



Kees van Berkel

Progress on Adaptive Survey Design at Statistics Netherlands in 2019

UNECE Workshop on Statistical Data Collection

14 – 16 October 2019, Switzerland, Geneva

Outline

1. Introduction
2. Methodology
3. Adaptive Survey Designs in 2018 and 2019
4. Dutch Health Survey 2019
5. Further activities

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



Assumptions:

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 .

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:

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

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

Methodology

Adaptive survey design

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 $CV(\rho) = S(\rho) / \bar{\rho}$.

Ideal situation: Little variation in response behaviour and a high average response rate.

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.

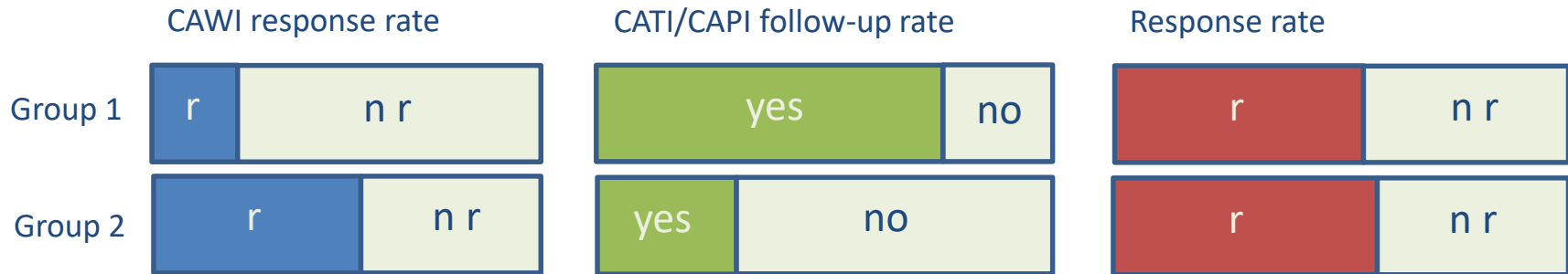


Methodology

Adaptive Survey Design

Observation strategy: CAWI → CATI/CAPI.

Features to adapt: CATI and CAPI follow-up.



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
2. Perception Survey

cawi → capi
cawi → cati/capi

2019

1. Health Survey
2. Perception Survey
3. Leisure Survey
4. Lifestyle Monitor

cawi → capi
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

urbanisation

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.



Target groups, Health Survey 2019

age	income	NL residents			migrants		
		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

Response rates

	CAWI	CAPI
min	3 15,7%	5 28,9%
max	4 51,4%	7 59,8%

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 $\bar{\rho} \geq \frac{9622}{18000} = 53.5\%$.



Health Survey

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	p cawi	f capi	p capi	p total
1	33	98	42	59
2	30	96	49	62
3	16 → 100	38	47	
4	51 → 61	43	64	
5	24 → 99	29	45	
6	38	98	40	61
7	30	79	60	62
8	43 → 68	52	62	
total	39	84	42	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	15.1
Yes	59.6	7.0	11.8



Effect of adaptive data collection on survey results?

This has been examined with the technique of bootstrapping:

Samples with replacement were drawn from the 2016-sample, with the correct numbers for cawi and capi per target group.

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 ↑

Further activities

1. Further reduction of CAPI.
2. Use of personal characteristics correlated with target variables in determining target groups.
3. Rethinking of post-survey adjustment.
4. Inclusion of mode-specific measurement biases.
5. Communication of shifts in survey results to the users.
6. Introduction of Adaptive Survey Design for more surveys.

End of talk