Reducing the bias of web survey based estimates

Discussion paper 07001

Jelke Bethlehem

Statistics Netherlands

Voorburg/Heerlen, January 2007

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

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

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg
The Netherlands

Printed by
Statistics Netherlands - Facility Services

Cover design
WAT ontwerpers, Utrecht

Information
E-mail: infoservice@cbs.nl
Via contact form: www.cbs.nl/infoservice

Where to order
E-mail: verkoop@cbs.nl

Internet
http://www.cbs.nl

Key figure: X-10
ISSN: 1572-0314
Production code: 6008307001

Statistics Netherlands

REDUCING THE BIAS OF WEB SURVEY BASED ESTIMATES

Summary: At first sight, web surveys seem to be an interesting and attractive means of data collection. They provide simple, cheap and fast access to a large group of people, and they offer all kinds of new possibilities such as use of sound, video and animation. However, there is another side to this coin. Due to methodological problems, outcomes of web surveys may be severally biased. Specific groups in the populations are under-represented among Internet users. Moreover, many web surveys apply self-selection of respondents instead of proper probability samples. This leads to a lack of representativity and thus to biased estimates. This paper describes some of the methodological problems, and explores the effect of various correction techniques (adjustment weighting and use of reference surveys) on the quality of web survey based estimates.

Keywords: web survey, online survey, bias, representativity, adjustment weighting, reference survey

1. Web surveys

1.1 Trends in data collection

Collecting data using a survey is a complex, costly and time-consuming process. Traditionally, surveys were carried out using paper forms (PAPI). One of the problems of this mode of data collection was that data usually contain many errors. Therefore, extensive data editing was required to obtain data of acceptable quality. Data editing activities often consume a substantial part of the total survey budget. Rapid developments in information technology in the last decades of the previous century made it possible to use microcomputers for computer-assisted interviewing (CAI). This type of data collection has three major advantages: (1) It simplifies work of the interviewers, because they do not have to pay attention any more to choosing the correct route through the questionnaire, (2) it improves the quality of the collected data, because answers can be checked and corrected during the interview, and (3) it considerably reduces time needed to process the survey data. Thus it improves the timeliness of survey results and it reduces survey costs. More on the benefits of CAI can be found in Couper et al. (1998).

The rapid development of the Internet in the last decade has lead to a new type of computer-assisted interviewing: Computer Assisted Web Interviewing (CAWI). The questionnaire is designed as a website, which is accessed by respondents. Web surveys are almost always self-administered: respondents visit the website, and complete the questionnaire by answering the questions. Not surprisingly, survey

organisations use, or consider using, web surveys. At first sight, web surveys seem to have some attractive advantages:

- 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;
- Web surveys offer new, attractive possibilities, such as the use of multimedia (sound, pictures, animation and movies);

Thus, web surveys seem to be a fast, cheap and attractive means of collecting large amounts of data. However, there are methodological problems. These problems are partly caused by using the Internet for selecting respondents, and partly by using the web as a measuring instrument. If these problems are not seriously addressed, web surveys may result in low quality data by which no proper inference can be made with respect to the target population of the survey.

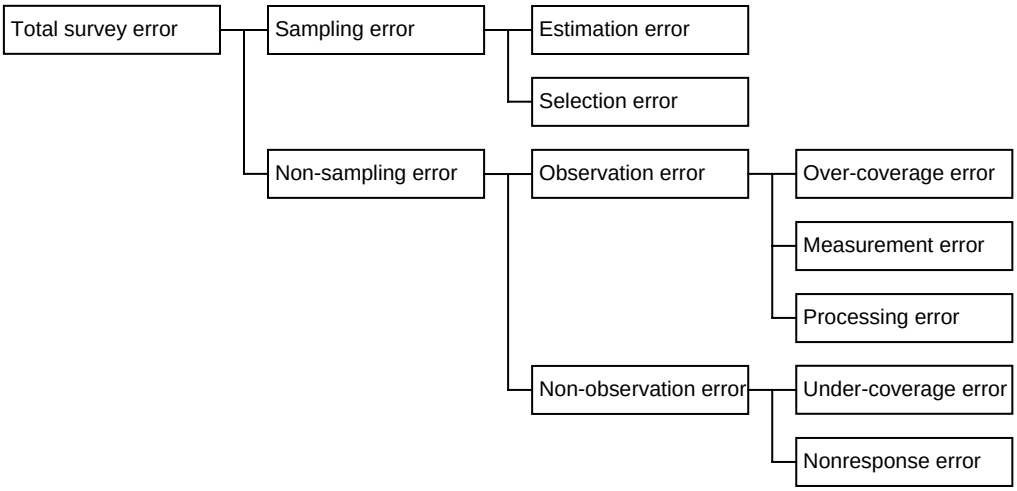
Cooper (2000) remarks that World Wide Web has made it possible for many organisations, other than traditional survey organisations, to conduct surveys. Several of these organisations are not aware of potential methodological risks involved in conducting web surveys. Therefore there are many bad surveys on the Internet. For respondents it is not always easy to distinguish the good from the bad. The ugly speak for themselves. The abundance of web surveys has a negative effect on response rates. Potential respondents are overwhelmed, and pull out. The effect is similar to that of telephone surveys, where the survey climate is spoiled by telemarketing activities.

This paper discusses some of the methodological issues that are specific for web surveys. Particularly, attention is paid to the effects of under-coverage and self-selection. Some theory is developed, and it is shown what the effects of some correction techniques can be. Practical implications are explored using data from a fictitious population.

1.2 About errors in surveys

One of the main objectives of a sample survey usually is to compute estimates of population characteristics. Such estimates will never be exactly equal to the population characteristics. There will always be some error. This error can have many causes. Bethlehem (1999) presents a taxonomy of possible causes. It is reproduced in figure 1.2.1. The taxonomy is a more extended version of the one given by Kish (1967).

Figure 1.2.1. Taxonomy of survey errors



The ultimate result of all these errors is a discrepancy between the survey estimate and the population characteristic to be estimated. This discrepancy is called the *total survey error*. Two broad categories can be distinguished contributing to this total error: sampling errors and non-sampling errors.

Sampling errors are introduced by the sampling design. They are due to the fact that estimates are based on a sample and not on a complete enumeration of the population. Sampling errors vanish if the complete population is observed. Since only a sample is available for computing population characteristics, and not the complete data set, one has to rely on estimates. The sampling error can be split in a selection error and an estimation error.

The *estimation error* denotes the effect caused by using a sample based on a random selection procedure. Every new selection of a sample will result in different elements, and thus in a different value of the estimator. The estimation error can be controlled through the sampling design. For example, by increasing the sample size, or by taking selection probabilities proportional to the values of some well chosen auxiliary variable, one can reduce the error in the estimate.

Estimation errors can have different effects in web surveys depending on the sampling mechanism used. If a proper sample is selected from a sampling frame that is independent of the Internet (e.g. a population register), the effect is the same as for other types of surveys. The estimation error can by quantified by applying probability theory. If sampling comes down to self-selection of respondents, there is also an estimation error, but it cannot be quantified since selection probabilities are unknown.

A *selection error* occurs when wrong selection probabilities are used. For example, true selection probabilities may differ from anticipated selection probabilities when elements have multiple occurrences in the sampling frame. Selection errors are hard to avoid without thorough investigation of the sampling frame.

A selection error can also occur in a web survey that is selected from a sampling frame. In the case of self-selection the selection probabilities are even unknown, so

there are no anticipated probabilities (unless a naive researcher assumes all selection probabilities are the same).

Non-sampling errors may even occur if the whole population is investigated. They denote errors made during the process of obtaining answers to questions asked. Non-sampling errors can be divided in observation errors and non-observation errors.

