Indicators for representative response

Barry Schouten (Statistics Netherlands)
Natalie Shlomo and Chris Skinner (University of Southampton)

www.risq-project.eu

Why indicators for representative response?

- Response rate is insufficient indicator of quality response
  - Response rate only limits maximal impact of nonresponse
  - Literature gives examples where increased response rate corresponded to increased nonresponse bias (Peytcheva & Groves 2009, Schouten, Cobben & Bethlehem 2009)

- There is a need for indicators that enable
  1. comparison of response quality in different surveys
  2. comparison of response quality over time in one survey
  3. monitoring of response quality during data collection
  4. optimization of data collection resources

What is representative response?

- Definition: *Response is representative with respect to X if the response propensities are constant for X.*

- R-indicator: *the variation in response propensities*

- Maximal bias: *the variation in response propensities divided by the response rate*

- R-indicator is overall measure allowing for comparison of different surveys when X is fixed. R takes values in [0,1].
- Maximal bias is worst case bias of response mean allowing for optimal allocation of resources.

What is representative response?

- Partial R-indicators decompose R-indicator based on the impact of single variables

  total variance = between variance + within variance

- Unconditional partial R-indicator for a single variable Z: *the between variance of response propensities*

- Conditional partial R-indicator for a single variable Z given X: *the within variation in response propensities given a stratification on X*

- Both type of indicators should ideally be close to 0 and allow for monitoring of data collection and resource allocation

Example 1: Comparing surveys

Full model and forward variable selection
A = gender, B = age x marital status, C = urbanization, D = house value, E = paid job,
F =household type and G = ethnic background

Health Survey 2005	*Consumer Satisfaction Survey 2005*
Full model = A+B+C+D+E+F+G R=0,808 CI=(0,794 – 0,823)	Full model = A+B+C+D+E+F+G 0,821 (0,807 – 0,834)
B R=0,855 CI=(0,840 – 0,870)	B R=0,846 CI=(0,832 – 0,860)
B+G R=0,829 CI=(0,814 – 0,842)	B+F R=0,832 CI=(0,818 – 0,846)
B+G+C R=0,817 CI=(0,803 – 0,832)	B+F+G R=0,828 CI=(0,814 – 0,842)
B+G+C+F R=0,812 CI=(0,797 – 0,828)	B+F+G+E R=0,825 CI=(0,812 – 0,840)
Final selection = B+G+C+F+E R=0,810 CI=(0,796 – 0,824)	Final selection = B+F+G+E+A R=0,824 CI=(0,810 – 0,838)

Example 2: Monitoring a survey (Consumer Satisfaction Survey 2005)

Full model:
gender + age x marital status + urban + house value + paid job + household type + ethnicity

