Bayesian stratified sampling for establishment surveys with uncertain design parameters

Jonathan Mendelson[1,2] and Michael Elliott[3]

FCSM, October 25, 2022

[1] Bureau of Labor Statistics
[2] University of Maryland
[3] University of Michigan

Bottom line up front

- Imprecisely estimated survey design parameters could harm sample efficiency
- There is a Bayesian approach to sample design, which accounts for this
- We identify the Bayesian optimal design in a particular establishment survey context
  - Outperforms Neyman-HT in simulation
  - Performs similarly or better than the main model-assisted approach considered

I. Introduction

Introduction

- Sample designs often assume population characteristics are known.
- In practice, some are typically estimated.
- *Example*. Optimal STSRS for estimating finite population mean via separate ratio model
  - Theory: $n_h \propto N_h S_{dh}/\sqrt{c_h}$ (Cochran, 1977)
    - S_{dh} is the stratum SD of a residual term
  - Practice: $n_h \propto N_h \hat{S}_{dh}/\sqrt{\hat{c}_h}$
- Typically, little attention is given to the effect of imperfect information on sample design.

In the above, $S_{dh}^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} \left(Y_{hi} - \frac{Y_h}{X_h} X_{hi} \right)^2$, and

N_h, n_h, c_h are the population size, sample size, and per-unit cost in stratum $h \in (1, \ldots, H)$.