Observation errors are one form of non-sampling errors. They refer to errors made during the process of obtaining and recording answers. An *over-coverage error* means that elements are included in the survey that do not belong to the target population. A *measurement error* occurs when a respondent does not understand a question, or does not want to give the true answer, or if the interviewer makes an error in recording the answer. Also, interviewer effects, question wording effects, and memory effects belong to this group of errors. A measurement error causes a difference between the true value and the value processed in the survey. A *processing* error denotes an error made during data processing, e.g. data entry.

Over-coverage errors and measurement erros can occur in web survey if a proper sampling frame is used. Over-coverage can be even more a problem if the web survey is based on self-selection. Then there is no control at all over who completes the questionnaire.

Non-observation errors are errors made because intended measurements cannot be carried out. *Under coverage* occurs when elements of the target population do not have a corresponding entry in the sampling frame. These elements can and will never be contacted. Another non-observation error is *nonresponse*. It is the phenomenon that elements selected in the sample do not provide the required information.

Under-coverage is a serious problem if the Internet is used as a sampling frame and the target population contains people without Internet. Web surveys also suffer from nonresponse. A web survey questionnaire is a form of self-administered questionnaire. Therefore, web surveys have a potential of high nonresponse rates. An additional source of nonresponse problems are technical problems of respondents having to interact with the Internet, see e.g. Couper (2000), Dillman and Bowker (2001), Fricker and Schonlau (2002), and Heerwegh and Loosveldt (2002). Slow modem speeds, unreliable connections, high connection costs, low-end browsers, and unclear navigation instructions may frustrate respondents. This often results in respondents discontinuing the completion of the questionnaire. In order to keep the survey response up to an acceptable level, every measure must be taken to avoid these problems. This requires a careful design of web survey questionnaire instruments.

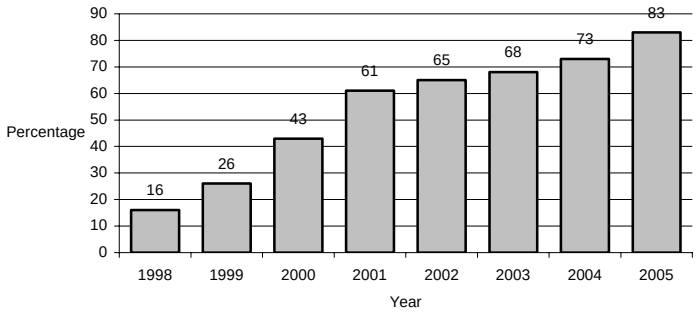
The taxonomy above makes clear that a lot can go wrong during a survey, and usually it does. Some errors can be avoided by taking preventive measures at the design stage. However, some errors will remain. The same applies to web surveys, and some problems are even more severe for web surveys. The next sections discuss coverage problems and selection problems for web surveys in more detail.

1.3 Coverage problems

The collection of all elements that can be contacted through the sampling frame is called the *frame population*. Since the sample is selected from the frame population, conclusions drawn from the survey data will apply to the frame population, and not necessarily to the target population. Coverage problems can arise when the frame population differs from the target population.

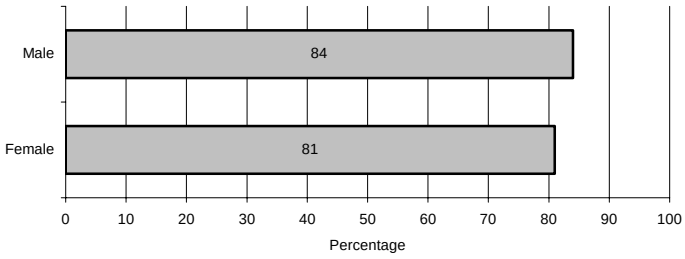
Under-coverage occurs when elements in the target population do not appear in the frame population. These elements have zero probability of being selected in the sample. Under-coverage can be a serious problem for Internet surveys. If the target population consists of all people with an Internet connection, there is no problem. However, usually the target population is wider than that. Then, under-coverage occurs due to the fact many people do not have access to the Internet.

Figure 1.3.1. Percentage of persons having Internet



In the Netherlands, the percentage of persons having an Internet connection at home increases from year to year, see figure 1.3.1 (source: www.cbs.nl). In seven years time, the percentage of Internet connections increased from 16% to 83%. Still, it is clear that not every household will have access to Internet in the near future.

Figure 1.3.2. Having Internet, by gender.



An analysis of data (source: www.cbs.nl) about Internet access in 2005 indicates that this is unevenly distributed over the population. Figure 1.3.2 shows the distribution by gender. Clearly, more males than females have access to the Internet.

Figure 1.3.3. Having Internet, by age.

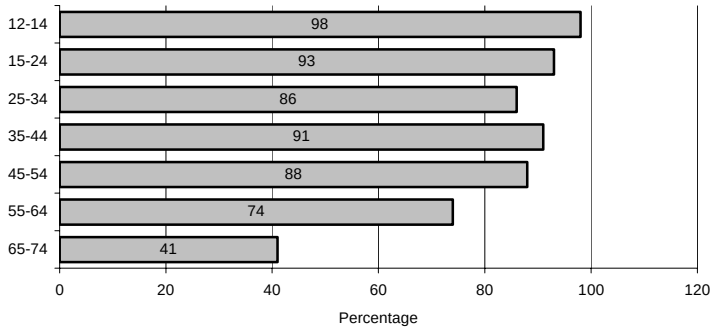
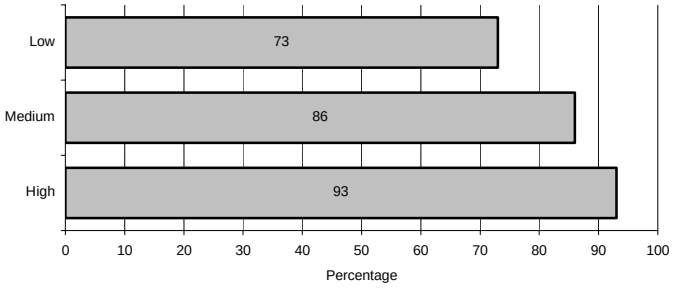


Figure 1.3.3 contains the percentage of people having Internet by age group (in 2005). The percentages of people having access to Internet at home decrease with age. Particularly, the people of age 55 and older will be very much under-represented when the Internet is used as a selection mechanism.

Figure 1.3.4 contains the percentage of people using the Internet by level of education (in 2005). It is clear that people with a higher level of education more frequently have Internet than people with a lower level of education.

Figure 1.2.4. Having Internet by level of education.



According to De Haan & Van ‘t Hof (2006) Internet access among non-native young people is much lower than among native young people: 91% of the young natives have access to Internet. This is 80% for young people from Surinam and Antilles, 68% for young people from Turkey and only 64% for young people from Morocco.

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).

It is clear that use of the Internet as a sampling frame will cause problems, because certain specific groups are substantially under-represented. Even if a proper probability sample is selected, the result will be a selective sample. Specific groups in the target population will not be able to fill in the (electronic) questionnaire form.

Note that ther is some similarity with CATI surveys in which telephone directories are used as a sampling frame. Here, people with out a phone are with an unlisted number will be excluded from the survey.

1.4 Selection problems

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 used, 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 surveys on the web are not based on probability sampling. A first group of web surveys consist of those surveys for which no selection at all takes place. The survey 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. The survey researcher is not in control of the selection process. These surveys will be called *self-selection surveys*.

At most, one could say that the target population of such a self-selection survey consist of people who have an Internet-connection and have a non-zero probability of visiting the website and participating in the survey. This is a not very well defined population and in most cases not the target population the researcher has in mind.

An additional problem of this type of web survey is that all selection probabilities are unknown. So, it is not possible to compute unbiased estimates for whatever the target population is.

A typical example of a self-selecting survey was the survey on the Dutch website *Kennisnet* (Knowledge net). This is a website for all those involved in education. More than 11,000 schools and other educational institutes use this website. The survey was an opinion poll for the general elections of 22 January 2003. Everybody, also those not involved in education, could participate in the poll. Table 1.4.1 contains both the official results (seats in parliament) of the election (column *Election*) and the results of this poll on the morning of the election day (column *Kennisnet*).

