Bayesian learning of design parameters for a new survey

Joep Burger*, Nino Mushkudiani*, Barry Schouten*†

*Statistics Netherlands

†Utrecht University

5th Workshop on Advances in Adaptive and Responsive Survey Design, 6–7November 2017, Ann Arbor, MI

Survey design

- Features
  - Incentive
  - Mode (Web, CATI, CAPI, mix)
  - …
- Uniform

- Adaptive

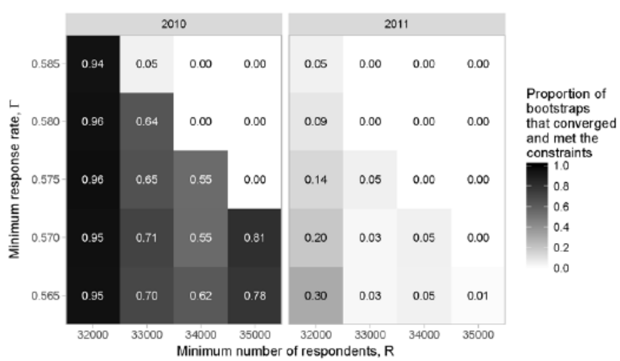
ASD

- Constrained optimization problem
  - allocation parameter $p_{s,i}$

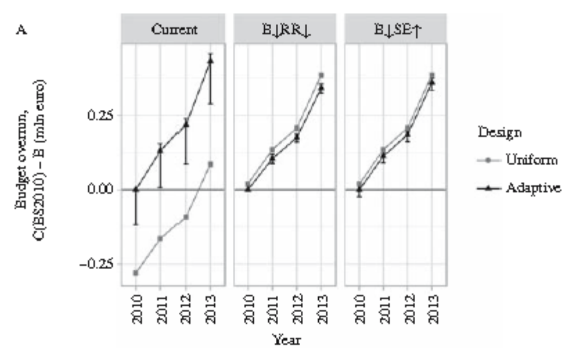
$$\max_{p_{s,i}} f\left(p_{s,i}\right)$$
$$\text{s.t.} \quad g\left(p_{s,i}\right) = G$$
$$h\left(p_{s,i}\right) \leq H$$

Sensitivity

- f, g, h also functions of design parameters
- ASD
  - fairly robust to imprecision
  - sensitive to realistic dynamics



ASD structure



ASD performance

(Burger et al. 2017)

Bayesian analysis

- Probability
  - Frequentistic: frequency in the long run
  - Bayesian: degree of belief
- $P(\theta|y) = P(\theta)\frac{P(y|\theta)}{P(y)}$
- Advantages
  - Include uncertainty about θ
  - Update prior knowledge with new survey data
- Bayesian ASD: Schouten et al. 2017

New survey

- Prior information $P(\theta)$?
  - Other surveys
  - Expert knowledge

Case study: EU-SILC

Case study: EU-SILC

- European Union Statistics on Income and Living Conditions (2003)
- 2016: redesign
- Per subprovince two-stage cluster sampling
  - PSU (municipality): $\pi_j = \frac{N_j}{N}$ (PPS)
  - SSU (16+): $\pi_{ij} = \frac{n_h}{r_h}$ ($i \in h$: income, hhsize, age)
- Web–CATI
- Experiment: 50% conditional incentive €10
- $n_1 = 16\text{k}, n_2 = 6\text{k}$

BADEN framework *light*

- Response propensity

$$\rho_i(s_{1,2}) = \rho_{1,i}(s_1) + \big(1 - \rho_{1,i}(s_1)\big)\,\rho_{2,i}(s_{1,2})$$
$$s_1 \in \{\mathrm{Web}^+, \mathrm{Web}^-\}$$
$$s_2 \in \{\mathrm{CATI}, s_\emptyset\}$$

- GLM → likelihood

$$\Phi^{-1}\big(\rho_{t,i}(s_{1,t})\big) = X_i\beta_t(s_t)$$

- Prior

$$\beta_t(s_t) \sim N\big(\mu(s_t), \Sigma(s_t)\big)$$

Prior information

- Other surveys
  - Labor Force Survey (134k)
  - Budget Survey (28k)
  - Housing Survey (78k)
  - Social Cohesion Survey (11k)
- Point estimates for $\rho_{1,i}(s_1)$ and $\rho_{2,i}(s_{1,2})$
- Distribution?

