How accurate are self-selection web surveys?

Jelke Bethlehem

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

Discussion paper (08014)

Statistics Netherlands

The Hague/Heerlen, 2008

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2005-2006	= 2005 to 2006 inclusive
2005/2006	= average of 2005 up to and including 2006
2005/’06	= crop year, financial year, school year etc. beginning in 2005 and ending in 2006
2003/’04–2005/’06	= crop year, financial year, etc. 2003/’04 to 2005/’06 inclusive

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

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Facility Services

Cover
TelDesign, Rotterdam

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

Where to order
E-mail: verkoop@cbs.nl
Telefax .. +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

Summary: A web survey seems to be an attractive means of collecting survey data, because it provides simple, cheap and fast access to a large group of people. However, there are pitfalls. Due to methodological problems, the quality of the outcomes of web surveys may be seriously affected. This paper addresses one of these problems, and that is self-selection of respondents. Self-selection leads to a lack of representativity and thus to biased estimates. The effect of self-selection on the distributional characteristics of estimators is described in detail. It is shown that the bias of estimators in self-selection surveys can be much larger than in surveys based on traditional probability samples. A simulation study also shows what can go wrong in a self-selection web survey. It is explored whether some correction techniques (adjustment weighting and use of reference surveys) can improve the quality of the outcomes. It turns out that there is no guarantee for success.

Keywords: web survey, online survey, self-selection, bias, representativity, adjustment weighting, reference survey

1. Trends in data collection

Collecting data using surveys is often a complex, costly and time-consuming process. Not surprisingly, continuous attempts have been made all through the history of survey research to improve timeliness and reducing costs, while at the same time maintaining a high level of data quality.

Developments in information technology in the last decades of the previous century made it possible to use microcomputers for data collecting. This led to the introduction of computer-assisted interviewing (CAI). Replacing the paper questionnaire by an electronic one turned out to have many advantages, among which were considerably shorter survey processing times and higher data quality. More on the benefits of CAI can be found in Couper et al. (1998).

The rapid development of the Internet has led to another new type of data collection: Computer Assisted Web Interviewing (CAWI). Such a web survey (also sometimes called online survey) is almost always self-administered: respondents visit a website, and complete the questionnaire by filling in a form on-line. Web surveys have some attractive advantages in terms of costs and timeliness:

- Now that so many people are connected to the Internet, a web survey is a simple means to get access to a large group of potential respondents;
- Questionnaires 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.

Thus, web surveys are a fast, cheap and attractive means of collecting large amounts of data. Not surprisingly, many survey organisations have implemented such surveys. However, the question is whether a web survey is also attractive from a quality point of view, because there are methodological problems. These problems are caused by using the Internet as a selection instrument for respondents.

This paper shows that the quality of web surveys may be seriously affected by these problems, making it difficult, if not impossible to make proper inference with respect to the target population of the survey. The two main causes of problems are under-coverage and self-selection. This paper focuses on self-selection problems and only briefly touches upon under-coverage. The effects of under-coverage are treated in more detail in Bethlehem (2007). Some theory about self-selection is developed in this paper. It explores to what extent weighting adjustment techniques can help to solve the problem. Practical implications are showed using data from a fictitious population.

2. Coverage and self-selection

Objective of a survey always is to collect information about a well-defined target population. To that end a sample is selected from this population. The methodology of survey sampling has been developed over a period of more than 100 years. It is based on the fundamental principle of probability sampling. Selecting random samples makes it possible to apply probability theory. Consequently, the accuracy of estimators can be quantified and controlled. The probability sampling principle has been successfully applied in official and academic statistics since the 1940's, and to a lesser extent also in more commercial market research.

At first sight, web surveys have much in common with other types of surveys. It is just another mode of data collection. Questions are not asked face-to-face or by telephone, but over the Internet. What is different for many web surveys, however, is that the principles of probability sampling have not been applied. Samples are not constructed by means of probability sampling but instead rely on *self-selection* of respondents. This can have a major impact on survey results.

There is also another methodological problem that web surveys share with surveys based on probability samples, and that is *under-coverage*. This problem occurs when elements in the target population do not appear in the sampling frame. Under-coverage can be a serious problem for web surveys. If the target population consists of all people with an Internet connection, there is no problem. However, usually the target population is wider than that. Then, under-coverage occurs due to the fact that still many people do not have access to the Internet.

Bethlehem (2007) describes the situation in The Netherlands with respect to Internet access. In the period from 1998 to 2006 the percentage of persons with Internet has increased from 16% to 85%. The question is whether this Internet population differs from the complete target population. The answer is yes in The Netherlands. Specific groups are substantially under-represented, like the elderly, the low educated, and the non-native part of the population. The results described above are in line with the findings of authors in other countries. See e.g. Couper (2000), and Dillman and Bowker (2001).

One could argue that this problem may disappear as the Internet penetration increases further. However, this is not evident. Bethlehem (2007) shows that the bias due to under-coverage of the estimator for the population mean $\bar{Y}$ of some variable Y is equal to

$$B(\bar{y}_I) = E(\bar{y}_I) - \bar{Y} = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N}(\bar{Y}_I - \bar{Y}_{NI}) . \tag{2.1}$$

The estimator $\bar{y}_I$ is the sample mean based on observations from just the Internet population. The means of Y in the Internet population and non-Internet population are denoted by $\bar{Y}_I$ and $\bar{Y}_{NI}$ respectively. Furthermore, N is the size of the total population and N_{NI} is the size of the non-Internet population.

The magnitude of this bias is determined by two factors. The first factor is the relative size N_{NI} / N of the population without Internet. The bias will decrease as a smaller proportion of the population does not have access to Internet. The second factor is the *contrast* $\bar{Y}_I - \bar{Y}_{NI}$ between the Internet-population and the non-Internet-population. The more the mean of the target variable differs for these two sub-populations, the larger the bias will be. An increased Internet coverage will reduce the bias because the factor N_{NI} / N is smaller. However, the contrast does not necessarily decrease as Internet coverage grows. It is even possible that the remaining hard-core group of people without Internet will be more and more deviant. This may cause the contrast to increase. So, taking into account the combined effect of both factors, there is no guarantee that increased Internet coverage will reduce the under-coverage bias.

3. Effect of self-selection

Horvitz and Thompson (1952) show in their seminal paper that unbiased estimates of population characteristics can be computed only if a real probability sample has been drawn, every element in the population has a non-zero probability of selection, and all these probabilities are known to the researcher. Furthermore, only under these conditions, the accuracy of estimates can be computed.

Many web surveys are not based on probability sampling. The survey questionnaire is simply put on the web. Respondents are those people who happen to have Internet, visit the website and decide to participate in the survey. These surveys are called *self-selection surveys*. The problem is that the survey researcher is not in control of the selection process. Selection probabilities are unknown and, moreover, they are considerably smaller than in traditional probability surveys. Therefore, no unbiased estimates can be computed nor can the accuracy of estimates be determined.