Table 1.4.1. The results of various opinion survey

	Election	Kennisnet	RTL4	SBS6	Nederland 1
Sample size		17,000	10,000	3,000	1,200
Seats in parliament:					
CDA (christian democrats)	44	29	24	42	42
LPF (populist party)	8	18	12	6	7
VVD (liberals)	28	24	38	28	28
PvdA (social democrats)	42	13	41	45	43
SP (socialists)	9	22	10	11	9
GL (green party)	8	26	9	6	8
D66 (liberal democrats)	6	4	7	5	6
Other parties	5	14	9	7	7
Mean Absolute Difference		12.5	5.3	1.8	0.8

The survey estimates were based on votes of approximately 17,000 people. No adjustment weighting was carried out. Although this is a large sample, it is clear that the survey results were no way near the true election results. The Mean Absolute

Difference (MAD) is equal to 12.5, which means that the estimated number of seats and the true number of seats differ on average by an amount of 12.5. This survey could certainly not be used for predicting election results.

Another example of a self-selection web survey was the election site of the Dutch Television channel RTL 4. It resembled to some extent the Kennisnet survey, but was targeted at a much wider audience. Again, the survey researcher had no control at all over who was voting. There was some protection, by means of cookies, against voting more than once. However, this also had the draw-back, that only one member of the family could participate.

Table 1.4.1 shows the survey results at noon on the day of the general elections (column *RTL4*). Figures were based on slightly over 10,000 votes. No weighting adjustment procedure was carried out. The results are better than that of the Kennisnet survey, since the MAD decreased from 12.5 to 5.3 However, deviations between estimates and true figures are still large, particularly for the large parties. Note that even a large sample size of over 10,000 people does not help to get accurate estimates.

A second group of web surveys consists of *access panels*. Such a panel is constructed by wide appeals on well-visited sites and Internet portals. At time of registration, basic demographic variables are asked. A large database of potential respondents is created in this way. For future surveys, samples are selected from this database. Only panel members can participate in these *web panel surveys*.

Also here the target population is unclear. One can only say that it consists of people who have an Internet-connection, who have a non-zero probability of being confronted with the invitation, and decide to participate in the panel.

Access panels have the advantages that values of basic demographic variables are available for all participants. So the distribution of these variables in the survey can be compared with their distribution in the population. Over- or under-representation of specific groups can be corrected by some kind of weighting adjustment technique. However, there is no guarantee that this leads to unbiased estimates.

An illustrative example of this approach was the general election survey of the Dutch commercial television channel *SBS6*. People were invited to participate in the survey. Those visitors of the site accepting the invitation were asked a number of questions related to socio-demographic characteristics and voting behaviour in the previous election. From the set of people that was obtained in this way, samples of size 3,000 were selected. Selection was carried out such that the sample was representative with respect to the social-demographic and voting characteristics. Selected people where asked for their voting behaviour in the coming election. Table 1.4.1 shows the results (column *SBS6*). The survey took place on the day before the general elections.

Although attempts have been made to create a representative sample, the results differ still from the final result. The MAD has decreased to 1.8, but is still substantial.

A better prediction would have been obtained with a true probability sample. Table 1.4.1 shows the results of a survey based on such a probability sample. It was carried out by the television channel *Nederland 1* in co-operation with the marketing agency *Interview-NSS*. A sample of size 1200 was selected by means of random digit dialling. The MAD has been reduced to 0.8.

The conclusion from the analysis above is that probability samples are a vital prerequisite for making proper inference about the target population of a survey. Even with a probability sample of only size 1,200 better results can be obtained than with a non-probability sample of size 10,000 or more.

Probability sampling has the additional advantages 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 considered to be representative for the Dutch population.

2. Sampling the Internet-population

2.1 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

$$Y = \frac{1}{N}\sum_{k=1}^{N} Y_k \qquad (2.1.1)$$

of the target variable Y.

The population U is divided into two sub-populations U_I of elements having access to Internet, and U_{NI} of elements not having access to the Internet. Associated with each element k is an indicator I_k, where $I_k = 1$ if element k has access to the Internet (and thus is an element of sub-population U_I), and $I_k = 0$ otherwise. The sub-population U_I will be called the *Internet-population*. Let

$$N_I = \sum_{k=1}^{N} I_k \qquad (2.1.2)$$

denote the size of sub-population U_I. Likewise, N_{NI} denotes the size of the sub-population U_{NI}, where $N_I + N_{NI} = N$.

The mean of the target variable for the elements in the Internet-population is equal to

$$\bar{Y}_I = \frac{1}{N_I} \sum_{k=1}^{N} I_k Y_k \tag{2.1.3}$$

2.2 A random sample from the Internet-population

The first situation to consider for a web survey is the more or less ideal case in which it is possible to select a random sample without replacement from the Internet-population. This would require a sampling frame listing all elements having access to the Internet. No such list exists, but there are ways to get close to such a situation. One way to do this is to select a random sample from a larger sampling frame (e.g. a population or address register), approach the selected people in a classical way (by mail, telephone, or face-to-face), and filter out only those people having access to the Internet. Next, selected people are provided with an Internet-address where they can fill in the questionnaire form. It is clear that initially such registers suffer from over-coverage, but with this approach every element in the Internet-population has a positive and known probability of being selected.

A random sample selected without replacement from the Internet-population can be represented by a series

$$a_1, a_2, ..., a_N \tag{2.2.1}$$

of N indicators, where the k-th indicator a_k assumes the value 1 if element k is selected, and otherwise it assumes the value 0, for k = 1, 2, ..., N. Note that always $a_k = 0$ for elements k outside the Internet-population.

The expected value $\pi_k = E(a_k)$ is the *first order inclusion probability* of element k. Horvitz and Thompson (1952) have shown that always an unbiased estimator of the population mean can be defined if all elements in the population have known, positive probability of being selected. The Horvitz-Thompson estimator for the mean of the Internet-population is defined by

$$\bar{y}_{HT} = \frac{1}{N_I} \sum_{k=1}^{N} a_k I_k \frac{Y_k}{\pi_k}, \tag{2.2.2}$$

Where by definition $Y_k / \pi_k = 0$ for all elements outside the Internet-population. In the case of a simple random sample from the Internet-population, all first order inclusion probabilities are equal to n / N_I. Therefore expression (2.2.2) reduces to

$$\bar{y}_I = \frac{1}{n} \sum_{k=1}^{N} a_k I_k Y_k \tag{2.2.3}$$

This estimator is an unbiased estimator of the mean $\bar{Y}_I$ of the Internet-population, but not necessarily of the mean $\bar{Y}$ of the target population. The bias is equal to

$$B(\bar{y}_{HT}) = E(\bar{y}_{HT}) - \bar{Y} = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N}(\bar{Y}_I - \bar{Y}_{NI}) \qquad (2.2.4)$$

The magnitude of this bias is determined by two factors. The first factor is the relative size N_{NI} / N of the sub-population without Internet. The bias will increase as a larger 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.

Presently, the size of the non-Internet population cannot be neglected in The Netherlands. Figure 1.2.1 shows that the percentage of people without Internet is rapidly decreasing, but still in the order of 20%.

Furthermore, there are substantial differences between these two sub-populations. The graphs in section 1.2 show that specific groups are under-represented in the Internet-population, e.g. the elderly, those with a low level of education, and ethnic minority groups. So, the conclusion is that generally a random sample from an Internet population will often not lead to unbiased estimates for the target population.

