The Impact of Nonresponse on Survey Quality

Jelke Bethlehem & Bart Bakker
Statistics Netherlands, The Hague, The Netherlands
E-mail: jbtm@cbs.nl

Abstract

Almost every survey suffers from nonresponse. Nonresponse rates are particularly high for voluntary surveys. The problem of nonresponse is that it affects the representativity of the survey results, and therefore causes estimates to be biased. Theoretically, it is possible to correct these estimates, but this requires sufficient auxiliary information. Unfortunately, such information is not always available. This papers discusses a number of issues and developments.

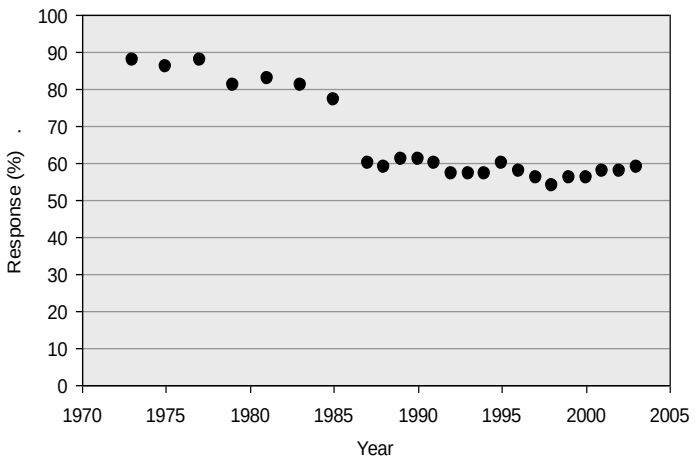
Keywords: Bias, Representativity, Response probability, R-indicator

1. The nonresponse problem

A survey is an important data collection instrument for official statistics. If the basic scientific principles of probability sampling are applied, and no other problems are encountered during the fieldwork, accurate estimates of population characteristics can be computed. These scientific principles mean that (1) a sample always has to be selected by means of probability sampling, (2) each element in the target population of the survey must have a positive (non-zero) probability of selection, and (3) the selection probabilities must be know for the responding elements. This was already described in the seminal paper by Horvitz & Thompson (1952). They show that under these three conditions always unbiased estimates can be computed, and also that the precision of the estimates can be determined.

Researchers should apply these principles in practice. There are, however, always practical problems when collecting data. One of these problems is nonresponse. It is the phenomenon that elements in the selected sample do not provide the requested information, or that the provided information is not usable. Nonresponse can have a negative impact on the quality of the survey. Particularly, for voluntary surveys, nonresponse rates can be high.

Figure 1. Response rate of the Dutch Labour Force Survey.



Nonresponse can have various causes. Usually, three causes are distinguished:

- *Non-contact*. Sample persons are not at home when a contact attempt is made.

- *Refusal*. Sampled persons refuse to participate, e.g. because they are not interested, they consider the survey as an intrusion of their privacy, or because they have no time.
- *Not-able*. An interview is not possible due to illness, mental or physical handicap, or language problems.

Refusal is usually the largest cause of nonresponse. If the survey is not mandatory, many people will simply refuse. They know there is no penalty for not participating. Response rates are low in The Netherlands. Figure 1 shows the trend of dropping response rates for the Labour Force Survey since the 1970's. The response rate is now around 60%, and major efforts are required to prevent it from dropping even more. Nonresponse problems are also becoming more severe in other countries.

The basic problem of nonresponse is not only that it reduces the amount of data that becomes available, but more importantly, that estimates of population characteristics may be biased, and therefore wrong conclusions are drawn from the survey. This happens if the nonresponse is *selective*, i.e. specific groups in the population are over-represented in the survey response, and other groups are under-represented.

2. The effect of nonresponse

There are various ways to describe the effects of nonresponse on estimators, see e.g. Bethlehem, Cobben & Schouten (2011). One way to do this is to use the Random Response Model. This model is based on the concept of the *response probability*. It assumes every element k in the population to have a certain, unknown probability ρ_k of response when invited to participate in a survey.

