Statistics
Netherlands

Discussion Paper

Solving the nonresponse problem with sample matching?

The views expressed in this paper are those of the author(s) and do not necesarily reflect the policies of Statistics Netherlands

2014 | 04

Jelke Bethlehem

14-02-2014

Solving the nonresponse problem with sample matching?

Jelke Bethlehem

Summary: There are various ways of selecting a sample for a survey, but over the years it has become clear that a scientifically sound way to do this is by means of a probability sampling. With increasing nonresponse rates, one may wonder, however, to what extent this paradigm can still be applied. Indeed, due to nonresponse, probability sampling surveys more and more resemble self-selection surveys. As a possible solution Douglas Rivers has proposed a different way of data collection, which is a form of sample matching. The idea is to select a probability sample from a sampling frame. The selected people are not asked to complete a questionnaire form, as this would lead to high nonresponse rates. Instead, people are located in a large (possibly not representative) web panel who resemble the selected persons. These panel members are invited to complete the questionnaire. The response rate will be high as the panel contains people who agreed to complete survey questionnaires regularly. This paper investigates the properties of sample matching in some more detail. Using simulated data, it is explored under which conditions sample matching may work.

Keywords: Sample matching, nonresponse, probability sampling, selfselection panel

Index

1. Introduction

1.1 Survey sampling

There are various ways of selecting a sample for a survey, but over the years it has become clear that a scientifically sound way to do this is by means of probability sampling. All elements (persons, households, businesses) in the target population must have a non-zero probability of selection. Moreover, all selection probabilities must be known.

Horvitz & Thompson (1952) show in their seminal paper probability sampling allows for computing accurate and unbiased estimates of population characteristics. And it is also possible to quantify the accuracy of these estimates, for example by means of a confidence interval. Neyman (1934) showed already long ago that other means of sampling (i.e. purposive sampling) may produce invalid estimates.

Survey methodology has been developed over a period of more than 100 years, starting with the work of Kiaer (1895). The paradigm of probability sampling has shown to work well in social research, official statistics and market research. It has allowed researchers to produce well-founded and reliable survey results.

There are developments, however, that make application of the probability sampling paradigm not so straightforward any more. The first development is increasing nonresponse rates. Selection in the sample not only depends on the known selection probabilities of the sample design, but also on unknown response probabilities. Some response probabilities are close to zero or may even equal zero. This affects the representativity of the survey response, and hence may lead to biased estimates of population characteristics.

Another development is the increased popularity of self-selection web surveys. These are surveys which are just put on the web. There is no sampling design. Everybody can complete the survey questionnaire. This comes down to a form of sampling in which all selection probabilities are unknown. The theory of Horvitz & Thompson cannot be applied. Therefore it is impossible to construct unbiased estimators.

This paper explores the use of matched samples as an alternative for estimators based on surveys suffering from a substantial amount of nonresponse.

1.2 Survey challenges

Nonresponse means that no information is obtained from a number of elements in the sample. The questionnaire forms remain completely empty. Nonresponse is often modelled by assigning an (unknown) probability of response to every element in the population. As a consequence, the probability of getting data from specific elements depends on two random processes: the (known) sample selection mechanism and the (unknown) response mechanism.

Since the probability of getting an observation is unknown, the theory of Horvitz & Thompson cannot be applied. Hence, it becomes impossible to compute unbiased estimates. Nonresponse rates have increased in many countries over time and this increases the risk of drawing wrong conclusions from a survey.

One can attempt to remove or reduce the bias due to nonresponse by applying some kind of weighting adjustment technique, see e.g. Bethlehem, Cobben & Schouten (2011). Another correction technique is to estimate the response probabilities and include them in the Horvitz-Thompson estimation procedure. There is, however, no guarantee that these correction techniques will fully remove the bias caused by nonresponse.

Another development in survey methodology is the emergence of web surveys. This type of survey quickly became very popular, particularly in the world of market research. This is not surprising, as a web survey is a simple means to get access to a large group of people.

Questionnaires of self-selection web surveys can be distributed at very low costs. No interviewers are needed, and there are no mailing and printing costs. Surveys can be launched very quickly. Little time is lost between the moment the questionnaire is ready and the start of the fieldwork. And online surveys offer new, attractive possibilities, such as the use of multimedia (audio, pictures, animation and video).

A serious problem of web surveys is often the lack of a proper sampling frame. Such frames may exist for a survey among employees of a large company or among students of a university, but there is no list of email-addresses of all people in the general population. There is also not something like Random Digit Dialling (RDD) which may be applied for telephone surveys. These frame problems cause many web surveys to be based on self-selection instead of probability sampling. Self-selection means that it is completely left to people to decide to participate in a survey. Respondents are those people who happen to have Internet, visit the website and decide to participate in the survey. Self-selection can be modelled by assigning every element in the target population a certain unknown participation probability. Since these probabilities are unknown, it is not possible to construct unbiased estimators for population characteristics. Self-selection web surveys may suffer from more problems: people who participate more than once, respondents from outside the target population, and groups of people who together attempt to manipulate the outcomes of the survey. For these reasons, self-selection surveys are considered out of the question for compiling accurate statistics about the general population by many survey researchers. Indeed, a special task force of AAPOR (American Association of for Public Opinion Research) concluded that "Researchers should avoid nonprobability online panels when one of the researcher objectives is to accurately estimate population values", see Baker et al. (2010).

1.3 Matched samples

With increasing nonresponse rates, one may wonder, however, to what extent this paradigm can still be applied. Due to nonresponse, probability sampling surveys more and more resemble self-selection surveys. Rivers (2007) states that many telephone polls in the United States have response rates of no more than 20 percent. And web surveys based on probability sampling have in many countries a nonresponse rate of at most 40%. This raises the fundamental question whether valid statistical inference is possible in such situations.

As a possible solution for the increasing nonresponse problems in probability-based surveys, Douglas Rivers has proposed a different way of data collection, which is a form of sample matching. His approach is described in, for example, Rivers (2007), Vavreck & Rivers (2008), and Rivers & Bailey (2009). Application of sample matching requires two ingredients:

- A *sampling frame* that covers the target population of the survey. A probability sample is selected from this frame. The sampling frame must contain a set of auxiliary variables, or it must be possible to uniquely link another data source of auxiliary variables (e.g. a register) to this frame.
- A *large panel*, also containing the set of auxiliary variables. There are no conditions imposed on recruitment for this panel. It may even be a self-selection panel.

