# Optimization of an adaptive survey design from a Bayesian perspective

Barry Schouten and Nino Mushkudiani

**JSM session Bayesian adaptive survey designs, July 30, 2017**

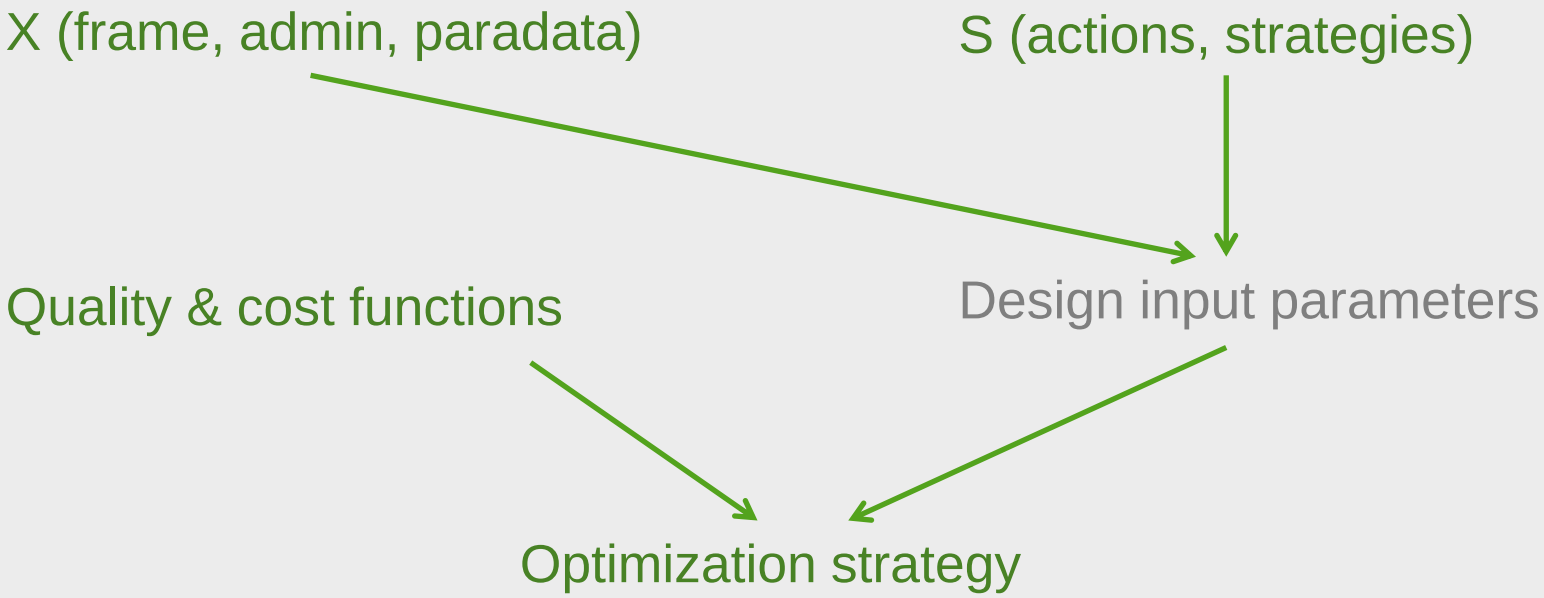
The Leverhulme Trust

Centraal Bureau voor de Statistiek

# Outline

- Optimization of adaptive survey design;
- Bayesian analysis framework;
- Health survey case study;
- Conclusions and future research;

# ASD optimization



# Optimization of ASD

Options to optimize and implement ASD:

1. Trial-and-error: Allocate strategies through a mix of expert knowledge and historic survey data;

2. Probability sampling with quota: Stop data collection, once specified stratum response rates are obtained;

3. Case prioritization: Order remaining nonrespondents based on their current response propensities or conditional bias and start with lowest propensities or largest conditional bias;

4. Mathematical optimization: Formulate ASD as a decision problem in which strategy allocation probabilities act as decision variables;

# Bayesian analysis

## Strategy

Regression coefficients and variances in contact, participation and costs models are assigned a distribution (prior) that is updated using survey data (posterior). Posterior is the new prior for a next round or wave.

## Elicitation of prior distribution parameters (hyperparameters):

- Expert knowledge;
- Historic survey data

## Numerical approximation of posterior distribution through MCMC;

Schouten, Muhkudiani, Shlomo, Durrant, Lundquist, Wagner (2017)

# Bayesian analysis

What is different in optimization?

