Using the CURIOS algorithm to manage the prioritization of CAPI surveys

Antoine Rebecq and Thomas Merly-Alpa

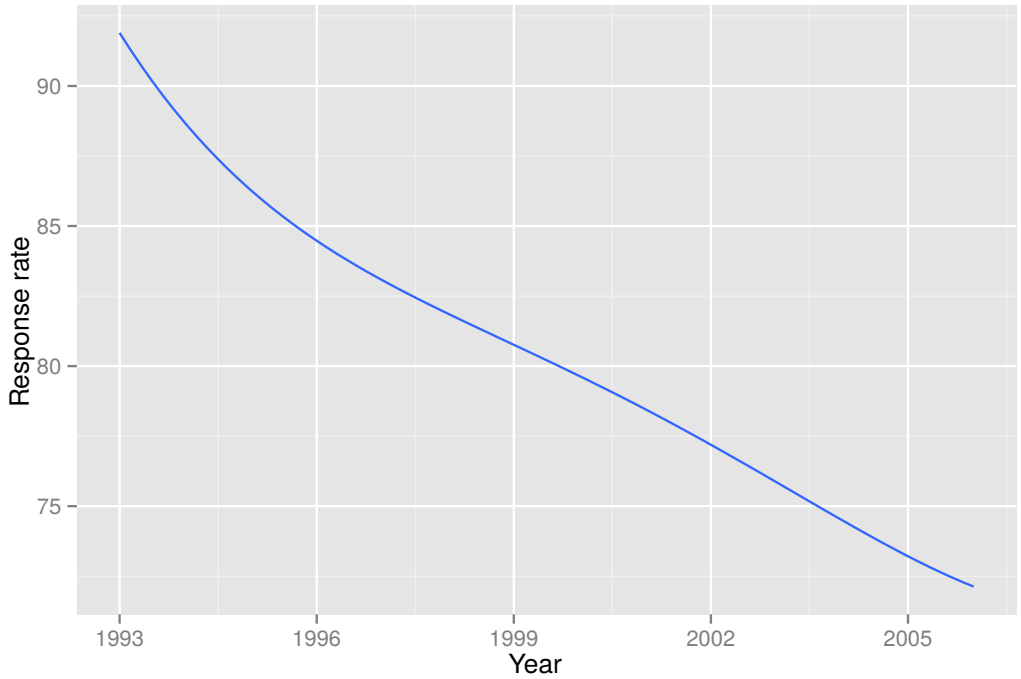
INSEE - France

September 15, 2015

Contents

Introduction

Response rates are dropping, and survey sampling theory does not account for non-response.



(Source : Survey of Labour and Income Dynamics (SLID) - Canada)

"General population" surveys

"General population" survey : A survey used to produce various results, using variables of interest that can be very wearkly correlated to one another.

- Example : Housing survey, European Social Survey, etc.
- Counter-example : Revenue/Wealth surveys (using Neyman/optimal allocation)

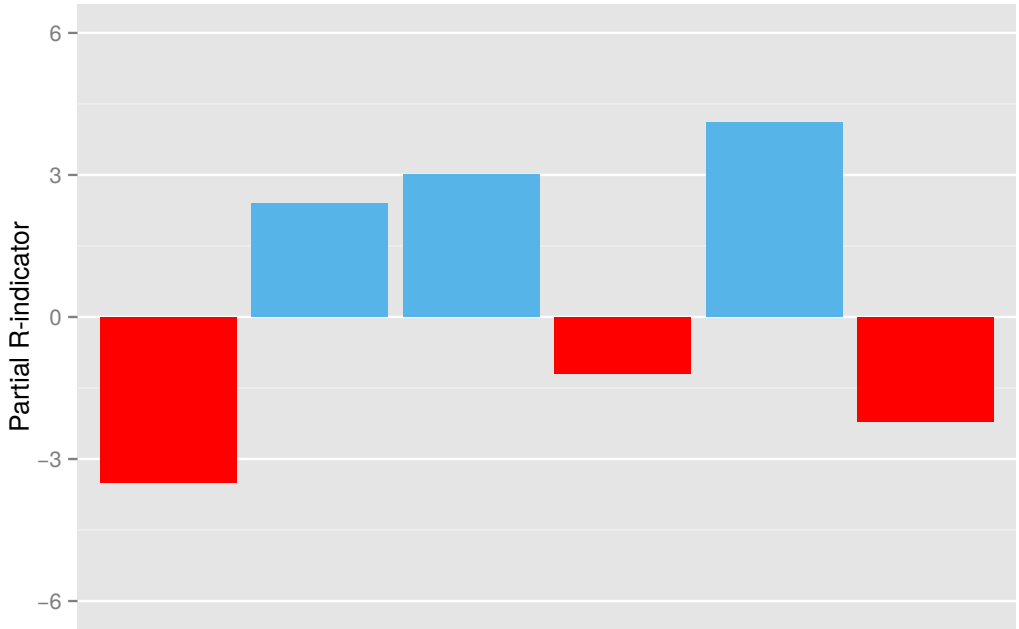
We select two features to assess the quality of a "general population" survey.

