On the Variability of Estimates based on Propensity Score Weighted Data from Web Panels

Annica Isaksson[1], Stig Danielsson[2] and Gösta Forsman[2]

[1]Statistics Sweden
SE-701 89 Örebro
e-mail: annica.isaksson@scb.se

[2]Department of Mathematics
Division of Mathematical Statistics
Linköping University
SE-581 83 Linköping
e-mail: stdan@mai.liu.se
gosfor@mai.liu.se

Report nr. 29 from the project "Modern statistical survey methods"
December 2004

Keywords: Variance estimation, propensity score adjustment

Abstract: This paper deals with variance estimation in Web surveys of the general population. Such surveys often utilize a panel of Web users, from which samples are selected for various surveys. Inference commonly suffers from considerable problems, including severe selection biases due to low Internet penetration in the population, and large nonresponse. Thus, good weighting procedures are badly needed. We restrict our attention here to an application of the 'propensity score weighting' procedure, in which a parallel telephone survey is used to estimate the propensities of being in the Web sample. The resulting weights may potentially reduce both selection bias and bias due to nonresponse. It is not obvious, however, how the estimator's variance should be estimated, and estimates are typically presented without adhesive uncertainty measures. This unsatisfying situation is the starting-point of our work. Since textbook variance formulae do not apply on the propensity score estimator, we try instead a model approach.

The support from the Bank of Sweden Tercentenary Foundation (Grant no 2000-5063) is gratefully acknowledged.

ON THE VARIABILITY OF ESTIMATES BASED ON PROPENSITY SCORE WEIGHTED DATA FROM WEB PANELS

Annica Isaksson[1], Stig Danielsson[2] and Gösta Forsman[2]
[1]Statistics Sweden, [2]Linköping University, Sweden

Keywords: Web surveys, propensity score adjustment, variance estimation.

1 Introduction

Consider a sample survey of the general population: the survey goal is to estimate a particular population entity, say, the population mean. Under most standard sampling designs, such as simple random sampling (SI) or stratified SI, this is a straightforward task, and suitable formulae are available in any textbook dealing with sampling (e.g., Cochran 1977; Särndal et al. 1992). These textbooks, however, rarely offer any advice on how to estimate the mean if a nonprobability procedure is used to select the sample. In such cases, the design-based theory does not hold, so inference must rely on model assumptions.

This paper will focus on Web surveys, which typically suffer from both a lack of appropriate sampling frames and the low penetration of Internet into the general population. In consequence, these surveys must often rely on volunteer panels. A simple estimator of the population mean, such as the sample mean, may suffer from severe selection bias if applied to such panel data. To avoid this, a model-based "propensity score estimator" has been proposed. Under ideal conditions this estimator would be free of selection bias. A remaining issue, dealt with in this paper, is how to estimate its variance.

1.1 The problem

Our starting point is a recurrent sample survey consisting of two parts characterized by the data collection means used—telephone (T) or Web (W). The Web is the main medium used for data collection, while data collection by telephone is rare, performed for the sole purpose of aiding estimation.

The parameter to be estimated is the population mean, $\bar{y}_U = \sum_{k \in U} y_k/N$, where U is the general population (of size N) and y_k is the fixed value of study variable y for individual $k \in U$.

The Web sample, s_W, is selected from U_W, a subset of U. In practice, we think of s_W as chosen from a volunteer panel of Internet users, possibly created by inviting visitors to popular Internet sites and portals. This corresponds to Type 3 in Couper's taxonomy of Web surveys (Couper 2000). The telephone sample, s_T, on the other hand, is an SI sample from U (to simplify matters, we assume that the frame population of the telephone survey coincides exactly with U). The sizes of s_W, $n = n_T + n_W$.

The problem is to estimate the mean of U from s_W, supported by s_T

1.2 Our approach

We deal with the situation described in section 1.1 by leaving the finite population framework and regarding y_k ($k \in U$) as a random variable associated with the kth individual (the actual y_k is taken as a realization of this random variable). Our viewpoint brings us to a model world sometimes referred to as a *superpopulation model* (see, e.g., Särndal et al. 1992, sec. 12.2); Cassel et al. 1977). The random variables, $y_1, ..., y_N$, are regarded as independently and identically distributed (iid) with a common mean, $E(y_k) = \mu$, and variance, $V(y_k) = \sigma^2$, for $k \in U$. From the general properties of a random sample (Casella and Berger 1990, theorem 5.2.2), the expectation and variance of $\bar{y}_U$ are then given by

$$E(\bar{y}_U) = \mu; \qquad V(\bar{y}_U) = \frac{\sigma^2}{N} \tag{1}$$

In this setting, the estimation problem discussed in section 1 is translated into the one of estimating μ and σ^2 from available data. To accomplish this, consider the following conditions, corresponding closely to those outlined in Rosenbaum and Rubin (1983).