- Uncertainty in design input parameters (contact, participation costs and propensities) can and needs to be accounted for;
- New cost and quality constraints such as maximum budget may not be exceeded with probability large than 10%;
- Identify designs that have potential but require more information;
- Search for global optima is computationally more burdensome;

Research questions:

1. How to perform optimization under a Bayesian analysis?
2. What is added value of Bayesian setting?

# Simulation – ASD for Dutch Health Survey

## Features

- Health survey: monthly, on-going person survey;
- Stratification based on age and personal income, i.e. static ASD;
- Three phases Web → F2F follow up → extended F2F follow-up, where phases 2 and 3 are optional;

## Optimization problem

- Maximize the expected response rate, subject to
- Number of respondents ≥ 1000;
- P[Costs per respondent > Cmax] < 10%
- P[R(age,income) < Rmin] < 10%

# Simulation – ASD for Dutch Health Survey

Simplifications

- Six strata based on age and personal income, {0-29,30-64,65+} x {≤1000 Euro, > 1000 Euro}, i.e. static ASD;
- 0-1 allocation probabilities of strata to phases 2 and 3, i.e. $3^6$ possible designs;

Optimization

- Posterior distributions of quality and cost indicators estimated by a Gibbs sampler, see Schouten et al (2017);
- Brute force, i.e. all possible designs evaluated;

# Simulation – ASD for Dutch Health Survey

Scenarios

- Prior: {non-informative, based on one month, based on six months};
- Making ASD decision after observing {one month, one quarter, one year} of data;

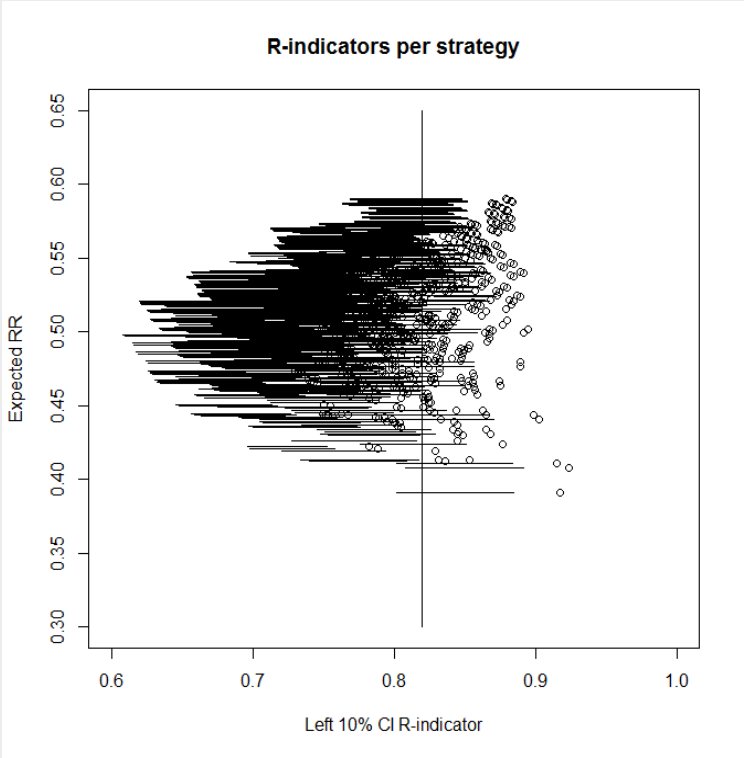
Specific research questions

1. How does prior information affect the optimization?
2. How does data sample size affect the optimization?
3. Are optimal ASD robust to prior and sample size?

