i,j \in S_{\nu 2}} \left| \frac{\pi_{\nu ij}}{\pi_{\nu i}\pi_{\nu j}} - 1 \right| = \mathcal{O}\left(N_\nu^{-1}\right), P_0 - \text{probability 1}$$

such that for some constants, $C_4, C_5 > 0$ and for N_ν sufficiently large,

$$|S_{\nu 1}| \leq C_4 N_\nu,$$

and

$$N_\nu \sup_{\nu} \max_{i,j \in S_{\nu 2}} \left| \frac{\pi_{\nu ij}}{\pi_{\nu i}\pi_{\nu j}} - 1 \right| \leq C_5,$$

(A5.2) (Special Case: Dependence restricted to countable blocks of bounded size)

Theorem (Updated)

Theorem

Suppose conditions hold. Then for sets $\mathcal{P}_{N_\nu} \subset \mathcal{P}$, *constants,* $K > 0$, *and* M *sufficiently large,*

$$\mathbb{E}_{P_0,P_\nu}\Pi^\pi\left(P : d^\pi_{N_\nu}(P,P_0) \geq M\xi_{N_\nu}|\mathbf{X}_1\delta_{\nu 1},\ldots,\mathbf{X}_{N_\nu}\delta_{\nu N_\nu}\right) \leq$$
$$\frac{16\gamma^2\left[\gamma\mathbf{C_2}+C_3\right]}{\left(Kf+1-2\gamma\right)^2 N_\nu\xi^2_{N_\nu}} + 5\gamma\exp\left(-\frac{Kn_\nu\xi^2_{N_\nu}}{2\gamma}\right), \tag{2}$$

which tends to 0 *as* $(n_\nu, N_\nu) \uparrow \infty$.

Note that $C_2 = C_4 + 1$. When $|S_{\nu 1}| \downarrow 0$ then $C_2 = 1$, leading to main result in Savitsky and Toth (2016).

Simulation Examples: Two (Three) Parameter Logistic Regression

- $$y_i \mid \mu_i \overset{\text{ind}}{\sim} Bern\left(F_l^{-1}(\mu_i)\right), \ i = 1, \ldots, N$$
- $$\boldsymbol{\mu} = -1.88 + 1.0\boldsymbol{x}_1 + 0.5\boldsymbol{x}_2$$
- The x_1 and x_2 distributions are $\mathcal{N}(0, 1)$ and $\mathcal{E}(r = 1/5)$ with rate r
- Size measure used for sample selection is $\tilde{\boldsymbol{x}}_2 = \boldsymbol{x}_2 - \min(\boldsymbol{x}_2) + 1$, *but* neither $\tilde{\boldsymbol{x}}_2$ or $\boldsymbol{x}_2$ are available to the analyst.