The "treatment assignment" of individual values of $k \in U$, here interpreted as the individual's possible inclusion in the Web panel, is indicated by the variable z_k:

$$z_k = \begin{cases} 1 & \text{if } k \in U_W \\ 0 & \text{if } k \notin U_W \end{cases} \tag{2}$$

The treatment assignment is assumed to be *strongly ignorable* given a random vector, $\boldsymbol{x}_k$, of covariates; that is, y_k $(k \in U)$ is conditionally independent of z_k given $\boldsymbol{x}_k$. It follows that the conditional expected value of y_k given $\boldsymbol{x}_k$, $E(y_k\,|\boldsymbol{x}_k)$, is independent of z_k. If treatment assignment is strongly ignorable given $\boldsymbol{x}_k$, it is strongly ignorable given any function of $\boldsymbol{x}_k$—any *balancing score*—such that $\boldsymbol{x}_k$). One implication of this is that the conditional expected value of y_k given $b(\boldsymbol{x}_k)$, $E(y_k\,|b(\boldsymbol{x}_k))$, is independent of z_k. The coarsest balancing score is the *propensity score*, $e\,(\boldsymbol{x}_k)$, defined as

$$e\,(\boldsymbol{x}_k) = \Pr\,(z_k = 1\,|\boldsymbol{x}_k\,) \tag{3}$$

the finest balancing score being $\boldsymbol{x}_k$ itself. Now assume that the propensity scores of all individuals in U are known. Theoretically, propensity scores may assume any values between 0 and 1. This limits their practical use somewhat, since the number of individuals having the same propensity score may be equal or close to zero. It seems plausible, however, for individuals having similar propensity scores to have similar conditional expected values. Thus, we assume that if U is divided into a large number, H, of classes, $U_1, ..., U_h, ..., U_H$, each containing individuals having similar propensity scores, then individuals within a class will share a common conditional mean and variance. Formally, we assume that

$$E(y_k\,|e(\boldsymbol{x}_k)) = \mu_h; \qquad V(y_k\,|e(\boldsymbol{x}_k)) = \sigma_h^2 \tag{4}$$

for all $k \in U_h$ $(h = 1, ..., H)$. Then, μ can be written as

$$\mu = \sum_{h=1}^{H} D_h \mu_h \tag{5}$$

where D_h denotes the probability that an individual, randomly selected from U, belongs to class U_h $(h = 1, ..., H)$.

Estimation of μ requires knowledge of the class membership of each individual $k \in s$. For $h = 1, ..., H$, let intersection $s_W \cap U_h$ be denoted s_{Wh} (of random size n_{Wh}), intersection $s_T \cap U_h$ be denoted s_{Th} (of random size n_{Th}), and union $s_{Wh} \cup s_{Th}$ be denoted s_h (of random size n_h). Assuming class membership to be known for sampled individuals, we propose the following sample-based estimates of D_h and μ_h. First, since s_T is chosen by SI, the distribution of s_T over classes is likely to resemble the corresponding population distribution over classes (if n_T is sufficiently large.) Thus, it makes sense to estimate D_h by $d_h = n_{Th}/n_T$. Second, since treatment assignment is strongly ignorable, estimation of the class means, μ_h, can be based solely on s_{Wh}. This motivates the estimation of μ_h by the class mean of the Web sample: $\bar{y}_{s_{Wh}} = \sum_{s_{Wh}} y_k/n_{Wh}$. The resulting estimator of μ is

$$\bar{y}_s = \sum_{h=1}^{H} d_h \bar{y}_{s_{Wh}}. \tag{6}$$

In practice, the propensity scores must be estimated for $k \in s$, which calls for some additional modeling. A strategy that lies near at hand is to formulate a logistic regression model for $e\,(\boldsymbol{x}_k)$ as a function of $\boldsymbol{x}_k$, and to estimate the propensity scores using this model. Then, sample s is divided into classes of similar estimated propensity scores. The sample classes should then coincide reasonably well with the classes in the population.

