R Indicators

Response rate and non-response bias

The response rate of a survey is frequently used as an indicator of the quality of the survey. However, it is well-known in the survey methodological literature that response rates by themselves can be poor indicators of non-response bias. Up until now no alternative indicators of non-response have been suggested that may be more useful to assess survey quality.

Non-response has two main consequences. First, it reduces the sample size. Therefore. it decreases the precision of the estimates. Second, it affects the sampling design. The probability to obtain an observation depends on both the selection probabilities specified in the sampling design and the unknown probabilities to respond (if selected). Just using selection probabilities in estimation procedures will lead to biased estimates.

The response rate is directly related to the precision. The higher the response rate, the higher the precision. The same does not hold for the bias. The bias does not only depend on the response rate, but also on the extent to which respondents on non-respondents differ. A higher response rate only minimises the maximal impact that non-response has under the worst-case scenario.

Example

The table below contains data from 1998 Dutch Integrated Survey on Household Living Conditions (POLS).

The fieldwork of the survey covered a period of two months. The first month was face-to-face. In the second month, non-respondents of the first month were once more approached by telephone. If no registered fixed phone number was available, no additional attempts were made.

The table contains response means and estimates after non-response correction after the first and second month of interviewing for five variables. The first three are target variables of the survey. The last two are derived from administrative data that are linked to the sample. These variables are treated as if they are target variables. Only values of respondents are used in estimation and non-response adjustment. Since complete sample means are know, it can be established how effective non-response adjustment is.

Variable	Response mean		Estimate	
	After 1 month	After 2 months	After 1 month	After 2 months
Employed	48.6%	50.4%	49.6%	50.6%
Owns house	63.0%	63.3%	59.1%	59.4%
Owns PC	59.6%	59.8%	57.3%	57.2%
Social allowance	10.5%	10.4%	11.6%	11.4%
Non-native	12.9%	12.5%	14.6%	14.4%

The complete sample means of Social allowance and of Non-native were 12.1% and 15.0%, respectively.

The example shows that after the first month response means are closer to the adjusted estimates. For the variables Social Allowance and Non-native it is clear that first month’s respondents represent the sample better than the overall response after the second month.

Furthermore, the non-response adjustment was not able to remove this difference. This difference is mainly caused by the composition of households with a registered fixed phone.

Increasing response rates may thus affect the representativity of the final response.

It can be shown that (under the Random Response Model) the bias of the response mean is equal to

$$R(\rho, Y)\ x\ S(\rho)\ x\ S(Y)\ /\ A(\rho)$$

in which $R(\rho,Y)$ is the correlation between response probabilities ρ and the values of the target variable (Y), $S(\rho)$ is the standard error of the response probabilities, $S(Y)$ is the standard error of the target variable (Y) and $A(\rho)$ is the average of the response probabilities.

This implies that low response rates have a negative impact on bias if there is a linear relation between the survey variable and response behaviour. We will use response probabilities to define the concept of representativity.

The concept of representativity

The concept of **representativity** is not one without debate. Kruskal and Mosteller (1979) give an extensive overview of the interpretations of this concept in the statistical and non-statistical literature. Here, we relate representativity to the missing-data-mechanism alone. We use the individual response probabilities $\rho[k]$, for $k = 1, 2, .., N$. We define the response data set to be**strongly representative** if all response probabilities are identical.

This is an attractive definition, but not useful is practice. We can not test whether our survey data set satisfies this definition, simply because we can not estimate the response probabilities. Therefore, we introduce a weaker definition that relates representativity to the available auxiliary information. A response data set is said to be **weakly representative** with respect to the sample and an auxiliary variable X if the average response probability is the same in each category of X.

The weak definition of representativity implies that the auxiliary variable and response behaviour are independent. This assumption can be tested with a chi-squared test.

The two definitions given here relate to the concept **Missing Completely At Random** (MCAR). A missing-data-mechanism is MCAR in case the probability to respond does not depend on the value of the target variable. Hence, in case the response dataset is strongly representative, the missing-data-mechanism is MCAR for any target variable. In case the response subset is weakly representative for X, then the missing-data-mechanism is MCAR for X but not necessarily for other variables.

Response probabilities can be used for the construction of indicators that measure divergence from representativity.

Measures of representativity

In practice response will hardly ever be fully representative with respect to all available auxiliary socio-demographic characteristics. In order to evaluate the quality of the response there is a need for indicators of how dissimilar the response data set is from the sample data set. One example of a **Representativity Indicator** (or: **R-indicator**) is given here.

This R-indicator is based on the standard deviation of estimated response probabilities. It is defined by

$$M(\rho)= 1 - 2\ S(\rho)$$

The response data set is representative if all response probabilities are equal. In this case the standard deviation is zero, and M(p) assumes the value 1.

The response data set is not representative if there is much variation in response probabilities. This is reflected by a large standard error. The maximum value the standard error can assume is 0.5. In this case the value of the R-indicator is equal to 0.

In practice, the values of the response probabilities are not known. Therefore, they are estimated using, for instance, a logistic or probit regression model.

Example

The R-indicator was used in an experiment with the Dutch Labour Force Survey (LFS). Non-respondents were re-approached with either a complete questionnaire form (call-back approach) or with a very short questionnaire (basic question approach). Both approaches aim at increasing response by following up of non-respondents.

The composite response rates (LFS plus call-back response and LFS plus basic-question response) were 76% and 77%, while for the LFS alone it was 62%. Hence, both approaches resulted in an substantial in crease in response. The interesting question is whether these approaches succeeded in also improving the composition of the response.

Response probabilities were estimated using a simple logistic regression model using all demographic and socio-economic auxiliary variables at hand. The values of the R-indicator M() are given in the table below.

Response	Rate	M
LFS	62%	0.79
LFS + basic question	76%	0.77
LFS + call-back	77%	0.85

The example shows that for the call-back approach representatvity is improved, while it is worse for the basic question approach. Apparently, it does not always pay (in terms of representativity) to put extra efforts in increasing response rates.

The project

It is the objective of the RISQ Project to develop effective Representativity Indicators. They should exploit the available auxiliary information as much as possible.

Furthermore, their behaviour will explored, both theoretically and in practical situations. It will be demonstrated how such indicators can be implemented and used in a practical survey and register environment. It will also be demonstrated that these Representativity Indicators can be used in several stages of the survey process:

- In the data collection phase to control the survey process in such a way that a representative sample is obtained;
- In the data collection phase to monitor the usefulness of subsequent data collection steps and in particular to target resources for collecting data in areas where divergence from representativity may be large;
- In the processing phase, to obtain more information about possible auxiliary variables for use in adjustment weighting;
- In the analysis phase, to investigate and compare the representativity of survey data sets.