Simulation Example: Three-Stage Sample

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

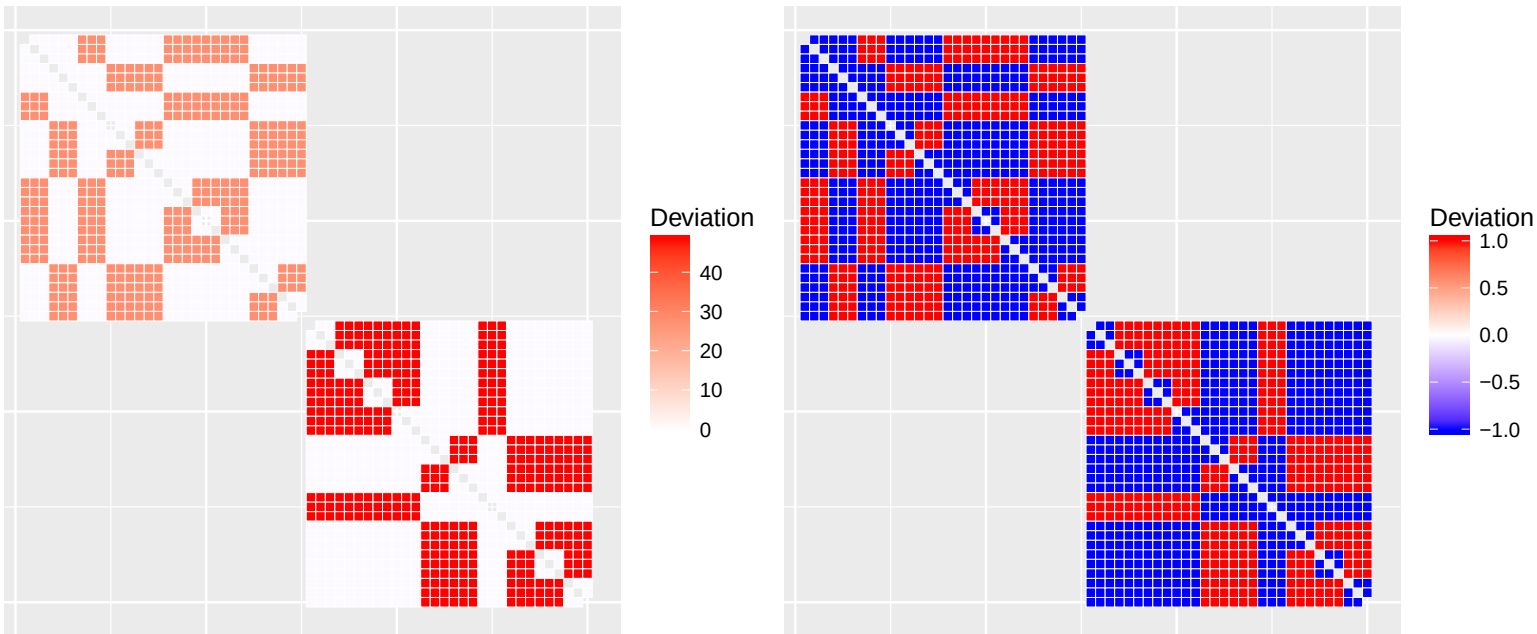


Figure: Factorization matrix $(\pi_{ij}/(\pi_i\pi_j) - 1)$ for two PSU's. Magnitude (left) and Sign (right). Systematic Sampling ($\pi_{ij} = 0$). Clustering and PPS sampling ($\pi_{ij} > \pi_i\pi_j$). Independent first stage sample ($\pi_{ij} = \pi_i\pi_j$)

Simulation Example: Three-Stage Sample (Cont)