2.3 Self-selection from the Internet-population

For many web-surveys no proper random sample has been selected from the Internet-population. These surveys rely on self-selection of respondents. Participation 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 Internet-population has unknown probability ρ_k of participating in the survey, for k = 1, 2, ..., N_I. The responding elements can be denoted by a series

$$r_1, r_2, ..., r_N \qquad (2.3.1)$$

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. Not that sampling without replacement is assumed. The expected value $\rho_k = E(r_k)$ will be called the *response propensity* of element k. For sake of convenience we have also introduced response propensities for non-Internet-population elements. By definition the values of all these probabilities are 0.

The realised sample size is equal to

$$n = \sum_{k=1}^{N} r_k \qquad (2.3.2)$$

A naive researcher assuming that every element in the Internet-population has the same probability of being selected in the sample, will use the sample mean

$$\bar{y}_S = \frac{1}{n}\sum_{k=1}^{N} r_k Y_k \tag{2.3.3}$$

as an estimator for the population mean. The expected value of this estimator is approximately equal to

$$E(\bar{y}_S) \approx \bar{Y}_I^* = \frac{1}{N_I \bar{\rho}}\sum_{k=1}^{N} \rho_k I_k Y_k \tag{2.3.4}$$

where $\bar{\rho}$ is the mean of all response propensities in the Internet-population, see e.g. Bethlehem (1988). Generally, the expected value of the sample mean is not equal to the population mean of the Internet-population. The only situation in which the bias vanishes is that in which all response propensities in the Internet-population are equal. In terms of nonresponse correction theory, this comes down to Missing Completely Missing At Random (MCAR).

Indeed, in this case, self-selection does not lead to an unrepresentative sample because all elements have the same selection probability. Bethlehem (1988) shows that the bias of the sample mean (2.3.3) can be written as

$$B(\bar{y}_S) = E(\bar{y}_S) - \bar{Y}_I \approx \bar{Y}_I^* - \bar{Y}_I = \frac{C(\rho, Y)}{\bar{\rho}}, \tag{2.3.5}$$

in which

$$C(\rho, Y) = \frac{1}{N_I}\sum_{k=1}^{N} I_k(\rho_k - \bar{\rho})(Y_k - \bar{Y}) \tag{2.3.6}$$

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

- The average response propensity. The more likely people are to participate in the survey, the higher the average response propensity 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 propensities, 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.

In many cases, the objective of the survey is not to estimate the mean of the Internet-population, but the mean of the total population, the target population. In this case the bias of the sample mean is equal to

$$B(\bar{y}_S) = E(\bar{y}_S) - \bar{Y} = E(\bar{y}_S) - \bar{Y}_I + \bar{Y}_I - \bar{Y} =$$
$$= \frac{N_{NI}}{N}(\bar{Y}_I - \bar{Y}_{NI}) + \frac{C(\rho, Y)}{\bar{\rho}} \qquad (2.3.7)$$

The bias now consists of two terms: a bias caused by interviewing just the Internet-population instead of the complete target population (under-coverage bias) and a bias caused by self-selection of respondents in the Internet-population (self-selection bias). Theoretically, it is possible that these two biases compensate one another. If people without Internet resemble people with Internet that are less inclined to participate, the combined effects will produce a larger bias. Practical experiences suggest that this may often be the case. For example, suppose Y is a variable measuring the intensity of some activity on the Internet (surfing, playing on-line games). Then a positive correlation between Y and response propensities is not unlikely. Also the mean of Y for the Internet-population will be positive while the mean of the non-Internet-population will be 0. So, both bias terms have a positive value.

3. Weighting adjustment

3.1 Why 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 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. To correct this, adjustment weights are computed. Weights are assigned to all records of observed elements. Estimates of population characteristics can now be obtained by using the weighted values instead of the unweighted values. Weighting adjustment is used to correct surveys that are affected by nonresponse, see e.g. Bethlehem (2002).

This section explores the possibility to reduce the bias of web survey estimates. For sake of convenience it is assumed that a simple random sample has been selected from the Internet-population. Section 3.2 analyses the effects of post-stratification, where weights are computed using the distribution of auxiliary variables in the complete population. Section 3.3 investigates the situation where the population distribution of auxiliary variables is estimated using data from a small, true probability sample (a so-called *reference survey*). Section 3.4 explores the

possibilities improving estimates by using data obtained from a simple random sample from the non-Internet-population. Finally, section 3.5 discusses propensity weighting, a weighting technique often applied by commercial market research agencies.

3.2 Post-stratification

Post-stratification is a well-known and often used weighting method. To carry out post-stratification, one or more qualitative auxiliary variables are needed. Here, only one such variable is considered. The situation for more variables is not essentially different. 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 sample of size n is selected from the Internet-population. If n_h denotes the number of sample elements in stratum h, then $n = n_1 + n_2 + ... + n_L$. The values of the n_h are the result of a random selection process, so they are random variables. Note that since the sample is selected from the Internet-population, only elements in the sub-strata $U_I \cap U_h$ are observed (for $h = 1, 2, \ldots, L$).

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

$$w_k = \frac{N_h / N}{n_h / n} \tag{3.2.1}$$

The simple sample mean

$$\bar{y}_I = \frac{1}{n} \sum_{k=1}^{N} a_k I_k Y_k \tag{3.2.2}$$

is now replaced by the weighted sample mean

$$\bar{y}_{I,PS} = \frac{1}{n} \sum_{k=1}^{N} a_k w_k I_k Y_k \tag{3.2.3}$$

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

$$\bar{y}_{I,PS} = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{y}_I^{(h)} = \sum_{h=1}^{L} W_h \bar{y}_I^{(h)} , \tag{3.2.4}$$

where $\bar{y}_I^{(h)}$ 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

$$E(\bar{y}_{I,PS}) = \frac{1}{N} \sum_{h=1}^{L} N_h E(\bar{y}_I^{(h)}) = \sum_{h=1}^{L} W_h \bar{Y}_I^{(h)} = \tilde{Y}_I , \tag{3.2.5}$$

where $\bar{Y}_I^{(h)}$ is the mean of the target variable in stratum *h* of the Internet-population. 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

$$B(\bar{y}_{I,PS}) = E(\bar{y}_{I,PS}) - \bar{Y} = \tilde{\bar{Y}}_I - \bar{Y} = \sum_{h=1}^{L} W_h(\bar{Y}_I^{(h)} - \bar{Y}^{(h)}) = \\ = \sum_{h=1}^{L} W_h \frac{N_{NI,h}}{N_h}(\bar{Y}_I^{(h)} - \bar{Y}_{NI}^{(h)}), \tag{3.2.6}$$

were $N_{NI,h}$ is the number of elements in stratum *h* of the non-Internet-population.

The bias will be small if there is (on average) no difference between elements with and without Internet within the strata. This is the case if there is a strong relationship between the target variable *Y* and the stratification variable *X*. 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. In nonresponse correction terminology, this situation comes down to Missing At Random (MAR).

In conclusion we can say the 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 be strongly correlated with all target variables.

Unfortunately, such variables are not very often 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

$$V(\bar{y}_{PS}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}^{(h)}). \tag{3.2.7}$$

Cochran (1977) shows that in the case of a simple random sampling from the complete population, this expression is equal to

$$V(\bar{y}_{PS}) = \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2 + \frac{1}{n^2}\sum_{h=1}^{L}(1-W_h)S_h^2, \tag{3.2.8}$$

where $f = n / N$ and S_h^2 is the variance in stratum *h*. If the strata are homogeneous with respect to *Y*, the variance of estimator will be small.

In the case of a simple random sample from the Internet-population, the variance of the estimator (3.2.4) becomes

$$V(\bar{y}_{I,PS}) = \sum_{h=1}^{L} W_h^2 \left(\frac{1}{nW_{I,h}} + \frac{1-W_{I,h}}{(nW_{I,h})^2} - \frac{1}{N_{I,h}} \right) S_{I,h}^2, \tag{3.2.9}$$

where $N_{I,h}$ is the size of stratum h in the Internet-population, $W_{I,h} = N_{I,h} / N_I$ and $S^2_{I,h}$ is the variance in stratum h of the Internet-population.

