Optimization of adaptive survey design under a Bayesian analysis

Barry Schouten (Utrecht University and Statistics Netherlands)
Nino Mushkudiani (Statistics Netherlands)

6th Workshop on Advances in Adaptive and responsive survey design
US Census Bureau, November 4-5, 2019

Context

Setting

1. Repeated general population survey
2. Relatively constant survey design (survey modes + major design features)
3. Rich administrative data and/or paradata
4. Changing response rates over time, especially in web and telephone
5. Desire to implement ASD accounting for uncertainty and time change

Approach taken:

Adopt Bayesian analysis using historic survey data to inform decisions under uncertainty

Conduct mathematical optimization

Context

Main challenges:

- Inclusion of time change, either by adding a hierarchical level for time to the Bayesian analysis, or by choosing the “right” period of historic data
- Optimization under a Bayesian analysis

Bayesian analysis of survey data collection

How is the analysis implemented?

1. Data collection is split in phases
2. Phase contact & participation rates modelled by probit regression
3. Phase contact & participation costs modelled by linear regression
4. Regression parameters receive prior distributions based on historic survey data and/or expert knowledge
5. Posterior distributions derived through Gibbs sampler with data augmentation (for the latent contact & participation variables)
6. Posterior distributions for quality and cost indicators come as important by-products

See Schouten, Mushkudiani, Shlomo, Durrant, Lundquist and Wagner (2018, JSSAM)

From monitoring to mathematical optimization

Assume a set of strategies $S = \{1,2, \dots, S\}$ and a stratification of the sample into $G = \{1,2, \dots, G\}$ strata.

Optimization amounts to choosing strategy allocation probabilities $p_g(s)$ such that a specified objective function is optimized under a number of specified constraints.

Example:

$$\min_{\{p_g(s)\}_{s,g}} CV(\rho_g)$$

such that

- $P[\sum_{g=1}^{G}\sum_{s=1}^{S} p_g(s)c_g(s) > 1.1B] \leq 0.05$
- $E\left(\sum_{s=1}^{S} p_g(s)\rho_g(s)\right) \geq R_g$

From monitoring to mathematical optimization

Challenges:

- Number of possible solutions (i.e. adaptive survey designs) grows very fast. When only 0-1 allocation probabilties are allowed, then the number of solutions is S^G;
- Under a Bayesian analysis, the objective function and constraints are random, i.e. the parameters need to be integrated out;

For example

$$\mathrm{E}\left(\sum_{s=1}^{S} p_g(s)\rho_g(s)\right) = \sum_{s=1}^{S} p_g(s) \int \rho_g(s;\theta)d\theta$$

Proposed optimization strategy

It is infeasible to evaluate the posterior distributions of the quality and cost indicators for all possible ASD solutions given that no explicit closed forms exist for these distributions.

Proposal: Include a pre-optimization step to limit the number of evaluations:

1. For each of the last B Gibbs sampler runs, the M solutions with highest objective function and satisfying all constraints are selected;
2. The posteriors for the selected solutions are evaluated based on all Gibbs sampler draws;
3. The optimal solution is searched from the selected solutions given the posterior distributions;

Case study – Dutch Health Survey

1. Monthly samples of about n=1200
2. Survey data available for January 2014 to July 2019
3. Strategies: web only, web + short F2F follow-up, web + extended F2F follow-up (i.e. leading to three phases)
4. Auxiliary data: socio-demographics, paradata web/F2F
5. Nine strata based on income and age: {No income and 1st quintile, 2nd to 4th quintile, 5th quintile} x {0-29, 30-59, 60+}
6. Make ASD decisions per quarter

$3^9 = 19629$ possible ASD solutions for 0-1 allocation probabilities

Case study – Dutch Health survey

ASD: Minimize coefficient of variation of response propensities for a number of relevant auxiliary variables given constraints on number of respondents, costs and F2F workloads

