Adaptive Survey Design at Statistics Netherlands

International Total Survey Error Workshop 2019

Kees van Berkel

10 – 12 June 2019, Bergamo, Italy

Outline

1. Introduction
   - why Adaptive Survey Design?
2. Methodology
   - stratification of target population
   - minimization upper limit bias
3. Adaptive Survey Designs in 2018 and 2019
4. Adaptive Survey Design in the Dutch Health Survey 2019
   - elaboration of the methodology

Introduction

Why Adaptive Survey Design?

Aim of Adaptive Survey Design:

to get a better balanced response by putting different effort in different groups of the population.

Adaptive Survey Design is effective in:

improving survey results, or reducing survey costs.

$$\hat{y}_{HT} = \sum_{k \in S} y_k / \pi_k$$

Methodology

Methodology

Adaptive survey design

1. The sample is a probability sample of size *n.*
2. Each person *k* in the population has a positive inclusion probability π_k.
3. Response follows the ‘Random response model’ in which person *k* responds with response probability ρ_k.
   Each ρ_k is only known to person *k*.

Methodology

Aim of survey: estimation of population means
for several target variables.

An estimator for the population mean $\bar{Y}$ of variable Y is the modified Horvitz-Thompson estimator:

$$\bar{Y}_{mHT} = \left(\sum_{k \epsilon r} Y_k / \pi_k \right) \Big/ \left(\sum_{k \epsilon s} 1/\pi_k \right).$$

In general this estimator is biased, unless all response probabilities ρ_k are equal.

Methodology

The bias can be approximated by $\frac{R(\rho,Y)\times S(\rho)\times S(Y)}{\bar{\rho}}$, with

Y : Target variable,
R : Pearson's correlation coefficient, $|R| \leq 1$,
S : Population standard deviation.

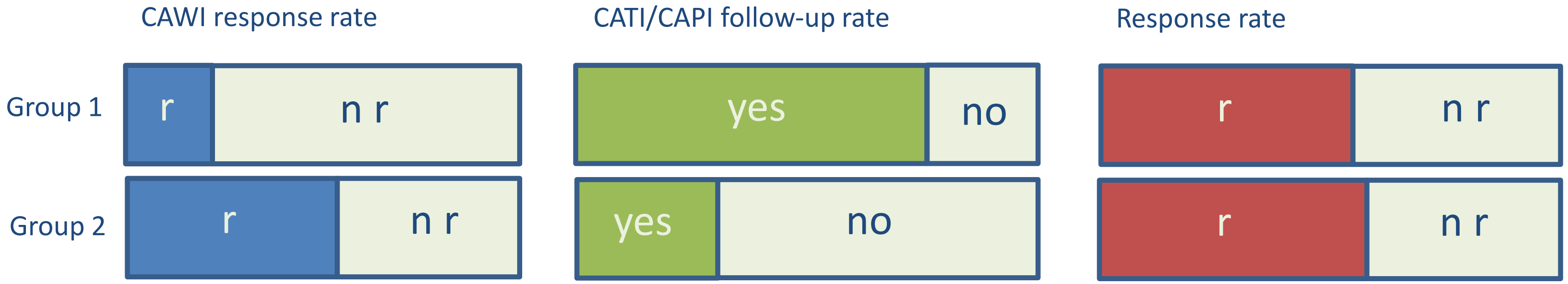
Upper limit for the bias: $\frac{S(\rho)\times S(Y)}{\bar{\rho}}$.

Aim: reduce bias by minimizing $CV(\rho) = S(\rho)/\bar{\rho}$.

Methodology

Observation strategy: CAWI → CATI/CAPI.

Features to adapt: CATI and CAPI follow-up.



Methodology

Determining target groups

People are divided into target groups based on personal characteristics, so that

- within each group: there is little variation in response behaviour per mode.
- between two groups: there is a big difference in response behaviour for at least one mode.

Methodology

Determining target groups

Clustering in two steps

1. Classification tree algorithm,
2. K-means clustering.

Methodology

Minimization problem

Minimize $CV(\rho)$ under constraints on

– budget,

– response numbers or rates,

– sample sizes per mode.

Solution: cawi sample size and inclusion probabilities,
cati and capi sampling fractions per target group,
estimate of $CV(\rho)$.