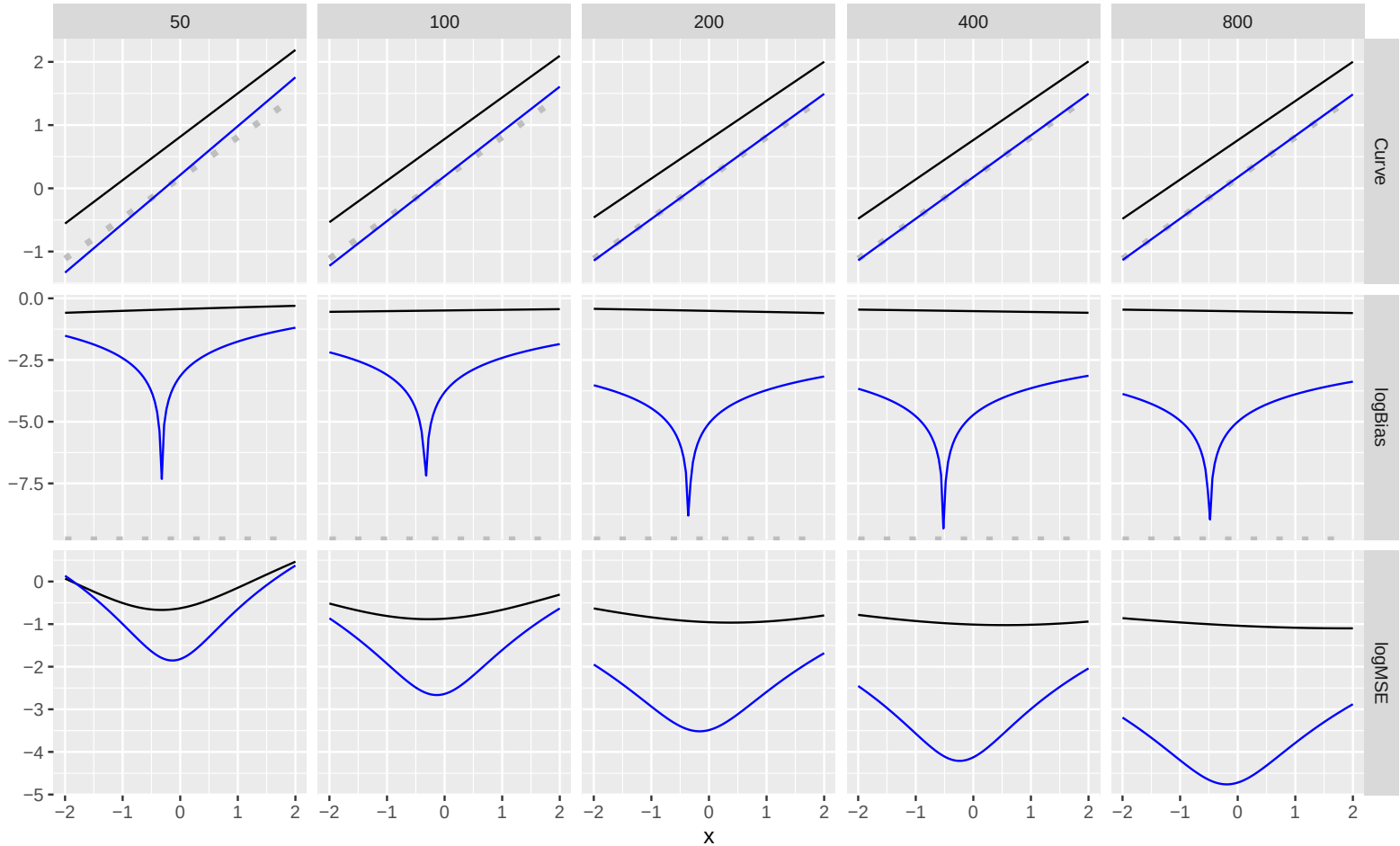


Figure: The marginal estimate of $\mu = f(x_1)$. **population curve**, sample with **equal weights**, and **inverse probability weights**. Top to bottom: estimated curve, log of BIAS, log MSE. Left to right: sample size (50 to 800).

Simulation Example: Sorted Partitions

- Sort population by $\tilde{\boldsymbol{x}}_2$
- Partition the population U into a "high" (U_1) group with the top $N/2$ and a "low" (U_2) group with the bottom $N/2$
- Only two possible samples of size $N/2$: U_1 and U_2.
- Assume an equal probability of selection of $1/2$.
  - $\pi_i = 1/2$, for all $i \in 1, \ldots, N$, and $\pi_{ij} = 1/2$ if $i \neq j \in U_k$, for $k = 1, 2$ and 0 otherwise.
  - Thus, the number of pairwise inclusions probabilities that do not factor ($\pi_{ij} \neq 1/4$) grows at rate $\mathcal{O}\left(N^2\right)$, *violating condition (A5.1)*.
- Alternatively, nest partitions within strata (every 50 sorted units) and sample indepedently across strata
  - factorization for all but $\mathcal{O}(N)$ pairwise inclusion probabilities *statisfies condition (A5.1)*

Simulation Example: Sorted Partitions (Cont)