The effects of self-selection can be illustrated using an example related to the general elections in The Netherlands in 2003. Various organisations made attempts to use opinion polls to predict the outcome of these elections. The results of these polls are summarised in table 3.1. *Politieke Barometer*, *Peil.nl* and *De Stemming* are opinion polls carried out by market research agencies. They are all based on samples from web panels (also called access panels). To reduce a possible bias, adjustment weighting has been carried out. The polls were conducted one day before the election. The Mean Absolute Difference indicates how big the differences (on average) are between the poll and the election results. Particularly, differences are large for the more volatile parties like PvdA, SP and the PVV.

DPES is the Dutch Parliamentary Election Study. The fieldwork was carried out by Statistics Netherlands in a few weeks just before the elections. The probability sampling principle has been followed here. A true (two-stage) probability sample was drawn. Respondents were interviewed face-to-face (using CAPI). The predictions of this survey are much better than those based on the online opinion polls.

Table 3.1. Dutch Parliamentary elections 2006.
Outcomes and the results of various opinion surveys

	Election result	**Politieke Barometer**	**Peil.nl**	**De Stemming**	**DPES 2006**
Sample size		1,000	2,500	2,000	2,600
Seats in parliament:					
CDA (christian democrats)	41	41	42	41	41
PvdA (social democrats)	33	37	38	31	32
VVD (liberals)	22	23	22	21	22
SP (socialists)	25	23	23	32	26
GL (green party)	7	7	8	5	7
D66 (liberal democrats)	3	3	2	1	3
ChristenUnie (christian)	6	6	6	8	6
SGP (christian)	2	2	2	1	2
PvdD (Animal party)	2	2	1	2	2
PVV (Conservative)	9	4	5	6	8
Other parties	0	2	1	2	1
Mean Absolute Difference		1.27	1.45	2.00	0.36

Probability sampling has the additional advantage that it provides protection against certain groups in the population attempting to manipulate the outcomes of the survey. This may typically play a role in opinion polls. Self-selection does not have this safeguard. An example of this effect could be observed in the election of the 2005 Book of the Year Award (Dutch: NS Publieksprijs), a high-profile literary prize. The winning book was determined by means of a poll on a website. People could vote for one of the nominated books or mention another book of their choice. More than 90,000 people participated in the survey. The winner turned out to be the new interconfessional Bible translation launched by the Netherlands and Flanders

Bible Societies. This book was not nominated, but nevertheless an overwhelming majority (72%) voted for it. This was due to a campaign launched by (among others) Bible societies, a Christian broadcaster and Christian newspaper. Although this was all completely within the rules of the contest, the group of voters could clearly not be considered to be representative of the Dutch population.

4. The theoretical framework

Let the target population of the survey consist of N identifiable elements, which are labelled 1,2,...,N. Associated with each element k is a value Y_k of the target variable Y. The aim of the web survey is assumed to be estimation of the population mean

$$\bar{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k \tag{4.1}$$

of the target variable Y.

Participation in a self-selection web-survey requires in the first place that respondents are aware of the existence of a survey (they have to accidentally visit the website, or they have to follow up a banner or an e-mail message). In the second place, they have to make the decision to fill in the questionnaire on the Internet. All this means that each element k in the population has unknown probability ρ_k of participating in the survey, for k = 1, 2, ..., N. The responding elements can be denoted by a series

$$r_1, r_2, ..., r_N \tag{4.2}$$

of N indicators, 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 $\rho_k = E(r_k)$ will be called the *response propensity* of element k.

The random variables r_1, r_2, …, r_N are independent. This sample selection process is a form of *Poisson sampling*. However in practical applications of Poisson sampling the selection probabilities are known, whereas they are unknown in a self-selection survey.

The number of respondents is equal to

$$n = \sum_{k=1}^{N} r_k \tag{4.3}$$

Note that n is a random variable. A naive estimator of the population mean is the sample mean

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{N} r_k Y_k \ . \tag{4.4}$$

This estimator implicitly assumes every element in the population to have the same probability of participating in the survey. Quantity (4.4) is the ratio of two random variables. It can be shown that its expected value is approximately equal to the ratio of the expected values of the both random variables. Hence

$$E(\bar{y}) » \bar{Y}^{*} = \frac{1}{N}\sum_{k=1}^{N}\frac{r_k}{\bar{r}}Y_k \tag{4.5}$$

where $\bar{\rho}$ is the mean of all response probabilities. Using an approach similar to Cochran (1977, p. 31), it can be shown that the variance of the sample mean is approximately equal to

$$V(\bar{y}) » \frac{1}{\left(N\bar{r}\right)^{2}}\sum_{k=1}^{N} r_k(1- r_k)(Y_k - \bar{Y}^{*})^{2} \tag{4.6}$$

Note that this expression for the variance does not contain a sample size (because no fixed size sample was drawn), but the expected sample size $N\bar{r}$. Not surprisingly, the variance decreases as the expected sample size increases.

Generally, the expected value of this sample mean is not equal to the population mean of the population. The only situation in which the bias vanishes is that in which all response probabilities in the Internet-population are equal. In terms of nonresponse correction theory, this comes down to Missing Completely At Random (MCAR). Indeed, in this case, self-selection leads to a representative sample because all elements have the same selection probability.

Bethlehem (1988) shows that the bias of the sample mean (4.4) can be written as

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} » \bar{Y}^{*} - \bar{Y} = \frac{C_{rY}}{\bar{r}}, \tag{4.7}$$

where

$$C_{rY} = \frac{1}{N}\sum_{k=1}^{N}(r_k - \bar{r})(Y_k - \bar{Y}) \tag{4.8}$$

is the covariance between the values of target variable and the response probabilities. The bias of the sample mean (as an estimator of the population mean) is therefore determined by two factors:

- The average response probability. The more likely people are to participate in the survey, the higher the average response probability will be, and thus the smaller the bias will be.
- The relationship between the target variable and response behaviour. The higher the correlation between the values of the target variable and the response probabilities, the higher the bias will be.

Three situations can be distinguished in which this bias vanishes:

1) All response probabilities are equal. Again, this is the case in the which the self-selection process can be compared with a simple random sample;

2) All values of the target variable are equal. This situation is very unlikely to occur. If this were the case, no survey would be necessary. One observation would be sufficient.
3) There is no relationship between target variable and response behaviour. It means participation does not depend on the value of the target variable. This corresponds to Missing Completely At Random (MCAR).