Estimator $\bar{y}_s$, being intuitively appealing, is already in use in various Web surveys (see, e.g., Terhanian, Smith, Bremer, and Thomas 2001). This paper uses the model assumptions to derive its expectation and variance, and, most importantly, to suggest an estimator of its variance.

2 The propensity score weighting procedure

The propensity score weighting procedure of interest in this paper comprises the following steps:

1. estimation of $e(\boldsymbol{x}_k)$ for each $k \in s$,
2. division of sample s into classes containing individuals having similar estimated values of $e(\boldsymbol{x}_k)$, and
3. estimation of μ.

This section discusses some features of steps 1 and 2.

In step 1, the propensity scores, $e(\boldsymbol{x}_k)$, are estimated by using the indicator variable, z_k, and the vector, $\boldsymbol{x}_k$, of the covariates, both of which are available for all $k \in s$. The covariates (sometimes referred to as "webographics") might concern lifestyle, attitudes, and self-perception. A standard logistic regression model for $e(\boldsymbol{x}_k)$ as a function of $\boldsymbol{x}_k$ is formulated (Neter, Kutner, Nachtsheim, and Wasserman 1996, eqn. 14.37; Manly 1994, eqn. 8.3), according to which values of z_k $(k \in U)$ are independent Bernoulli random variables having conditional expected values:

$$E(z_k \,|\boldsymbol{x}_k\,) = e(\boldsymbol{x}_k) = \frac{\exp\left(\boldsymbol{\beta}'\boldsymbol{x}_k\right)}{1+\exp\left(\boldsymbol{\beta}'\boldsymbol{x}_k\right)} \tag{7}$$

where

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}; \quad \boldsymbol{x}_k = \begin{bmatrix} 1 \\ x_{1k} \\ \vdots \\ x_{p-1,k} \end{bmatrix}$$

If s is an SI sample from U, then—as in type (1) in Manly (1994, p. 120)—application of logistic regression is straightforward, and $e(\boldsymbol{x}_k)$ is estimated by

$$\hat{e}(\boldsymbol{x}_k) = \frac{\exp\left(\mathbf{b}'\boldsymbol{x}_k\right)}{1+\exp\left(\mathbf{b}'\boldsymbol{x}_k\right)} \tag{8}$$

where $\mathbf{b}$ is a vector of maximum likelihood (ML) estimates of $\beta_0, \beta_1, ..., \beta_{p-1}$. In our case, s_W and s_T are lumped together to form s. As shown in Seber (1984, p. 312) and discussed in Manly (1994, sec. 8.10), when applied to lumped data, the model in equation (7) needs modification. In our setting, intercept β_0 should be reduced by

$$\log_e\left[\frac{n_W\,(1-P_W)}{n_T P_W}\right], \tag{9}$$

where $P_W = N_W/N$ is the Web panel fraction of the total population (the ML estimate of β_0 must of course be adjusted correspondingly).

Next, the total sample, s, is divided into weighting classes containing individuals having similar estimated propensity scores. In the literature, one sometimes finds the recommendation to form several (around five) groups, and to make them of equal size in terms of n_{Th}. This recommendation is based on an early paper by Cochran (1968), in which subclassification by a single covariate is considered. In our setting, division of the sample should aim to form groups of individuals having similar propensity scores; if so, it makes no sense to create groups of equal size. Instead, the group members' closeness in terms of $\hat{e}(\boldsymbol{x})$ is crucial.

3 Statistical modeling

To derive the statistical properties of $\bar{y}_s$, we use statistical models for y_k and the vector $\boldsymbol{n}_T = (n_{T1}, ..., n_{Th}, ..., n_{TH})$. In this section, our models are formulated, and the corresponding statistical properties of $\bar{y}_s$ are investigated. Please note, however, that several potential sources of bias and variance are ignored, including

- the choice of x variables included in the logistic regression model,
- the fit of the logistic regression model, and
- the division of the Web sample into classes by $\hat{e}(\boldsymbol{x})$ instead of $e(\boldsymbol{x})$.

As will soon be discussed, we also ignore the randomness of $n_{W1}, ..., n_{WH}$.

Our approach relies on the following random models for y_k and $\boldsymbol{n}_T$.

Model $\mathbf{m_1}$
Conditional on $e(\boldsymbol{x})$, the study variable values y_k for $k \in s_h$, $h = 1, ..., H$, are iid random variables with expectation $E_{m_1}(y_k) = \mu_h$ and variance $V_{m_1}(y_k) = \sigma_h^2$.