Equivalence principle

1. **minimize the dispersion of the non-response adjusted weights**.
   - Avoid influential individuals
   - More precise econometric studies

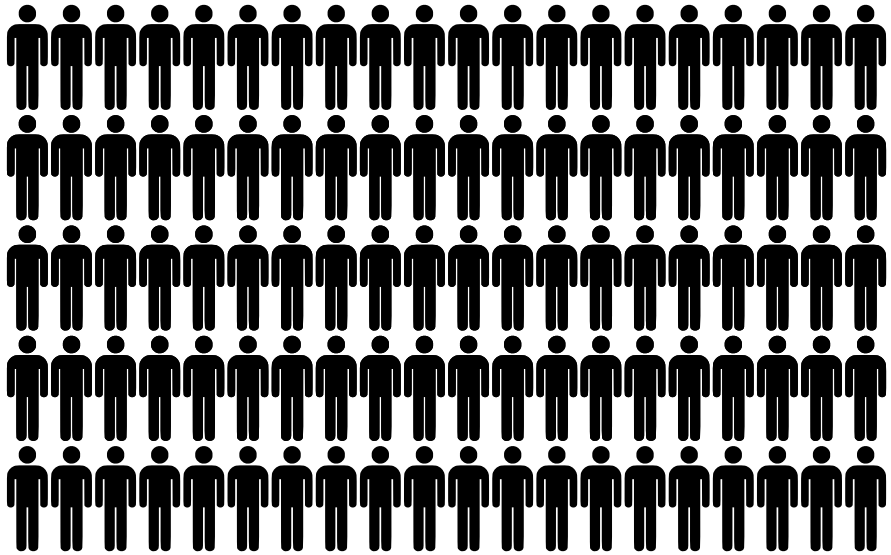
Balance principle

2. **Balance the profile of the respondants**, so that the structure of the final sample is the same as the initial sample.
We use R-indicators (Schouten, 2009).



Usual sampling method

The population is



Selecting 20 people (whatever the sampling design) gives us a sample :

Prioritizing units

Among the selected individuals, some (in red) have a particular behaviour in terms of non-response.

A way to identify these people is to use the R-indicators.