Prior *distribution*

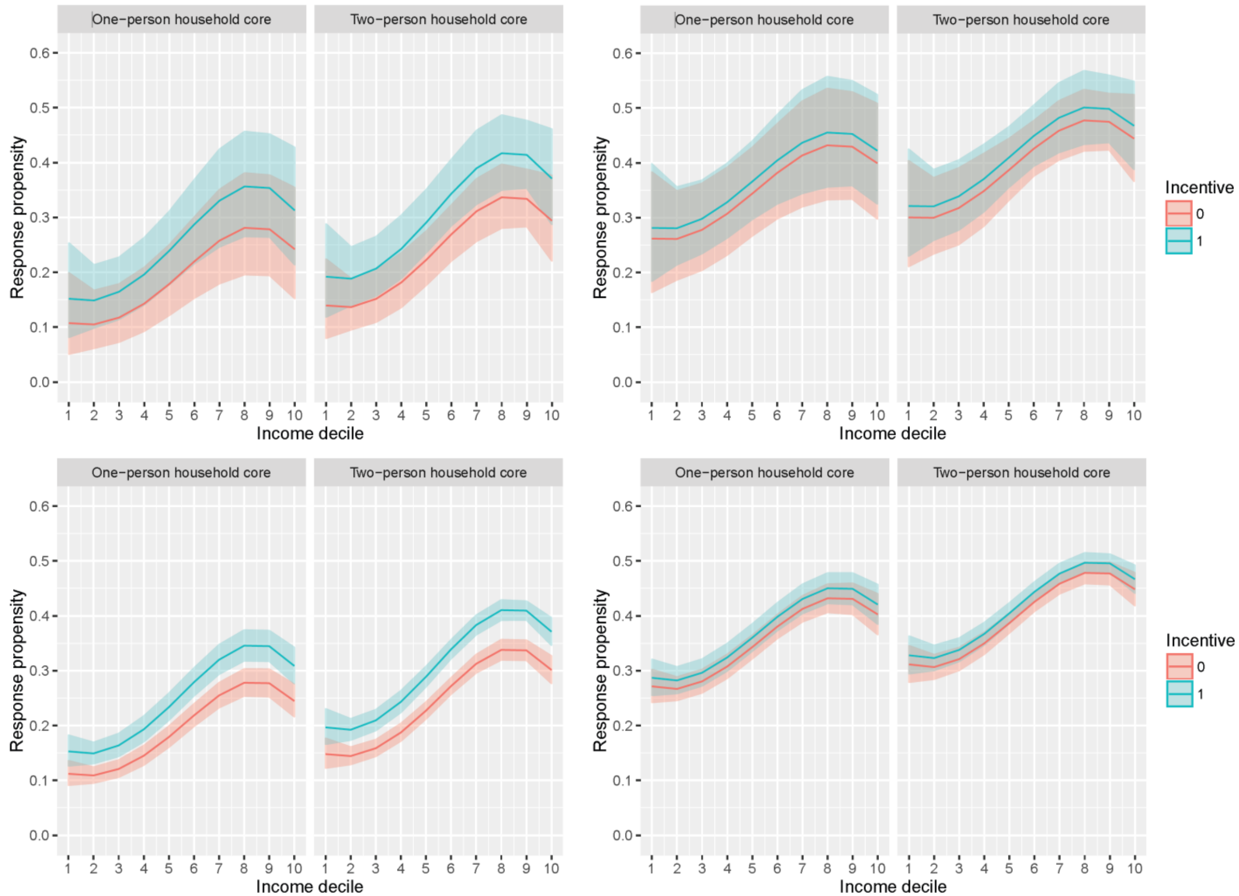
- Simulate sample of size n
- Stratify: $n_g = \frac{N_g}{N} n$ (g: income_10 × hhsize_2)
- Assign incentive: $\mathrm{Binom}(n, 0.5)$
- Link response propensities $\rho_{t,i}\left(s_{1,t}\right)$
- For $b = 1, \ldots, 100$ iterations
  - Draw response $U_{t,b} \sim \mathrm{Binom}\left(n, \rho_{t,i}\left(s_{1,t}\right)\right)$
  - Estimate $\beta_{t,b}$: $\Phi^{-1}\left(P\left(U_{t,b,i} = 1\right)\right) = X_i \beta_{t,b}$
- $\mu(s_t), \Sigma(s_t)$