From model m_1 (and the general properties of a random sample), the conditional expectation and variance of $\bar{y}_{s_{Wh}}$ (conditional on $e(\boldsymbol{x})$ and n_{Wh}) are $E_{m_1}(\bar{y}_{s_{Wh}}) = \mu_h$ and $V_{m_1}(\bar{y}_{s_{Wh}}) = \sigma_h^2/n_{Wh}$, respectively. Also, $\bar{y}_{s_{Wh}}$ and $\bar{y}_{s_{Wi}}$ $(h, i = 1, ..., H; i \neq h)$ are independent. Since s_{Wh} is not a probability sample, the statistical properties of n_{Wh} are unknown. Therefore, throughout our analysis, we condition on n_{Wh}.

Model $\mathbf{m_2}$
Each individual $k \in s_T$ is independently assigned membership in one of H classes. For each individual, the probability of being assigned to class h is D_h. Thus, the random vector, $\boldsymbol{n}_T$, has a *multinomial distribution* with n_T trials, H possible outcomes, and cell probabilites $D_1, ..., D_H$.

Under model m_2, the marginal distribution of n_{Th} $(h = 1, ..., H)$ is binomially distributed with parameters n_T and D_h. It follows that the expectation and variance of n_{Th} are $E_{m_2}(n_{Th}) = n_T D_h$ and $V_{m_2}(n_{Th}) = n_T D_h (1 - D_h)$, respectively.

In addition to models m_1 and m_2, we assume that $\bar{y}_{s_{Wh}}$ and d_h $(h = 1, ..., H)$ are independent. This makes sense, since they are based on two different data sets, selected independently of U.

4 Statistical properties of $\bar{y}_s$

The expectation and approximate variance of $\bar{y}_s$, based on the models formulated in sec. 3, are given in theorem 4.1. The theorem is proved in the appendix.

Theorem 4.1 *Under model m_1 and m_2, the estimator $\bar{y}_s$ is model unbiased for μ. The variance of $\bar{y}_s$ is given by*

$$V_{m_1 m_2}(\bar{y}_s) = V_1 + V_2 \qquad (10)$$

where

$$V_1 = \frac{1}{n_T} \sum_{h=1}^{H} \left[D_h (\mu_h - \mu)^2 + D_h (1 - D_h) \frac{\sigma_h^2}{n_{Wh}} \right]$$

and

$$V_2 = \sum_{h=1}^{H} D_h^2 \frac{\sigma_h^2}{n_{Wh}}.$$

We construct an estimator of $V_{m_1 m_2}(\bar{y}_s)$ using the "method of moments" (Casella and Berger 1990, ch. 7). In this context, it means that we replace the unknown model parameters in the variance expression with their sample analogues. This gives the estimator

$$\hat{V}(\bar{y}_s) = \hat{V}_1 + \hat{V}_2 \qquad (11)$$

where

$$\hat{V}_1 = \frac{1}{n_T} \sum_{h=1}^{H} \left[d_h \left(\bar{y}_{s_{Wh}} - \bar{y}_s\right)^2 + d_h (1 - d_h) \frac{s_{Wh}^2}{n_{Wh}} \right],$$

$s_{Wh}^2 = \sum_{k \in s_{Wh}} \left(y_k - \bar{y}_{s_{Wh}}\right)^2 / (n_{Wh} - 1)$, and

$$\hat{V}_2 = \sum_{h=1}^{H} d_h^2 \frac{s_{Wh}^2}{n_{Wh}}.$$

The method of moments is intuitively rather than theoretically motivated. In consequence, there is no guarantee that $\hat{V}(\bar{y}_s)$ is model unbiased for the true variance.

5 Simulation

In this section, we will familiarize ourselves with $\bar{y}_s$ and $\hat{V}(\bar{y}_s)$ through a simulation. We will create an artifical target population, draw a large number of independent samples from the same, and use these samples to investigate the estimators' statistical properties.

5.1 Creation of the target population

An artificial target population U of $N = 50,000$ elements is constructed as follows.

Covariates: We simulate N values of a bivariate standard normal distribution:

$$(X_1, X_2) \sim N(\mathbf{0}, \mathbf{\Sigma})$$

with covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

and N values of the Bernoulli-distributed variable:

$$X_3 \sim Be\left(\frac{\exp(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2)}{1 + \exp(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2)} \right).$$

Class index h	Conditions		
	x_1	x_2	x_3
1	≤ 0	≤ 0	0
2	≤ 0	> 0	0
3	> 0	≤ 0	0
4	> 0	> 0	0
5	≤ 0	≤ 0	1
6	≤ 0	> 0	1
7	> 0	≤ 0	1
8	> 0	> 0	1

Table 1: Division into classes.

