ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2018)
80, *Part* 1, *pp.* 33–56

Statistical inference based on randomly generated auxiliary variables

Barry Schouten

Statistics Netherlands, The Hague, and Utrecht University, The Netherlands

[Received February 2015. Final revision June 2017]

Summary. In most real life studies, auxiliary variables are available and are employed to explain and understand missing data patterns and to evaluate and control causal relationships with variables of interest. Usually their availability is assumed to be a fact, even if the variables are measured without the objectives of the study in mind. As a result, inference with missing data and causal inference require some assumptions that cannot easily be validated or checked. In this paper, a framework is constructed in which auxiliary variables are treated as a selection, possibly random, from the universe of variables on a population. This framework provides conditions to make statistical inference beyond the traces of bias or effects found by the auxiliary variables themselves. The utility of the framework is demonstrated for the analysis and reduction of non-response in surveys. However, the framework may be more generally used to understand the strength of association between variables. Important roles are played by the diversity and diffusion of the population of interest, features that are defined in the paper and the estimation of which is discussed.

Keywords: Causal inference; Independent variable; Missing data; Non-response; Surveys

1. Introduction

There have been two crucial and very influential developments in statistical theory over the last half-century: missing data inference and causal inference. The well-known missing data mechanisms missingness completely at random, missingness at random and missingness not at random (Little and Rubin, 2002) and variants of them (see Seaman *et al.* (2013)) appear frequently in the literature. They provide sufficient and necessary conditions to separate the confounding of selection and measurement, or selection and treatment, in case part of the data is missing. These conditions are formulated in terms of the variables of interest and variables that are auxiliary to the study. The theory of causal inference, e.g. Rubin (2005), Heckman (2008), Pearl (2009) and Robins and Hernán (2009), provides tools to investigate the presence and absence of causation, making use of graphical representations of the variables and labelling them, for instance, as instrumental, backdoor or frontdoor variables.

However, both statistical inference with missing data and causal inference treat the variables that are studied in the data as fixed and given, and the generation of the variables themselves is not modelled. Clearly, associations between variables are modelled extensively, but the way that they arise and are measured or observed is essentially left open. As a consequence, sufficient conditions are available to ignore missing data, but one may fail to come up with variables that actually satisfy these conditions or motivate why they should hold. See, for instance,

Address for correspondence: Barry Schouten, Methodology and Quality Division, Statistics Netherlands, PO Box 24500, Den Haag 2490 HA, The Netherlands.
E-mail: bstn@cbs.nl

 1369–7412/18/80033

Molenberghs *et al.* (2008) for a discussion on the missingness at random and missingness not at random assumptions and the more general discussion on enriched data through coarsening in Molenberghs *et al.* (2012).

Causal inference theory provides structured methodology on how to evaluate causal relationships and to free the estimation of causal effects from confounding, but it does not tell us why causal relationships are absent or present. A good example is the exclusion restriction in sample selection models, which states that at least one variable should be excluded, i.e. an instrumental variable, to guarantee identifiability of correlations between error terms.

This paper is motivated by the conviction that the nature of the variables themselves and the way in which they are generated need to be modelled to understand and evaluate the validity of assumptions underlying statistical inference. The paper models the generation of variables and the associations between them. In the model, an important role is played by the diversity and diffusion of populations with respect to so-called variable-generating distributions. Diversity is defined as the maximal resolution of such a distribution, whereas diffusion is defined in terms of the relative bin sizes. It is shown that diversity and diffusion are important properties of a population that appear in association measures.

Four research questions are discussed.

(a) Can a sensible framework be constructed that captures random generation of auxiliary variables?
(b) What implications follow from the framework about the associations between variables?
(c) Can the diversity and diffusion of a population for a variable-generating distribution be estimated?
(d) Can the framework assist in checking the validity of statistical inference?