Priors

$t = 1$ $t = 2$

$n = 1000$

$n = 10{,}000$



Posterior distribution

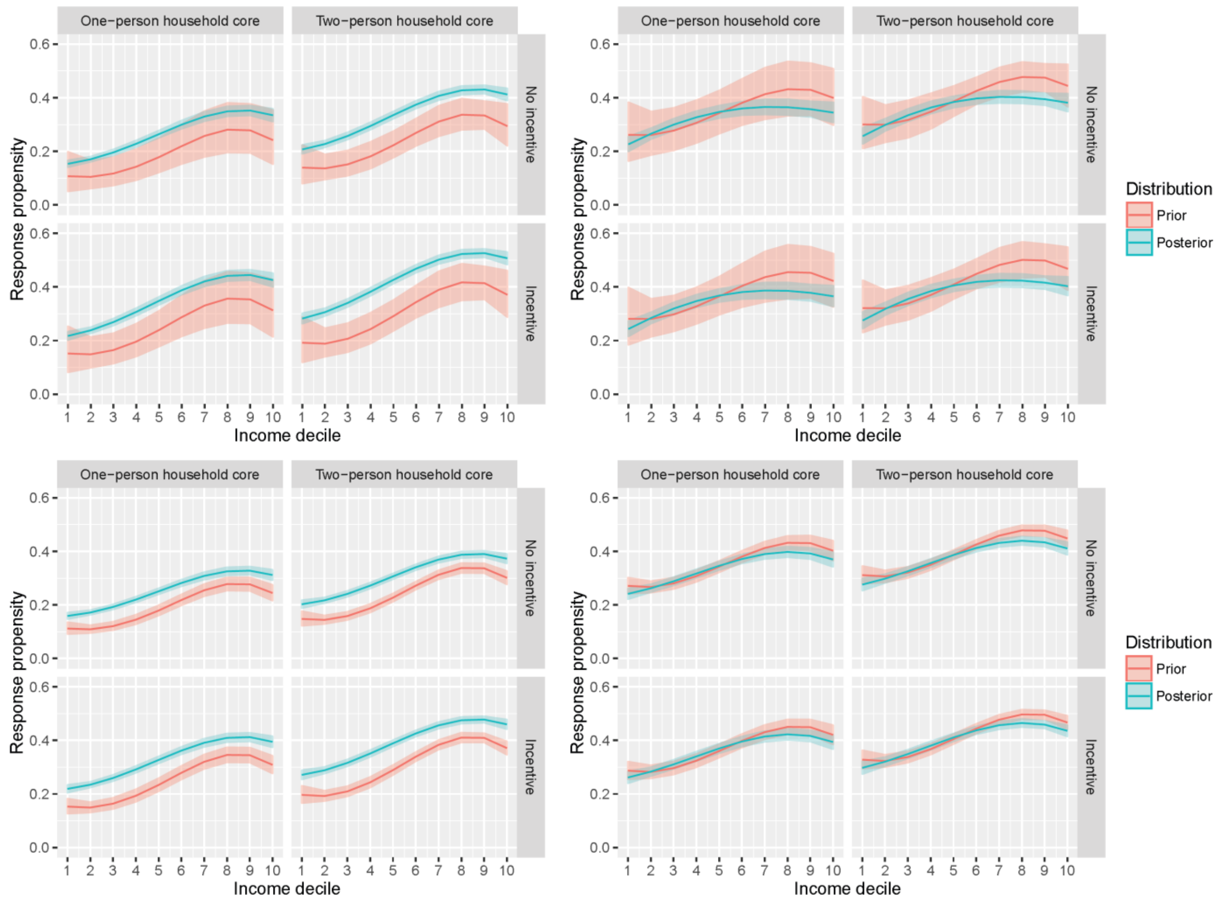
- Data: $U_{t,i} = \begin{cases} 1 & \text{if } Z_{t,i} > 0 \\ 0 & \text{if } Z_{t,i} \leq 0 \end{cases}$
- Gibbs sampling
  - Draw $Z_{t,i}$ from truncated $N(X_i \mu_t, 1)$
  - $\mu_{\text{full}}(s) = \Sigma_{\text{full}}(s)\left(\left(\Sigma(s)\right)^{-1}\mu(s) + X'Z_t\right)$
  - $\Sigma_{\text{full}}(s) = \left(\left(\Sigma(s)\right)^{-1} + X'X\right)^{-1}$
  - Draw $\beta_t(s)$ from $N\left(\mu_{\text{full}}(s), \Sigma_{\text{full}}(s)\right)$
  - 10,000 iterations