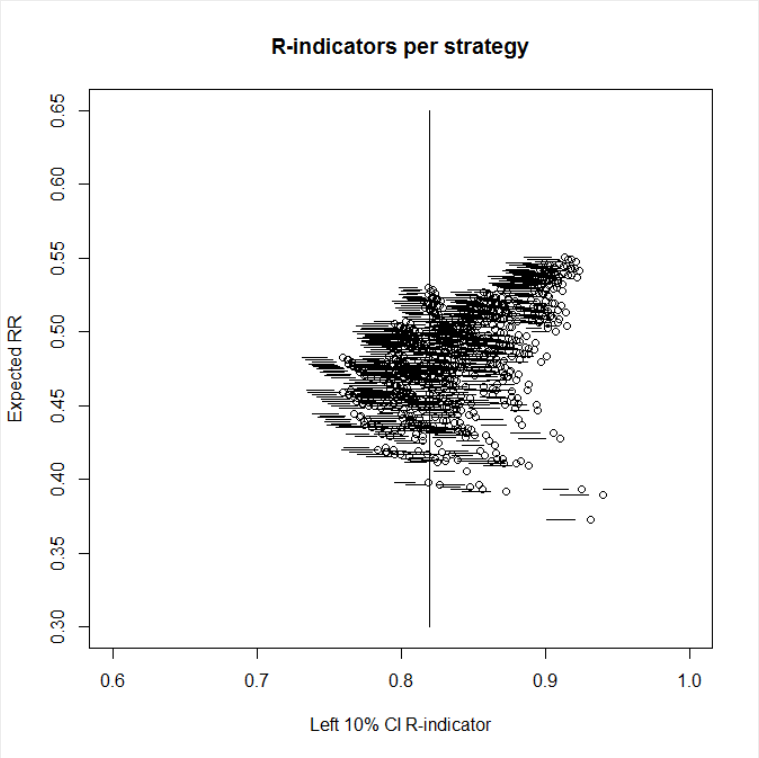
# Simulation results – posterior R-indicator