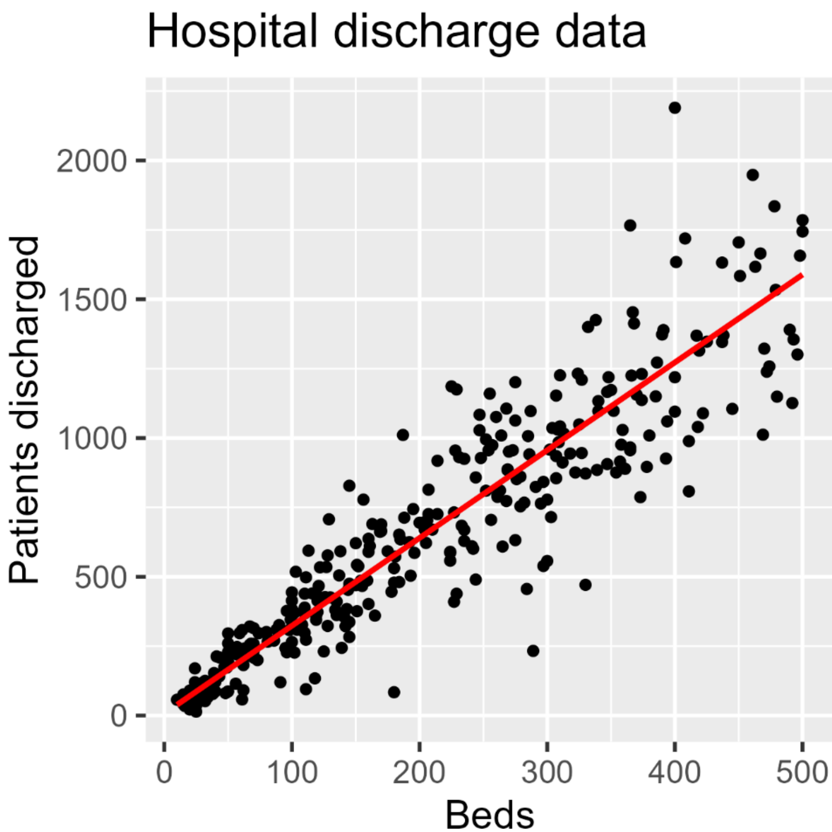
Selected Bayesian design literature

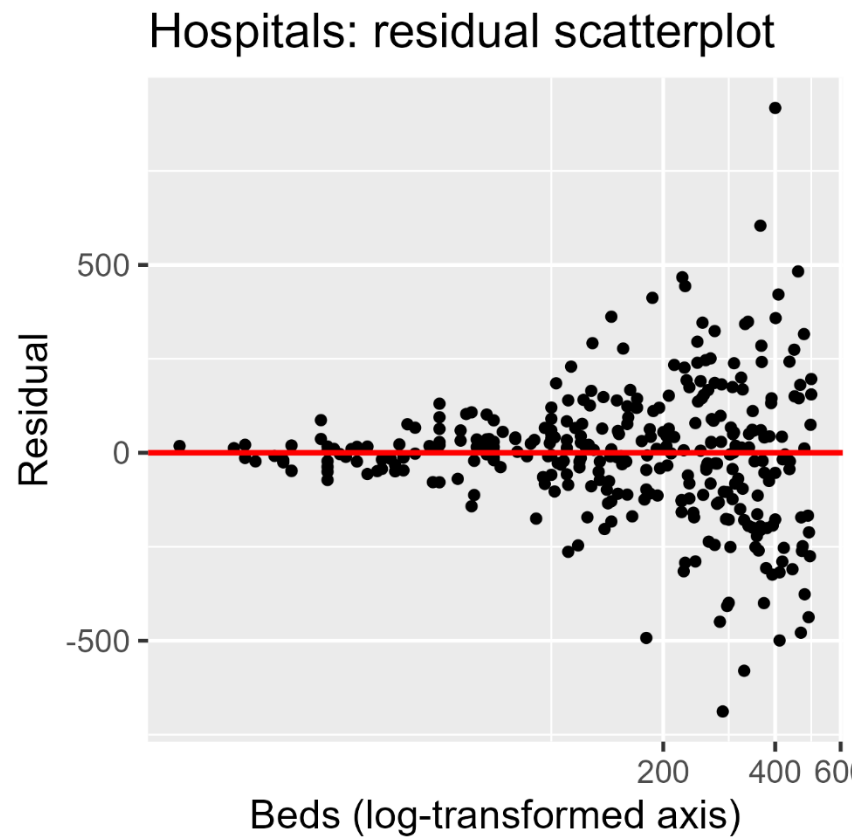
- Bayesian optimal experimental design (Lindley, 1972) can be applied to STSRS sample allocation
  - Flexible approach; accommodates uncertainty
- Draper & Guttman (1968) consider continuous data
  - Assumes use of pilot study data
  - Special case leads approximately to Neyman allocation
  - However, D&G assume fixed strata means and variances
- Rao & Ghangurde (1972) consider categorical data
  - Assumes Dirichlet-multinomial model
  - Applicability for continuous, skewed distributions?

Heteroscedasticity and design

- Consider $\{X_i, Y_i; i = 1, \ldots, N\}$, where
  - $Y_i = \beta X_i + \varepsilon_i$
  - $\mathrm{E}_M(\varepsilon_i) = 0$
  - $\mathrm{Var}_M(\varepsilon_i) = \sigma^2 X_i^b$; known b, $\{X_i > 0\}$
  - Independent ε_i's
- "b" (coefficient of heteroscedasticity) can meaningfully affect optimal allocation
  - PPS/GREG strategy: $\pi_i \propto X_i^{b/2}$ (e.g., SSW, 1992)

Heteroscedasticity, visualized





Data source: National Hospital Discharge Survey of 1968 (via PracTools)

- See Henry & Valliant (2009) for more real examples

Bayesian decision theory for optimal experimental design

- Lindley (1972) treats as a two-part decision:
  - Choose the experiment, $e \in E$ (e.g., $e = \{n_h\}$)
    - This results in the sample (data), $x \in X$
  - Translate the data into a terminal decision
    - For example, compute estimate $\hat{\theta}$ for parameter $\theta \in \Theta$ (e.g., finite population mean)
- Define a loss function of the form $L\left(\hat{\theta}, \theta, e, x\right)$
- Lindley suggests finding optimal $\hat{\theta}, e$ via

$$\min_{e} \int_{X} \left(\min_{\hat{\theta}} \int_{\Theta} L(\hat{\theta}, \theta, e, x) p(\theta|x, e) p(x|e) d\theta \right) dx$$

In the above, $p(\theta|x, e)$ is the posterior pdf and $p(x|e)$ is a conditional pdf.

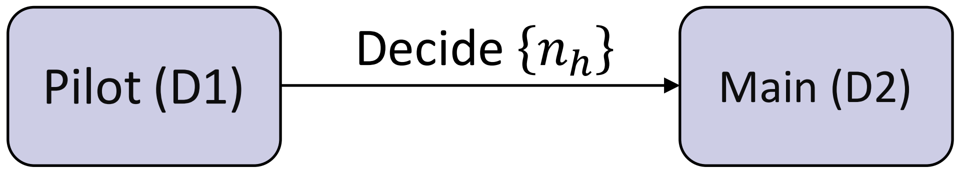
Our research

- We consider optimal STSRS design while accounting for heteroscedastic errors and uncertain design parameters
  - We aim for weaker assumptions than some previous Bayesian work
  - We accommodate uncertain design parameters via Bayesian decision theoretic formulation