The idea is to select a probability sample from the sampling frame. The selected people are not asked to complete a questionnaire form, as this would lead to high nonresponse rates. Instead, people are located in the panel who resemble the selected persons. These panel members are invited to complete the questionnaire. The response rate will be high as the panel contains people who agreed to complete survey questionnaires regularly.

The main issue of sample matching is to find people in the panel who resemble the selected people in the sampling frame as much as possible. The auxiliary variable variables are used for this.

Rivers tested his approach in the Cooperative Congressional Election Study (CCES). He concluded that sample matching is capable of removing nonresponse bias in this survey. He also concluded that his estimates were less biased than those of surveys based on random digit dialling recruitment.

This paper investigates the properties of sample matching in some more detail. Using simulated data, it is explored under which conditions sample matching may work. Chapter 2 presents the mathematical framework. It shows what the effects of nonresponse are on the bias of estimates. It also describes how self-selection surveys produce biased estimates. Chapter 3 is about the basics of sample matching. It stresses the important role of auxiliary variables in matching persons in the sampling frame to persons in the panel. Chapter 4 gives an account of a small simulation study. It shows that sample matching is no guarantee for success. Chapter 5 compares sample matching with some more traditional sampling approaches: selecting a stratified sample from the panel and selecting a simple random sample from the panel followed by post-stratification. The final chapter 6 discusses the advantages and disadvantages of the various approaches.

2. Basic concepts

2.1 Sampling from a finite population

Let the finite *target population U* of the survey consist of a set of N identifiable elements, which are labelled 1, 2, …, N. Associated with each element k is an unknown value Y_k of the *target variable*. There are many population characteristics that may be of interest to a researcher, but here it is assumed that objective of the sample survey is estimation of the population mean

$$\bar{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k \quad . \tag{2.1.1}$$

To estimate this population parameter, a probability sample of size n is selected without replacement. The sample can be represented by a set a_1, a_2, …, a_N of indicators. The k-th indicator a_k assumes the value 1 if element k is selected in the sample, and otherwise it assumes the value 0. The *first-order inclusion probability* of element k is defined by $\pi_k = E(a_k)$, for $k = 1, 2, …, N$.

The *Horvitz-Thompson estimator* is defined by

$$\bar{y}_{HT} = \frac{1}{N}\sum_{k=1}^{N} a_k \frac{Y_k}{\pi_k} \quad . \tag{2.1.2}$$

It can be shown that this is an unbiased estimator of the population mean (because $E(a_k) = \pi_k$). In the case of simple random sampling all first-order inclusion probabilities are equal: $\pi_k = n/N$. Then the Horvitz-Thompson estimator reduces to the sample mean

$$\bar{y}_{HT} = \frac{1}{n}\sum_{k=1}^{N} a_k Y_k \quad . \tag{2.1.3}$$

2.2 Nonresponse in a probability sample

Suppose there is nonresponse in the survey. It is assumed that each element k in the population has a certain, unknown probability ρ_k of response. If element k is selected in the sample, a random mechanism is activated that results with probability ρ_k in response and with probability 1 - ρ_k in nonresponse. Under this model, a set of response indicators R_1, R_2, …, R_N can be introduced, where $R_k = 1$ if the corresponding element k responds, and where $R_k = 0$ otherwise. So, $P(R_k = 1) = \rho_k$, and $P(R_k = 0) = 1 - \rho_k$.

The survey response only consists of those elements k for which $a_k = 1$ and $R_k = 1$. Hence, the number of available cases is equal to

$$n_R = \sum_{k=1}^{N} a_k R_k \quad , \tag{2.2.1}$$

Note that this realized sample size is a random variable. Likewise, the number of non-respondents is equal to

$$n_{NR} = \sum_{k=1}^{N} a_k (1 - R_k), \tag{2.2.2}$$

where $n = n_R + n_{NR}$. The values of the target variable only become available for the n_R responding elements. The mean of these values is denoted by

$$\bar{y}_R = \frac{1}{n_R}\sum_{k=1}^{N} a_k R_k Y_k \ . \qquad (2.2.3)$$

Bethlehem (2009) shows that the expected value of the response mean is approximately equal to

$$E(\bar{y}_R) \approx \frac{1}{N}\sum_{k=1}^{N} \frac{\rho_k}{\bar{\rho}} Y_k \ , \qquad (2.2.4)$$

where

$$\bar{\rho} = \frac{1}{N}\sum_{k=1}^{N} \rho_k \qquad (2.2.5)$$

is the mean of all response probabilities in the population. Expression (2.2.4) shows that, generally, the expected value of the response mean is unequal to the population mean to be estimated. Therefore, this estimator is biased. Bethlehem, Cobben & Schouten (2011) show that this bias is approximately equal to

$$B(\bar{y}_R) = E(\bar{y}_R) - \bar{Y} \approx \frac{R_{\rho Y} S_{\rho} S_Y}{\bar{\rho}} \ , \qquad (2.2.6)$$

where $R_{\rho Y}$ is the correlation between the response probabilities and the values of the target variable, S_Y is the standard deviation of the variable Y, and S_ρ is the standard deviation of the response probabilities. From this expression of the bias a number of conclusions can be drawn:

- The bias is large if there is a strong correlation between the target variable of the survey and the response behaviour. There is no bias if there is no correlation.
- The bias is large if the variation of the values of the response probabilities is large. There is no bias if all response probabilities are equal.
- The bias is large if the response probabilities are small.

It is clear that the validity of the survey outcomes depends on both the response rate of the survey and the composition of the response. A small response rate combined with a response lacking representativity may easily lead to wrong conclusion due to biased estimates.

2.3 A self-selection sample

Self-selection means that the researcher is not in control of the sample selection process. He just makes the survey questionnaire available, and waits and sees what happens. A typical example is a web survey, where everyone can complete the questionnaire on the internet. Also people outside the target population of the survey can do it. It is sometimes even possible to fill in the questionnaire more than once.

Participation in a self-selection web survey requires that respondents are aware of the existence of the survey. Moreover, they must have access to the Internet, they have to visit the website (for example by following up a banner, an e-mail message, or a commercial on radio or TV), and they have to decide to fill in the questionnaire This means that each element k in the population has unknown probability ρ_k of participating in the survey, for $k = 1, 2, ..., N$.