Study

- Different choices of B and M
- Two historic data settings for a new quarter:
  1. Only last quarter is used to derive a prior;
  2. All previous quarters are used to derive a prior;

Case study – Example

- $B = 250$ and $M = 1$
- Setting: All available previous quarters are used to derive a prior

Q2 2014 optimal designs (5 designs selected)

Web-F2F	Web-F2F+	Web-F2F	Web-F2F+	Web-F2F+	Web-F2F	Web-F2F+	Web-F2F	Web-F2F
Web-F2F	Web-F2F+	Web-F2F	Web-F2F+	Web-F2F+	Web-F2F	Web-F2F+	Web-F2F+	Web-F2F
Web-F2F	Web-F2F+	Web-F2F	Web-F2F+	Web-F2F+	Web-F2F	Web-F2F+	Web-F2F+	Web-F2F+
Web-F2F	Web-F2F+	Web-F2F	Web-F2F+	Web-F2F+	Web-F2F+	Web-F2F+	Web-F2F	Web-F2F
Web-F2F	Web-F2F+	Web-F2F	Web-F2F+	Web-F2F+	Web-F2F+	Web-F2F+	Web-F2F+	Web-F2F

Q2 2015 optimal designs (2 designs selected)

Web-F2F	Web-F2F+	Web-F2F	Web-F2F+	Web-F2F+	Web-F2F	Web-F2F	Web-F2F	Web-F2F
Web-F2F	Web-F2F+	Web-F2F	Web-F2F+	Web-F2F+	Web-F2F	Web-F2F+	Web-F2F+	Web-F2F

Case study – Example

- $B = 250$ and $M = 1$
- Setting: All previous quarters are used to derive a prior

Q2 2014 posteriors for five best solutions

RR	R-indicator	CV	RR SD	CV SD
0.643	0.915	0.067	0.0074	0.0105
0.650	0.924	0.059	0.0074	0.0101
0.651	0.919	0.062	0.0074	0.0101
0.648	0.903	0.075	0.0074	0.0107
0.655	0.913	0.067	0.0073	0.0102

Q2 2015 posteriors for two best solutions

RR	R-indicator	CV	RR SD	CV SD
0.638	0.940	0.047	0.0035	0.0049
0.643	0.939	0.047	0.0035	0.0050

Case study – general findings

- Impact of number of Gibbs draws B is very modest
- Number of best solutions M has some impact for short historic time periods, but vanishes when historic periods grow longer
- Number of best solutions decreases quickly for accumulating historic survey data

Discussion: Main effect strategy allocation probabilities

Proposed optimization strategy is not fully satisfactory, because the number of strata needs to be limited, and, consequently, the number of included variables is as well.

Proposal:

- Let $\tilde{p}_x(s;\beta,x) = \beta^T x + \varepsilon$ be an allocation "tendency" with x a vector of auxiliary variables and ε a $N(0,1)$ random variable.
- Analogous to probit regression, unit i is assigned to strategy s when $\tilde{p}_x(s;\beta,x_i)$ is larger than zero, i.e. $p_x(s;x) = 1 - \Phi\,(\text{-}\beta^T x)$

Discussion: Main effect strategy allocation probabilities

In case of the example

$$\min_{\beta} CV(\rho_X)$$

such that

- $P[\sum_i \sum_{s=1}^{S} p(s;\beta,x_i)c(s;x_i) > 1.1B] \leq 0.05$
- $E\left(\sum_i \sum_{s=1}^{S} p(s;\beta,x_i)\rho(s;x_i)\right) \geq R$

The size of β can be relatively small, allowing for a considerable reduction in the optimization complexity for ASD.

Discussion

- How to avoid overly complex ASD optimization problems under Bayesian analysis of survey data collection?
- Given prior-posterior distributions, optimization problems may have very different objectives and constraints? e.g.
  - Probability that a budget overrun of more than 10% occurs must be smaller than 5%
  - Probability that response rate exceeds a threshold is at least 90%
- Would “main effect” allocation probabilities be an option to further explore?