First wave

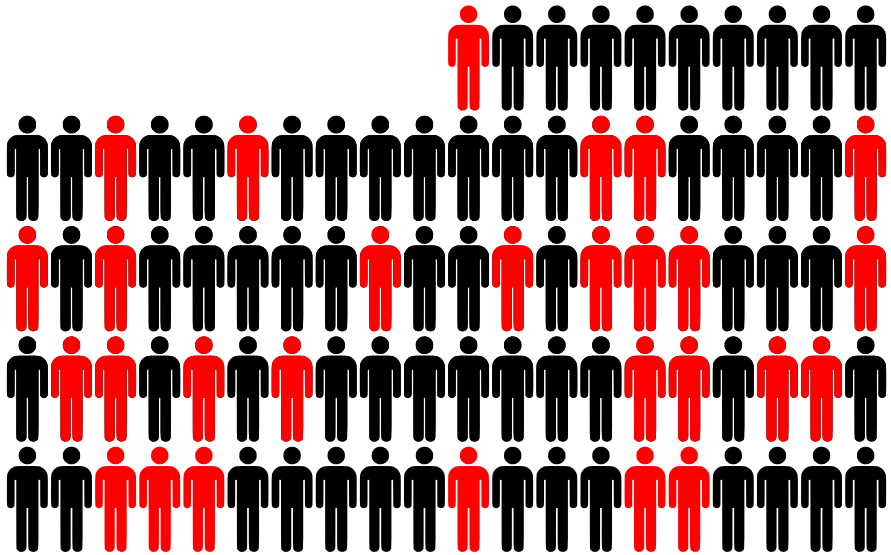
Instead of "just" selecting the "regular" sample :

We divide the sample into two waves :

At the end of the first wave, the CURIOS algorithm (using R-indicators, learning on the response phenomenon, Monte-Carlo, etc.) identifies the profiles that need prioritizing.

Back to the population

We identify these profiles in our remaining population

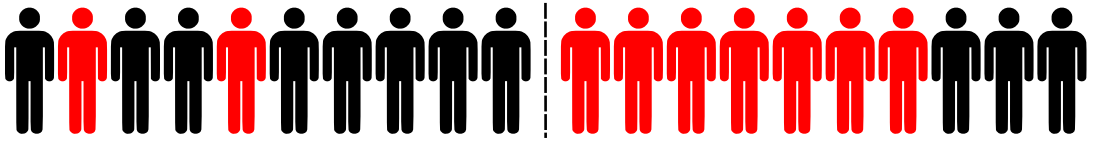


and we select them (same sampling design but different probabilities) 

Two-wave process

So, instead of giving the classic sample :

We divide the sampling design into two waves :



The CURIOS algorithm

The second wave sample is computed by minimizing the following program :

$$\arg\min_{S_2} \ \mathbb{E}\left[\Sigma(w_{NRA}) + \lambda \cdot \Gamma(S)\right]$$

Γ and $\Sigma(w_{NRA})$ are constructed using learning and Monte-Carlo techniques.

Two-wave sampling and the CURIOS algorithm

In Canada (Stat Can), CATI surveys are managed in real-time. We can't do this in the context of CAPI because it's much more difficult to manage follow-ups.

National survey on Education and Professional Qualification

Survey led roughly every 7 years since 1968, used for very diverse studies, such as :

- Social mobility
- Professional mobility
- Studies on professional status of immigrants
- ...

We try to design a somehow "optimal" sampling design for the second wave of the 2015 survey.

National survey on Education and Professional Qualification

For example, let's focus on two variables : age and housing type. Preliminary analysis (R-indicator). **Nothing is inferred about estimators at this point** :

Age	R-indicator	Response rate
23-33	-0.0366	60,5%
33-45	-0.0133	71,2%
45-55	0.0140	74,7%
55-65	0.0322	76,3%

Housing type	R-indicator	Response rate
House	0.0405	76,2%
Flat	-0.0554	61,9%

National survey on Education and Professional Qualification

Variable	Modality	Variation wave 2 (pct)	Variation total sample (pct)
Age	23-33	+ 13.4	+ 2.0
	33-45	+ 7.8	+ 1.2
	45-55	- 6.3	- 1.0
	55-65	- 10.6	- 1.8
Housing type	House	- 7.0	- 1.1
	Flat	+ 12.7	+ 2.0

The "specificity" objective

This method also works for "specific purpose" surveys (when the sample design focuses on the measure of one variable, e.g. Neyman allocation)

The "specificity" objective

We add the "specificity term" to our optimization problem (distance to the initial sample design)

$$\arg\min_{S_2} \ \mathbb{E}\left[\Sigma(w_{CNR}) + \lambda_1 \cdot \Gamma(S) + \lambda_2 \mathcal{d}(S_2, S_{initial})\right]$$

"Specific purpose" surveys