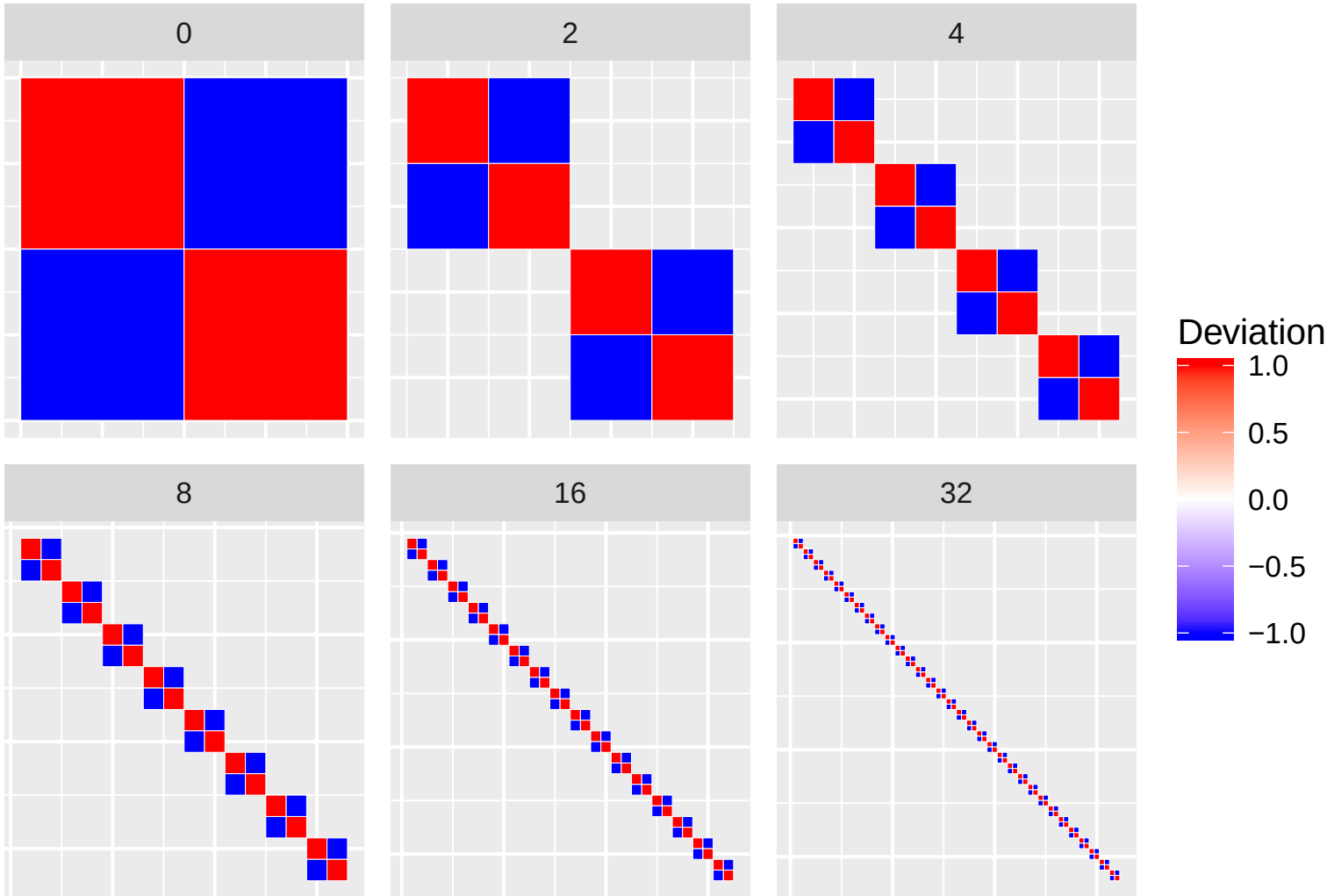


Figure: Matrix $\{i, j\}$ of deviations from factorization $(\pi_{ij}/(\pi_i\pi_j) - 1)$ for an equal probability dyadic partition design by number of strata (0 to 32). Empty cells correspond to 0 deviation (factorization).

Simulation Example: Sorted Partitions (Cont)

