Assessing and improving calibration weighting of web surveys using the R-indicator

*Rong Wei, Van L Parsons, and Yulei He

Division of Research and Methodology

National Center for Health Statistics, CDC

2023 FCSM Research & Policy Conference

October 24, 2023

Outline

- Motivation
- Introduction:
    1. “Representativeness” by Schouten, Cobben and Bethlehem (SCB original paper 2009)
    2. Adapt SCB’s “Representativeness” to the Web survey/Benchmark case in this study
- Method
- Results
- Discussion
- Summary
- References

Motivation

- Quickly produced commercial panel-based web surveys have been developed to complement the ability of the federal statistical system to provide health information about the U.S. population.

- Despite their great potential, statistical inferences based on these web surveys might be subject to potential bias compared with traditional, high-quality household surveys.

- To mitigate these biases, units from the web survey are usually calibrated to external controls by reweighting samples, often using a benchmark survey of high quality.

- We propose to use an adaptation of the ***R-indicator***, originally suggested as a measure of quantifying “representativeness” of survey response, to assess and improve the quality of **calibration weighting**. This metric can be effectively used to identify possible calibration variables and compare alternative weighting strategies.

Introduction: “Representativeness” using a metric R-indicator by SCB, 2009

Use a definition proposed from Schouten, Cobben and Bethlehem (Survey Methodology 2009), i.e., **SCB’s** **response *vs.* non-response** context, indicators for the representativeness of survey response are defined as

Definition of “Representativeness” of response:

For a population: X = covariate with H categories h=1 to H

$$\rho_{hu} = P(\text{ unit u is a response} \mid u \in h)$$

“representativeness” for class covariates requires $\bar{\rho}_1 = \bar{\rho}_2 = \ldots = \bar{\rho}_H$

In words: Nonresponse mechanism is **M**issing **C**ompletely **A**t **R**andom with respect to X.

“Representativeness” cont.

A metric defining constancy of response propensities in the population:

$$S^2(\rho_x) = \sum_{u=1}^{N} \frac{1}{N}(\rho_{u,x} - \bar{\rho}_x)^2$$ where, $$\bar{\rho}_x = \sum_{u=1}^{N} \frac{1}{N}\rho_{u,x}$$

the ***R-indicator*** is $R(\rho_x) = 1 - 2\sqrt{S^2(\rho_x)}$ and $0 \le R(\rho_x) \le 1$

- “Not representative” as $R(\rho_x)$ approaches 0
- “Representative” as $R(\rho_x)$ approaches 1

This generic population metric can be adapted to sampling situations.

Introduction: Adapt SCB response/non-response "representative" metrics to Web Panel surveys / Benchmark survey "representative" metrics

We have two independent designs, ***Benchmark*** and ***Web*** over the population.

Usually, the ***Benchmark*** design (National Health Interview Survey, or NHIS) is a well-established population survey considered to be of high quality with pre-specified calibrated weights.

The ***Web*** design will correspond to a web survey with its own pre-specified calibrated weights.

The two questions we are addressing are:

1. Are ***Web*** and ***Benchmark*** "representing" the same population?
2. Can ***Web's*** weighting be modified for better representation?

Method: Use a 2-survey *R-indicator* for assessment of *Web* with regard to *Benchmark*

1. Pool the samples from ***Web*** and ***Benchmark***.

2. Prediction only, so no variance structures used.

3. Scale the weights so that the sum of the ***Web*** weights = sum of the ***Benchmark*** weights. The weighted proportion of each sample to the weighed total is ½.

4. Propensity estimation.

   Model $y_u \sim f(x_u)$, on the ***X*** (usually logistic regression)

$y=1$ if a unit *u* in the pooled sample is in ***Web***

$y=0$ if a unit *u* in the pooled sample is in ***Benchmark***

X : important covariates (domains)

Method Cont.

5. For unit u with covariate x_u the prediction is $\hat{\rho}_{u|x} = \hat{P}(\text{ unit u is from web} \mid x)$

the mean prediction is $\hat{\bar{\rho}} = \sum_{u=1}^{n_w+n_B} \hat{\rho}_{ux} w_u$

the distance of predictions from the mean is $\hat{S}^2(\hat{\rho}_X) = \sum_{u=1}^{n_w+n_B} w_u (\hat{\rho}_{ux} - \hat{\bar{\rho}})^2$ (weights scaled to 1)

Web's is "Representative" if the $\hat{\rho}_{u|x}$'s are roughly constant or if $\hat{S}^2(\hat{\rho}_X)$ is small.