3.3 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). Under the assumption of no nonresponse, 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 unbiasedness of the reference survey in improving estimates.

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

$$\bar{y}_{I,RS} = \sum_{h=1}^{L} \frac{m_h}{m} \bar{y}_I^{(h)} \tag{3.3.1}$$

where $\bar{y}_I^{(h)}$ is the web survey based estimate for the mean of stratum h of the Internet-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

$$E(\bar{y}_{I,RS}) = E_I E_P(\bar{y}_{I,RS} \mid I) = E_I\left(\sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_I^{(h)}\right) = \sum_{h=1}^{L} W_h \bar{Y}_I^{(h)} = \tilde{Y}_I, \tag{3.3.2}$$

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 (3.2.4). The bias of this estimator is equal to

$$B(\bar{y}_{I,RS}) =_I E(\bar{y}_{I,RS}) - \bar{Y} = \tilde{Y}_I - \bar{Y} = \sum_{h=1}^{L} W_h \left(\bar{Y}_I^{(h)} - \bar{Y}^{(h)}\right) =$$
$$= \sum_{h=1}^{L} W_h \frac{N_{NI,h}}{N_h} \left(\bar{Y}_I^{(h)} - \bar{Y}_{NI}^{(h)}\right). \tag{3.3.3}$$

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. This implies that if the stratum means for the Internet-population and for the target population do not differ much, this results in a small bias. So, using a reference survey with the proper auxiliary variables can substantially reduce the bias of web survey estimates.

Note that the expression for the bias of the reference survey estimator is equal to that of the post-stratification estimator. 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

$$V(\bar{y}_{I,RS}) = \frac{1}{m} \sum_{h=1}^{L} W_h \left(\bar{Y}_I^{(h)} - \tilde{Y}_I\right)^2 + \frac{1}{m} \sum_{h=1}^{L} W_h (1 - W_h) V(\bar{y}_I^{(h)}) + \sum_{h=1}^{L} W_h^2 V(\bar{y}_I^{(h)}) \tag{3.3.4}$$

The proof is given in the appendix. The quantity $\bar{y}_I^{(h)}$ is measured in the web survey. Therefore its variance $V(\bar{y}_I^{(h)})$ will be of the order $1/n$. 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 $1/mn$, and the third term of order $1/n$. Since 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. So, the large number of observations in the web survey does not help to produce accurate estimates. One could say that the reference survey approach reduces the bias of estimates at the cost of a higher variance. See section 4 for an example showing this effect.

3.4 Sampling the non-Internet population

The fundamental problem of web surveys is that persons without Internet are excluded from the survey. This problem could be solved by selecting a stratified sample. The target population is assumed to consist of two strata: the Internet-population U_I of size N_I and the non-Internet-population U_{NI} of size N_{NI}.

To be able to compute an unbiased estimate, a simple random sample must be selected from both strata. The web survey provides the data about the Internet-stratum. If this is a random sample with equal probabilities, the sample mean

$$\bar{y}_I = \frac{1}{n}\sum_{k=1}^{N} a_k I_k Y_k \tag{3.4.1}$$

is an unbiased estimator of the mean of the Internet-population.

Now suppose a random sample (with equal probabilities) of size m is selected from the non-Internet-stratum. Of course, there is no sampling frame for this population. This problem can be avoided by selecting a sample from the complete target population (a reference survey) and only using people without Internet access. Selected people with Internet access can be added to the large web sample, but this will have no substantial effect on estimators. The sample mean of the non-Internet sample is denoted by

$$\bar{y}_{NI} = \frac{1}{m}\sum_{k=1}^{N} b_k (1 - I_k) Y_k \ , \tag{3.4.2}$$

where the indicator b_k denotes whether or not element k is selected in the reference survey, and

$$m = \sum_{k=1}^{N} b_k (1 - I_k) \, . \tag{3.4.3}$$

The stratification estimator is now defined by

$$\bar{y}_{ST} = \frac{N_I}{N}\bar{y}_I + \frac{N_{NI}}{N}\bar{y}_{NI} \, . \tag{3.4.4}$$

This is an unbiased estimator for the mean of the target population. Application of this estimator assumes the size N_I of the Internet-population and the size N_{NI} of the non-Internet-population to be known. The variance of the estimator is equal to

$$V(\bar{y}_{ST}) = \left(\frac{N_I}{N}\right)^2 V(\bar{y}_I) + \left(\frac{N_{NI}}{N}\right)^2 V(\bar{y}_{NI}) \, . \tag{3.4.5}$$

The variance of the sample mean in the Internet stratum is of order $1/n$ and the variance in the non-Internet stratum is of order $1/m$. Since m will be much smaller than n in practical situation, and the relative sizes of the Internet-population and the non-Internet-population de not differ that much, the second term will determine the magnitude of the variance. So the advantages of the large sample size of the web survey are for a great part lost by the bias correction.

3.5 Propensity weighting

Propensity weighting is used by several commercial market research organisations to correct for a possible bias in their web surveys. The original idea behind propensity weighting goes back to Rosenbaum & Rubin (1983, 1984). They developed a technique for comparing two populations. They attempt to make the two

populations comparable by simultaneously controlling for all variables that were thought to explain the differences.

In the case of a web survey, there are also two populations: those who participate in the web survey (if asked), and those who will not participate.

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 estimating the probability (propensity score) of participating, given the values of the explanatory variables.

Application of propensity weighting assumes some kind of random process determining whether or not someone participates in the web survey. Each element k in the population has a certain, unknown probability ρ_k of participating, 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 $\rho(X)$ is the conditional probability that a person with observed characteristics X participates, i.e.

$$\rho(X) = P(R = 1 \mid X)$$

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 that was introduced in section 3.2. The propensity score is often modelled using a logit model:

$$log\left(\frac{\rho(X_k)}{1-\rho(X_k)}\right) = \alpha + \beta_k' X_k + \varepsilon_k$$

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?”.

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 to allow for precise estimation. A small sample size results in large standard errors of estimates. This is similar to the situation described in section 3.3.

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).

4. A simulation study

4.1 A fictitious population

To explore how effective correction techniques can be, a simulation study was carried out. A fictitious population was constructed. The relationships between variables involved was such that it could resemble a real life situation. With respect to the Internet-population, both Missing At Random (MAR) and Not Missing At Random (NMAR) occurred. The characteristics of various estimators (before and after correction) were computed, either analytically or based on a large number of simulations.

A fictitious population of 30,000 individuals was constructed. There were six variables:

- Age in three categories: Young (with probability 0.40), Middle aged (with probability 0.35) and Old (with probability 0.35).
- Level of education in three categories: Low (with probability 0.35), Middle (with probability 0.40) and High (with probability 0.35).

- Ethnic background in two categories: Native (with probability 0.85) and Non-native (with probability 0.15).
- Having access to Internet with two categories Yes and No. The probability of having access to Internet depended on the two variables Age and Ethnic background. For natives, the probabilities were 0.90 (for Young), 0.70 (for Middle aged) and 0.50 (for Old). So, Internet access decreases with age. For Non-natives, these probabilities were 0.20 (for Young), 0.10 (for Middle aged) and 0.00 (for Old). These probabilities reflect the much lower Internet access among non-natives.
- Frequency of downloading music from the Internet (per week). Of course, this frequency is 0 for those without Internet access. If someone had Internet access, the frequency depended on age. For young people, the frequency was a randomly selected integer from the interval [0, 20]. For middle age people, the interval was [0, 10] and for the elderly [0, 5]. So, downloading music decreased with age.
- Frequency of reading newspapers (per week). This frequency only depended on age. For young people, the frequency was a randomly selected integer from the interval [0, 3]. For middle age people, the interval was [0, 7] and for the elderly [4, 7]. So, reading newspapers increased with age.