II. Problem set-up and Bayesian analysis

Problem set-up

- Study design:



  - Pilot is only used for designing main study
  - Strata defined upfront
- Model: $Y_{hi} = \alpha_h X_{hi} + \varepsilon_{hi}$, where $\varepsilon_{hi} \overset{ind}{\sim} N\left(0, v_h X_{hi}^b\right)$
  - Known $X_{hi} > 0$; known b
- Prior (diffuse): $\pi\left(\left\{\alpha_h, \frac{1}{v_h}\right\}\right) \propto \prod_{h=1}^{H} v_h$

Overview: our Bayesian decision theoretic analysis for the finite population mean

1. Objective: $L\left(\bar{Y}, \hat{\hat{\bar{Y}}}, e, D2\right) = \left(\hat{\hat{\bar{Y}}} - \bar{Y}\right)^2$
   - Minimized when $\hat{\hat{\bar{Y}}} = \mathrm{E}(\bar{Y}|D2, e, b)$
2. Posterior loss is $\mathrm{Var}(\bar{Y}|D2, e, b)$
   - Apply Ericson (1969) to obtain
3. Preposterior analysis: average over future data $(D2|D1)$
   - Consider uncertainty with respect to:
     - Posterior for parameters given pilot, $\{\alpha_h, v_h\}|D1$
     - Sample indicators, $\{s_{2h}\}$
     - Model uncertainty given above, $D2|(\alpha_h, v_h, D1, s_{2h})$
   - Results provided in paper
4. Optimize via mathematical programming

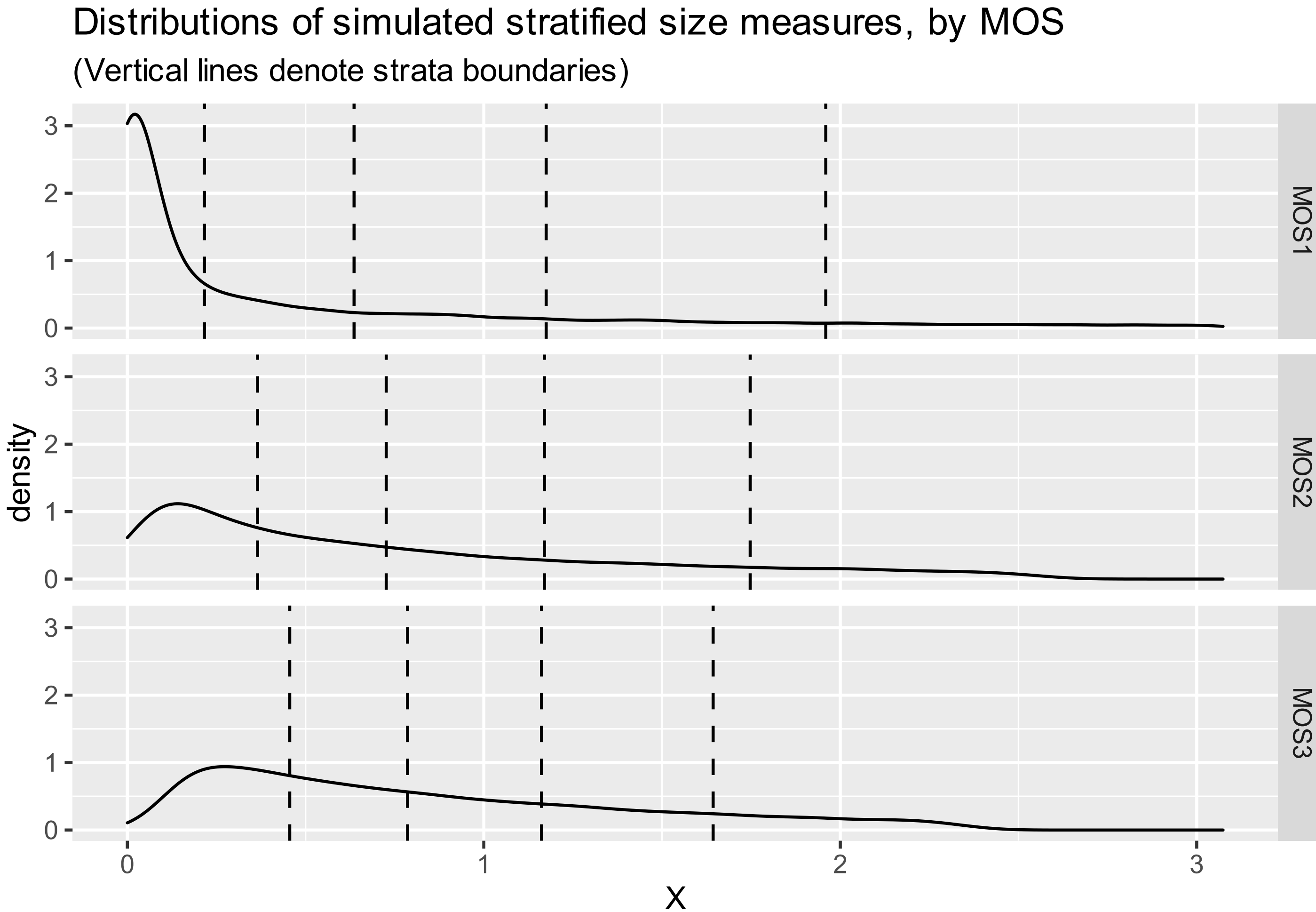
Above, $e = \{n_h\}$, $\bar{Y} = \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi}$, and $s2_h$ refers to the main study sample in stratum h.