It is assumed in this section that there are no under-coverage problems. So in principle everyone has a nonzero probability of participating in the survey. The responding elements are denoted by a set of indicators $R_1, R_2, ..., R_N$, where the k-th indicator R_k assumes the value 1 if element k participates, and otherwise it assumes the value 0, for k = 1, 2, ..., N. The expected value $\rho_k = E(R_k)$ is the participation probability of element k. The realized sample size is denoted by

$$n_S = \sum_{k=1}^{N} R_k \ . \tag{2.3.1}$$

Lacking any knowledge about the values of the response probabilities, a naive researcher would implicitly assume all these probabilities to be equal. In other words: simple random sampling is assumed. Consequently, the sample mean

$$\bar{y}_S = \frac{1}{n_S}\sum_{k=1}^{N} R_k Y_k \tag{2.3.2}$$

is used as an estimator for the population mean. Bethlehem (2009) shows that the expected value of this estimator is approximately equal to

$$E(\bar{y}_S) \approx \frac{1}{N}\sum_{k=1}^{N} \frac{\rho_k}{\bar{\rho}} Y_k \tag{2.3.3}$$

where $\bar{\rho}$ is the mean of all response probabilities. Note that this expression is similar to expression (2.2.4). The only difference is that the ρ_k are response probabilities (after sample selection) in expression (2.2.4) and direct participation probabilities in expression (2.3.3). The bias of estimator (2.3.2) is approximately equal to

$$B(\bar{y}_S) = E(\bar{y}_S) - \bar{Y} \approx \frac{R_{\rho Y} S_\rho S_Y}{\bar{\rho}} , \tag{2.3.4}$$

where $R_{\rho Y}$ is the correlation between the participation probabilities and the values of the target variable, S_Y is the standard deviation of the variable Y, and S_ρ is the standard deviation of the participation probabilities. From this expression of the bias a number of conclusions can be drawn:

- The bias is larger if the correlation between the target variable of the survey and the participation behaviour is stronger. There is no bias if there is no correlation.
- The bias is larger if the variation of the values of the participation probabilities is larger. There is no bias if all participation probabilities are equal.
- The bias is larger if the participation probabilities are smaller.

The participation probabilities are low and unknown for self-selection surveys. Therefore, there is a substantial risk of large biases.

2.4 Nonresponse bias and self-selection bias

It is clear from expression (2.2.4) that, generally, the expected value of this sample mean is not equal to the population mean. One situation in which the bias vanishes is that in which all response probabilities are equal. In terms of the theory of missing data, this comes down to Missing Completely At Random (MCAR). This is the situation in which the cause of missing data is completely independent of all variables measured in the survey. For more information on MCAR and other missing data mechanisms, see Little & Rubin (2002).

Note that also for a self-selection survey, the case of MCAR does not lead to an unrepresentative sample because all elements have the same participation probability.

As was mentioned earlier, the expression for the bias in case of nonresponse is similar to that for self-selection surveys. In practical situations however, their values are not the same. For example, the probability samples for surveys of Statistics Netherlands have response rates of around 60%. This means that the average response probability is 0.6. There have been self-selection web surveys in the Netherlands with large samples. An example is *21minuten.nl*. Approximately 170,000 people completed the questionnaire in 2006. Assuming the target population to consist of all Dutch citizens from the age of 18, the average response probability was 170,000 / 12,800,000 = 0.0133. This is a much lower value than the 0.6 of probability sampling based surveys. This means there is a risk of a much larger bias in a self-selection survey.

From expression (2.2.6) or (2.3.4) an upper bound for the bias can be computed. Given the mean response probability $\bar{\rho}$, there is a maximum value the standard deviation S_ρ of the response probabilities cannot exceed:

$$S(\rho) \leq \sqrt{\bar{\rho}(1-\bar{\rho})} \ . \qquad (2.4.1)$$

This implies that in the worst case S_ρ assumes its maximum value if the correlation coefficient $R_{\rho Y}$ is equal to either +1 or -1. Then the absolute value of the bias will be

$$|B_{MAX}| = S_Y \sqrt{\frac{1}{\bar{\rho}} - 1} \quad . \qquad (2.4.2)$$

In case of a survey based on probability sampling with a response rate of around 60%, the maximum absolute bias is therefore equal to 0.816×S_Y. In case of a self-selection survey with a size 170,000 from a population of size 12,800,00, the maximum absolute bias is 8.619×S_Y. This is more than 10 times as large.

Note that in case the target variable is an indicator variable (it only assumes the value 0 and 1), also an upper bound can be computed for S_Y. In this case $S_Y^2 = \bar{Y}(1-\bar{Y}) \leq 1/4$. Hence, some impression of the potential damage of nonresponse can be obtained.

3. Sample matching

3.1 The principle of sample matching

A probability sample has the advantage that the researcher is in control of the sampling mechanism, but it has the disadvantage of possibly high nonresponse rates. A self-selection web panel has the advantage that it contains individuals with high response probabilities (because all members agreed to regularly participate in surveys), but the disadvantage of an unknown selection mechanism. Sample matching as proposed by Rivers (2007) can be applied to combine the positive aspects of both approaches.

As described in section 1, the first ingredient of sample matching is a sampling frame that covers the target population of the survey. Applying the principles of probability sampling, a simple random sample is selected from the sampling frame. The selected individuals are not asked to complete the questionnaire. Therefore, nonresponse problems are avoided.

The second ingredient for sample matching is a large panel. This panel may even be a 'dirty panel', i.e. panel recruitment may have been based on self-selection. What counts is that the panel members have high response rates when asked to participate in a survey.

Sample matching comes down to locating individuals in de panel that look as much as possible like the individuals who have been selected from the sampling frame. These panel members are asked to complete the survey questionnaire. So the selected people from the sampling frame are replaced by similar looking people from the panel. One could also see it as a form of donor imputation: the unknown values of the variables for a sample person are imputed by taking the values of these variables from a similar person in the panel.

To be able to find similar looking persons in the panel, a set of auxiliary variables is required. The values of these variables should be available, both in the sampling frame and in the panel. It is assumed that all auxiliary variables are categorical variables, i.e. their values are labels that divide the population in groups.

Suppose, the frame and the panel have three auxiliary variables in common: age (in 10 age categories), gender (2 categories) and marital status (4 categories). If a married male with age between 30 and 35 years is selected from the sampling frame, then we have to find married male in the same age category in the panel.