This produces two continuous and one discrete covariate. The model parameters are set to $\rho = .5$, $\gamma_0 = 0$, and $\gamma_1 = \gamma_2 = 1$.

Division into classes: We use the realized values on the covariates to partition U into $H = 8$ classes, $U_1, ..., U_h, ..., U_H$, in accordance with Table 1. The realized size of class U_h is denoted N_h.

Study variable: For class U_h $(h = 1, ..., 8)$, we simulate N_h values of a study variable as

$$Y_h \sim N\left(\mu_h, \sigma_h^2\right)$$

where $\mu_1 = -0.4$ and $\mu_h = \mu_{h-1} + 0.1$ for $h > 1$, and $\sigma_h = \sqrt{\sigma_h^2} = \lambda_0 + \lambda_1 |\mu_h|$. The model parameters are set to $\lambda_0 = \lambda_1 = 1$. Note that in this way, we obtain different study variable means for different classes, larger means for larger values on the covariates, and variances proportional to the level of the means.

Treatment assignment: For class U_h $(h = 1, ..., 8)$, we simulate N_h values of the Bernoulli variable:

$$Z_h \sim Be(\theta_h)$$

where $\theta_1 = 0.1$ and $\theta_h = \theta_{h-1} + 0.1$ for $h > 1$. In this way, the treatment assignment is dependent on all auxiliary variables (through the forming of the classes). Furthermore, treatment assignment is strongly ignorable in the sense discussed in section 1.2.

5.2 Sampling from the artifical population

From the population, $R = 10,000$ independent samples $s_{(1)}, ..., s_{(r)}, ..., s_{(R)}$ are drawn. Each sample, $s_{(r)}$, is the union of $s_{T(r)}$ and $s_{W(r)}$, where $s_{T(r)}$ is an SI sample from U, and $s_{W(r)}$ an SI sample from U_W. Throughout, the sizes of $s_{T(r)}$ and $s_{W(r)}$ are $n_{T(r)} = 1000$ and $n_{W(r)} = 5000$, respectively.

5.3 Estimation and results

In the estimation, the class membership of each sampled individual is assumed to be known. Thus, we limit our attention to the favorable case in which there is no uncertainty in the division of $s_{(r)}$ into classes. For $r = 1, ..., R$, we calculate a propensity score estimate $\bar{y}_{s(r)}$ in accordance with equation (6). In addition, we calculate the variance estimates $\hat{V}_{1(r)}$ and $\hat{V}\left(\bar{y}_{s(r)}\right)$ in accordance with the formulae for $\hat{V}_1$ and $\hat{V}(\bar{y}_s)$, respectively, in equation (11). Averages of the sample estimates are calculated as

$$\bar{\bar{y}}_{s(r)} = \frac{1}{R}\sum_{r=1}^{R} \bar{y}_{s(r)}; \qquad \bar{\hat{V}}_1 = \frac{1}{R}\sum_{r=1}^{R} \hat{V}_{1(r)};$$

$$\bar{\hat{V}}\left(\bar{y}_{s(r)}\right) = \frac{1}{R}\sum_{r=1}^{R} \hat{V}\left(\bar{y}_{s(r)}\right) \tag{12}$$

and an approximation of the true variance of $\bar{y}_s$ as

$$V\left(\bar{y}_s\right) = \frac{1}{N}\sum_{r=1}^{R}\left(\bar{y}_{s(r)} - \bar{\bar{y}}_{s(r)}\right)^2. \tag{13}$$

Figure 1 shows the frequency distribution of the estimated relative bias in $\bar{y}_s$, $\left(\bar{y}_{s(r)} - \mu\right)/\mu$. On average, the relative bias is very close to zero:

$$\frac{\bar{\bar{y}}_{s(r)} - \mu}{\mu} = -.005.$$

This result was as expected, since the artificial population is constructed in accordance with our model assumptions.