III. Simulation

Simulation design: compare strategies across a series of artificial populations

- "Strategy" denotes allocation + estimator
- We generated $P = 90$ bivariate populations, and applied each strategy $R = 1000$ times
- For population p, simulation r:
  - Draw an equally allocated pilot sample of $m = 75$ units
  - For strategy a:
    - Allocate and draw a main study sample of $n = 500$ units
    - Obtain point estimate and 95% CI
- Compare strategies' RMSE, bias, and CI coverage/width
  - For instance: $rmse\left(\hat{Y}_{(p,a)}\right) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{Y}_{(p,a)}^{(r)} - Y_{(p)}\right)^2}$

We considered three size measures



We considered 30 structures for $Y_{hi}|X_{hi} \sim N(\alpha_h X_{hi}, v_h X_{hi}^b)$

- 5 levels of b considered: $b \in \{0, 0.5, 1, 1.5, 2\}$
- 6 choices of $\{\alpha_h, v_h\}$, where $\{v_h\}$ were chosen as to approximately yield target correlations

Scenario	α_1	α_2	α_3	α_4	α_5	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
1. Baseline	1	1	1	1	1	0.7	0.7	0.7	0.7	0.7
2. Lower correlations	1	1	1	1	1	0.5	0.5	0.5	0.5	0.5
3. Higher correlations	1	1	1	1	1	0.9	0.9	0.9	0.9	0.9
4. Increasing correlations, fixed slopes	1	1	1	1	1	0.5	0.6	0.7	0.8	0.9
5. Fixed correlations, decreasing slopes	1.4	1.2	1	0.8	0.6	0.7	0.7	0.7	0.7	0.7
6. Increasing correlations, decreasing slopes	1.4	1.2	1	0.8	0.6	0.5	0.6	0.7	0.8	0.9

We compared several strategies

- We focused on three main strategies:
  - Neyman plug-in/HT estimator (N-HT)
  - Cochran plug-in/separate ratio estimator (C-SR)
  - Bayesian allocation/prediction estimator (B-P)
- We also considered three rule-of-thumb allocations suggested or implied by Cochran for different levels of b (used SR estimator for each)
  - $n_h \propto N_h$
  - $n_h \propto N_h\sqrt{\bar{X}_h}$
  - $n_h \propto N_h\bar{X}_h$

Simulation results: main strategies

- The three main strategies:
  - were approximately unbiased; and
  - had near-nominal coverage for 95% CIs.
- Therefore, we focused on analyzing RMSE
  - Findings on RMSE were paralleled by analogous findings for CI relative width

B-P consistently outperformed N-HT

- Use of N-HT led to RMSE 11%–175% higher than B-P for individual populations (MOS1 pops displayed below)
- Results varied greatly by assumptions for $f(Y_{hi}|X_{hi})$
  - Compare 2nd and 3rd data columns below

Relative increase in RMSE from N-HT versus B-P (among MOS1 pops)