Downloading music and reading newspapers were used as target variables. If a web surveys is carried out, downloading music will suffer from Not Missing At Random (NMAR). There is a direct relationship between this target variable and missingness due to having no access to Internet. So estimates should be biased and correction techniques should not work.

Reading newspapers suffers from Missing At Random (MAR). There is a direct relationship between this target variable and the auxiliary variable age. And there is a direct relationship between age and missingness due to having no Internet access. Estimates will be biased, but correction using Age should help to reduce a bias.

4.2 Simulation results

Tables 4.2.1 and 4.2.2 contain the simulation results for the target variable downloading music. The population mean of this variable is 4.512. So on average 4.5 times a week people download music from the Internet.

The characteristics of the sampling distribution were estimated either analytically or by repeating the sample selection process 10,000 times.

Columns 2 and 3 of table 4.2.1 show that in the ideal situation of drawing a simple random sample from the target population, the sample mean is an unbiased estimator. As can be expected, the standard error of the estimator decreases with an increasing sample size.

Table 4.2.1. Simulation results for downloading music – part 1

	Target population		Internet population		Weighting by age	
Sample size	Mean	Standard error	Mean	Standard error	Mean	Standard error
500	4.512	0.249	7.090	0.246	6.385	0.179
1000	4.512	0.174	7.090	0.172	6.385	0.125
1500	4.512	0.141	7.090	0.138	6.385	0.100
2000	4.512	0.121	7.090	0.118	6.385	0.086
2500	4.512	0.107	7.090	0.014	6.385	0.076
3000	4.512	0.097	7.090	0.094	6.385	0.068

Columns 4 and 5 show what happens if a simple random sample is selected just from the Internet-population. The sample mean is now an unbiased estimator for the mean in the Internet-population. Apparently this mean is much larger (7.090) than the mean in the target population (4.152). So a web survey results in an estimate that is substantially biased.

Also here the standard error is smaller as the sample size increases. This makes things worse with respect to confidence intervals. In the case of drawing a simple random sample from the target population, the sample mean has (approximately) a normal distribution. Then the 95% confidence interval for the population mean is equal to

$$I = \left((\bar{y} - 1.96 \times S(\bar{y}) \, ; \, \bar{y} + 1.96 \times S(\bar{y}) \right), \tag{4.2.1}$$

where $S(\bar{y})$ is the standard error of the sample mean. The probability that this interval contains the true value, is by definition (approximately) equal to

$$P(\bar{Y} \in I) = 0.95 \, . \tag{4.2.2}$$

This is the *confidence level* of the confidence interval. In the case of selecting a simple random sample from the Internet-population, the sample mean $\bar{y}_I$ has to be used. If we denote this confidence interval by I_I, Bethlehem and Kersten (1985) show that the confidence level is now equal to

$$P(\bar{Y} \in I_I) = \Phi\left(1.96 - \frac{B(\bar{y}_I)}{S(\bar{y}_I)} \right) - \Phi\left(-1.96 - \frac{B(\bar{y}_I)}{S(\bar{y}_I)} \right), \tag{4.2.3}$$

in which Φ is the standard normal distribution function, and $B(\bar{y}_I) / S(\bar{y}_I)$ is the *relative bias* of the estimator. The confidence reaches its maximum of 0.95 if the relative bias is equal to 0. As the relative bias increases, the confidence level decreases. For example, for a relative bias of 1 (the bias is equal to the standard error), the confidence level is only 0.83.

Going back to the Internet survey in table 4.2.1, the relative bias for a sample of size 500 is equal to (7.090 – 4.512) / 0.249 = 10.353. So, the bias is more than 10 times the standard error. This comes down to a confidence level of 0.00. This means that the confidence interval will (almost) never contain the true population value. For a sample of size 3,000 the bias remains the same while the standard error reduces to

0.097. The relative bias is now equal to 26.577, which means that the bias is more than 26 times the standard error. So, the situation is even worse.

Sometimes advocates of web surveys claim that due to their large sample size web surveys just have to be representative. The results in table 4.2.1 confirm once more that it does not help to increase the sample size to reduce the bias.

Columns 6 and 7 in table 4.2.1 show the effect of applying post-stratification by age class. The expected value of the estimator reduces from to 7.090 to 6.385, but it is still far from the true value 4.512. In fact, the bias of the unweighted sample mean is caused by two variables. The most important one is having Internet access, but also age has some effect. Weighting reduces only that part of the bias that is caused by age.

Note that the standard error of the weighted estimate is somewhat smaller than that of the unweighted estimate. This is caused by the fact that the apparently the strata are a little more homogeneous with respect to music download behaviour than the Internet population as a whole.

Table 4.2.2 contains the result of applying the reference survey approach to correct for the bias. Simulations have been carried out for reference surveys of sizes 100, 200 and 400.

Table 4.2.2. Simulation results for downloading music – part 2

	Reference survey Size = 100		Reference survey Size = 200		Reference survey Size = 400	
Sample size	Mean	Standard error	Mean	Standard error	Mean	Standard error
500	6.385	0.359	6.385	0.283	6.385	0.237
1000	6.385	0.335	6.385	0.252	6.385	0.199
1500	6.385	0.326	6.385	0.241	6.385	0.185
2000	6.385	0.322	6.385	0.235	6.385	0.177
2500	6.385	0.319	6.385	0.232	6.385	0.172
3000	6.385	0.317	6.385	0.230	6.385	0.169

It is clear from the table that a reference survey is not able to remove the bias. The expected value of the estimator is 6.385. This value is identical to that of the post-stratification estimator. This will come as no surprise as both estimators theoretically have the same expected value.

The standard error of the reference survey estimate is substantially larger than that of the post-stratification estimator. This was also to be expected. It was shown in section 3.3 that the standard error of the reference survey estimate is mainly determined by the size of the reference survey sample. For example, the standard error for a web survey of size 500 combined with a reference survey sample size of 100 is equal to 0.359. If the sample size of the web survey is increased from 500 to 3,000 the standard error only reduces form 0.359 to 0.317

A standard error of approximately 0.317 could also have been obtained by drawing a simple random sample of size 300 from the target population. The conclusion can be

that a combination of a web survey and a reference survey of a total size of 3000 + 100 = 3100 cases will have the same accuracy as a simple random sample of size 300. From the viewpoint of accuracy, the reference survey approach is certainly not efficient.

Note that a standard error of 0.230 for a reference survey size of 200 comes down (in terms of accuracy) to a simple random sample of a size of approximately 600. And a standard error of 0.169 for a reference survey size of 400 comes down (in terms of accuracy) to a simple random sample of size 1200.

The simulation results show that for a variable that suffers from Not Missing At Random (NMAR) estimates will be biased and correction techniques will not be able to remove this bias.

Applying the stratification approach for any target variable directly related to the Interenet (like downloading music) does not make sense, of course. Any estimate for the non-Internet-population will always have the value 0 because all values of the target variable are equal to 0 in that stratum. An estimate for the target population can simply be computed by multiplying an estimate for the mean in the Internet-population by N_I / N. And the variance of this new estimator is obtained by multiplying the variance of Internet-population estimator by $(N_I / N)^2$.

The second target variable to be analysed is reading newspapers. Tables 4.2.3 and 4.2.4 contain the results for this variable. The population mean is 3.205. So on average people read a newspaper 3.2 times a week.

Columns 2 and 3 of table 4.2.3 show that in the ideal situation of drawing a simple random sample from the target population, the sample mean is an unbiased estimator. As can be expected, the standard error of the estimator decreases with an increasing sample size.