Assuming the objective of the survey is to estimate the population mean of a variable Y, and a simple random sample has been selected to do so, the response mean is not an unbiased estimator. The bias is approximately equal to

$$B(\bar{y}_R) = \frac{R_{Y\rho} S_Y S_\rho}{\bar{\rho}} \tag{2.1}$$

This is an important expression as it gives insight in the factors determining the magnitude of the bias. A first factor contributing to the bias is the correlation $R_{Y\rho}$ between the values of the target variable of the survey and the response probabilities. There are ample examples of surveys where such a relationship exists. One example is the Dutch Labour Force Survey, where unemployed have a lower response probability than the employed. The stronger the correlation is, the larger the bias will be. Unfortunately, this quantity cannot be computed in practice, since the values of Y are unknown for the nonrespondents.

A second factor contributing to the bias is the average response probability $\bar{\rho}$. This quantity can be estimated unbiasedly by the response rate of the survey. A low response rate will lead to a large bias. This shows the importance of high response rates. Generally, interviewer-assisted surveys (such as CAPI and CATI) have higher response rates than self-administered surveys (web and mail).

A third factor contributing to the bias is the standard deviation S_ρ of the response probabilities. The more the response probabilities vary in magnitude, the larger the bias will be. This is not surprising as groups with small response probabilities are under-represented in the survey response, and groups with large response probabilities are over-represented, thereby affecting the representativity of the response.

The response probabilities are unknown in practice. This makes it impossible to compute the bias in (2.1). However, it is possible to compute the worst case. The

upper bound for the bias is equal to

$$|B(\bar{y}_R)| \leq B_{MAX} = S_Y \sqrt{\frac{1}{\bar{\rho}} - 1} \tag{2.2}$$

It shows there is a larger bandwidth for the nonresponse bias as the response rate decreases.

3. Nonresponse and representativity

The quality of the survey response is partly determined by the response rate and partly by the variation of the response probabilities. The response rate is available, and often used as a quality indicator of the survey response. What is needed is an indicator for the amount of variation of the response probabilities.

The response probability is a theoretical concept that cannot be observed. What can be observed is the value of the indicator R_k, which assumes the value 1 if element k responds (with probability ρ_k), and otherwise assumes the value 0 (with probability 1-ρ_k). The idea is to estimate the response probabilities using the available information. Schouten, Cobben & Bethlehem (2009) propose an indicator for representativity which they call the *R-indicator*. It is defined by

$$R = 1 - 2S_\rho, \tag{3.1}$$

where S_ρ is the standard deviation of the (estimated) response probabilities. R is equal to 1 if all response probabilities are equal. This is the case of complete representativity. The smaller the value of R, the larger the lack of representativity.

An indicator like the R-indicator adds something extra. There are ample examples of survey situations in which increasing the response rate also increases the lack of representativity, and therefore does not help to reduce the nonresponse bias. Schouten, Cobben & Bethlehem (2009) describe a test with the Labour Force Survey in which nonrespondents were re-approached, see table 3.1. They were asked to just answer a few basic questions. With this Basic Question Approach, the response rate could be increased from 62% to 76%. However, the value of the R-indicator dropped from 0.80 to 0.78. So, more was not better. The composition of the response deteriorated.

Table 3.1. Response rate and R-indicator in the Labour Force Survey

Response	Response rate	R-indicator
Main	62.2%	80.1%
Main + Basic Question Approach	75.6%	78.0%

We can conclude that all kinds of attempts to increase the response rate are only meaningful if also the representativity of the response is improved. A focus on 'low hanging fruit' (i.e. people with high response probabilities) is wrong. One should concentrate on getting response from people with low response probabilities. Such an approach will increase their response probabilities, and consequently reduce the variation of the response probabilities.

4. Estimating response probabilities

The concept of response probability is a theoretical one, but also an attractive one, as one has an intuitive feeling for them. For conducting a nonresponse analysis, response probabilities have to be estimated. This is usually done with a logit model. The advantage of this model is that estimated probabilities are always between 0 and 1.

It is assumed that all relevant covariates (required to explain the values of the response probabilities) are included in the model. A drawback of this modeling approach is that the individual values of the covariates in the model are required for both respondents and nonrespondents. Often such information is not available. Sometimes, it can be retrieved from the sampling frame. For example, if the sampling frame is a population register, demographic variables like gender, age, marital status, ethnic background and region of residence can be used. The question may be raised, however, if these variables are sufficient for explaining response behaviour.

