Social Statistics Away Day
May 16th 2024

What is Sample Representativeness?

Natalie Shlomo
Natalie.Shlomo@manchester.ac.uk

Sample Representativeness

- Bias of the sample mean assuming SRS: $Bias(\bar{y}_r) = (1 - \bar{R})(\bar{Y}_r - \bar{Y}_{nr})$
- Response rate not sufficient as a quality indicator to capture impact of nonresponse
- Bias also depends on the contrast between respondents and non-respondents with respect to a target variable
- Develop indicators on whether group of respondents represent complete sample

.

Project RISQ
(Representativeness Indicators for Survey Quality, www.risq-project.eu)

Aim:

- Compare response to surveys that share same target population
- Compare response to a survey longitudinally
- Monitor response during data collection,
- Control response by adaptive survey designs (Schouten, Peytchev, Wagner ,2018)

What is representativeness?

- Representativeness defined by individual response probabilities that need to be estimated
- Let ρ_X denote the response propensity function for variable X and $\rho_X(\mathrm{x})$ the probability that a population unit with value $X = x$ will respond to the survey
- Two definitions for representativeness of survey response:

Definition: A response to a survey is representative with respect to X when response propensities are constant, $\rho_X(\mathrm{x})$ is a constant function.

Definition: A response to a survey is conditional representative with respect to X given Z when conditional response propensities are constant for X, $\rho_{XZ}(\mathrm{x},\mathrm{z}) = \rho_Z(\mathrm{z})$ for all x.

- The two definitions can be measured for any auxiliary vectors X and Z so need a distance function: $d(\rho_1, \rho_2) = \sqrt{\frac{1}{N}\sum_U(\rho_{1,i} - \rho_{2,i})^2}$
- Assess whether data collection succeeded in obtaining a balanced response for a set of pre-selected variables X that are available before and during data collection (and does not include a target variable Y)

Sample Representativeness

- Representativeness indicator (R-indicator) is the transformed distance between ρ_X and the constant response rate and defined as: $R = 1 - 2S(\rho_X)$
- Value of 1 is full representativeness and (close to) 0 the largest possible deviation from representative response (note: the maximum standard deviation for the binomial distribution is 0.5).

Denoting ρ_i for $\rho_X(\mathrm{x})$, the R-indicator is estimated by:

$R = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{n} d_i(\rho_i - \bar{\rho})^2}$ where d_i is the survey design weight and

$\bar{\rho} = \frac{1}{N}\sum_{i=1}^{n} d_i\rho_i$ overall response rate

Sample Representativeness

- Unconditional Partial R-indicator with respect to one variable Z defined as $P_u(Z) = S(\rho_Z)$
- Assume Z has H categories and let $\Delta_{h,i}$ the 0-1 indicator function for $Z = h$ so $n_h = \sum_{i=1}^{n} d_i \, \Delta_{h,i}$. Let $\overline{\rho_h}$ be the mean of response probabilities for category h in Z

Estimated by $P_u(Z) = \sqrt{\frac{1}{N}\sum_{h=1}^{H} n_h(\overline{\rho_h} - \bar{\rho})^2}$

- It holds that $P_u(Z) \leq 0.5$, the larger the value the stronger the impact of variable Z on lack of representativeness
- Unconditional categorical-level Partial R-indicator with respect to one category h of Z

Estimated by $P_u(Z,h) = \sqrt{\frac{n_h}{N}}(\overline{\rho_h} - \bar{\rho})$

- Negative signed $P_u(Z,h)$ is a lack of representativeness sand positive signed $P_u(Z,h)$ over-representativeness
- Can build profiles of individuals to target data collection

Sample Representativeness

- Conditional Partial R-indicator measures relative importance of a variable conditional on all other variables in the response model. As such conditional partial R-indicators attempt to isolate the part of the deviation of representative response that is attributable to a variable alone.
- Checks whether high unconditional partial R-indicator is still high conditional on other variables
- Define cross-classification of all model variables, with the exception of variable Z
- Cross-classification results in L cells: $s_1, s_2, \dots, s_L$ and denote n_l weighted sample size in cell l for $l = 1,2, \dots L$

Estimated by $P_c(Z) = \sqrt{\frac{1}{N}\sum_{l=1}^{L}\sum_{i\in s_l} d_i(\rho_i - \overline{\rho_h})^2}$

- $P_c(Z) \leq 0.5$, the larger the value the stronger the impact of variable Z on lack of representativeness conditional on all other variables
- Conditional categorical-level Partial R-indicator with respect to one category h of Z

Estimated by $P_c(Z,h) = \sqrt{\frac{1}{N}\sum_{l=1}^{L}\sum_{i\in s_l} d_i\Delta_{h,i}(\rho_i - \overline{\rho_h})^2}$

R-indicators
Software and User Manual

www.risq-project.eu

R-code – currently being re-instated on the website but can be found at various websites:
https://www.practicalsignificance.com/posts/r-indicators-in-r/
https://github.com/addinall/RISQ

Manual: https://hummedia.manchester.ac.uk/institutes/cmist/risq/RISQ-manual-v21.pdf

R-code includes sample size bias corrections and confidence intervals for R-indicators, and also produces the CVs of response propensities: $\frac{S(\rho_X)}{\bar{\rho}}$) and their confidence intervals (useful for data collection monitoring)