Adaptive Survey Designs 2018 and 2019

Adaptive Survey designs Implemented by Statistics Netherlands

2018

1. Health Survey	cawi → capi
2. Perception Survey	cawi → cati/capi

2019

1. Health Survey	
2. Perception Survey	
3. Leisure Survey	cawi → capi
4. Lifestyle Monitor	cawi → cati/capi

Adaptive Survey Design

in the Dutch Health Survey 2019

Health Survey

features

– aim: describing developments in health, medical care and lifestyle

– target population: people living in the Netherlands

– sampling design: simple random sample of people

– observation strategy: CAWI → CAPI

– desired number of respondents: 9500 per year

Health Survey

Determining target groups

The main personal characteristics used in determining the target groups are

- ethnicity
- urbanization
- age
- income
- ethnicity of parents
- marital status
- educational level
- gender
- place in household
- type of household
- wealth
- home ownership

Dataset: Health Survey, January – June 2018.

Health Survey

Determining target groups

Step 1.

The classification tree algorithm in the R package rpart determined three variables, and merged categories.

- Ethnicity(2): NL residents, migrants.
- Age(6): 0-11, 12-24, 25-44, 45-64, 65-74, 75+.
- Income(3): quintiles 1-2, 3-4, 5-10.

Step 2.

K-means clustering produced eight target groups.

Health Survey

Target groups

Partition of the population into target groups, Health Survey 2019

		NL residents			migrants		
age	income	1	2	3-5	1	2	3-5
0-11		2	8	8	7	8	8
12-24		1	6	8	3	1	1
25-44		2	6	1	3	5	5
45-64		1	6	4	5	5	5
65-74		6	8	4	1	5	4
75+		1	2	8	7	1	4

Health Survey

Target groups

Response rates per target group

Ordered by CAWI response rate

group	%r cawi	%r capi
3	15.7	38.3
5	23.8	28.9
2	29.9	49,0
7	30.2	59.8
1	32.6	41.8
6	37.7	40.2
8	42.8	51.6
4	51.4	43.4
total	38.6	42.3

Ordered by CAPI response rate

group	%r cawi	%r capi
5	23.8	28.9
3	15.7	38.3
6	37.7	40.2
1	32.6	41.8
4	51.4	43.4
2	29.9	49,0
8	42.8	51.6
7	30.2	59.8
total	38.6	42.3

Health Survey

Minimization problem

Minimize $CV(\rho) = S(\rho)/\bar{\rho}$ under constraints

1. CAWI sample size ≤ 18000.
2. Expected response size ≥ 9622.
3. CAPI sample size = 8040.
4. One CAPI sampling fraction per target group.

From constraints 1 and 2 it follows that $\bar{\rho} \geq \frac{9622}{18000} = 53.5\%$.

Minimization problem

Problem is solved with the R package Alabama.

The package uses the Augmented Lagrangian Adaptive Barrier Minimization Algorithm for optimizing smooth nonlinear functions with constraints.

The algorithm may end up in a local minimum, so different starting values were used and the best solution was selected.

Health Survey

Solution

group	n cawi	r cawi	%r cawi	n elig	n capi	f capi	r capi	%r capi	r total	%r total
1	3475	1133	33	2272	2228	98	931	42	2064	59
2	947	283	30	644	619	96	303	49	587	62
3	572	90	16	468	468	100	179	38	269	47
4	4193	2154	51	1977	1209	61	524	43	2678	64
5	1698	404	24	1256	1238	99	358	29	762	45
6	1189	448	38	718	701	98	282	40	730	61
7	231	70	30	157	123	79	74	60	144	62
8	3830	1639	43	2126	1454	68	750	52	2389	62
total	16135	6221	39	9617	8040	84	3401	42	9622	60

Health Survey

Solution

Quality indicators

Adaptive Survey Design	$\bar{\rho}$	$S(\rho)$	$CV(\rho) = \frac{S(\rho)}{\bar{\rho}}$
	%		
No	63.6	9.6	15.12
Yes	59.6	7.0	11.75

Effect of adaptive data collection on survey results?

This has been examined for the 2018 design, with the technique of bootstrapping.

> Samples with replacement were drawn from the 2016-sample, with the correct numbers for cawi en matching numbers per target group for capi.
>
> Estimates were made for the core variables of the Health Survey.

Health Survey

Survey results

Most of the survey results with adaptive data collection do not differ much from those without adaptation.

The greatest shifts:

1. Use of non-prescribed medicine ↑
2. Psychologically unhealthy ↓
3. Smoking non-western migrants ↑
4. Smoking western people ↓

Discussion
Questions

Discussion

Questions

1. How to reduce mode-specific measurement bias?

2. Is bias more a selection problem?

3. How relevant is the nonresponse bias?

4. Is adaptive survey design more effective in reducing costs than in improving accuracy?

End of talk