Table 4.2.3. Simulation results for reading newspapers – part 1

	Target population		Internet population		Weighting by age	
Sample size	Mean	Standard error	Mean	Standard error	Mean	Standard error
500	3.205	0.101	2.834	0.097	3.205	0.074
1000	3.205	0.071	2.834	0.067	3.205	0.051
1500	3.205	0.057	2.834	0.054	3.205	0.041
2000	3.205	0.049	2.834	0.046	3.205	0.035
2500	3.205	0.044	2.834	0.041	3.205	0.031
3000	3.205	0.039	2.834	0.037	3.205	0.028

Columns 4 and 5 show what happens if a simple random sample is selected just from the Internet population. The sample mean is now an unbiased estimator for the mean in the Internet-population. Apparently this mean (2.834) is smaller than the mean in the target population (3.205). So a web survey would result in an estimate that is substantially biased.

Also here the standard error is smaller as the sample size increases. This makes things worse with respect to confidence intervals. The relative bias for a sample of size 500 is equal to (2.834 – 3.205) / 0.101 = -3.673. So, the bias is almost 4 times the standard error. This comes down to a confidence level of 0.04 (instead of 0.95). This means that the confidence interval will probably not contain the true population value. For a sample of size 3,000 the bias remains the same while the standard error reduces to 0.039. The relative bias is now equal to -9.513, which means that the bias is more than 9 times the standard error, which comes down to a confidence level of 0.

Columns 6 and 7 in table 4.2.3 show the effects of applying post-stratification by age class. The expected value of the estimator changes from 2.834 to 3.205. The bias is completely removed, as could be expected. The bias was caused by a relationship between having access to Internet and age, which is a case of MAR. So, correction using age should indeed remove the bias.

Note that the standard error of the weighted estimate is somewhat smaller than that of the unweighted estimate. This is caused by the fact that the apparently the age strata are a little more homogeneous with respect to reading newspapers than the Internet population as a whole.

Table 4.2.4 contains the result of applying the reference survey approach to correct for the bias. Simulations have been carried out for reference surveys of sizes 100, 200 and 400.

Table 4.2.4. Simulation results for reading newspapers – part 2

	Reference survey Size = 100		Reference survey Size = 200		Reference survey Size = 400	
Sample size	Mean	Standard error	Mean	Standard error	Mean	Standard error
500	3.205	0.176	3.205	0.135	3.205	0.109
1000	3.205	0.168	3.205	0.124	3.205	0.095
1500	3.205	0.165	3.205	0.120	3.205	0.090
2000	3.205	0.163	3.205	0.118	3.205	0.087
2500	3.205	0.162	3.205	0.117	3.205	0.086
3000	3.205	0.162	3.205	0.116	3.205	0.084

Clearly, a reference survey is here able to remove the bias. The expected value of the estimator is 3.205. This value is identical to that of the post-stratification estimator. Again, this will come as no surprise as both estimators theoretically have the same expected value.

The standard error of the reference survey estimate is also here substantially larger than that of the post-stratification estimator. For example, the standard error for a web survey of size 500 combined with a reference survey sample size of 100 is equal to 0.176. If the sample size of the web survey is increased from 500 to 3,000 the standard error only reduces form 0.175 to 0.162.

A standard error of 0.162 could also have been obtained by drawing a simple random sample of size 200 from the target population. The conclusion can be that a

combination of a web survey and a reference survey of a total size of 3000 + 100 = 3100 cases will have the same accuracy as a simple random sample of size 200. So, the reference survey approach produces here unbiased estimates, but it is not efficient in terms of accuracy.

Note that a standard error of 0.116 for a reference survey size of 200 comes down (in terms of accuracy) to a simple random sample of size 400. And a standard error of 0.084 for a reference survey size of 400 comes down (in terms of accuracy) to a simple random sample of size 800.

Table 4.2.5 contains the results of applying the stratification approach to the target variable reading newspapers. It is assumed that the sizes N_I en N_{NI} of the Internet-population and the non-Internet-population are known.

As can be expected, stratification leads to unbiased estimators. The unbiased estimators for both strata can be combined into an unbiased estimator for the target population. The standard error of the estimator depends on the sizes of the samples in both strata.

Table 4.2.5. Simulation results for reading newspapers – part 3

	Non-Internet sample size = 100		Non-Internet sample size = 200		Non-Internet sample size = 400	
Internet sample size	Mean	Standard error	Mean	Standard error	Mean	Standard error
500	3.205	0.103	3.205	0.084	3.205	0.074
1000	3.205	0.093	3.205	0.072	3.205	0.059
1500	3.205	0.089	3.205	0.067	3.205	0.053
2000	3.205	0.087	3.205	0.065	3.205	0.050
2500	3.205	0.086	3.205	0.063	3.205	0.048
3000	3.205	0.085	3.205	0.062	3.205	0.047

Comparing the standard errors in table 4.2.5 with those in table 4.2.3 shows that weighting by age results in more precise estimators than stratification. Apparently, the various age classes are more homogeneous with respect to the values of the target variable than just the Internet and non-Internet-population.

Comparing the standard errors in table 4.2.5 with those in table 4.2.4 shows that the reference survey approach produces less precise estimates than stratification. Having to estimate the sizes of the age strata introduces a lot of extra uncertainty.

It should be noted that stratification can only be applied if all target variables are measured in both the Internet-population and non-Internet-population. For the reference survey approach only the weight variable (age in this example) has to be measured. So, the reference survey questionnaire can be much smaller.

The simulation results show that for a variable that suffers from Missing At Random (MAR) estimates will be biased and correction techniques will be able to remove this bias. In the case of a reference survey the bias is removed at the cost of a much lower accuracy.

5. Discussion and conclusions

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

Selecting a probability sample requires a sampling frame. The Internet is not an ideal sampling frame. It suffers from under-coverage. Certain groups in the population are under-represented, e.g. the elderly, low educated and non-natives. Therefore, estimates will often be biased and correction techniques are required to remove this bias. Unfortunately, correction techniques will be effective only if not having access to the Internet can be seen as Missing At Random (MAR).

It should be noted that also other modes of data collection have their coverage problems. For example, a CATI survey requires a sampling frame consisting of telephone numbers. Statistics Netherlands uses only fixed-line listed telephone numbers for this, as well as listed mobile numbers. Only between 60% and 70% of the people in the Netherlands have a listed phone number, see Cobben (2004). The other 30% to 40% have a non-listed fixed-line number or just a mobile phone. All in all about 70-75% of the Dutch are traced down by either fixed-line or mobile phone.

The under-coverage problem will become even more severe over time. This is due to the popularity of mobile phones and the lack of lists of mobile phone numbers, see e.g. Kuusela (2003). The situation is improving for surveys using the Internet as a sampling frame. In many countries there is a rapid rise in households having Internet access. This percentage of households with Internet is now over 80% in The Netherlands, and it keeps growing. So one might expect that in the near future web survey coverage problems will be less severe.

Unbiased estimators for population characteristics can only be constructed if all elements in the population have a known and positive probability of being selected. This is not always the case for web surveys. Commercial market research agencies in The Netherlands have carried out an analysis of all their major online panels, see Van Ossenbruggen et al. (2006). It turned out that most of these panels are based on self-selection of respondents. The researchers concluded that panel members differ substantially from other people, and that therefore most of these panels cannot be considered representative for the population.

Can a web survey be an alternative for a CAPI or CATI survey? Coverage problems may solve itself in the future, but there are also other aspects to consider. With respect to data collection, there is a substantial difference between CAPI and CATI on the one hand and web surveys on the other. Interviewers carry out the fieldwork in CAPI and CATI surveys. They are important in convincing people to participate in the survey, and they also can assist in completing the questionnaire. There are no interviewers in a web survey. It is a self-administered survey. Therefore quality of collected data may be lower due to higher nonresponse rates and more errors in

answering questions. According to De Leeuw & Collins (1997) response rates tend to be higher if interviewers are involved. However, response to sensitive questions is higher without interviewers. Currently, little is known about the quality web survey data as compared to CAPI or CATI survey data.