SCB form: ***R-indicator*** $\hat{R}(\hat{\rho}_X) = 1 - 2\sqrt{\hat{S}^2(\hat{\rho}_X)}$ in range $[0,1]$

$\hat{R}(\hat{\rho}_X) \approx 1$ interpreted as ***Web*** and ***Benchmark*** are "equally representative" with respect to X

Features of the R-indicator

1. For the ***Web*** and ***Benchmark*** surveys, the initial weights can be considered as survey adjusted weights. They may be pre-adjusted for non-response and calibrated to external controls.

2. The *R-indicator*, $\hat{R}(\rho_x)$ and the form $\hat{S}^2(\rho_x)$ are equivalent metrics, with the latter form targeting 0 as an indication of representativeness. The latter form is more amenable to explaining features of the metric.

3. The scaling of the two survey's weights to sum to ½ makes the *R-indicator* a useful metric to evaluate different weighting methods on the ***Web*** in relation to the ***Benchmark***. Deviations of $\hat{\rho}_{W,x}$ and $\hat{\rho}_{B,x}$ from 0.50 over all observations are main components of the *R-indicator*.

4. If x_1 and x_2 are two sets of covariates and $x_{12} = (x_1, x_2)$ is the combined set then $\hat{R}(\rho_{x_{12}}) \leq \hat{R}(\rho_{x_1})$, *i.e.*, adding more covariates to the model decreases the *R-indicator*.

General application to determine impact of survey weights and covariates on survey ’s “representativeness”

Pre-release, consider the Web survey as open to survey calibration methods.

Determine a weighting method that achieves some degree of “representativeness” with a Benchmark survey.

Start with w1=1 for raw assessment and

w2= Web provided weight (possibly complex strategy)

Select re-calibration weighting methods w3, …, wk (may include w2 population controls along with additional controls based on benchmark variables.

Select assessment covariates (can be different from calibration controls)

Evaluate the *R-indicator* by weight and assessment covariates.

General application: cont.

Create a weighting method / assessment covariate-vector table with different "representative" covariate groupings. Determine a weighting method that meets the Web survey's objectives (subjective).

$$\begin{bmatrix} D \quad \text{weight option} & | & \text{covariate option} & | & \text{covariate option} \\ & | & x_1 & | & x_2 \\ & | & \text{R-indicator} & | & \text{R-indicator} \\ w_1 & | & R(x_1) & | & R(x_2) \\ w_2 & | & R(x_1) & | & R(x_2) \\ w_3 & | & R(x_1) & | & R(x_2) \end{bmatrix}$$

Example: the 2019 NHIS* serves as the Benchmark survey while the RANDS 4** is the Web survey

Survey	Weight system	Weight calibration variables
NHIS (n=31,997)	NHIS Final calibrated Weight	Census provided demographic variables
RANDS 4 (n=3,442)	Unit Weight	No
	AmeriSpeak Weight	Census provided demographic variables^
	Candidate re-weightings	
	Calibwgt5	By raking: 5 demographic variables^: gender, age, race/ethnicity, education, Census region
	Calibwgt9	By raking: 9 variables: variables from 5-variable calibration plus marital status, income, and selected health outcomes (asthma, diabetes)

***National Health Interview Survey (NHIS)** which is based on a personal interview with weighting which includes nonresponse adjustment and raking to US population totals.

****RANDS 4** is a web-panel survey (conducted by NORC) based on AmeriSpeak with weights adjusted to US population totals.

^ common demographics may vary in definition by AmeriSpeak and candidate re-weightings.

