Measuring Survey Quality through Representativity Indicators using Sample and Population based Information

Chris Skinner, Natalie Shlomo,
Barry Schouten, Jelke Bethlehem
Li-Chun Zhang,
Representativity Indicators for Survey Quality

collaboration between:

national statistical institutes of Netherlands, Norway and Slovenia

& universities of Leuven & Southampton
Representatvity Indicators

• quality indicators for survey non-response

• to supplement response rate

• to measure how well respondents represent population

• tools for use at different stages of survey process (data collection+)
Aim of paper

• to consider estimation of representativity indicators using either:
  • sample-based information (microdata for both respondents and nonrespondents), or
  • population-based information (microdata for respondents and aggregate data for population)
How to define representativity indicator?

two approaches

• both based on idea of response propensity
Response propensity

Idea: R-indicator measures homogeneity of response propensities:

\[ \rho_i = \text{probability of response for unit } i \text{ given values of auxiliary variables} \]
R-indicator

\[ R(\rho) = 1 - 2S(\rho) \]

where
\[ S(\rho) = \sqrt{\frac{1}{N-1} \sum_{U} (\rho_i - \bar{\rho}_U)^2} \]

- \( R(\rho) = 1 \) if \( \rho_i \) constant
- \( R(\rho) = 0 \) if \( S(\rho) = 0.5 \)

Schouten, Cobben & Bethlehem (2009, *Survey Methodology*)
Alternative Indicator

\[ Q^2(\rho) = \left[ \sum_U \rho_i \right]^{-1} \left[ \sum_U \rho_i (\phi_i - \bar{\phi})^2 \right] \]

where \( \phi_i = \rho_i^{-1} \)

proposed by Särndal and Lundström (2008, JOS) in the context of selecting auxiliary variables for weighting adjustment
Estimated R-indicator

Sample-based – auxiliary variables recorded for whole sample

(1) estimate response propensities using e.g. logistic regression

\[ \hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{N-1} \sum_s d_i (\hat{\rho}_i - \hat{\rho}_U)^2} \]
Estimated R-indicator

**Population-based** – auxiliary variables only measured on respondents and in aggregated form for population

(1) Estimate response propensities using ordinary least squares

\[ \hat{\rho}_i^{OLS} = x_i \left[ N(S_{xx} + \bar{x}x') \right]^{-1} \sum_r d_r x_i \]

if population covariance matrix \( S_{xx} \) known. Estimate \( S_{xx} \) from respondents if only population mean mean vector \( \bar{x} \) known
Estimated R-indicator

Population based (continued)

\[ \hat{R}_r(\rho) = 1 - 2 \sqrt{\frac{1}{N-1} \sum_r d_r \hat{\rho}_i^{-1} (\hat{\rho}_i - \hat{\rho}_r)^2} \]
Simulation Study

• Samples from Israel census data on 753000 individuals
• ‘realistic’ sampling with fractions 1:50, 1:100, 1:200
• ‘realistic’ non-response based on type of locality, household size, children in household – overall response rate 82%
Simulation Means of $\hat{R}(\rho)$ for Sample (S), Population with Known Covariance matrix (PC) and Population with Unknown Covariance Matrix (PUC) of Auxiliary Variables – True Model

![Graph showing the simulation means of \( \hat{R}(\rho) \) for different models and sample sizes. The graph includes lines for True Model S based, True Model PC based, and True Model PUC based, with y-axis values ranging from 0.8 to 0.88 and x-axis values ranging from 0.5% to 2.0%.]
Simulation Means of $\hat{R}(\rho)$ for Sample (S), Population with Known Covariance matrix (PC) and Population with Unknown Covariance Matrix (PUC) of Auxiliary Variables - Less Complex Model
$\hat{R}(\rho)$ for Sample (S), Population with Known Covariance Matrix (PC) and Population with Unknown Covariance Matrix (PUC) of Auxiliary Variables – True Model
\( \hat{R}(\rho) \) for Sample (S), Population with Known Covariance Matrix (PC) and Population with Unknown Covariance Matrix (PUC) of Auxiliary Variables – Less Complex Model
Simulation Means of $q^2$ for Sample (S) and Population (P) Based Auxiliary Variables – True Model
Simulation Means of $q^2$ for Sample (S) and Population (P) Based Auxiliary Variables - Less Complex Model
$q^2$ for Sample (S) and Population (P) Based Auxiliary Variables
- True Model
Conclusions

• representativity indicators defined in terms of response propensities
• can estimate accurately given either sample-based or population-based information
• have also applied to surveys from RISQ countries
• and have examined bias-corrected estimators and estimators of standard errors of estimators