Statistics Netherlands has, fortunately, access to many more covariates. The Social Statistics Database (SSD) is an integrated system of social statistics. It contains a wide range of information on every person in The Netherlands. Among the variables in the database are variables on demography, geography, income, labour, health and social protection. See Bethlehem, Cobben & Schouten (2011) for more background information.

An analysis of the estimated response probabilities can provide insight in the nonresponse mechanism. Bethlehem & Schouten (2011) estimated the response probabilities for one of the surveys of Statistics Netherlands. The covariates in the logit model were gender, age, marital status, ethnic background, size of household, type of household, listed phone number, employment status, region, and degree of urbanization. Figure 4.1. shows the distribution of these probabilities. It is clear that they have a substantial variation. So there is a risk of biased estimates.

Figure 4.1. Distribution of the estimated response probabilities

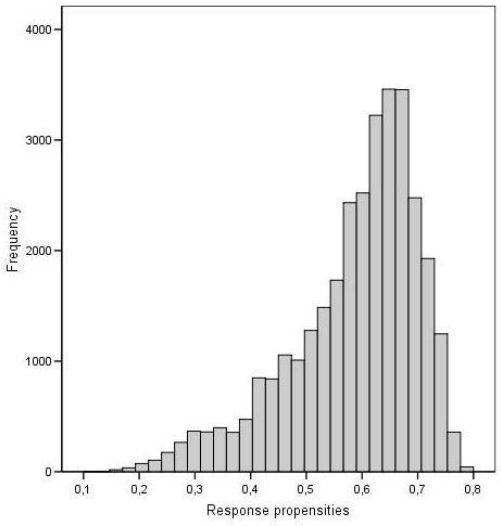
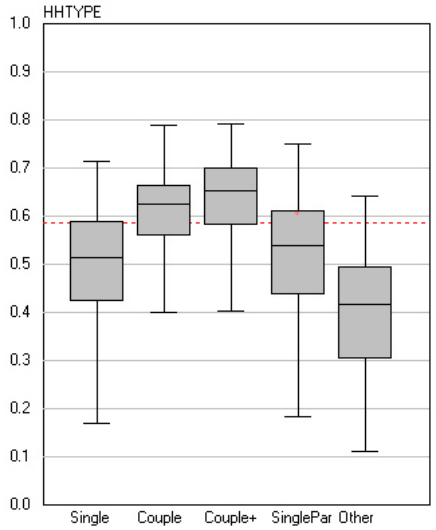


Figure 4.2. Relation between response probabilities and type of household



Further insight can be obtained by looking at the distribution of response probabilities within the categories of auxiliary variables. Figure 4.2 shows an example where box plots were made for various types of household. The graph shows, for example, that single persons have lower response probabilities than couples with children (denoted by Couple+). See also Bethlehem (2012).

In the ideal situation, there is no variation within the categories, and all variation is between categories. In this case, the auxiliary variable is able to explain response behaviour completely. Unfortunately, there is both within and between variation in figure 4.2. This means that the variable is able to explain at least some response behaviour, but not all of it.

5. Reducing the nonresponse bias

Nonresponse is becoming more and more a serious problem in survey research. It is not easy, if not impossible, to avoid nonresponse from happening in the field. So it is always necessary to apply some kind of correction technique. Most correction techniques apply a form of adjustment weighting. The basic idea of weighting is to assign adjustment weights to respondents that correct for under- or over-representation of specific groups. Bethlehem, Cobben & Schouten (2011) describe various weighting techniques, like post-stratification, generalized regression estimation, raking ratio estimation, and weighting based on response probabilities.

All these correction techniques can only be applied effectively if proper auxiliary variables are available. These are variables that have been measured in the survey, and for which the distribution in the population (or complete sample) is available. Preferably, the individual values of these variables for the nonrespondents are available.

Not every auxiliary variable is effective in a weighting procedure. Useful auxiliary variables must satisfy two conditions:

- It has explanatory power as a covariate in the logit model for estimating the response probabilities;

- It has explanatory power as a covariate in a model explaining the behaviour of the target variables of the survey.