Left 10% of posterior distribution for R-indicator against response rate

Non-informative x month data



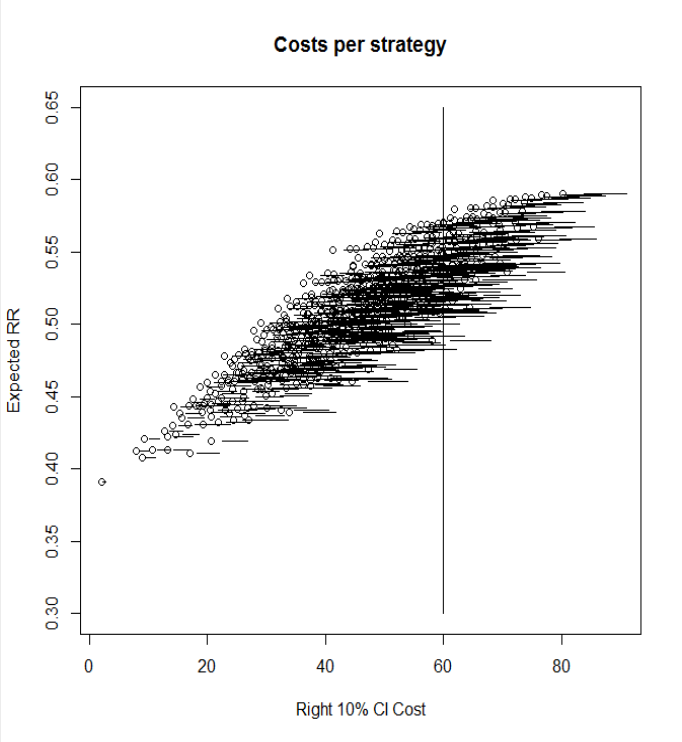

Six months x 12 months data



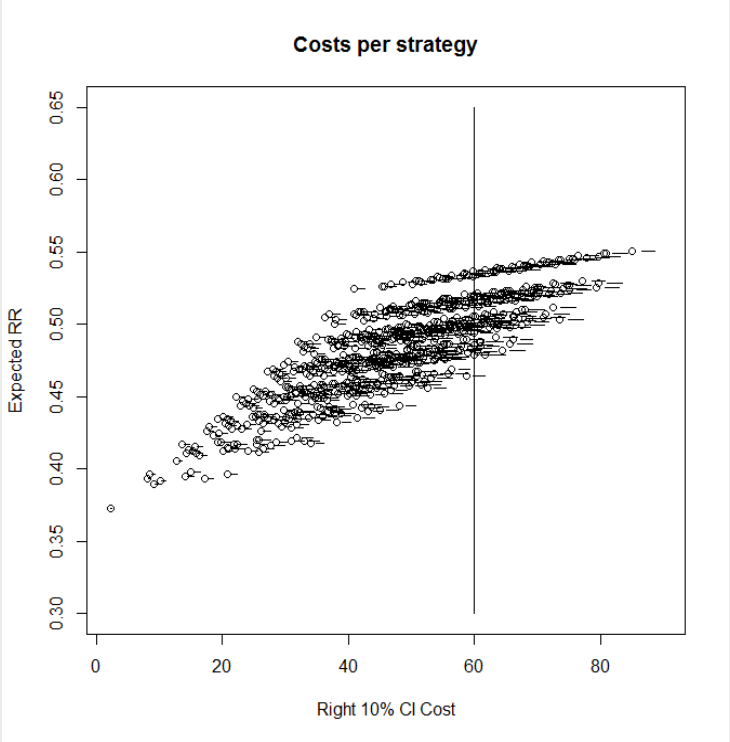
# Simulation results – posterior costs

Right 10% of posterior distribution for costs against response rate

Non-informative x month data




Six months x 12 months data



# Simulation results – possible designs

Proportion of possible designs that satisfies the cost and R constraints

C < 60 Euro per respondent, R > 0.82

| Data | Prior | | |
|---|---|---|---|
| | Noninformative | One month | Six months |
| Month | 10% | 16% | 8% |
| Quarter | 20% | 16% | 20% |
| Year | 27% | 36% | 39% |

C < 80 Euro per respondent, R > 0.82

| Data | Prior | | |
|---|---|---|---|
| | Noninformative | One month | Six months |
| Month | 18% | 31% | 22% |
| Quarter | 31% | 32% | 34% |
| Year | 43% | 53% | 55% |

# Simulation results – optimal designs

Optimal allocations to the six age – income strata for all scenarios

C < 60 Euro per respondent, R > 0.82

| Data | Prior | | |
|---|---|---|---|
| | Noninformative | One month | Six months |
| Month | 2, 3, 3, 2, 2, 2 | 2, 3, 2, 2, 2, 2 | 3, 2, 2, 2, 2, 2 |
| Quarter | 3, 2, 2, 2, 3, 2 | 2, 2, 2, 3, 3, 2 | 2, 3, 2, 2, 2, 2 |
| Year | 2, 3, 2, 2, 2, 2 | 3, 2, 2, 2, 2, 3 | 2, 3, 2, 3, 2, 2 |

C < 80 Euro per respondent, R > 0.82

| Data | Prior | | |
|---|---|---|---|
| | Noninformative | One month | Six months |
| Month | 3, 3, 2, 2, 3, 3 | 3, 3, 2, 2, 3, 2 | 3, 3, 2, 3, 2, 3 |
| Quarter | 3, 3, 2, 2, 3, 3 | 3, 3, 2, 2, 2, 3 | 3, 3, 2, 2, 3, 3 |
| Year | 3, 3, 2, 2, 3, 3 | 3, 3, 2, 2, 3, 3 | 3, 3, 2, 2, 3, 3 |

# Computation times

- Gibbs sampler per scenario very fast, even for larger models including many variables. All nine scenarios can be done in a few hours;
- Brute force optimization is doable for three designs and six strata, about 15 minutes on standard 32-bit machine;
- For more strategies and/or strata, the computation times quickly become infeasible:
  - 10 strata is 81 times longer;
  - 5 design options is approximately 21 times longer;

# Conclusions

- Computation times: Under modest numbers of strata, computation times quickly become infeasible, even under 0-1 allocations;
- As expected: Number of eligible designs increases with sample size and decreases with prior variance;
- Subtle variations between optimal designs possible, even for relatively informative priors and larger sample sizes, i.e. response rate is a smooth function of allocation probabilities;

# Discussion

- Bayesian analysis framework has attractive features as uncertainty is accounted for and a wider set of quality-cost functions can be included, but;
- Needs clever optimization strategies;

Future study:

- How to optimize for many strata and/or design options?
- How to combine optimization and learning/updating?
- How to employ the Bayesian analysis framework in other optimization strategies?
- Should we extend to allocation probabilities in [0,1] interval?

# Model - survey design parameters

Three types of survey design parameters are sufficient to compute most quality and cost functions:

- $\rho_i(s)$ : Response propensity for a unit and strategy;
- $C_i(s)$ : Costs for a unit and strategy;
- $D_i(y, s)$ : Method effect on $y$ for a unit and strategy;

For interviewer modes, response propensities and costs are detailed for different types of nonresponse.

Design parameters are modelled through generalized linear models using a selection of the available covariates.

# Model – quality and cost functions

Example ($d_i$= sample inclusion weight):

- Response rate: $RR(s) = \frac{1}{N}\sum_{i=1}^{n} d_i \rho_i\,(s)$
- Total costs: $B(s) = \sum_{i=1}^{n} c_i\,(s)$
- R-indicator of response propensities for relevant X

$$R(X,s) = 1 - 2\sqrt{\frac{1}{N}\sum_{i=1}^{n} d_i(\rho_i\,(s) - RR(s))^2}$$