Posteriors

$t = 1$ $t = 2$

$n = 1000$

$n = 10{,}000$



Quality indicators

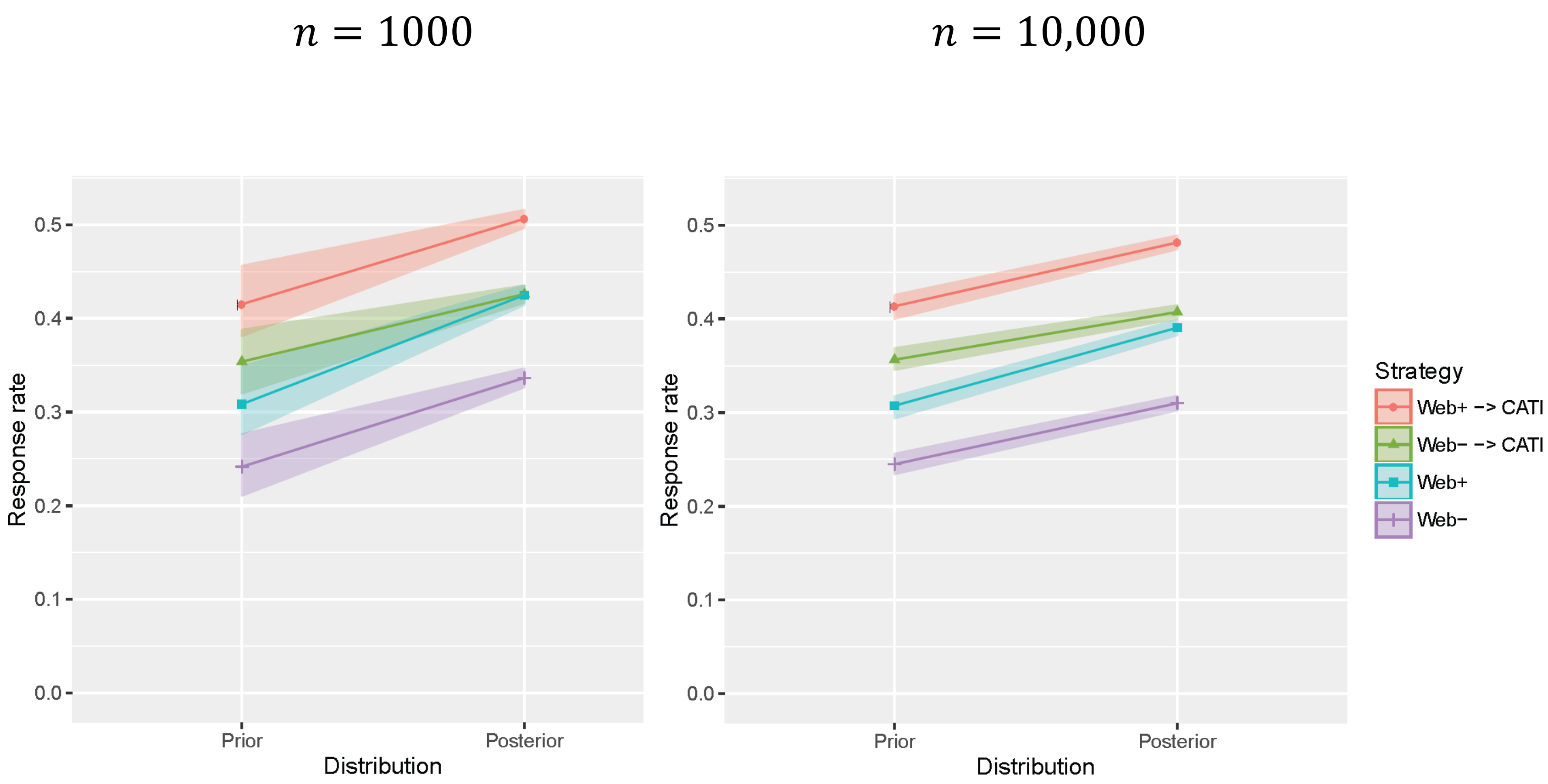
- Response rate

$$RR(s_{1,2}) = \frac{1}{\sum_{i=1}^{n} d_i} \sum_{i=1}^{n} d_i \rho_i(s_{1,2})$$