It is important to realize at the design stage of a survey, that auxiliary variables will be needed for obtaining acceptable estimates of population characteristics, and therefore they must be measured in the survey.

Nonresponse bias correction will only be effective if all relevant auxiliary variables are used in the weighting model. Often only some demographic variables are available. Sometimes, socio-economic variables can be retrieved from administrative registers. One can also think about using paradata, i.e. data about the survey process and observations made by interviewers. Unfortunately, practical experience shows that the available auxiliary variables are often not sufficient to remove the bias completely.

There is some anecdotal evidence that the choice of auxiliary variables is more important than the choice of the specific correction technique used (post-stratification, generalized regression estimation, raking ratio estimation, or response probability weighting). As long as the right variables are included, the bias will be reduced.

Traditional surveys are designed such that each sample person receives the same treatment. The researcher is not in control of the response probabilities. Their values may vary considerably as is shown in figure 4.1. This contributes to the nonresponse bias. Increasing the sample size does not help as the response probabilities remain unchanged. Another approach could be to re-approach the nonrespondents. Then there is the risk that interviewers go for the ‘low hanging fruit’: they approach nonrespondents with (in their view) the highest probability of success. This approach increases the response probabilities of those with high response probabilities, but does nothing about those with low response probabilities. A more effective approach would be to concentrate on persons with low response probabilities.

This calls for a new approach of survey design in which not every selected person obtains the same treatment. This is the field of *adaptive survey design*, see Wagner (2008). Adaptive survey design assumes that different people may receive different treatment. Treatments are defined before the survey starts. It is possible that treatments are adapted during the fieldwork as new information is collected. The treatment

assigned to a person may depend on information that is already available at the start of the fieldwork. Such information may be found in the sampling frame (e.g. demographic variables) or in administrative sources that can be linked to the sample. This is not always possible. Initial treatment may also be adapted using the paradata that is collected during the fieldwork. Such information may include observations by interviewers.

The idea is to estimate response probabilities using the available auxiliary variables, and to give special treatment to those with low response probabilities. The effect of the treatment should be such that the low response probabilities are increased. This reduces the variation of the response probabilities, and therefore also the nonresponse bias.

It may not be so easy to implement adaptive survey designs in practice. The auxiliary information needed to estimate response probabilities may not be available or is insufficient. Moreover, it also complicates the organization of the fieldwork, as it may lead to unanticipated changes in the middle of the data collection process.

6. Conclusion

Notwithstanding all efforts made in the design stage of a survey, researchers may encounter serious problems collecting data in the field. Nonresponse is one such problem. It may affect the quality of the outcomes of the survey in a serious way. As it is very difficult, if not impossible, to reduce nonresponse in the field, efforts should concentrate on reducing the nonresponse bias by applying some correction technique. This will only be successful if sufficient, relevant auxiliary information is available.

More and more survey organizations are confronted with budget cuts and pressure to reduce survey costs. This causes a shift from interviewer-assisted surveys (CAPI and CATI) to self-administered surveys, in particular web surveys. Experience shows that response rates for web surveys based on probability sampling are low, typically not more than 40%. Moreover, the different causes of nonresponse cannot be distinguished anymore. This will make nonresponse bias correction even more difficult in the future.

References

Bethlehem, J.G (2012), *Using Response Probabilities for Assessing Representativity*. Discussion Paper 201212, Statistics Netherlands, The Hague/Heerlen, The Netherlands.

Bethlehem, J.G., Cobben, F. & Schouten, B. (2011), *Handbook of Nonresponse in Household Surveys*. John Wiley & Sons, Hoboken, NJ.

Bethlehem, J.G. & Schouten, B. (2011), *Nonresponse Analysis with the R-indicator*. Technical Report, Statistics Netherlands, Division of Methodology and Quality.

Horvitz, D. G. & Thompson, D. J. (1952), *A generalization of sampling without replacement from a finite universe*. Journal of the American Statistical Association, 47, pp. 663–685.

Schouten, B., Cobben, F. & Bethlehem, J.G. (2009), Measures for the Representativeness of Survey Response. *Survey Methodology* 36, pp. 101-113.

Wagner, J. (2008), *Adaptive Survey Design to Reduce Nonresponse Bias*. PhD Thesis, University of Michigan, Ann Arbor, MI