Optimal Stratification in Bayesian Adaptive Survey Designs

Yongchao Ma\textsuperscript{1}  Nino Mushkudiani\textsuperscript{2}  Barry Schouten\textsuperscript{1,2}

\textsuperscript{1}Utrecht University
\textsuperscript{2}Statistics Netherlands

European Survey Research Association 2021
Outline

1 Introduction
   - Stratification in Adaptive Survey Design
   - Research Question

2 Methodology

3 A Case Study
   - Dutch Health Survey
   - Stratification
   - Optimization
   - Determine Optimal Stratification

4 Discussion
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 (e.g., population register)
- Balance the responses over strata defined by auxiliary variables (e.g., age groups)
- **Assumption**: selected auxiliary variables are related to the target survey variables
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
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 $\hat{Y}$
  - Response to Web $\sim \hat{Y}$

- Response $X$
  - Response to Web $\sim X$

- Cost $X$
  - Number of visits $\sim X$
Stratification based on the Predicted Probabilities of Success of Survey Variables.

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Smoking probabilities</th>
<th>Health probabilities</th>
<th>Obesity probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (5841)</td>
<td>$\geq 0.21$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 (720)</td>
<td>$&lt; 0.21$</td>
<td>$&lt; 0.56$</td>
<td></td>
</tr>
<tr>
<td>3 (1370)</td>
<td>$&lt; 0.21$</td>
<td>$\geq 0.86$</td>
<td>$&lt; 0.06$</td>
</tr>
<tr>
<td>4 (626)</td>
<td>$\geq 0.13 &amp; &lt; 0.21$</td>
<td>$\geq 0.86$</td>
<td>$\geq 0.06$</td>
</tr>
<tr>
<td>5 (371)</td>
<td>$\geq 0.08 &amp; &lt; 0.13$</td>
<td>$\geq 0.86$</td>
<td>$\geq 0.06$</td>
</tr>
<tr>
<td>6 (188)</td>
<td>$&lt; 0.08$</td>
<td>$\geq 0.86$</td>
<td>$\geq 0.06$</td>
</tr>
<tr>
<td>7 (1240)</td>
<td>$\geq 0.16 &amp; &lt; 0.21$</td>
<td>$\geq 0.56 &amp; &lt; 0.86$</td>
<td></td>
</tr>
<tr>
<td>8 (825)</td>
<td>$&lt; 0.16$</td>
<td>$\geq 0.56 &amp; &lt; 0.63$</td>
<td></td>
</tr>
<tr>
<td>9 (2016)</td>
<td>$&lt; 0.16$</td>
<td>$\geq 0.63 &amp; &lt; 0.86$</td>
<td></td>
</tr>
</tbody>
</table>

*Note:* Stratum size in parentheses. Total sample size is 13197.
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
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
Winner: $\hat{Y}$
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