With only three variables it will not be difficult to find a similar person in the panel. It is even not unlikely there are many persons within the same age category, with the same gender and marital status. In such a situation a procedure is needed to select a person from the group of similar looking persons. One approach is to select persons at random from the group.

People with the same gender, age and marital status may still differ in many other respects. The group is not so homogeneous. If the groups are more homogeneous, persons will be more similar. Such homogeneous groups can be created by using more auxiliary variables. This is only possible of these variables are available in both the sampling frame and the panel. Moreover, the more variables are used for matching, the larger the panel must be. For example, in The Netherlands one could consider using the variables municipality (400 categories), degree of urbanisation (5 categories), marital status (4 categories), type of household (5 categories), size of household (5 categories), age group (5 categories), level of education (5 categories), and ethnic background (5 categories). All together there are 25,000,000 different combinations (assuming that every combination is possible). To find these different people in the panel, the panel should at least contain 25,000,000 members. This is a very large panel. If the panel is not so large, it is necessary to adapt the matching procedure. One way out could, for example, be to use less variables, but then matched persons will be less similar.

3.2 Estimation

Suppose, there is a target population of size N. A simple random sample of size n is to be selected (without replacement) from this population. Assuming a perfect sampling frame, this frame also consists of N records.

Suppose a set of auxiliary variables is used to match records from the frame with records from the panel. The number of groups obtained by crossing these variables is denoted by L. The number of elements in group h is denoted by N_h, for h = 1,2, ..., L. Hence, $N_1 + N_2 + \ldots + N_L = N$. The number of sample elements in group h is denoted by n_h, for h = 1, 2, ..., L, where $n_1 + n_2 + \ldots + n_L = n$. The n_h are random variables. The expected value of the sample proportion n_h/n is equal to the population proportion N_h/N: $E(n_h/n) = N_h/N$.

For each of the n_h selected persons in group h, corresponding persons have to be found in the panel. Suppose the panel consists of m persons. The set of auxiliary variables can also be used to divide the panel in L groups. Let m_h be the number of persons in group h, with $m = m_1 + m_2 + \ldots m_L$. We now consider three situations.

The first situation is that the group in the panel contains more people than the sample from the group in the frame: $n_h \leq m_h$. Then n_h persons can be found to randomly draw n_h persons without replacement from the m_h persons in the panel.

The second situation is that the group in the panel contains less people than the sample from the group in the frame: $n_h \geq m_h$. Then it is not possible any more to draw a random sample of size n_h without replacement from the group in the panel. One solution is to draw the sample with replacement. Consequently, some persons in the group will be selected more than once. One has to decide what to do with these multiple selected persons. Interview them multiple times? Or copy there completed questionnaire form? Another solution is to use less auxiliary variables for matching, with the risk that less similar persons will be matched.

The third situation is that the group in the panel contains no people at all ($m_h = 0$). Then it is not possible to draw one or more persons. A way out could be to define some kind of distance function based on less variables than the complete set of auxiliary variables. This may come down to drawing persons from a less detailed stratification.

For the sake of convenience it is assumed that the panel is so large that persons can be obtained from the group in the panel by selecting a random sample without replacement (the first situation).

De n_h values of the target variable that are observed in group h of the panel, are denoted by $y_{h1}, y_{h2}, \ldots, y_{hn_h}$. The sample mean is equal to

$$\bar{y}_{SM} = \frac{1}{n}\sum_{h=1}^{L}\sum_{i=1}^{n_h} y_{hi} \tag{3.2.1}$$

Assuming the panel to be a self-selection panel, and applying the theory in section 2.3, the expected value of this quantity is equal to

$$E(\bar{y}_{SM}) = E_F E_P(\bar{y}_{SM} \mid F) = \frac{1}{N}\sum_{h=1}^{L}\frac{1}{\bar{\rho}_h}\sum_{k=1}^{N_h}\rho_{hk}Y_{hk} \ , \tag{3.2.2}$$

where E_F refers to the expectation over the sampling distribution in the sampling frame, and E_P refers to the expectation over the distribution in the panel. Furthermore, ρ_{hk} is the participation probability of person k in group h, and Y_{hk} is the value of the target variable of this person. Finally,

$$\bar{\rho}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} \rho_{hk} \tag{3.2.3}$$

is the mean of the participation probabilities of all people in group h, for $h = 1, 2, \ldots, L$. It is clear from expression (3.2.2) that the estimator is not unbiased. The bias of this estimator is equal to

$$B(\bar{y}_{SM}) = E(\bar{y}_{SM}) - \bar{Y} = \frac{1}{N} \sum_{h=1}^{L} N_h \frac{R_{\rho Y,h} S_{\rho,h} S_{Y,h}}{\bar{\rho}_h}\,. \tag{3.2.4}$$

$R_{\rho Y,h}$ is correlation between the participation probabilities and the values of the target variable in group h, $S_{Y,h}$ is standard deviation of the target variable in group h, and likewise $S_{\rho,h}$ is the standard deviation of the participation probabilities in group h.

The bias in (3.2.4) will vanish if there is no correlation between the participation probabilities and the target variable within the groups: $R_{\rho Y,h} = 0$. This comes down to the Missing at Random (MAR) assumption in the theory of missing data. See, for example, Little & Rubin (2002). It means the set of auxiliary variables used to link the frame to the panel must contain all variables needed to explain the participation behaviour.

The bias in (3.2.4) also vanishes if participation probabilities of all elements in the population are the same. In this case the standard deviation of the participation probabilities is equal to 0: $S_{\rho,h} = 0$. It means that the members of the panel can be seen as a simple random sample. Hence, the panel allows for unbiased estimation.

The bias in (3.2.4) is small if the participation probabilities are large. Unfortunately, this is often not the case in panels. Participation probabilities can be very low.

4. A simulation study

4.1 The population

To investigate the properties of sample matching, a small simulation study was conducted. A target population was created consisting of 100,000 persons. There were two auxiliary variables: age (in 3 categories, young, middle-aged and old) and level of education (in 2 categories, low and high).

The target variable was a dummy variable indicating voting intentions for the next election. The variable could assume two values: 0 (no) and 1 (yes). The population percentage of voters was 45.9%. To create this variable, persons in the population were assigned voting probabilities that ranged from 0.2 for low educated young people to 0.8 for high educated elderly. Voting increased with age and level of education.