Expression (4.7) for the bias of the sample mean can be rewritten as

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} \approx \bar{Y}^* - \bar{Y} = \frac{R_{\rho Y} S_\rho S_Y}{\bar{\rho}}, \tag{4.9}$$

in which $R_{\rho Y}$ is the value of correlation between the target variable and the response probabilities, S_ρ is the standard deviation of the response probabilities, and S_Y is the standard deviation of the target variable. Given the mean response probability $\bar{\rho}$, there is a maximum value the standard S_ρ cannot exceed:

$$S_\rho \le \sqrt{\bar{\rho}(1-\bar{\rho})}\,. \tag{4.10}$$

This implies that in the worst case (S_ρ assumes it maximum value and the correlation $R_{\rho Y}$ is equal to either +1 or -1) the absolute value of the bias will be equal to

$$\left|B_{\max}(\bar{y})\right| = S_Y \sqrt{\frac{1}{\bar{\rho}} - 1}\,. \tag{4.11}$$

Bethlehem (1988) shows the formula (4.7) also applies in the situation in which a probability sample has been drawn, and subsequently non-response occurs during the fieldwork.

Consequently, expression (4.11) provides a means to compare potential biases in various survey designs. For example, regular surveys of Statistics Netherlands are all based on probability sampling. Their response rates are around 70%. This means the absolute maximum bias is equal to 0.65 × $S_{y.}$ One of the largest web surveys in The Netherlands is *21minuten.nl*. This survey is supposed to supply answers to questions about important problems in Dutch society. It is a self-selection web survey. Within a period of six weeks in 2006 about 170,000 people completed the online questionnaire. The target population of this survey was not defined, as everyone could participate. If it is assumed the target population consists of all Dutch from the age of 18, the average response probability is equal to 170,000 / 12,800,000 = 0.0133. Hence, the absolute maximum bias is equal to 8.61 × S_y. It can be concluded that the bias of a large web survey can be a factor 13 larger than bias of a smaller probability survey.

5. Weighting adjustment

Weighting adjustment is a family of techniques that attempt to improve the quality of survey estimates by making use of auxiliary information. *Auxiliary information* is defined here as a set of variables that have been measured in the survey, and for which information on their population distribution is available. By comparing the population distribution of an auxiliary variable with its sample distribution, it can be assessed whether or not the sample is representative for the population (with respect to this variable). If these distributions differ considerably, one must conclude that the sample is selective.

Note that for a probability sample in which non-response has occurred, it is also possible to use the distribution of the auxiliary variables in the complete sample instead of their population distribution. Such information can sometimes be retrieved from the sampling frame. This situation does not apply to self-selection samples as there is no sampling frame.

To correct for a lack of representativity, adjustment weights can be computed. Weights are assigned to records of all respondents. Estimates of population characteristics can now be obtained by using weighted values instead of the unweighted values. Weighting adjustment is often used to correct surveys that are affected by nonresponse, see e.g. Bethlehem (2002). This section explores the possibility to reduce the bias of self-selection web survey estimates by applying post-stratification. This is a well-known and often used weighting technique.

To carry out post-stratification, one or more qualitative auxiliary variables are needed. Here, only one such variable is considered. The extension to more variables is essentially the same. Suppose, there is an auxiliary variable *X* having *L* categories. So it divides the target population into *L* strata. The strata are denoted by the subsets U_1, U_2, ..., U_L of the population *U*. The number of target population elements in stratum U_h is denoted by N_h, for *h* = 1, 2, ..., *L*. The population size *N* is equal to $N = N_1 + N_2 + ... + N_L$. This is the population information assumed to be available.

Suppose a self-selection sample is selected from the Internet-population. If n_h denotes the number of respondents in stratum *h*, then $n = n_1 + n_2 + ... + n_L$. The values of the n_h are the result of a Poisson sampling process, so they are random variables.

Post-stratification assigns identical adjustment weights to all elements in the same stratum. The weight w_k for a respondent *k* in stratum *h* is equal to

$$w_k = \frac{N_h / N}{n_h / n} \qquad (5.1)$$

The simple sample mean

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{N} r_k Y_k \qquad (5.2)$$

is now replaced by the weighted sample mean

$$\bar{y}_{PS} = \frac{1}{n}\sum_{k=1}^{N} w_k r_k Y_k \qquad (5.3)$$

Substituting the weights and working out this expression leads to the post-stratification estimator

$$\bar{y}_{PS} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h = \sum_{h=1}^{L} W_h \bar{y}_h \ , \qquad (5.4)$$

where $\bar{y}_h$ is the sample mean in stratum *h* and $W_h = N_h \ / \ N$ is the relative size of stratum *h*. The expected value of this post-stratification estimator is equal to

$$E(\bar{y}_{PS}) = \frac{1}{N}\sum_{h=1}^{L} N_h E(\bar{y}_h) = \sum_{h=1}^{L} W_h \bar{Y}_h^* = \tilde{\bar{Y}}^* \ , \qquad (5.5)$$

where

$$\bar{Y}_h^* = \frac{1}{N_h}\sum_{k=1}^{N_h} \frac{r_{k,h}}{\bar{r}_h} Y_{k,h} \qquad (5.6)$$

is the weighted mean of the target variable in stratum *h*. The subscript *k*,*h* denotes the *k*-th element in stratum *h*, and $\bar{r}_h$ is the average response probability in stratum *h*.

Expression (5.6) is the analogue of expression (4.5), but now computed for stratum *h*. Generally, this mean will not be equal to the mean $\bar{Y}_h$ of the target variable in stratum *h* of the target population. The bias of this estimator is equal to

$$\begin{aligned} B(\bar{y}_{PS}) &= E(\bar{y}_{PS}) - \bar{Y} = \tilde{\bar{Y}}^* - \bar{Y} = \sum_{h=1}^{L} W_h (\bar{Y}_h^* - \bar{Y}_h) = \\ &= \sum_{h=1}^{L} W_h \frac{R_{rY,h} S_{r,h} S_{Y,h}}{\bar{r}_h}, \end{aligned} \qquad (5.7)$$

were the subscript *h* indicates that the respective quantities are computed just for stratum *h* and not for the complete population.

This bias will be small if

- The response probabilities are similar within strata;
- The values of the target variable are similar within strata;
- There is no correlation between response behaviour and the target variable within strata.

These conditions can be realised if there is a strong relationship between the target variable *Y* and the stratification variable *X*. Then the variation in the values of *Y* manifests itself between strata but not within strata. In other words, the strata are homogeneous with respect to the target variable. Also if the strata are homogeneous with respect to the response probabilities, the bias will be reduced. In nonresponse correction terminology, this situation comes down to Missing At Random (MAR).

In conclusion it can be said that application of post-stratification will successfully reduce the bias of the estimator if proper auxiliary variables can be found. Such variables should satisfy three conditions:

- They have to be measured in the survey;
- Their population distribution (N_1 ,N_2, ..., N_L) must be known;
- They must produce homogeneous strata.

Unfortunately, such variables are rarely available, or there is only a weak correlation.