- Coefficient of variation

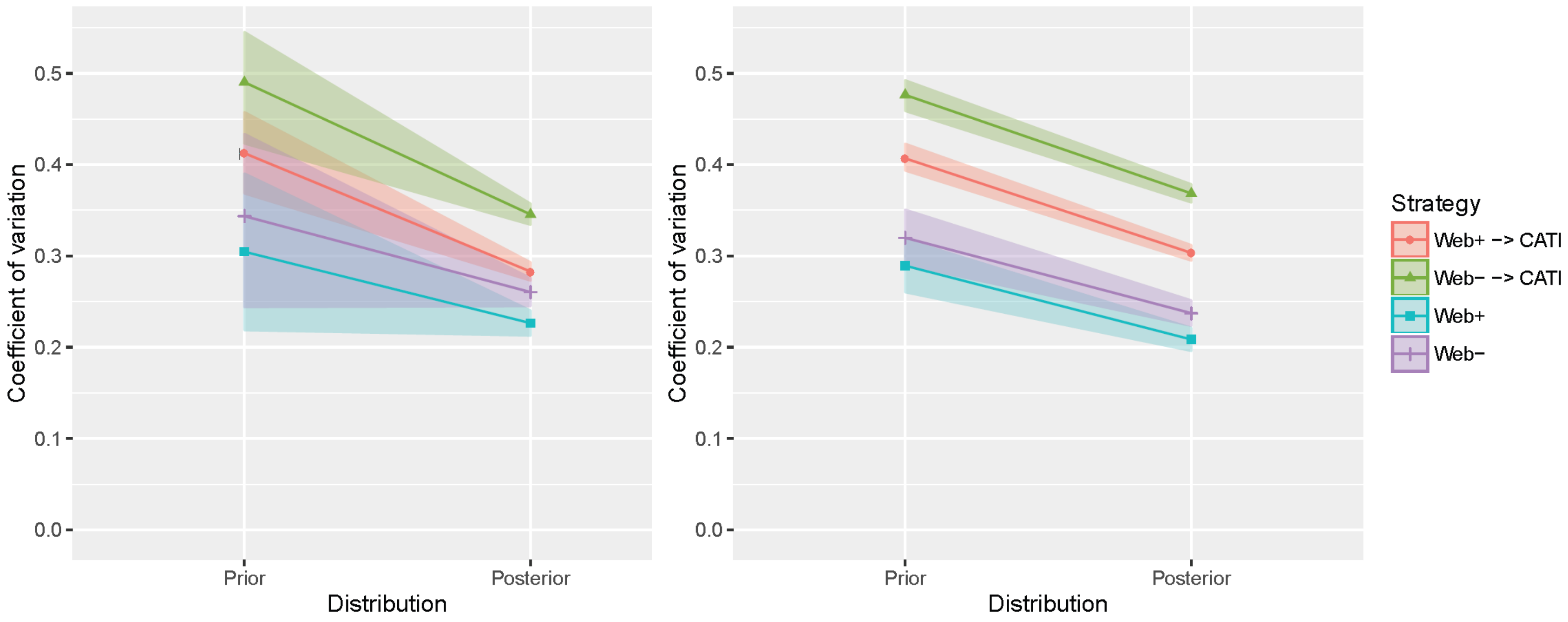
$$CV(X, s_{1,2}) = \frac{\sqrt{\frac{1}{\sum_{i=1}^{n} d_i} \sum_{i=1}^{n} d_i \left(\rho_i(s_{1,2}) - RR(s_{1,2})\right)^2}}{RR(s_{1,2})}$$