Fig. 2 shows the frequency distribution of the ratio $\hat{V}_{1(r)}/\hat{V}\left(\bar{y}_{s(r)}\right)$. On average,

$$\frac{\bar{\hat{V}}_1}{\bar{\hat{V}}\left(\bar{y}_{s(r)}\right)} = .642,$$

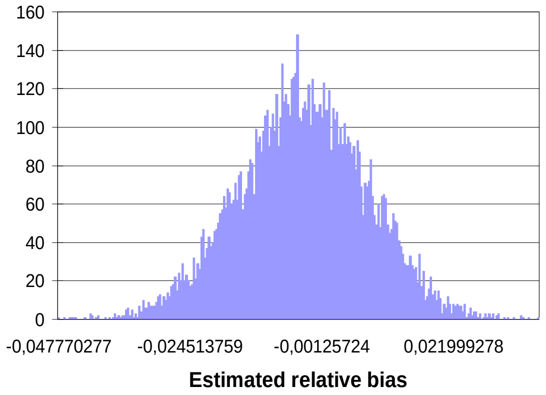

Figure 1: Frequency distribution of $\left(\bar{y}_{s(r)} - \mu\right)/\mu$.

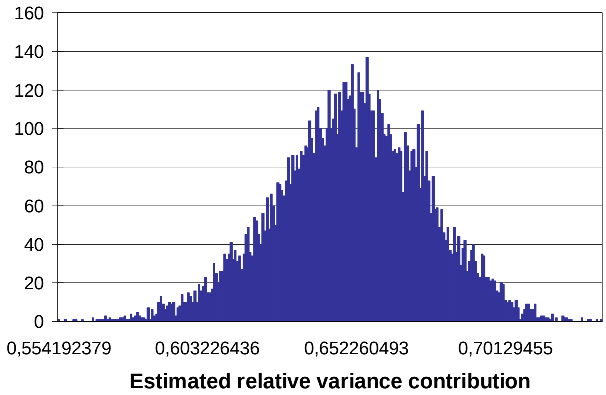

Figure 2: Frequency distribution of $\hat{V}_{1(r)}/\hat{V}\left(\bar{y}_{s(r)}\right)$.

which illustrates that the term V_1 may represent a large proportion of the total variance.

The relative bias of $\overline{\hat{V}}\left(\bar{y}_{s(r)}\right)$, finally, is approximately given by

$$\frac{\overline{\hat{V}}\left(\bar{y}_{s(r)}\right) - V\left(\bar{y}_s\right)}{V\left(\bar{y}_s\right)} = .113,$$

indicating that the suggested variance estimator is quite conservative.

6 Conclusions and final remarks

The propensity score estimator has developed from statistical practice and its needs, rather than as a theoretical exercise. This is probably the reason why its theoretical motivation is not made entirely clear in the literature. In this paper, we have formulated a simple (ideal) model world, in which the propensity score estimator of the population mean is unbiased for the same. In this setting, it is straightforward to develop an expression for the estimator's variance. By replacing unknown entities in the variance formula with their sample counterparts, we arrive at an intuitive variance estimator.

Our variance expression consists of two terms, V_1 and V_2, the second of which resembles the variance of a poststratified estimator. One might feel tempted to confine oneself to estimating V_2. It is, however, easy to conceive of situations in which this would lead to serious underestimation of the total variance; for instance, if the class means differ greatly, or if the telephone sample is small.

In the simulation study, we have made sure that the propensity score estimator really is unbiased for the true mean if the model assumptions hold. We have demonstrated that the variance term, V_1, may represent a large proportion of the total variance, and discovered that our variance estimator is likely slightly to overestimate the true variance.

In the simulation, the propensity score of each sampled individual was known; in reality, however, they must be estimated from the sample data. Further simulations are necessary to investigate the impact of this additional step on the propensity score estimator. The consequences of deviating from the strong ignorability assumption also remain to be investigated.

7 Acknowledgement

The financial support of this work by the Bank of Sweden Tercentenary Foundation (Grant no. 2000-5063) is gratefully acknowledged.

8 References

Agresti, A. (1990), *Categorical Data Analysis,* New York: Wiley.

Casella, G., and Berger, L. (1990), *Statistical Inference,* Belmont, CA: Duxbury Press.

Cassel, C.-M., Särndal, C.-E., and Wretman, J. H. (1977), *Foundations of Inference in Survey Sampling,* New York: Wiley.

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley.

Cochran, W. G. (1968), The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies, *Biometrics*, 24, 295–313.

Couper, M. P. (2000), Web Surveys: A Review of Issues and Approaches, *Public Opinion Quarterly*, 64, 464–494.

Manly, B. F. (1994), *Multivariate Statistical Methods* (2nd ed.), London: Chapman & Hall.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996), *Applied Linear Statistical Models* (4th ed.), Chicago: Irwin.