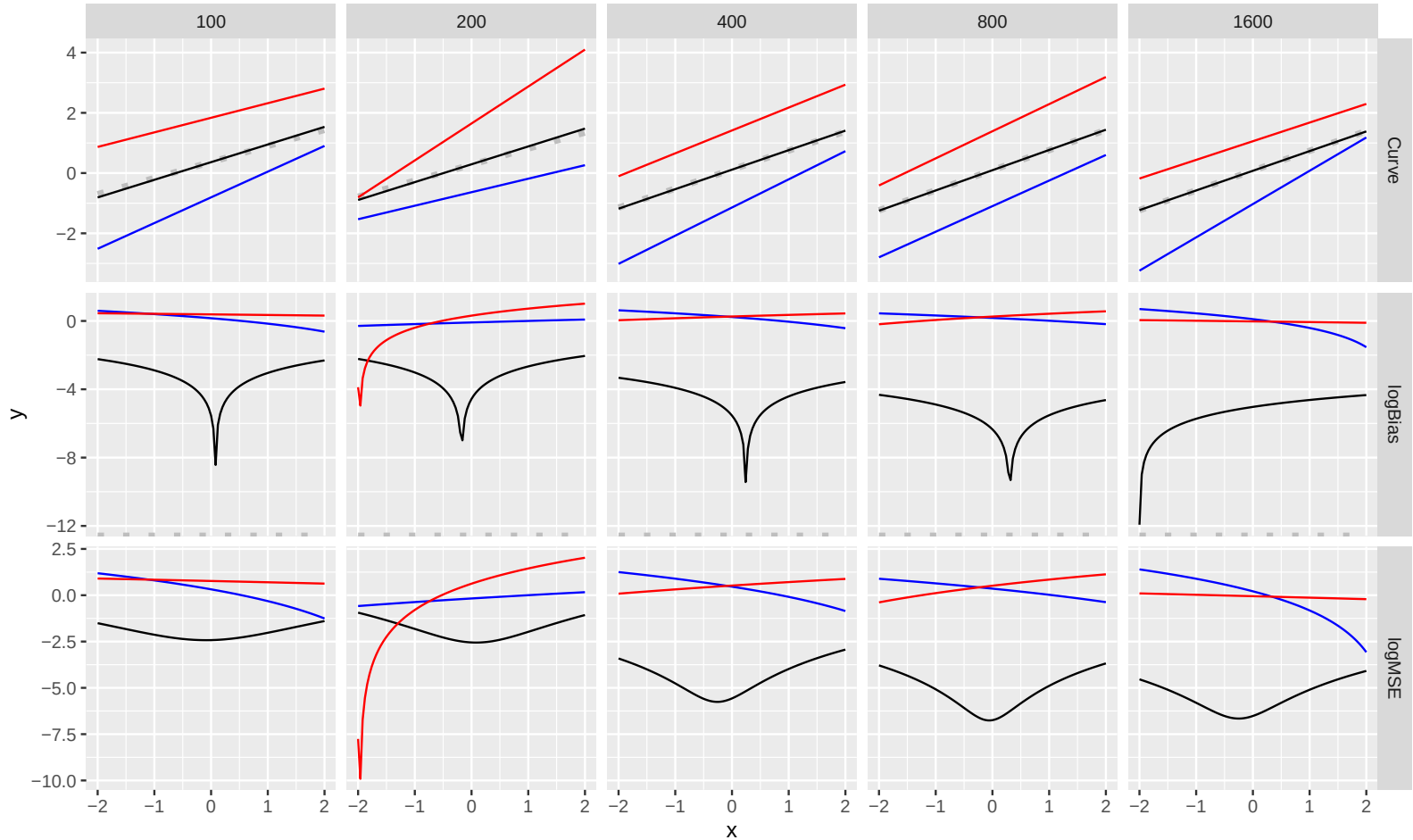


Figure: The marginal estimate of $\mu = f(x_1)$ Compares the **(true) population curve** to binary partition (**high** and **low**) and the **stratified partition** sample. Top to bottom: estimated curve, log of absolute bias, log of mean square error. Left to right: population size (100 to 1600).

Conclusions and Future Research

- **Motivation**: lack of theory to justify consistent estimation for complex, multistage cluster designs such as the NSDUH
  - Approximate or asymptotic factorization of joint sampling probabilities *exclude* these designs
- **Results**: New theoretical conditions now allow for realistic dependence between sampling units.
  - Dependence between units within a cluster is *unrestricted* if **cluster size is bounded** and **dependence between clusters attenuates**. This dependence can be positive (joint selection) or negative (mutual exclusion).
  - Units sorted along a gradient can now be fully justified if the sampling occurs independently **across strata or within clusters**.
- **Current Research**:
  - Impact of survey design on posterior interval estimation.
  - Correcting for mispecification of sampling design efficiency (effective sample size)

References

Ghosal, S. and van der Vaart, A. (2007), 'Convergence rates of posterior distributions for noniid observations', *Ann. Statist.* **35**(1), 192–223.
URL: *http://dx.doi.org/10.1214/009053606000001172*

Morton, K. B., Aldworth, J., Hirsch, E. L., Martin, P. C. and Shook-Sa, B. E. (2016), Section 2, sample design report, *in* '2014 National Survey on Drug Use and Health: Methodological Resource Book', Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Rockville, MD.

Novelo, L. L. and Savitsky, T. (2017), 'Fully Bayesian Estimation Under Informative Sampling', *ArXiv e-prints* .
URL: *https://arxiv.org/abs/1710.00019*

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

Williams, M. R. and Savitsky, T. D. (2017), 'Bayesian Pairwise Estimation Under Dependent Informative Sampling', *ArXiv e-prints* .
URL: *https://arxiv.org/abs/1710.10102*

Yi, G. Y., Rao, J. N. K. and Li, H. (2016), 'A Weighted Composite Likelihood Approach for Analysis of Survey Data under Two-level Models', *Statistica Sinica* **26**, 569–587.