Response rate



Coefficient of variation

$n = 1000$ $\quad$ $n = 10{,}000$



Conclusions

- Bayesian approach logical
- BADEN framework general enough
- New survey: prior influential
- Reasonable ball park
- Conditional incentive
  - Higher RR
  - Lower CV

- CATI follow-up
  - Higher RR
  - Higher CV

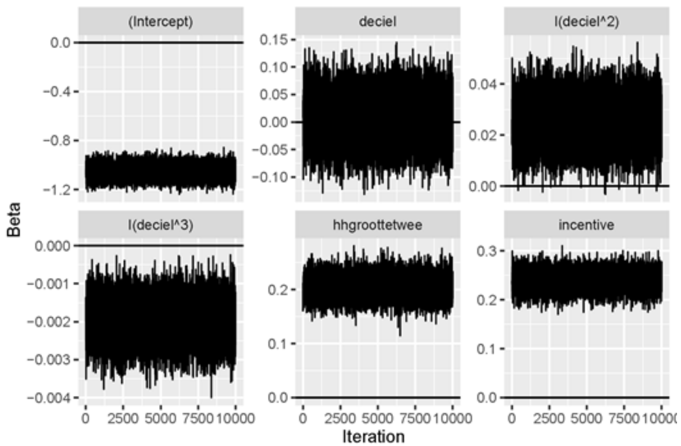
Future

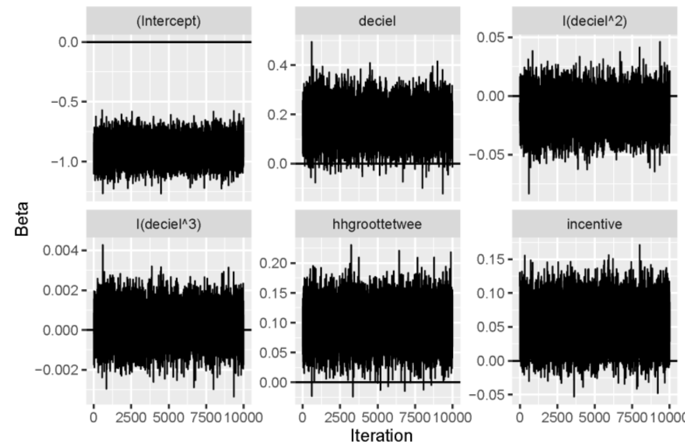
- Paradata
- Other design parameters
  – Costs
  – Measurement effect
- Other quality indicators
- Optimization

Convergence

$t = 1$ $\qquad$ $t = 2$

$n = 1000$





$n = 10{,}000$