Each person in the population was assigned a participation probability for the panel. These probabilities varied between 0.01 and 0.15. Participation probabilities also increased by age and level of education. Table 4.1 contains an overview of voting and participation probabilities.

Using these probabilities, a panel was selected from the population. The resulting size was 7458 persons. Not surprisingly, low educated young people were under-represented and high educated elderly were over-represented.

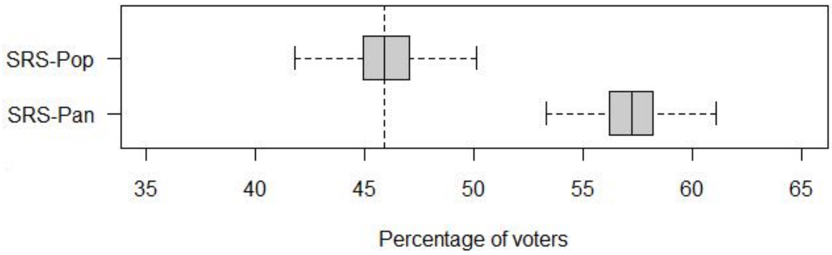
The percentage of voters in the panel is 57.2%, which is higher than the percentage of voters in the population (45.9%). The bias is 11.3%. Since the complete population is known, the components of bias expression (2.3.4) can be computed. The correlation coefficient between the target variable (voting intention) and participation behaviour is equal to $R_{\rho Y}$ = 0.368. So there is correlation. The standard deviation of the participation probabilities is equal to S_ρ = 0.0461 and the standard deviation of the target variable is S_Y = 0.498. Finally, the mean participation probability is $\bar{\rho} =$ 0.0745. Substitution of these quantities in expression (2.3.4) results in a bias of 0.113, which corresponds to 11.3%.

Table 4.1. Voting probabilities and participation probabilities in the generated population

Voting probabilities			Participation probabilities		
Age	Education		Age	Education	
	Low	High		Low	High
Young	0.20	0.40	Young	0.01	0.09
Middle	0.40	0.60	Middle	0.05	0.12
Old	0.60	0.80	Old	0.09	0.15

To explore the behaviour of various estimators, samples of size 1,000 were drawn from this population. This was repeated 10,000 times for each estimator. The resulting distributions of the estimators are displayed by means of box plots. Figure 4.1 compares random samples from the target population with random samples from the panel.

Figure 4.1. Random samples from the population and from the panel

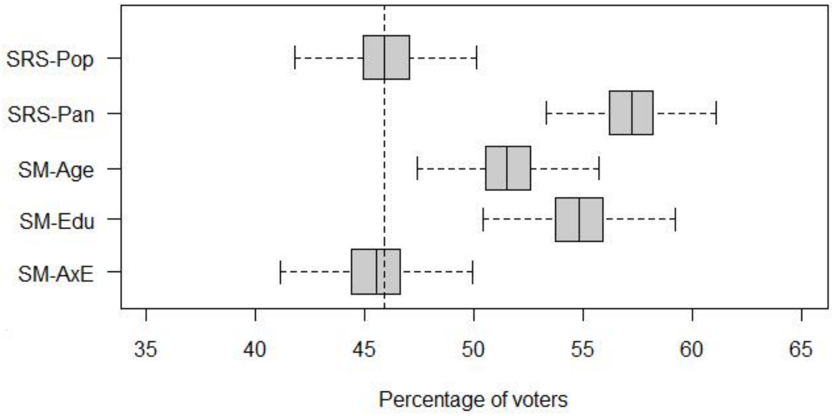


The distribution of the sample mean for random samples from the population is displayed in the upper part of the graph (SRS-Pop). The box plot shows there is a symmetrical distribution around the value to be estimated (45.9%, represented by the vertical dashed line). This indicates an unbiased estimator. The distribution of the sample mean for samples from the panel is displayed in the lower part of the graph (SRS-Pan). The whole distribution has shifted to the right. The values of the estimator are systematically too high. Their average is 57.2%. This indicates a biased estimator. This is caused by the fact that the panel is not representative for the population.

Now the effect of sample matching is explored. Each sample from the frame is matched to the panel, and the sample average is computed for the linked persons. Three situations are considered. In the first one, only the auxiliary variable age (in 3 categories) is used for matching. In the second one, only the variable education (in two categories) is used. And in the third one, both age and educated are used for matching.

The set of matching variables only takes a limited amount of values. So, it will not be a problem to find matching persons in the panel. In case of more candidates, a random person is taken. The distributions of the resulting estimators are shown in figure 4.2.

Figure 4.2. The distribution of matched sample estimators



The distribution for the matched sample estimator based on just the variable age is denoted by SM-Age. The estimator is biased, but the bias is smaller than that of the sample mean of samples from the panel (SRS-Pan). Still, all possible values of the estimator are significantly too high. The expected value of the estimator is 51.5% which means a bias of 51.5% – 45.9% = 5.6%.

The distribution for the matched sample estimator based on just the variable education is denoted by SM-Edu. This estimator performs even worse than that of SM-Age. The bias of the SM-Edu estimator is smaller than that of samples from the panel, but it is still very large. The expected value of the SM-Edu estimator is equal to 54.8%, which comes down to a bias of 54.8% - 45.9% = 8.9%.

The bias vanishes when both variables age and education are used for sample matching. The distribution of the resulting estimator is denoted by SM-AxE. It is not surprising that the bias vanishes, because all variables used for sample matching also explain participation behaviour for the panel. If, for example, just the variable age would have been used for matching, there still are differences between young people in the panel and outside the panel. High educated young people are over-represented in the panel. Therefore their voting probability is also higher. If both variables age and education are used for sample matching, there are no differences any more between members of groups inside and outside the panel. They have the same voting probabilities.

The results of the simulations are summarized in table 4.2. The conclusion can be drawn that sample matching will not always remove the bias of estimators. This depends on the set of variables that is used for matching the sampling frame and the panel. Sample matching is only successful if all explanatory variables are used that are required to explain the participation behaviour of the panel members. If not all relevant variables are used, only part of the bias will be removed.