	Scenario for $\{\rho_h, \alpha_h\}$					
	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes
b = 0	65%	32%	175%	72%	69%	77%
b = 0.5	44%	18%	138%	44%	51%	36%
b = 1	46%	24%	127%	35%	40%	28%
b = 1.5	52%	24%	139%	41%	51%	38%
b = 2	64%	44%	159%	75%	68%	62%

B-P did about as well or better than C-SR, with marked differences across populations

- B-P showed the greatest advantage for a subset of MOS1 scenarios (top and bottom rows below)
- In contrast, differences were fairly muted for most MOS2 and MOS3 populations, which had less skewness

Relative increase in RMSE from C-SR versus B-P (among MOS1 pops)						
	Scenario for $\{\rho_h, \alpha_h\}$					
	1. Baseline	2. Lower corrs	3. Higher corrs	4. Inc. corrs	5. Dec. slopes	6. Inc. corrs, dec. slopes
b = 0	19%	10%	18%	32%	29%	40%
b = 0.5	7%	3%	0%	2%	12%	7%
b = 1	6%	7%	3%	1%	4%	3%
b = 1.5	11%	9%	5%	5%	6%	5%
b = 2	23%	21%	23%	25%	24%	29%

B-P sometimes produced more stable allocations than the main alternatives

- Differences in allocations' stability were starkest for MOS1, $b = 2$ pops, for instance:

Allocation summary statistics by allocation and stratum Population 25 (MOS1, b=2, baseline ρ_h, α_h scenario)						
	Neyman		Cochran		Bayesian	
h	$\mathrm{E}(n_h)$	$sd(n_h)$	$\mathrm{E}(n_h)$	$sd(n_h)$	$\mathrm{E}(n_h)$	$sd(n_h)$
1	149	48	135	50	112	19
2	90	20	92	22	98	17
3	76	16	80	19	85	15
4	85	17	85	19	91	15
5	101	22	107	24	114	19

Performance was mixed for rule of thumb strategies

- C-SR and B-P strategies, which incorporate pilot data for allocation, consistently did as well or better than the RT-SR strategies
  - $n_h \propto N_h \bar{X}_h$ performed quite badly in some situations (e.g., RMSE 82%–204% higher than B-P for b = 2, MOS1 populations)
  - In contrast, $n_h \propto N_h \sqrt{\bar{X}_h}$ had reasonable performance for a subset of the $b = 1$ scenarios (depending on the α_h and ρ_h)

IV. Application

We applied our methods to analyzing tax returns of public charities

- Source: IRS Form 990 data (National Center for Charitable Statistics [NCCS], Urban Institute)
  - Analyzed 140,858 domestic operating public charities meeting inclusion criteria
  - X = log revenue, 2008
  - Y = log revenue, 2013
- Unstratified MCMC analysis yielded $\hat{b} = 0.55$ and 95% CI of (0.25, 0.66)

NCCS application (continued)

- We formed 24 strata based on nonprofit sector (8 groups) by revenue class (3 groups)
- Methods paralleled earlier simulation
  - $R = 10{,}000$ equally allocated pilots of 360 units used to design main studies of 1,800 units
  - Compared RMSE, relative bias, CI properties

C-SR and B-P again outperformed N-HT

- C-SR and B-P offered substantial reduction in RMSE than N-HT
- All three methods were approximately unbiased and had near-nominal CI coverage

Table. NCCS Simulation Results

Strategy	Relative RMSE	1000*RelBias	CI Coverage (%)	1000*CI RelWidth
N-HT	1.427	-0.00	94.7	6.48
C-SR	1.014	-0.01	94.8	4.57
B-P	1.000	-0.00	95.5	4.63

Note: RMSE is displayed relative to that of the B-P strategy.

V. Discussion

We provided a Bayesian approach to sample design for our problem

- We considered STSRS designs for establishments
  - Allow for heteroscedastic errors → improved realism
  - Problem formulated via Bayesian decision theory
  - We derived the approximate expected posterior variance, which is then minimized
- We assessed performance via simulation to artificial and real data
  - The proposed B-P strategy provided substantial gains versus design-based approach
  - B-P strategy did as well or better than the main model-assisted approach considered

Potential future directions

- Consider other population structures, including those not following our model
- Compare to additional sampling strategies
- Extend to scenarios where “b” is unknown
- Consider other loss functions
- Identify other ways to express prior knowledge (e.g., in absence of pilot)

Comments? Questions?

Jonathan Mendelson
PhD candidate, JPSM
Research Statistician, BLS (OSMR/BSRC)
mendelson.jonathan@bls.gov