Rosenbaum, P. R., and Rubin, D. B. (1983), The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, 70, 41–55.

Ross, S. M. (1997), *Introduction to Probability Models* (6th ed.), San Diego, CA: Academic Press.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer.

Seber, G. (1984), *Multivariate Observations*, New York: Wiley.

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

A Proof of Theorem 3.1

We start with the expectation. Using the conditional independency of $\bar{y}_{s_{Wh}}$ and d_h,

$$E_{m_1m_2}(\bar{y}_s) = \sum_{h=1}^{H} E_{m_2}(d_h)\, E_{m_1}(\bar{y}_{s_{Wh}}) = \sum_{h=1}^{H} D_h\mu_h.$$

Now let us turn to the variance. From the general properties of the variance of a sum of random variables (see Ross 1997, eqn. 2.16),

$$\begin{aligned} V_{m_1m_2}(\bar{y}_s) &= \sum_{h=1}^{H_s} V_{m_1m_2}(d_h\bar{y}_{s_{Wh}}) \\ &\quad +2\sum_{h=1}^{H_s}\sum_{i<h} Cov_{m_1m_2}(d_h\bar{y}_{s_{Wh}}, d_i\bar{y}_{s_{Wi}}) \\ &= V_1 + V_2 \end{aligned}$$

where $Cov_{m_1m_2}(d_h\bar{y}_{s_{Wh}}, d_i\bar{y}_{s_{Wi}})$ is the covariance between $d_h\bar{y}_{s_{Wh}}$ and $d_i\bar{y}_{s_{Wi}}$.

Consider any term $V_{m_1m_2}(d_h\bar{y}_{s_{Wh}})$ in V_1. Since d_h and $\bar{y}_{s_{Wh}}$ are independent,

$$\begin{aligned} V_{m_1m_2}(d_h\bar{y}_{s_{Wh}}) &= \left[E_{m_1}(\bar{y}_{s_{Wh}})\right]^2 V_{m_2}(d_h) \\ &\quad + \left[E_{m_2}(d_h)\right]^2 V_{m_1}(\bar{y}_{s_{Wh}}) \\ &\quad + V_{m_2}(d_h)\, V_{m_1}(\bar{y}_{s_{Wh}}) \\ &= \mu_h^2 \frac{D_h(1-D_h)}{n_T} + D_h^2\frac{\sigma_h^2}{n_{Wh}} \\ &\quad + \frac{D_h(1-D_h)}{n_T}\frac{\sigma_h^2}{n_{Wh}}, \end{aligned}$$

and

$$\begin{aligned} V_1 &= \sum_{h=1}^{H}\left[\mu_h^2\frac{D_h(1-D_h)}{n_T} + D_h^2\frac{\sigma_h^2}{n_{Wh}}\right. \\ &\quad \left. + \frac{D_h(1-D_h)}{n_T}\frac{\sigma_h^2}{n_{Wh}}\right]. \end{aligned}$$

Now consider any covariance term in V_2:

$$\begin{aligned} & Cov_{m_1m_2}(d_h\bar{y}_{s_{Wh}}, d_i\bar{y}_{s_{Wi}}) \\ =\; & E_{m_1m_2}(d_h\bar{y}_{s_{Wh}} d_i\bar{y}_{s_{Wi}}) \\ & - E_{m_1m_2}(d_h\bar{y}_{s_{Wh}})\, E_{m_1m_2}(d_i\bar{y}_{s_{Wi}}) \\ =\; & E_{m_1m_2}(d_h\bar{y}_{s_{Wh}} d_i\bar{y}_{s_{Wi}}) \\ & - D_h\mu_h D_i\mu_i. \end{aligned}$$

By use of conditioning,

$$\begin{aligned}
& E_{m_1 m_2}(d_h \bar{y}_{s_{Wh}} d_i \bar{y}_{s_{Wi}}) \\
= \; & E_{m_1}[\bar{y}_{s_{Wh}} \bar{y}_{s_{Wi}} E_{m_2}(d_h d_i \,|m_1\,)] \\
= \; & E_{m_1}\left\{ \bar{y}_{s_{Wh}} \bar{y}_{s_{Wi}} \left[\frac{1}{n_T^2} Cov_{m_2}(n_{Th} n_{Ti} \,|m_1\,) \right.\right. \\
& \left.\left. + \; E_{m_2}(d_h \,|m_1\,) \, E_{m_2}(d_i \,|m_1\,) \right] \right\} \\
= \; & E_{m_1}\left\{ \bar{y}_{s_{Wh}} \bar{y}_{s_{Wi}} \left[\frac{1}{n_T^2} Cov_{m_2}(n_{Th} n_{Ti} \,|m_1\,) \right.\right. \\
& \left.\left. + \; D_h D_i \right] \right\}
\end{aligned}$$