It can be shown that, in general, the variance of the post-stratification estimator is equal to

$$V(\bar{y}_{PS}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) \, . \tag{5.8}$$

In the case of a self-selection web survey, the variance $V(\bar{y}_h)$ of the sample mean in a stratum is the analogue of variance (4.6) but restricted to observations in that stratum. Therefore, the variance of the post-stratification estimator is approximately equal to

$$V(\bar{y}_{PS}) = \sum_{h=1}^{L} W_h^2 \frac{1}{\left(N_h \bar{\rho}_h\right)^2} \sum_{k \in U_h}^{N} \rho_k (1-\rho_k)(Y_k - \bar{Y}_h^*)^2 \, . \tag{5.9}$$

This variance is small if the strata are homogeneous with respect to the target variable. So, a strong correlation between the target variable Y and the stratification variable X will reduce both the bias and the variance of the estimator.

6. Weighting adjustment with a reference sample

The previous section showed that post-stratification can be an effective correction technique provided auxiliary variables are available that have a strong correlation with the target variables of the survey. If such variables are not available, it might be considered to conduct a *reference survey*. This reference survey is based on a small probability sample, where data collection takes place with a mode different from the web, e.g. CAPI (Computer Assisted Personal Interviewing, with laptops) or CATI (Computer Assisted Telephone Interviewing). The reference survey approach has been applied by several market research organisations, see e.g. Börsch-Supan et al. (2004) and Duffy et al. (2005).

Under the assumption of full response, or ignorable nonresponse, this reference survey will produce unbiased estimates of quantities that have also been measured in the web survey. Unbiased estimates for the target variable can be computed, but due to the small sample size, these estimates will have a substantial variance. The question is now whether estimates can be improved by combining the large sample size of the web surveys with the unbiased estimates of the reference survey.

To explore this, it is assumed that one qualitative auxiliary variable is observed both in the web survey and the reference survey, and that this variable has a strong correlation with the target variable of the survey. Then a form of post-stratification can be applied where the stratum means are estimated using web survey data and the stratum weights are estimated using the reference survey data. This leads to the post-stratification estimator

$$\bar{y}_{RS} = \sum_{h=1}^{L} \frac{m_h}{m} \bar{y}_h \qquad (6.1)$$

where $\bar{y}_h$ is the web survey based estimate for the mean of stratum *h* of the target population (for *h* = 1, 2, ..., *L*), and m_h / *m* is the relative sample size in stratum *h* for the reference sample (for *h* = 1, 2, ..., *L*). Under the conditions described above the quantity m_h / *m* is an unbiased estimate of $W_h = N_h / N$.

Let *I* denote the probability distribution for the web survey and let *P* be the probability distribution for the reference survey. Then the expected value of the post-stratification estimator is equal to

$$\begin{aligned} E(\bar{y}_{RS}) &= E_I E_P(\bar{y}_{RS} \mid m_1, m_2, ..., m_L) = E_I\left(\sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h \right) = \\ &= \sum_{h=1}^{L} W_h \bar{Y}_h^* = \tilde{Y}^* \end{aligned}, \qquad (6.2)$$

where $W_h = N_h / N$ is the relative size of stratum *h* in the target population. So, the expected value of this estimator is identical to that of the post-stratification estimator (5.4). The bias of this estimator is equal to

$$\begin{aligned} B(\bar{y}_{RS}) &= E(\bar{y}_{RS}) - \bar{Y} = \tilde{Y}^* - \bar{Y} = \sum_{h=1}^{L} W_h (\bar{Y}_h^* - \bar{Y}_h) = \\ &= \sum_{h=1}^{L} W_h \frac{R_{rY,h} S_{r,h} S_{Y,h}}{\bar{r}_h}, \end{aligned} \qquad (6.3)$$

A strong relationship between the target variable and the auxiliary variable used for computing the weights means that there is little or no variation of the target variable within the strata. Consequently, the correlation between target variable and response behaviour will be small, and the same applies to the standard deviation of the target variable. So, using a reference survey with the proper auxiliary variables can substantially reduce the bias of web survey estimates.

Note that the bias of the reference survey estimator is equal to that of the post-stratification estimator, see expression (5.6). An interesting aspect of the reference survey approach is that any variable can be used for adjustment weighting as long as it is measured in both surveys. For example, some market research organisations use ‘webographics’ or ‘psychographic’ variables that divide the population in 'mentality groups'. People in the same groups have more or less the same level of motivation and interest to participate in such surveys. Effective weighting variables approach the MAR situation as much as possible. This implies that within weighting strata there is no relationship between participating in a web survey and the target variables of the survey.

It can be shown that if a reference survey is used, the variance of the post-stratification estimator is equal to

$$V(\bar{y}_{RS}) = \frac{1}{m}\sum_{h=1}^{L} W_h\left(\bar{Y}_h^* - \tilde{Y}^*\right)^2 + \frac{1}{m}\sum_{h=1}^{L} W_h\left(1-W_h\right)V(\bar{y}_h) + \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) \qquad (6.4)$$

The proof is given in the appendix. The quantity $\bar{y}_h$ is measured in the web survey. Therefore its variance $V(\bar{y}_h)$ will be at most of the order $1/E(n) = 1/(N\bar{\rho})$). This means that the first term in the variance of the post-stratification estimator will be of the order 1/*m*, the second term of order 1/(*mE(n))*, and the third term of order 1/*E(n)*. Since *E*(*n)* will generally be much larger than *m* in practical situations, the first term in the variance will dominate, i.e. the (small) size of the reference survey will determine the accuracy of the estimates.

Moreover, since strata are based on groups of people with the same psychographics scores, and target variables may very well be related to the psychographic variables, the stratum means $\bar{Y}_h^*$ may vary substantially. This also contributes to a large value of the first variance component.

The conclusion is that a large number of observations in the web survey do not help to produce accurate estimates. The reference survey approach may reduce the bias of estimates, but it does so at the cost of a higher variance.

The effectiveness of a survey design is sometimes also indicated by means of the *effective sample size*. This is the sample size of a simple random sample of elements that would produce an estimator with the same precision. Use of a reference survey implies that the effective sample size is much lower than the size of the web survey. See section 8 for an example showing this effect.

7. Propensity weighting

Propensity weighting is used by several market research organisations to correct for a possible bias in their web surveys, see e.g. Börsch-Supan et al. (2004) and Duffy et al. (2005). The original idea behind propensity weighting goes back to Rosenbaum & Rubin (1983, 1984).

Propensity scores are obtained by modelling a variable that indicates whether or not someone participates in the survey. Usually a logistic regression model is used where the indicator variable is the dependent variable and attitudinal variables are the explanatory variables. These attitudinal variables are assumed to explain why someone participates or not. Fitting the logistic regression model comes down to estimating the probability (propensity score) of participating, given the values of the explanatory variables.