Table 4.2. Summary of the simulation results
(10,000 simulations of samples of size 1,000)

Estimation	Expected value	Bias	Standard error
Simple random sample from the population	45.93	-0.01	1.57
Simple random sample from the panel	57.21	11.27	1.43
Sample matching using age	51.52	5.58	1.55
Sample matching using education	54.79	8.85	1.58
Sample matching using age and education	45.50	-0.44	1.58

5. Other estimation approaches

5.1 Stratified sampling

The matched sample estimator as defined by expression (3.2.1) can also be written as

$$\bar{y}_{SM} = \sum_{h=1}^{L} \frac{n_h}{n} \bar{y}_h \,, \tag{5.1.1}$$

where $\bar{y}_h$ is the mean of the n_h observations in group h. The quantity n_h/n is an unbiased estimator of the quantity N_h/N. Therefore, a different estimator can be obtained by replacing n_h/n by N_h/N in expression (5.1.1). This results in

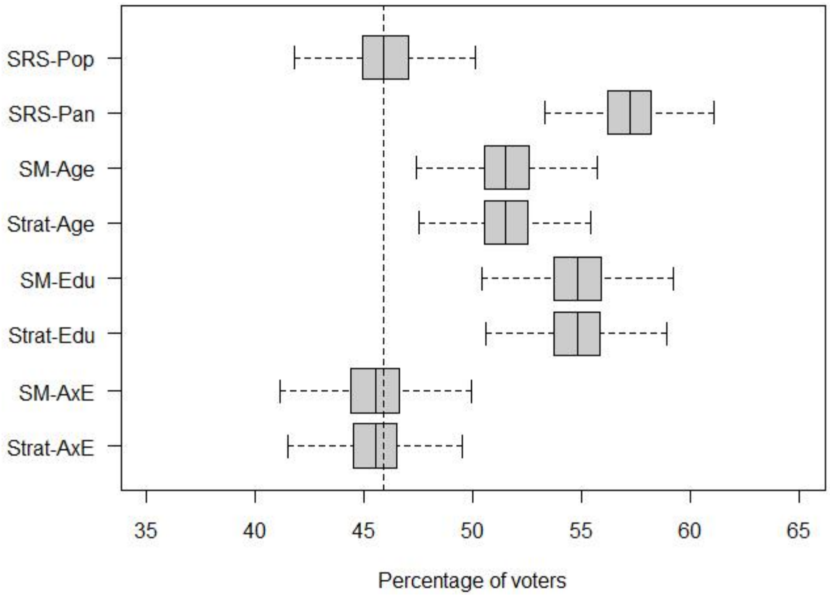
$$\bar{y}_{ST} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h \,. \tag{5.1.2}$$

This is the estimator that would have been obtained for stratified sampling from the frame with proportional allocation, i.e. the sample sizes n_h in the strata are proportional to the sizes N_h of the strata in the population. In practice this would mean using the sampling frame to determine the stratum sizes, and then selecting a stratified sample from the panel.

To explore the behaviour of estimator (5.1.2), the simulation study in section 4 was continued. Stratified samples of size 1,000 were drawn from the panel. This was repeated 10,000 times. The resulting distributions of the estimators are displayed by means of box plots. Figure 5.1 compares the distributions of the three matched sample estimators with the corresponding stratification estimators.

The performance of the stratification estimator with respect to bias reduction is the same as that of the matched sample estimator. The bias reduction of the sample matching using the variable age (denoted by SM-Age) is of the same magnitude as the bias reduction of stratification by age (denoted by Strat-Age). Similarly, the bias reduction of SM-Edu is the same as that of Strat-Edu. Finally, both SM-AxE and Strat-AxE are capable of removing the bias completely.

Figure 5.1. The distribution of stratification estimators



The results of the simulation are summarized numerically in table 5.1. The table confirms the conclusions based on the graph: The matched sample estimator and the stratification estimator have the same behaviour with respect to reducing the bias.

Table 5.1. Comparing sample matching with stratification (10,000 simulations of samples of size 1,000)

Estimation	Expected value	Bias	Standard error
Sample matching using age	51.52	5.58	1.55
Stratification using age	51.51	5.57	1.50
Sample matching using education	54.79	8.85	1.58
Stratification using education	54.79	8.85	1.57
Sample matching using age and education	45.50	-0.44	1.58
Stratification using age and education	45.51	-0.43	1.45

Note that the stratification estimators have slightly smaller standard errors. The reason is that the number of observations per group is fixed. In contrast, the number of observations per group for the matched sample estimators are the result of selecting a random sample from the sampling frame. This introduces an extra source of variation.

5.2 Post-stratification estimation

One of the obvious approaches to obtain less biased estimators from a self-selection panel would be to draw a simple random sample from the panel, followed by a bias reduction attempt by means of post-stratification. So the panel itself is used as a sampling frame and the original sampling frame serves to compute the population distribution of weighting variables. The expression for the post-stratification estimator becomes

$$\bar{y}_{PS} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{L} \frac{N_h}{N} \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \, , \qquad (5.1.3)$$

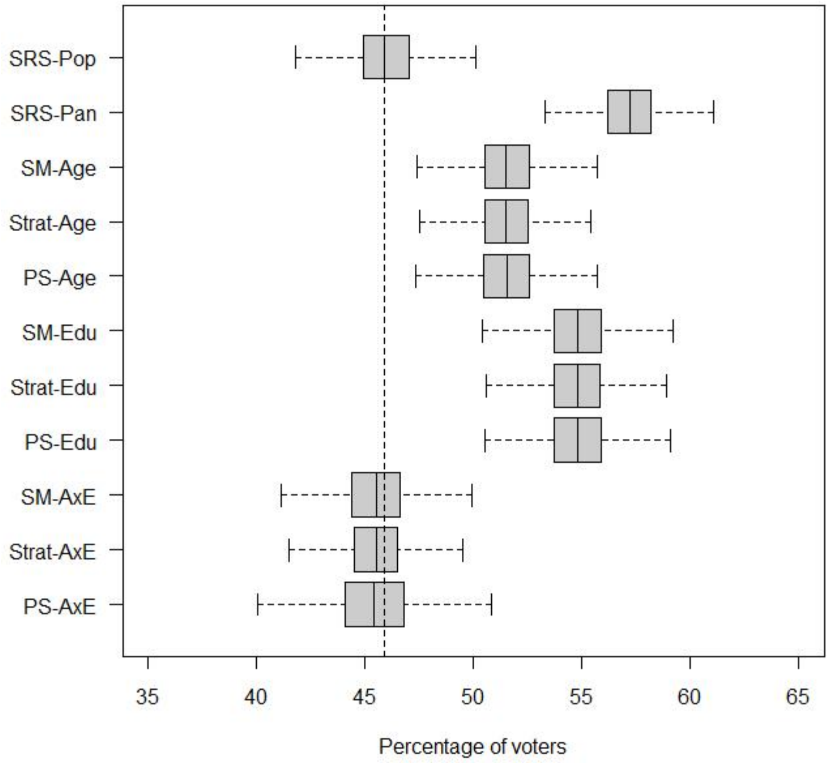
where the values y_{hi} are obtained by means of a simple random sample from the panel. Expression (5.1.3) resembles expression (5.1.2). The difference is, however, that in the case of stratification a stratified sample is selected from the panel and in case of post-stratification, a simple random sample is selected from the panel.