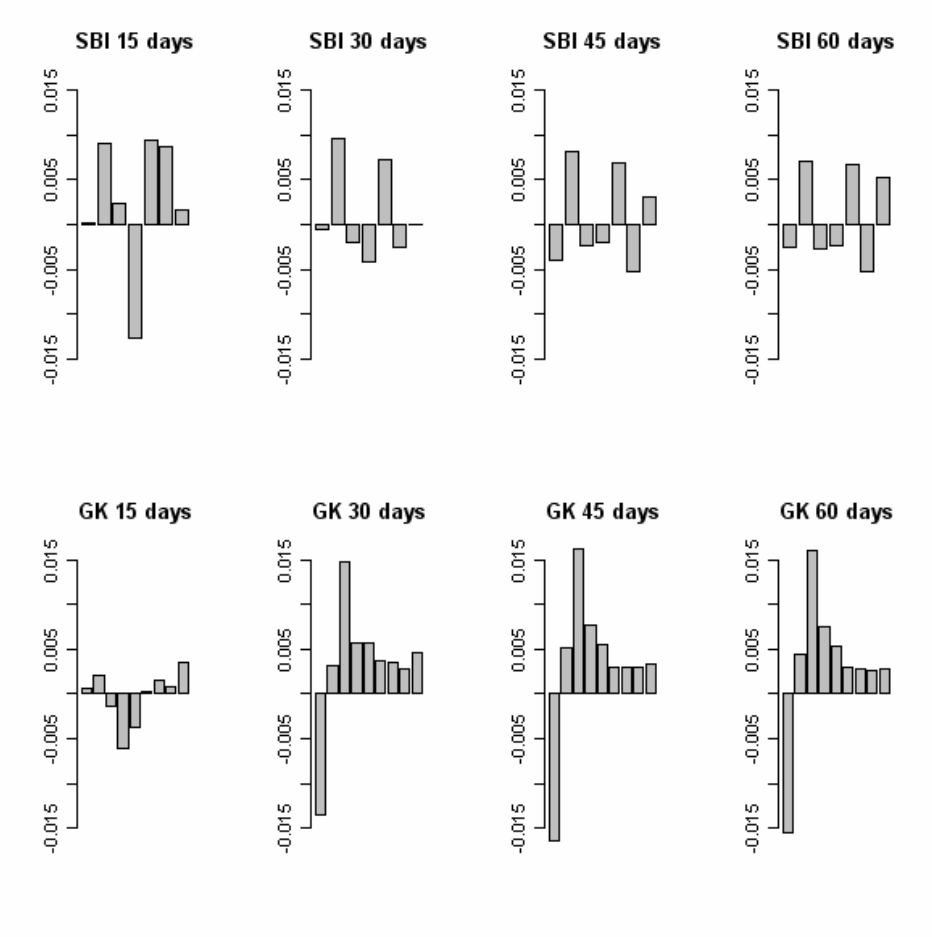
		eligible	contact	cooper-ation	response	eligible	contact	cooper-ation	response
		Univariate				**Multivariate**			
urban density	*N*	*8,9*	*14,8*	*14,7*	*22,6*	*1,6*	*9,5*	*10,5*	*9,3*
Very strong >2500	2844	-0,5	-12,4	-8,2	-16,9	0,0	4,9	0,2	2,4
Strong 1500 - 2500	3960	0,3	2,0	2,5	2,8	0,0	0,9	3,5	4,0
Medium 1000 - 1500	3455	0,1	2,1	5,1	5,7	0,1	0,3	0,7	0,5
Little 500 - 1000	3692	1,7	1,2	4,6	5,7	0,1	0,3	0,2	0,1
Not < 500	3913	1,3	7,3	-2,4	3,3	0,0	2,6	6,3	1,6
unknown	43	-8,6	-0,5	-9,6	-11,9	0,0	0,0	0,0	0,0
gender	*N*	*3,4*	*22,4*	*50,6*	*61,7*	*2,8*	*8,4*	*6,7*	*11,8*
male(s)	2058	-4,3	-19,4	-17,0	-31,9	0,4	3,7	1,1	5,9
mixed	11419	3,4	10,9	29,1	36,1	0,2	1,0	0,8	2,3
female(s)	3826	-0,2	-2,8	-37,7	-38,5	0,1	1,8	0,4	2,2
*mean worth of houses in neighbourhood (*100*	*N*	*13,8*	*15,2*	*29,2*	*32,2*	*4,5*	*9,0*	*7,8*	*9,0*
unknown	84	-13,1	-13,3	-12,2	-11,1	0,2	0,1	0,1	0,0
0-75	180	0,0	1,6	-9,8	-8,5	0,1	1,2	1,1	0,1
75-100	731	-2,8	-4,0	-14,7	-18,8	0,6	0,2	1,4	3,0
100-150	3625	-1,0	-3,1	-12,2	-14,6	0,3	0,4	0,2	0,1
150-200	5022	0,5	4,6	2,1	4,7	0,1	2,3	0,3	0,7
200-250	3317	1,8	0,6	11,1	11,3	0,2	0,6	1,5	1,0
250-300	1982	0,9	0,7	8,2	8,5	0,0	0,6	0,4	0,1
300-400	1938	1,3	0,8	6,4	6,9	0,0	1,1	0,2	0,1
400-500	641	-0,3	-1,3	-1,3	-2,2	0,1	1,3	0,6	2,4
500 and more	405	1,6	0,2	-1,2	-0,1	0,5	0,5	0,4	0,6

Example 3: Monitoring a survey in time (*Short Term Statistics 2006)*

Survey		*X = Business size + type*				*X = Business size x VAT + type*			
		15days	*30days*	*45days*	*60days*	*15days*	*30days*	*45days*	*60days*
Industry	R	0,921	0,933	0,940	0,942	0,905	0,918	0,931	0,933
	CI	0,913-0,928	0,927-0,940	0,935-0,944	0,938-0,946	0,897-0,913	0,913-0,922	0,926-0,935	0,928-0,938
	Bias	8,1%	4,2%	3,5%	3,3%	9,7%	5,2%	4,1%	3,8%
Retail	R	0,961	0,946	0,940	0,941	0,881	0,879	0,883	0,890
	CI	0,954-0,967	0,940-0,952	0,935-0,945	0,936-0,946	0,873-0,888	0,873-0,886	0,876-0,889	0,883-0,896
	Bias	3,9%	3,5%	3,5%	3,3%	12,0%	7,7%	6,8%	6,2%

Example 3 - continued

Unconditional partial indicators for business type (SBI) and business size (GK)



Important side remarks

- *Dependence on external information*: No statement about the representativeness of response is possible without information that is auxiliary to the survey.
- *Dependence on sample size:* The strength of any statement about the nature of response to a survey will depend on the sample size.
- *Non-response adjustment*: Indicators are not designed for the selection of weighting variables but for the evaluation and enhancement of response.
- *Less is better?*: One may always attain representative response with respect to some standard by simply erasing (or sub-sampling) overrepresented groups in the response.

RISQ papers and software

- Detailed papers on
  - Estimation of (partial) R-indicators
  - How to use R-indicators
  - Applications in fieldwork monitoring and responsive design
- Software release 1
  - SAS and R, and R-cockpit
  - Manual
  - Test file

available at www.risq-project.eu

Release 2 (nov ’10): confidence intervals partial R-indicators
Release 3 (apr ’11): population-based R-indicators