Each person *k* in the population is assumed to have a certain, unknown probability ρ_k of participating in the survey, for *k* = 1, 2, .., *N*. Let r_1, r_2, …, r_N denote indicator variables, where $r_k = 1$ if person k participates in the survey, and $r_k = 0$ otherwise. Consequently, $P(r_k = 1) = \rho_k$.

The *propensity score ρ(X)* is the conditional probability that a person with observed characteristics X participates, i.e.

$$\rho(X) = P(r = 1 \mid X) \qquad (7.1)$$

It is assumed that within the strata defined by the values of the observed characteristics X, all persons have the same participation propensity. This is the Missing At Random (MAR) assumption. The propensity score is often modelled using a logit model:

$$\log\left(\frac{\rho(X_k)}{1-\rho(X_k)}\right) = \alpha + \beta' X_k \qquad (7.2)$$

The model is fitted using Maximum Likelihood estimation. Once propensity scores have been estimated, they are used to stratify the population. Each stratum consists of elements with (approximately) the same propensity scores. If indeed all elements within a stratum have the same response propensity, there will be no bias if just the elements in the Internet population are used for estimation purposes. Cochran (1968) claims that five strata are usually sufficient to remove a large part of the bias. The market research agency Harris Interactive was among the first to apply propensity score weighting, see Terhanian et al. (2001).

To be able to apply propensity score weighting, two conditions have to be fulfilled. The first condition is that proper auxiliary variables must be available. These are variables that are capable of explaining whether or not someone is willing to participate in the web survey. Variables often used measure general attitudes and behaviour. They are sometimes referred to as ‘webographic’ or ‘psychographic’ variables. Schonlau et al. (2004) mention as examples “Do you often feel alone?” and “On how many separate occasions did you watch news programs on TV during the past 30 days?”.

It should be remarked that attitudinal questions are much less reliable than factual questions. Respondents may never have thought of the topics addressed in attitudinal questions. They have to make up their mind at the very moment the question is asked. Their answers may be depend on their current circumstances, and may very over time. Therefore, attitudinal question may be subject to substantial measurement errors.

The second condition for this type of adjustment weighting is that the population distribution of the webographic variables must be available. This is generally not the case. A possible solution to this problem is to carry out an additional reference survey. To allow for unbiased estimation of the population distribution, the reference survey must be based on a true probability sample from the entire target population.

Such a reference survey can be small in terms of the number of questions asked. It can be limited to the webographic questions. Preferably, the sample size of the reference survey should be large enough to allow for precise estimation. A small sample size results in large standard errors of estimates.

Schonlau et al. (2004) describe the reference survey of Harris Interactive. This is a CATI survey, using random digit dialling. This reference survey is used to adjust several web surveys. Schonlau et al. (2003) stress that the success of this approach depends on two assumptions: (1) the webographics variables are capable of explaining the difference between the web survey respondents and the other persons in the target population, and (2) the reference survey does not suffer from non-ignorable nonresponse. In practical situations it will not be easy to satisfy these conditions.

It should be noted that from a theoretical point of view propensity weighting should be sufficient to remove the bias. However, in practice the propensity score variable will often be combined with other (demographic) variables in a more extended weighting procedure, see e.g. Schonlau (2004).

8. A simulation study

To explore the effects of self-selection and correction techniques, a simulation study was carried out. A fictitious population was constructed. For this population, voting intentions for the next general elections were simulated and analysed. The relationships between variables involved were modelled somewhat stronger than they probably would be in a real life situation. Effects are therefore more pronounced, making it clearer what the pitfalls are.

The characteristics of estimators (before and after correction) were computed based on a large number of simulations. First, the distribution of the estimator was determined in the ideal situation of a simple random sample from the target population. Then, it was explored how the characteristics of the estimator change if self-selection is applied. Finally, the effects of weighting (post-stratification and reference survey) were analysed.

A fictitious population of 100,000 individuals was constructed. There were five variables:

- The variable *Internet* indicates how active a person is on the internet. There are two categories. Very active users and more passive users. The population consists for 1% of active users and for 99% of passive users. Active users have a response propensity of 0.99 and passive users have a response propensity of 0.01.

- The variable *Age* in three categories: young, middle aged and old. The active Internet users consist for 60% of young people, for 30% of middle aged people and only for 10% of old people. The age distribution for passive Internet users is 40% young, 35% middle aged and 25% old. So, typically younger people are more active internet users.

- Will vote for the National Elderly Party (NEP). The probability to vote for this party only depends on age. Probabilities are 0.00 (for Young), 0.30 (for Middle aged) and 0.60 (for Old).
- Will vote for the New Internet Party (NIP). The probability to vote for this party depends both on age and use of Internet. For active Internet users, the probabilities were 0.80 (for young), 0.40 (for middle aged) and 0.20 (for old). All probabilities were equal to 0.10 for passive Internet users. So, for active users voting decreases with age. Voting probability is always low for passive users.

Figure 8.1. Relationships between variables

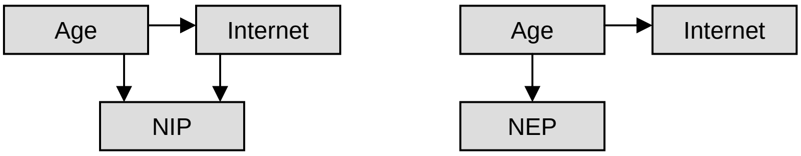


Figure 8.1 shows the relationships between the variables in a graphical way. The decision not to participate in a self-selection survey can be seen as a form of nonresponse. The theory on nonresponse (see for example Little & Rubin, 2002) distinguishes three nonresponse generating mechanisms:

- *Missing Completely At Random* (MCAR). There is no relationship at all between the mechanism causing missingness and target variables of the survey. This is the ideal situation. The mechanism only leads to a reduced number of observations. Estimators will not be biased.
- *Missing At Random* (MAR). There is an indirect relationship between the mechanism causing missingness and the target variables of the survey. The relationship runs through a third variable, and this variable is measured in the survey as an auxiliary variable. In this case estimates are biased, but it is possible to correct for this bias. For example, if the auxiliary variable is used to construct strata, there will be no bias within strata, and the post-stratification will remove the bias.
- *Not Missing At Random* (NMAR). There is a direct relationship between the mechanism causing missingness and target variables of the survey. This is the worst case. Estimators will be biased and it is not possible to remove this bias.

The variable NEP (National Elderly Party) suffers from missingness due to MAR in the experiment. There is direct relationship between voting for this party and age, and also there is a direct relationship between age and the probability to participate in the survey. This will cause estimates to be biased. It should be possible to correct for this bias by weighting using the variable age.