The panel is not representative for the population. Since the sample for post-stratification is selected from this panel, one can expect serious representativity problems. This was already shown in figure 4.1. Therefore, weighting adjustment is badly needed in an attempt to reduce the bias of estimators.

To explore the behaviour of estimator (5.1.3), the simulation study was continued. Samples of size 1,000 were drawn from the panel. This was repeated 10,000 times. The resulting distributions of the post-stratification estimators are displayed by means of box plots. Figure 5.2 compares the distributions of various matched sample estimators, with the corresponding stratification and post-stratification estimators.

The performance of the post-stratification estimator with respect to bias reduction is the same as that of the stratification estimator and the matched sample estimator. The magnitude of the bias reduction does not depend on the type of estimator but on the variables used. There is a limited bias reduction if only the variable education is used. There is more bias reduction for the variable age. And the bias is completely removed when both variables are used.

Figure 5.1. The distribution of matched sample estimators, stratification estimators and post-stratification estimators.



Note that the variance of the post-stratification estimator for age by education (denoted by PS-AxE) is clearly larger than that of the corresponding matched sample estimator (MS-AxE) and stratification estimator (Strat-AxE). This is caused by a large variation in the weights produced by post-stratification. The weights vary from 0.5 (for the high-educated elderly) to 8.2 (for the low-educated young people). A large variation in weights may lead to unstable estimates and an increased variance.

Table 5.2. Comparing sample matching with stratification and post-stratification (10,000 simulations of samples of size 1,000)

Estimation	Expected value	Bias	Standard error
Sample matching using age	51.52	5.58	1.55
Stratification using age	51.51	5.57	1.50
Post-stratification using age	51.53	5.59	1.57
Sample matching using education	54.79	8.85	1.58
Stratification using education	54.79	8.85	1.57
Post-stratification using education	54.81	8.87	1.61
Sample matching using age and education	45.50	-0.44	1.58
Stratification using age and education	45.51	-0.43	1.45
Post-stratification using age and education	45.46	-0.48	2.00

The results of all simulations are summarized in table 5.2. It is clear that the properties of the three types of estimators are more or less the same in this simulation study. There is no reason to think that it would be different in other situations.

The conclusion can be that if a panel (possibly constructed by means of self-selection) is available and the population distributions of the set of relevant auxiliary variables, one could simply apply post-stratification. Sample matching would not produce better estimates in terms of bias. Stratified sampling results in somewhat smaller standard errors than Sample matching.

6. Sample matching with a probability panel

It was assumed in the previous sections that the panel was a ‘dirty panel’. Members were recruited by means of self-selection and therefore had unknown selection probabilities. As a result, estimates were biased and this bias could only be removed by using the proper auxiliary variables.

What if the panel is based on probability sampling? What will be the effect on sample matching? Focusing on web panels, there are several general population web panels based on probability sampling. One example is the LISS Panel (Longitudinal Internet Studies for the Social Sciences). It contains approximately 5,000 households. This panel was set up in 2006 by CentERdata, a research institute in The Netherlands. Objective of the LISS Panel is to provide a laboratory for the development and testing of new, innovative research techniques, while collecting data for the scientific community. The panel is based on a true probability sample of households drawn from the population register of The Netherlands. See also Scherpenzeel (2008). Another example is the Knowledge Panel in the U.S (www.knowledgenetworks.com). The panel consists of approximately 50,000 adult members and 3,000 teenagers (with consent of their parents). This panel exists since 1999. Recruitment used to be based on Random Digit Dialling (RDD), but nowadays an address based sampling frame is used.

Recruiting a sample of people for a web panel requires substantial effort. There is no sampling frame with email addresses . So usually some other mode of data collection is used for recruitment, like mail, CATI, CAPI, or a combination of these modes. This is expensive and time-consuming. See, for example, Cobben & Bethlehem (2013). This may be the reason that probability panels often are not very large. This limits the possibilities of use of such a panel in a sample matching approach.

Another problem of probability-based web panels is that they suffer from nonresponse, and this may affect the representativity of the panel. Particularly in the recruitment phase, nonresponse rates may be high. It is therefore of vital importance to correct estimates by applying some kind of weighting adjustment technique.

The properties of sample matching were studied assuming a perfect probability panel was available. The simulation experiment was continued where samples were selected from a sampling frame and sample persons were matched to persons in the panel. To that end a panel of size 7458 persons was created by randomly selecting persons from the population of size 100.000. The population was the same as the one used for the simulation experiments in sections 4 and 5. Also the size of the probability panel was the same as that of the self-selection panel (7458 persons).

For the three estimation approaches (sample matching, stratified sampling and post-stratification), 10,000 samples of size 1,000 were selected. The properties of the resulting estimators are summarized in table 6.1.

It is clear that all estimators are (almost) unbiased. There is a tiny bias, but this caused by the fact that the mean of the probability panel is not exactly equal to the population mean. This is the margin of error caused by selecting a random sample.

The results in table 6.1 are not very surprising. The panel can already be seen as a simple random sample from the population. So, estimates based on a random sample from the panel will be unbiased. Sampling matching cannot improve these already unbiased estimators.

The situation would be different if the probability panel was affected by selective nonresponse. This would result in unknown selection probabilities. Basically, this would mean going back to the situation of a self-selection panel. Sample matching could be applied here provided the proper auxiliary variables are available for matching.