where $Cov_{m_2}(n_{Th} n_{Ti} \,|m_1\,)$ is the covariance between n_{Th} and n_{Ti}. Since n_{Th} and n_{Ti} belong to a multinomial distribution, from Agresti (1990, p. 44),

$$Cov_{m_2}(n_{Th} n_{Ti} \,|m_1\,) = -n_T D_h D_i,$$

and we arrive at

$$\begin{aligned}
& E_{m_1 m_2}(d_h \bar{y}_{s_{Wh}} d_i \bar{y}_{s_{Wi}}) \\
= \; & E_{m_1}\left\{ \bar{y}_{s_{Wh}} \bar{y}_{s_{Wi}} \left[D_h D_i \left(1 - \frac{1}{n_T} \right) \right] \right\} \\
= \; & D_h D_i \left(1 - \frac{1}{n_T} \right) E_{m_1}(\bar{y}_{s_{Wh}} \bar{y}_{s_{Wi}}) \\
= \; & D_h D_i \left(1 - \frac{1}{n_T} \right) \mu_h \mu_i.
\end{aligned}$$

Thus, V_2 is given by

$$\begin{aligned}
V_2 \; = \; & 2 \sum_{h=1}^{H} \sum_{i<h} \left[D_h D_i \left(1 - \frac{1}{n_T} \right) \mu_h \mu_i \right. \\
& \left. - \; D_h \mu_h D_i \mu_i \right] \\
= \; & -\frac{2}{n_T} \sum_{h=1}^{H} \sum_{i<h} D_h D_i \mu_h \mu_i \\
= \; & \frac{1}{n_T} \left[\sum_{h=1}^{H} \mu_h^2 D_h^2 - \left(\sum_{h=1}^{H} \mu_h D_h \right)^2 \right].
\end{aligned}$$

Finally, we add V_1 and V_2:

$$\begin{aligned}
V_1 + V_2 \; = \; & \sum_{h=1}^{H} \left[\mu_h^2 \frac{D_h (1 - D_h)}{n_T} + D_h^2 \frac{\sigma_h^2}{n_{Wh}} \right. \\
& \left. + \; \frac{D_h (1 - D_h)}{n_T} \frac{\sigma_h^2}{n_{Wh}} \right] \\
& + \frac{1}{n_T} \left[\sum_{h=1}^{H} \mu_h^2 D_h^2 - \left(\sum_{h=1}^{H} \mu_h D_h \right)^2 \right] \\
= \; & \frac{1}{n_T} \left[\sum_{h=1}^{H} \mu_h^2 D_h - \sum_{h=1}^{H} \mu_h^2 D_h^2 \right. \\
& + D_h (1 - D_h) \frac{\sigma_h^2}{n_{Wh}} + \sum_{h=1}^{H} \mu_h^2 D_h^2 \\
& \left. - \left(\sum_{h=1}^{H} \mu_h D_h \right)^2 \right] + \sum_{h=1}^{H} D_h^2 \frac{\sigma_h^2}{n_{Wh}}
\end{aligned}$$

$$\begin{aligned}
= \; & \frac{1}{n_T} \left\{ \left[\sum_{h=1}^{H} \mu_h^2 D_h - \left(\sum_{h=1}^{H} \mu_h D_h \right)^2 \right] \right. \\
& \left. + \; D_h (1 - D_h) \frac{\sigma_h^2}{n_{Wh}} \right\} \\
& + \sum_{h=1}^{H} D_h^2 \frac{\sigma_h^2}{n_{Wh}} \\
= \; & \frac{1}{n_T} \sum_{h=1}^{H} D_h \left[\left(\mu_h - \bar{\mu}_{(y|e)} \right)^2 + (1 - D_h) \frac{\sigma_h^2}{n_{Wh}} \right] \\
& + \sum_{h=1}^{H} D_h^2 \frac{\sigma_h^2}{n_{Wh}}
\end{aligned}$$

which equals the stated expression for $V_{m_1 m_2}(\bar{y}_s)$.