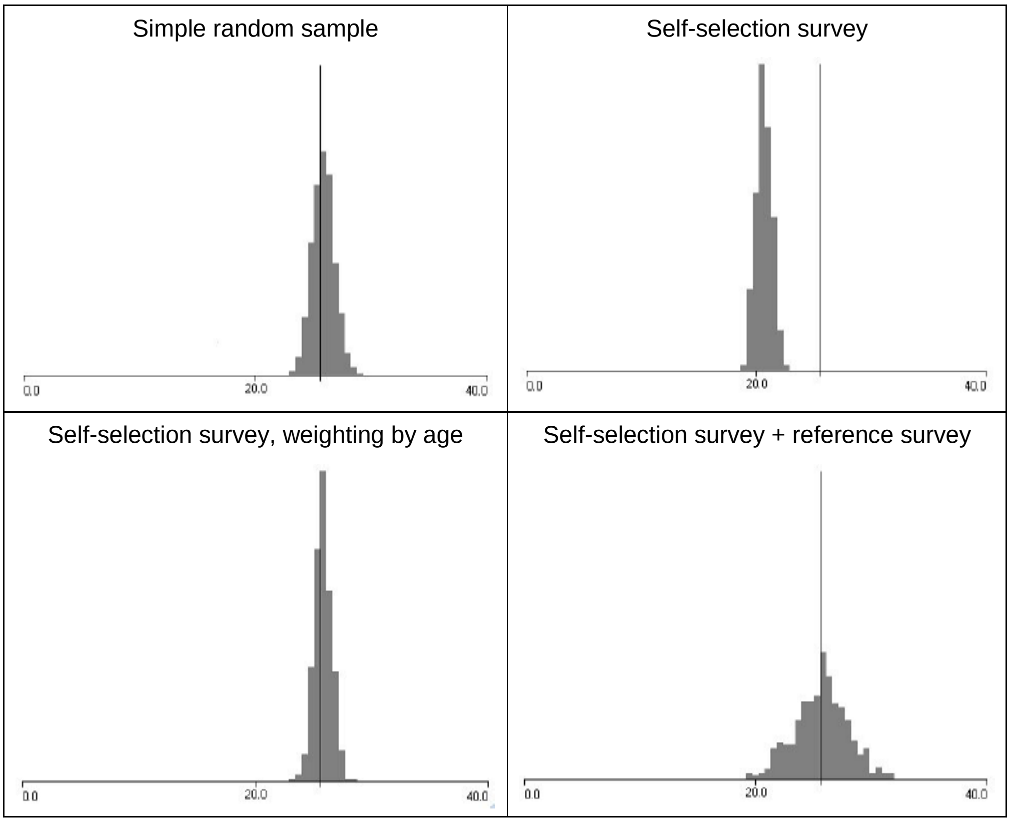
The variable NIP (National Internet Party) suffers from NMAR. There exists a direct relationship between voting for this party and the response probability. Estimates will be biased, and there is no correction possible.

The distribution of estimators for the percentage of voters for both parties was determined in various situations by repeating the selection of the sample 500 times. The average response probability in the population is 0.01971. Therefore, the expected sample size in a self-selection survey is equal to 1971.

Figure 8.2 contains the results for the variable NEP (votes for National Elderly Party). The upper-left graph shows the distribution of the estimator for simple random samples of size 1971 from the target population. The vertical line denotes the population value to be estimated (25.6%). The estimator has a symmetric distribution around this value. This is a clear indication that the estimator is unbiased.

The upper-right graph shows what happens if samples are selected by means of self-selection. The shape of the distribution remains more or less the same, but the distribution as a whole has shifted to the left. All values of the estimator are systematically too low. The expected value of the estimator is only 20.5%. The estimator is biased. The explanation of this bias is simple: Relative few elderly are active Internet users. Therefore, they are under-represented in the samples. These are typically people who will vote for the NEP.

Figure 8.2. Results of the simulations for variable NEP



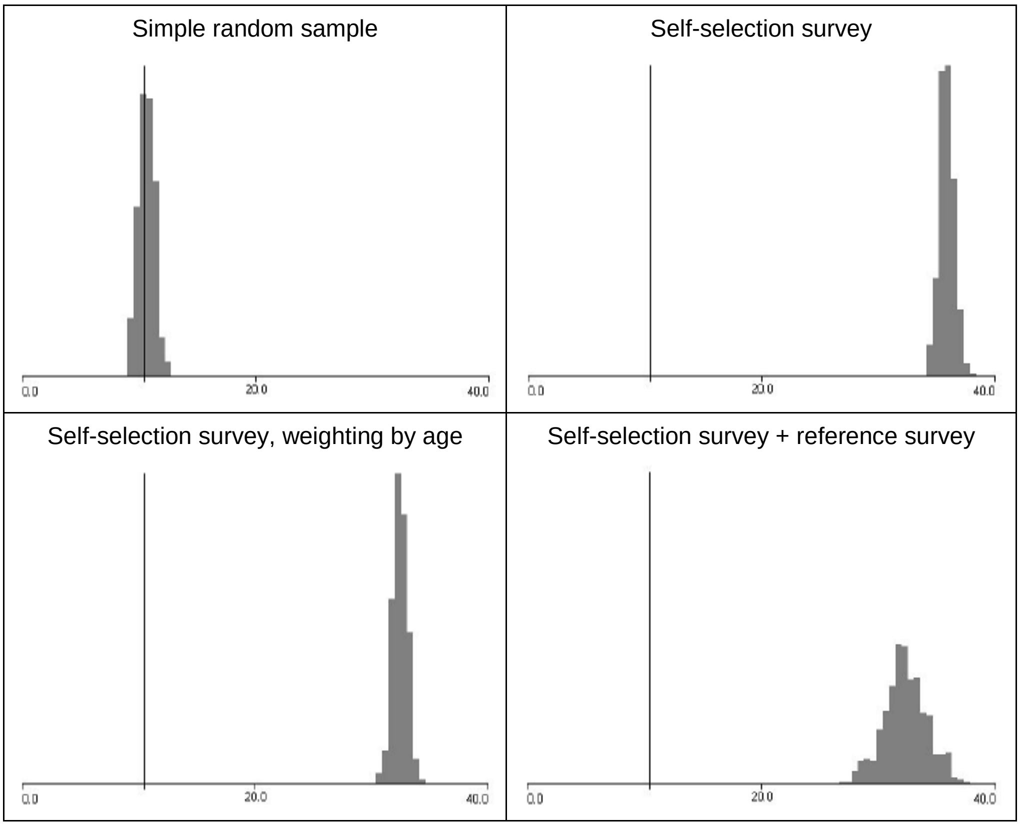
The lower-left graph in figure 8.2 shows the distribution of the estimator in case of post-stratification by age. The bias is removed. This is possible because this is a case of Missing At Random (MAR).