CAPI and CATI are both a form of computer assisted interviewing (CAI). CAI has the advantage that error checking can be implemented. It means that answers to questions are checked for consistency. Errors can be detected during the interview, and therefore also corrected during the interview. It has been shown, see e.g. Couper et al. (1998), that CAI can improve the quality of the collected data. The question is now whether error checking should be implemented in a web survey? What happens when respondents are confronted with error messages? Maybe they just correct their mistakes, but it may also happen that they will become annoyed and stop answering questions. There may be a trade-off here between nonresponse and data quality. Further research should make clear what the best approach is.

The reference survey is proposed as one way to remove the bias of estimates in web surveys. One of the advantages of a reference survey is that auxiliary variables can be used for weighting that are highly correlated with either target variables or missingness. Therefore correction will be 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. One attractive characteristic of a web survey is that it is rather easy to collect a large amount of data. If a reference survey is used, the large sample size of the web survey does not imply a high precision. So one may wonder whether it is still worth while to carry out a web survey.

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.

If a reference survey is conducted, stratified estimation may be an option. The Internet-population is one stratum and the non-Internet-population is another stratum. In principle this results in unbiased estimates. The drawback is that the complete questionnaire has to be used in the survey of the non-Internet-population. If the reference survey is used for weighting purposes, only relevant weighting variables should be measured in both surveys. This reduces the reference survey in size and costs, and also the nonresponse may be less of a problem if a very short questionnaire is used.

Given the analysis in this paper, one can say that a web survey based on self-selection and correction by means of a reference survey is not a reliable and cost-effective data collection instrument. This does not mean it is completely useless. When given a sound basis, e.g. using probability sampling and more developed

correction techniques, web surveys hold a promise for deployment in an official statistics environment. This makes web surveys an interesting and worth while topic for future research.

6. Acknowledgements

The author wishes to thank Paul Knottnerus and Steven de Bie for valuable contributions to earlier versions of this paper.

7. References

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

Bethlehem, J.G. (1999), Cross-sectional Research. In: H.J. Adèr and G.J. Mellenbergh, *Research Methodology in the Social, Behavioural & Life Science*. Sage Publications, London, pp.110-142.

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. and Kersten, H.M.P. (1985), On the treatment of non-response in sample surveys. *Journal of Official Statistics* 1, pp 287-300.

Cobben, F. (2004), Nonresponse correction techniques in household surveys at Statistics Netherlands: a CAPI-CATI comparison. Technical report, Statistical Netherlands, Methods and Informatics Department, Voorburg, The Netherlands.

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., O'Reilly, J.M. (eds.) (1998), *Computer Assisted Survey Information Collection*. Wiley, New York.

De Haan, J. & Van 't Hof, C. (2006), *Jaarboek ICT en samenleving, de digitale generatie*. Sociaal en Cultureel PlanBureau, Den Haag.

De Leeuw, E. & Collins M. (1997), Data collection methods and survey quality. In: Lyberg, L., Biemer, P., Collins, M., De Leeuw, E., Dippo, C., Schwarz, N. &

Trewin, D. (eds.), *Survey Measurement and Process Control*. Wiley, New York, pp. 199-220.

Dillman, D A. 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.

Fricker, R. and Schonlau, M. (2002), Advantages and disadvantages of Internet research surveys: Evidence from the literature. *Field Methods* 15, pp. 347-367.

Heerwegh, D. and Loosveldt, G. (2002), An evaluation of the effect of response formats on data quality in web surveys. Paper presented at the International Conference on Improving Surveys, Copenhagen, 2002.

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

Kish, L. (1967), *Survey Sampling*. Wiley, New York, USA.

Kuusela, V. (2003), Mobile phones and telephone survey methods. In R. Banks, J. Currall, J. Francis, L. Gerrard, R. Kahn, T. Macer, M. Rigg, E. Ross, S. Taylor & A. Westlake (Eds.), *ASC 2003 - The impact of new technology on the survey process. Proceedings of the 4th ASC international conference,* pp. 317-327. Chesham Bucks, UK: Association for Survey Computing (ASC).

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.

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.

Van Ossenbruggen, R., Vonk, T. & Willems, P. (2006), Online panels, goed bekeken. *Clou* 24, pp. 28-34.

Appendix A. Proof of theorem

Theorem 1. *The variance of the estimator*

$$\bar{y}_{I,RS} = \sum_{h=1}^{L} \frac{m_h}{m} \bar{y}_I^{(h)}$$

is equal to

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

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

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

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

$$V\left(\frac{m_h}{m} \bar{y}_I^{(h)}\right) = E_P\left(V_I\left(\frac{m_h}{m} \bar{y}_I^{(h)} \mid P\right)\right) + V_P\left(E_I\left(\frac{m_h}{m} \bar{y}_I^{(h)} \mid P\right)\right) =$$
$$= E_P\left(\left(\frac{m_h}{m}\right)^2 V\left(\bar{y}_I^{(h)}\right)\right) + V_P\left(\frac{m_h}{m} \bar{Y}_I^{(h)}\right) =$$
$$= \left(\frac{W_h(1 - W_h)}{m} + W_h^2\right) V\left(\bar{y}_I^{(h)}\right) + \frac{W_h(1 - W_h)}{m}\left(\bar{Y}_I^{(h)}\right)^2.$$

The covariance term $C\left(\frac{m_h}{m} \bar{y}_I^{(h)}, \frac{m_g}{m} \bar{y}_I^{(g)}\right)$ can be written as

$$C\left(\frac{m_h}{m}\bar{y}_I^{(h)},\frac{m_g}{m}\bar{y}_I^{(g)}\right)=$$
$$=E_P\left(C_I\left(\frac{m_h}{m}\bar{y}_I^{(h)}\frac{m_g}{m}\bar{y}_I^{(g)}\mid P\right)\right)+C_P\left(E_I\left(\frac{m_h}{m}\bar{y}_I^{(h)}\mid P\right),E_I\left(\frac{m_g}{m}\bar{y}_I^{(g)}\mid P\right)\right).$$

By means of conditioning on the realised numbers of observations in the strata, it can be shown that

$$E_I\left(\bar{y}_I^{(h)}\bar{y}_I^{(g)}\mid P\right)=E_I\left(\bar{y}_I^{(h)}\mid P\right)E_I\left(\bar{y}_I^{(g)}\mid P\right)=\bar{Y}_I^{(h)}\bar{Y}_I^{(g)}.$$

Therefore

$$C_I\left(\frac{m_h}{m}\bar{y}_I^{(h)}\frac{m_g}{m}\bar{y}_I^{(g)}\mid P\right)=\frac{m_h}{m}\frac{m_g}{m}C_I\left(\bar{y}_I^{(h)},\bar{y}_I^{(g)}\mid P\right)=0.$$

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

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

Since

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

and

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

if follows that

$$V(\bar{y}_{I,RS})=V\left(\sum_{h=1}^{L}\frac{m_h}{m}\bar{y}_I^{(h)}\right)=$$
$$=\frac{1}{m}\sum_{h=1}^{L}W_h(1-W_h)V\left(\bar{y}_I^{(h)}\right)+\frac{1}{m}\sum_{h=1}^{L}W_h(1-W_h)\left(\bar{Y}_I^{(h)}\right)^2+$$
$$+\sum_{h=1}^{L}W_h^2V\left(\bar{y}^{(h)}\right)+\frac{1}{m}\left[\sum_{h=1}^{L}W_h^2\left(\bar{Y}_I^{(h)}\right)^2-\left(\sum_{h=1}^{L}W_h\bar{Y}_I^{(h)}\right)^2\right]=$$
$$=\frac{1}{m}\sum_{h=1}^{L}W_h\left(\bar{Y}_I^{(h)}-\tilde{\bar{Y}}_I\right)^2+\frac{1}{m}\sum_{h=1}^{L}W_h(1-W_h)V(\bar{y}_I^{(h)})+\sum_{h=1}^{L}W_h^2V(\bar{y}_I^{(h)}).$$

This completes the proof.