Table 6.1. Comparing sample matching with stratification and post-stratification for a probability panel (10,000 simulations of samples of size 1,000)

Estimation	Expected value	Bias	Standard error
Simple random sampling from the population	45.93	-0.01	1.57
Simple random sampling from the panel	45.84	-0.10	1.46
Sample matching using age	45.59	-0.35	1.57
Stratification using age	45.59	-0.35	1.49
Post-stratification using age	45.60	-0.34	1.39
Sample matching using education	45.93	-0.01	1.57
Stratification using education	45.92	-0.02	1.54
Post-stratification using education	45.92	-0.02	1.42
Sample matching using age and education	45.69	-0.25	1.57
Stratification using age and education	45.65	-0.29	1.42
Post-stratification using age and education	45.66	-0.28	1.33

7. Conclusions

This paper explored the possibilities of sample matching as a technique to avoid the problems caused by high nonresponse rates in probability surveys. The analysis was restricted to the situation in which on the one hand a sampling frame was available that covered the population completely, and on the other hand, there was a large self-selection panel.

The idea of sample matching is to select a random sample from the sampling frame, and to match the sample persons to members of the panel. Instead of approaching the sample persons in the frame, the linked panel members are asked to complete the questionnaire. The response rate will be high as the panel contains people who have agreed to participate in surveys regularly.

Frame persons are linked to panel persons by means of a set of auxiliary variables. The more variables are available for matching, the more similar the linked persons will be. It will, however, also be more difficult to find matching persons in the panel. The panel needs to be large for this. Only two auxiliary variables were used in the simulation experiment in this paper: age and education. This makes it easy to find matching persons, but these persons may still differ in many other respects from the sampled persons.

This paper treats sample matching as a form of weighting adjustment. However, it can also be seen as a form of imputation. The data of a person selected from the frame is not used. So on could consider it to be used. Instead, data of another person (one in the panel) is substituted, which is a form of donor imputation.

Application of sample matching requires a sufficiently large set of auxiliary variable to be available in both the sampling frame and the panel. Moreover, the values of the auxiliary variables must have been measured in the same way. Particularly if the mode of data collection for the frame is different from that of the panel, there may be mode effects, i.e. if the same question is asked in a different mode, a different answer is given. As a consequence, records are not matched that should have been matched.

Application of sample matching is no guarantee for success. This technique will only be able to totally remove the nonresponse bias if the set of auxiliary variables is capable of explaining the participation behavior completely. If the set of auxiliary variables only explains part of the participation behavior, the bias will be reduced but does not vanish. This is clearly shown in the simulation experiments, where using either age or education reduces, but not removes, the bias. The bias vanishes completely only if both variables are used.

Auxiliary variables in surveys are often categorical variables. As a consequence, sample matching comes down to stratifying the panel and selecting the right number of persons from each stratum. The number of persons per stratum is defined by the sample numbers per stratum in the frame. The simulation study shows that the sample fractions per stratum in the frame can be replaced by the corresponding population fractions. The behavior of the estimator with respect to bias reduction will remain the same. In fact, this reduces sample matching to selecting a stratified sample (with proportional allocation) from the frame.

Another approach could be to draw a simple random sample from the panel, and then to apply post-stratification using the auxiliary variables, where the population distribution of the auxiliary variables is obtained from the sampling frame. The simulation study shows that this estimation approach has the same properties with respect to bias reduction as matched samples and stratified sampling. It should be noted that post-stratification of the panel sample may result in larger variances, since the sample may not be representative giving rise to weights with substantial variation.

Due to self-selection, the panel may be very skew. Specific groups may be substantially under-represented. If the number of persons in such a group is very small, these persons run the risk of being selected every time a new sample is selected. They may not like this, which could result in nonresponse.

The overall conclusion may be that sample matching has no substantial advantages over stratified sampling and post-stratification estimation. All three approaches reduce the bias in the same way. The amount of bias reduction only depends on the auxiliary variables used, and not on the specific method.

In many situations it will probably be simpler to just draw a simple random sample from the panel, followed by post-stratification. The population distribution of the auxiliary variables can be obtained from a sampling frame, but it may also be obtained from other sources like a population register.

8. References

Baker, R., Blumberg, S.J., Brick, J.M., Couper, M.P., Courtright, M., Dennis, J.M., Dillman, D., Frankel,, M.R., Garland, P.,Groves, R.M., Kennedy, C., Krosnick, J., Lavrakas, P.J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R.K. & Zahs, D. (2010). Research Synthesis: AAPOR Report on Online Panels. *Public Opinion Quarterly* 74, pp. 711–781.

Bethlehem, J.G (2009), *Applied Survey Methods, A Statistical Perspective*. John Wiley & Sons, Hoboken, NJ.

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

Cobben, F. & Bethlehem, J.GH. (2013), *Web Panels for Official Statistics*. Discussion paper 201307, Statistics Netherlands, The Hague, The Netherlands.

Kiaer, A. N. (1895), Observations et Expériences Concernant des Dénombrements Représentatives. *Bulletin of the International Statistical Institute*, IX, Book 2, pp. 176-183.

Little, R.J.A & Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Second edition. New York: John Wiley & Sons.

Rivers, D. (2007), *Sampling for Web Surveys*. Paper presented at the Joint Statistical Meetings, Section on Survey Research Methods, Salt Lake City, Utah.

Rivers, D & Bailey, D. (2009), *Inference from Matched Samples in the 2008 U.S. National Elections*. Paper presented at the 64th Annual Conference of the American Association for Public Opinion Research, Hollywood, Florida.

Scherpenzeel, A. (2008), An Online Panel as a Platform for Multi-Disciplinary Research. In: Stoop, I. & Wittenberg, M. (eds.), *Access Panels and Online Research, Panacea or Pitfall?* Aksant, Amsterdam.

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

Vavreck, L. & Rivers, D. (2008), The 2006 Cooperative Congressional Election Study. *Journal of Elections* 18, pp.355-366.

Explanation of symbols

.	Data not available
*	Provisional figure
**	Revised provisional figure (but not definite)
x	Publication prohibited (confidential figure)
–	Nil
–	(Between two figures) inclusive
0 (0.0)	Less than half of unit concerned
empty cell	Not applicable
2013–2014	2013 to 2014 inclusive
2013/2014	Average for 2013 to 2014 inclusive
2013/'14	Crop year, financial year, school year, etc., beginning in 2013 and ending in 2014
2011/'12–2013/'14	Crop year, financial year, etc., 2011/'12 to 2013/'14 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures..

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress: Statistics Netherlands, Grafimedia
Printed by: Statistics Netherlands, Grafimedia
Design: Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
verkoop@cbs.nl
Fax +31 45 570 62 68

60083 201404 X-10