Post-stratification by age can only be applied if the distribution of age in the population is known. If this is not the case, one could consider to conduct a small (*m* = 100) reference survey, in which this population distribution is estimated unbiasedly. The lower-right graph in figure 8.2 shows what happens in this case. The bias is removed but at the cost of a substantial increase in variance.

Figure 8.3 shows the results for the variable NIP (votes for New Internet Party). The upper-left graph shows the distribution of the estimator for simple random samples of size 1971 from the target population. The vertical line denotes the population value to be estimated (10.5%). Since the estimator has a symmetric distribution around this value, it is clear that the estimator is unbiased.

The upper-right graph shows what happens if samples are selected by means of self-selection. The distribution has shifted to the right considerably. All values of the estimator are systematically too high. The expected value of the estimator is now 35.6%. The estimator is severely biased. The explanation of this bias is straightforward: voters for the NIP are over-represented in the self-selection samples.

Figure 8.3. Results of the simulations for variable NIP



Simple random sample

Self-selection survey

Self-selection survey, weighting by age

Self-selection survey + reference survey

The lower-left graph in figure 8.3 shows the effect of post-stratification by age. Only a small part of the bias is removed. Weighting is not successful. This is not surprising as there is a direct relationship between voting for the NIP and use of Internet. This is a case of NMAR.

Also in this case one can consider conducting a small reference survey if the population distribution of age is not available. The lower-right graph in figure 8.3 shows what happens in this case. Only a small part of the bias is removed and at the same time there is a substantial increase in variance.

The following conclusion can be drawn from this simulation study:

- If Missing At Random (MAR) or Not Missing At Random (NMAR) applies to survey participation, estimates based on a self-selection web survey will be biased;
- There is no guarantee that weighting will remove the bias. This correction technique will only work in case of Missing At Random (MAR), and the proper auxiliary variables are used for weighting;
- A reference survey will only be effective in removing the bias if Missing At Random (MAR) applies, and the proper auxiliary variable are measured;
- Use of a small reference survey will always substantially increase the variance of estimators.

9. Discussion and conclusions

This paper discussed some of the methodological problems caused by self-selection in web surveys. The underlying question is whether such a survey can be used as a data collection instrument for making valid inference about a target population. Costs and timeliness are important arguments in favour of web surveys. However, this paper concentrated on quality aspects like unbiasedness and accuracy of estimates.

It was shown that self-selection can cause estimates of population characteristics to be biased. This seems to be similar to the effect of nonresponse in traditional probability sampling based surveys. However, it was shown that the bias in self-selection surveys can be substantially larger. Depending on the response rate in a web survey, the bias can in a worst case situation even be more than 13 times as large.

Weighting techniques (including propensity weighting) can help to reduce the bias, but only if the sample selection mechanism satisfies the Missing at Random (MAR) condition. This is a strong assumption. It requires weighting variables that show a strong relationship with the target variables of the survey and the response probabilities. Often such variables are not available.

Sometimes a reference survey is used as a means to obtain the proper weighting variables. Indeed, this approach can be successful if such variables can be measured both in the web survey and in the reference survey. There are some reports that webographics variables seem to work well. These attitudinal or lifestyle variables seem to be capable of explaining response behaviour. They measure activities of respondents (e.g. reading) and perceptions about possible violations of privacy.

Schonlau et al. (2007) show that use of these webographics variables in propensity weighting may work, but not always.

One of the advantages of a reference survey is that the best auxiliary variables can be selected for weighting. This will make correction more effective. A disadvantage of a reference survey is that it results in large standard errors. So a reference survey reduces the bias at the cost of a loss in precision. The effective sample size will be smaller than the realised sample size. For example, the effective sample size in the simulation was just under 300. So a web survey of size almost 2000 produced estimates with a precision that also could have been obtained with a simple random sample of size 300.

Psychographic variables are used as auxiliary variables in some web surveys. Although they may be correlated with the target variables of the survey, they are attitudinal variables, and therefore may be subject to large measurement errors.

The reference survey only works well if it is a real probability sample without nonresponse, or with ignorable nonresponse (MCAR). This condition may be hard to satisfy in practical situations. Almost every survey suffers from nonresponse. If reference survey estimates are biased due to nonresponse, the web survey bias is replaced by a reference survey bias. This does not really help to solve the problem.

Reference surveys will be carried out in a mode other than CAWI. This means there may be mode effects that have an impact on estimates. Needless to say that a reference survey will dramatically increase survey costs.

10. References

Bethlehem, J.G. (1988), Reduction of the nonresponse bias through regression estimation. *Journal of Official Statistics* 4, pp. 251-260.

Bethlehem, J.G. (2002), Weighting Nonresponse Adjustments Based on Auxiliary Information. In: Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A. (Eds): *Survey Nonresponse*. Wiley, New York.

Bethlehem, J.G. (2007), *Reducing the bias of web survey based estimates*. Discussion paper 07001, Statistics Netherlands, Voorburg/Heerlen, The Netherlands.

Börsch-Supan, A., Elsner, D., Faßbender, H., Kiefer, R., McFadden, D. and Winter, J. (2004), *Correcting the participation bias in an online survey*. Report, University of Munich, Germany.

Cochran, W.G. (1968), The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24, pp. 205-213.

Cochran, W.G. (1977), *Sampling Techniques*, Third Edition. John Wiley & Sons (1977).

Couper, M.P. (2000), Web surveys: A review of issues and approaches. *Public Opinion Quarterly* 64, pp. 464-494.

Couper, M.P., Baker, R.P., Bethlehem, J.G., Clark, C.Z.F., Martin, J., Nicholls II, W.L. and O’Reilly, J.M. (eds.) (1998), *Computer Assisted Survey Information Collection*. Wiley, New York.

Dillman, D A. and Bowker, D. (2001), The web questionnaire challenge to survey methodologists. In: Reips, U.D. and Bosnjak, M. (eds.), *Dimensions of Internet Science*, Pabst Science Publishers, Lengerich, Germany.

Duffy, B, Smith, K., Terhanian, G. and Bremer, J (2005), Comparing data from online and face-to-face surveys. *International Journal of Market Research* 47, pp. 615-639.

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

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

Rosenbaum, P.R. & Rubin. D.B. (1983), The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, pp. 41-55.

Rosenbaum, P.R. & Rubin. D.B. (1984), Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79, pp. 516-524.

Schonlau, M., Fricker, R.D. & Elliott, M.N. (2003), *Conducting Research Surveys via E-mail and the Web*. Rand Corporation, Santa Monica, CA.

Schonlau, M., Zapert, K., Payne Simon, L., Haynes Sanstad, K., Marcus, S., Adams, J. Kan, H., Turber, R. & Berry, S. (2004), A Comparison between responses from propensity-weighted web survey and an identical RDD survey. *Social Science Computer Review* 22, pp. 128-138.

