Optimal Stratification in Bayesian Adaptive Survey Designs

Yongchao Ma[1] Nino Mushkudiani[2] Barry Schouten[1,2]

[1]Utrecht University
[2]Statistics Netherlands

European Survey Research Association 2021

Universiteit Utrecht

Outline

Stratification in Adaptive Survey Design

- Why?
  - Different data collection strategies are effective for different groups of people
  - Assigning the right strategies to the right people
  - Identify groups of people with different preferences for being approached
- How?
  - Prior to the start of data collection, stratification is based on
    - historic survey data
    - fully observed auxiliary data (eg. population register)
  - Balance the responses over strata defined by auxiliary variables (eg. age groups)
  - **Assumption**: selected auxiliary variables are related to the target survey variables

Research Question

- How to stratify the target population into subgroups effectively and efficiently?

Methodology

- Stratify the target population directly on the target survey variables
- How?
  - Predict target survey variables by fully observed auxiliary data
  - Clustering by Classification and Regression Tree (CART)
- Strata are directly related to target survey variables.
- Balancing the responses over these strata directly improves the survey estimates

Dutch Health Survey

- Data collected from April 2017 to March 2018 selected
- Sample size: 13197
- Strategies
  - Web only
  - Web + short F2F follow-up (at most 3 visits)
  - Web + extended F2F follow-up (more than 3 visits)

Dutch Health Survey

- Target survey variables (dichotomized) Y
  - Self-perceived health
  - Smoking
  - Obesity
- Auxiliary variables X
  - Age
  - Sex
  - Income level
  - Migration status
  - Marital status
  - Urbanisation level of the residential neighbourhood
  - Household type
  - Education level
  - Whether they received rent benefit

Stratification

- Response $\hat{Y}$
- ResponseX
- CostX

Response to Web $\sim \hat{Y}$

Response to Web $\sim X$

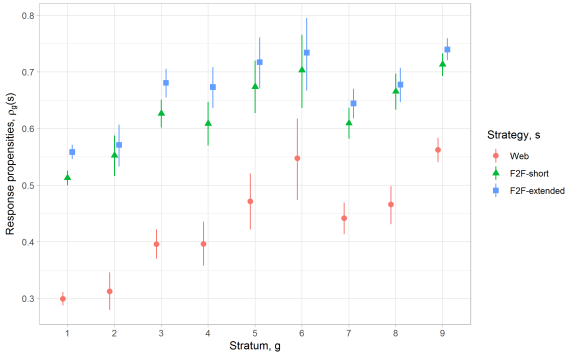
Number of visits $\sim X$

Stratification - Response $\hat{Y}$

Stratification based on the Predicted Probabilities of Success of Survey Variables.

Stratum	Smoking probabilities	Health probabilities	Obesity probabilities
1 (5841)	≥ 0.21		
2 (720)	< 0.21	< 0.56	
3 (1370)	< 0.21	≥ 0.86	< 0.06
4 (626)	≥ 0.13 & < 0.21	≥ 0.86	≥ 0.06
5 (371)	≥ 0.08 & < 0.13	≥ 0.86	≥ 0.06
6 (188)	< 0.08	≥ 0.86	≥ 0.06
7 (1240)	≥ 0.16 & < 0.21	≥ 0.56 & < 0.86	
8 (825)	< 0.16	≥ 0.56 & < 0.63	
9 (2016)	< 0.16	≥ 0.63 & < 0.86	

Note: Stratum size in parentheses. Total sample size is 13197.



Optimization - Response $\hat{Y}$

- Minimize the coefficient of variation (CV) of response propensities subject to the constraints on response rate (RR) and cost per respondent (B)
- Set the response rate at 50% and the cost per respondent at €42
- $3^9 = 19683$ possible solutions for 0/1 allocation probabilities
- Evaluated the posteriors of each solution to search for the optimal solution

Optimization - ResponseX & CostX

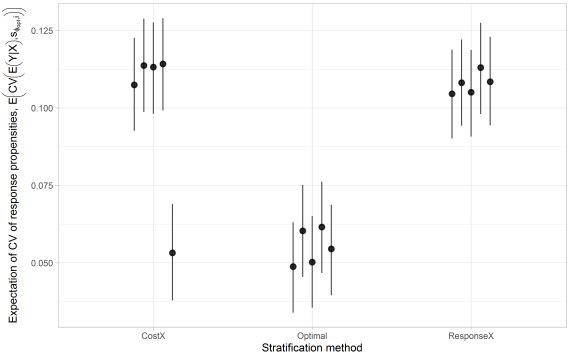
- Same steps for
  - estimating response propensities and costs
  - optimizing from the possible solutions

Determine Optimal Stratification

- Select top 5 optimal solutions based on each stratification
- Evaluate their coefficient of variation (CV) of **individual response propensities with respect to predicted survey variables**
  - Solution that incurs the minimum CV is the optimal solution
  - Corresponding stratification is subsequently the optimal stratification

Determine Optimal Stratification

- Winner: Response $\hat{Y}$



Discussion

- When the predictive power of X is very low...
- Strategy-dependent measurement error of the survey variables
- Compare all the stratification methods in one go