Candidate variables (x) used in logistic models: Pr(Web=1|x) ~ Bx

Variable	Number of Categories	Category group
Gender	2	Male, female
Age group	3	18 - 44, 45 - 64, 65+
Race/ethnicity	4	Hispanic, NH white, NH black, NH other
Education	3	<=High school, some college, >=Bachelor
Region	4	Northeast, Midwest, South, West
Marital status	2	Married, not married
Income	2	<$50,000, $50.000+
Asthma (ever)	2	Yes, no
Diabetes (ever)	2	Yes, no
Health status	2	Fair/poor, good+
Anxiety (severe)	2	Yes, no
Depression (severe)	2	Yes, no

Result: logistic model: Pr(Web=1|x) ~ Bx for single health outcome(x)

Single x	Unit-weight	NORC Weight	NCHS Calibrated Weight
Asthma*	0.743	0.911	0.935
Diabetes*	0.746	0.926	0.958
Health status**	0.618	0.704	0.808
Anxiety**	0.648	0.739	0.889
Depression**	0.649	0.742	0.891

*Weighted regression using *Calibwgt5* ; **Weighted regression using *Calibwgt9*.

Results: NCHS Calibrated Weights improved the *Web* survey's "representativeness" with higher *R-indicators*

Results: logistic model: Pr(Web=1|x) ~ Bx for multiple outcomes: health status + asthma + diabetes + depression + anxiety

	Unit Weight	NORC Weight	Calibwgt5	Calibwgt9
R-indicator	0.655	0.755	0.760	0.916

Results: NCHS Calibrated Weights improved the Web survey's "representativeness" with higher *R-indicators*

Impact of survey weights and covariates on survey ’s “representativeness”, cont.

1. Impact from survey weight in propensity score (PS) logistic regression: we used different weighting strategies to improve the R-indicator, i.e., we compared R-indicators with different weights included in PS models.

2. Impact from covariates included in PS logistic regression: For point estimates, target health outcomes might vary with R-indicator computing PS models.

Summary

- The *R-indicator* : used to assess the “representativeness” of a web-panel based health survey as compared to the NHIS (benchmark).
- The metric can be used to evaluate possible weighting strategies and select covariates common to both surveys.
- In our case study example, the *R-indicators* helped improve calibration reweighting when compared to the web survey’s weight.
- *R-indictors* on periodic web-panels may suggest:

1. Additional weight calibrations are needed;
2. Design feature changes from previous survey *R-indicator* assessments;
3. New non-sampling issues.

Other studies on R-indicators

- Schouten *et al*. (2012). *R-indicators* can be applied to establish the quality of register data.

- Roberts *et al.* (2020) Case study using data from the Swiss European Social Survey and nonresponse follow-up survey indicated that a validation of *R-indicator* depends on the auxiliary data used in *R-indicator* estimation.

- Michael *et al.* (2022) studied “universal adaptability”, which focused on target-independent inference that competes with propensity scoring.

References

- Schouten B, Cobben F and Bethlehem J 2009 Indicators for the representativeness of survey response. Survey Methodology, June 2009 101 Vol. 35, No. 1, pp. 101-113 https://www.researchgate.net/publication/267836796
- Rosenbaum PR, Rubin DB, 1983. Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. Journal of the Royal Statistical Society. Series B (Methodological) , 1983, Vol. 45, No. 2 (1983), pp. 212-218 https://www.jstor.org/stable/2345524
- Rosenbaum PR, Rubin DB, 1983. The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41-55.
- Roberts C, Vandenplas C, and Herzing JME, 2020. A Validation of R-Indicators as a Measure of the Risk of Bias using Data from a Nonresponse Follow-Up Survey. Journal of Official Statistics, Vol. 36, No. 3, 2020, pp. 675–701. http://dx.doi.org/10.2478/JOS-2020-0034
- Barry Schouten, Jelke Bethlehem, Koen Beullens, ØyvinKleven, Geert Loosveldt, Annemieke Luiten, Katja Rutar, Natalie Shlomo and Chris Skinner 2012 Evaluating, Comparing, Monitoring, and Improving Representativeness of SurveyResponse Through R-Indicators and Partial R-Indicators International Statistical Review(2012), 80, 3, 382–399 doi:10.1111/j.1751-5823.2012.00189.x
- Michael P, Kim MP, Kern C, Goldwassera S, Kreutere F, Reingoldg O, 2022. Universal adaptability: Target-independent inference that competes with propensity scoring. PNAS 2022 Vol. 119 No. 4, https://doi.org/10.1073/pnas.2108097119

Thank you!

Rong Wei: rrw5@cdc.gov