Schonlau, M., Van Soest, A. and Kapteyn, A. (2007), Are “webographic” or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? *Survey Research Methods*, Vol. 1, No. 3, pp. 155-163.

Terhanian, G., R. Smith, J. Bremer, and R. K. Thomas (2001), Exploiting Analytical Advances: Minimizing the Biases Associated with Internet-Based Surveys of Non-Random Samples. *ARF/ESOMAR: Worldwide Online Measurement*, ESOMAR Publication Services, Vol. 248, 2001, pp. 247–272.

Appendix A. Proof of theorem

Theorem 1. *The variance of the estimator*

$$\bar{y}_{RS} = \sum_{h=1}^{L} \frac{m_h}{m} \bar{y}_h$$

is equal to

$$V(\bar{y}_{RS}) = \frac{1}{m}\sum_{h=1}^{L} W_h \left(\bar{Y}_h^* - \tilde{Y}^*\right)^2 + \frac{1}{m}\sum_{h=1}^{L} W_h \left(1 - W_h\right) V(\bar{y}_h) + \sum_{h=1}^{L} W_h^2 V(\bar{y}_h)$$

Proof: It is assumed that the vector (m_1, m_2, ..., m_L)′ follows a multinomial distribution with probabilities W_1, W_2, ..., W_L. Consequently, the variance of m_h / m is equal to W_h (1 - W_h) / m. Since elements are selected without replacement in the reference survey, (m_1, m_2, ..., m_L)′ formally follows a multivariate hypergeometric distribution. However, for a sample from a large population both distributions are approximately equal.

Also note that the random variables m_h and $\bar{y}_h$ are independent, because they are computed using data from different (i.e. independent) surveys.

The variance of the estimator can be written as

$$V(\bar{y}_{RS}) = V\left(\sum_{h=1}^{L} \frac{m_h}{m} \bar{y}_h\right) =$$

$$\sum_{h=1}^{L} V\left(\frac{m_h}{m}\bar{y}_h\right) + 2\sum_{h=1}^{L}\sum_{g=h+1}^{L} C\left(\frac{m_h}{m}\bar{y}_h, \frac{m_g}{m}\bar{y}_g\right).$$

Let *I* denote the probability distribution for the web survey and *P* the distribution for the reference survey. The variance in the first component is now equal to

$$V\left(\frac{m_h}{m}\bar{y}_h\right) = E_P\left(V_I\left(\frac{m_h}{m}\bar{y}_h \mid P\right)\right) + V_P\left(E_I\left(\frac{m_h}{m}\bar{y}_h \mid P\right)\right) =$$

$$= E_P\left(\left(\frac{m_h}{m}\right)^2 V\left(\bar{y}_h\right)\right) + V_P\left(\frac{m_h}{m}\bar{Y}_h^*\right) =$$

$$= \left(\frac{W_h(1 - W_h)}{m} + W_h^2\right) V\left(\bar{y}_h\right) + \frac{W_h(1 - W_h)}{m}\left(\bar{Y}_h^*\right)^2.$$

The covariance term $C\left(\frac{m_h}{m}\bar{y}_h, \frac{m_g}{m}\bar{y}_g\right)$ can be written as

$$C\left(\frac{m_h}{m}\bar{y}_h, \frac{m_g}{m}\bar{y}_g\right) =$$

$$= E_P\left(C_I\left(\frac{m_h}{m}\bar{y}_h, \frac{m_g}{m}\bar{y}_g \mid P\right)\right) + C_P\left(E_I\left(\frac{m_h}{m}\bar{y}_h \mid P\right), E_I\left(\frac{m_g}{m}\bar{y}_g \mid P\right)\right).$$

Due to the nature of Poisson sampling, the stratum mean $\bar{y}_h$ and $\bar{y}_g$ are independent. Therefore

$$C_I\left(\frac{m_h}{m}\bar{y}_h, \frac{m_g}{m}\bar{y}_g \mid P\right) = \frac{m_h}{m}\frac{m_g}{m} C_I\left(\bar{y}_h, \bar{y}_g \mid P\right) = 0.$$

Because (m_1, m_2, …, m_L)' follows a multinomial distribution, the covariance of m_h and m_g is equal to $C(m_h, m_g) = -mW_hW_g$. Therefore

$$C_P\left(E_I\left(\frac{m_h}{m}\bar{y}_h \mid P\right), E_I\left(\frac{m_g}{m}\bar{y}_g \mid P\right)\right) = C_P\left(\frac{m_h}{m}\bar{Y}_h^*, \frac{m_g}{m}\bar{Y}_g^*\right) = -\frac{1}{m}W_hW_g\bar{Y}_h^*\bar{Y}_g^*.$$

Since

$$\sum_{h=1}^{L} V\left(\frac{m_h}{m}\bar{y}_h\right) = \sum_{h=1}^{L}\left(\frac{W_h(1-W_h)}{m}V\left(\bar{y}_h\right) + \frac{W_h(1-W_h)}{m}\left[\bar{Y}_h^*\right]^2 + W_h^2 V\left(\bar{y}_h\right)\right)$$

and

$$2\sum_{h=1}^{L}\sum_{g=h+1}^{L} C\left(\frac{m_h}{m}\bar{y}_h, \frac{m_g}{m}\bar{y}_g\right) = -\frac{2}{m}\sum_{h=1}^{L}\sum_{g=h+1}^{L} W_hW_g\bar{Y}_h^*\bar{Y}_g^* =$$

$$= \frac{1}{m}\left[\sum_{h=1}^{L} W_h^2\left(\bar{Y}_h^*\right)^2 - \left(\sum_{h=1}^{L} W_h\bar{Y}_h^*\right)^2\right]$$

if follows that

$$V(\bar{y}_{RS}) = V\left(\sum_{h=1}^{L}\frac{m_h}{m}\bar{y}_h\right) =$$

$$= \frac{1}{m}\sum_{h=1}^{L} W_h(1-W_h)V\left(\bar{y}_h\right) + \frac{1}{m}\sum_{h=1}^{L} W_h(1-W_h)\left(\bar{Y}_h^*\right)^2 +$$

$$+ \sum_{h=1}^{L} W_h^2 V\left(\bar{y}_h\right) + \frac{1}{m}\left[\sum_{h=1}^{L} W_h^2\left(\bar{Y}_h^*\right)^2 - \left(\sum_{h=1}^{L} W_h\bar{Y}_h^*\right)^2\right] =$$

$$= \frac{1}{m}\sum_{h=1}^{L} W_h\left(\bar{Y}_h^* - \tilde{Y}^*\right)^2 + \frac{1}{m}\sum_{h=1}^{L} W_h(1-W_h)V(\bar{y}_h) + \sum_{h=1}^{L} W_h^2 V(\bar{y}_h).$$

This completes the proof.