To answer the last research question, the framework is applied and demonstrated in the setting of missing data. Over recent decades interest in statistical data has increased strongly, which went parallel to a very strong increase in computational power and to the computerization of society. Data collection is costly and missing data are difficult to avoid. Therefore, in many statistical areas modelling of missing data is a key endeavour, and it seems to become even more important with the interest in 'big data'. Without a complete theory about the causes for the missing data, however, it must be accepted that the available auxiliary variables do not guarantee a missingness at random mechanism. The framework that is presented here gives conditions to extrapolate the traces of bias found by auxiliary variables to other variables, i.e. to missingness not at random mechanisms. The original motivation for this paper came from the pursuit to reduce the effect of missing data in surveys through so-called adaptive survey designs (Schouten *et al.*, 2013; Wagner *et al.*, 2013; Särndal and Lundquist, 2014). In these designs, data collection strategies (i.e. treatments) are adapted to auxiliary information that becomes available before or during data collection. The designs assume that detectable bias due to non-response is a signal of even larger biases on variables of interest to the survey. Typically, the proportion of explained variation in non-response by such variables is quite low, and the designs are often criticized for removing non-response bias during the data collection stage that could equally well be removed in the estimation or adjustment stage. It is explained in this paper that the theoretical results provide conditions for the efficacy of such designs to remove bias, even after adjustment. However, the results can be applied more generally to evaluate any non- (fully) randomized data collection.

In Section 2, the conceptual framework is laid out. In Section 3, the estimation of diversity is discussed. Section 4 elaborates the framework to missing data in surveys. To demonstrate the

utility of the framework and to answer the last research question, in Section 5, the missing data in an on-line panel are evaluated. Section 6 ends with a discussion.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from `www.risq-project.eu`.

2. A framework for the generation of variables on a population

This section sets the basic theory and constructs. In Section 2.1, notation is introduced and the generation of variables on a population is modelled. In Section 2.2, population diversity and diffusion are defined. In Section 2.3, two theorems are presented under the framework where variables are treated as randomly generated. In Section 2.4, the expected number of categories for randomly generated variables is elaborated.

2.1. The generation of variables on a population

The objective of this section is to model associations between variables on a population. Although the framework could (and should) be extended to variables with a continuous measurement level, the attention is limited to categorical variables. An extension to continuous variables is not straightforward and may complicate the interpretation of the framework that is presented here, which by itself represents a conceptual change in thinking about statistical data. Furthermore, in many settings of interest such variables are coded as categorical variables.

Suppose that there is a population of interest of size N on which measurements can be made by using a set of potential instruments and that the measurements are termed variables once they are stored. Suppose that the variables, as selected by the measurer, are drawn randomly from the pool of potential variables according to some probability distribution.

The number of variables that can be formed on a population can be very large, whereas the number of available variables in a data set is typically relatively small. As a result, it is pointless, or even meaningless, to attempt to construct various families of variable-generating distributions and to derive empirically to what family a set of variables belongs. Here, it is shown that two subclasses of such distributions, uniform grouping and clustered grouping, may be sufficiently general.

2.1.1. Random-grouping distributions

An instrument is a random grouping of population units. Let s_i be the indicator representing to what group unit i is assigned, and let $s = (s_1, s_2, \ldots, s_N)^{\mathrm{T}}$ be the vector of indicators (with T for transpose). Let C be the (random) number of groups or categories of the resulting variable. Let $p(C, s)$ represent a random-grouping probability distribution defined on $\{1\} \times \{1\}^N \cup \{2\} \times \{1,2\}^N \cup \{3\} \times \{1,2,3\}^N \cup \ldots \cup \{N\} \times \{1,2,\ldots,N\}^N$. Let $A_1, A_2, \ldots, A_C$ denote the groups and let $\Delta_{i,c}$ be the 0–1 indicator for the event $s_i = c$. Finally, let $C_{\max}$ be the smallest c with $p(C > c) = 0$. The resulting groups of population units represent a variable, say Z.

Multiple instruments, labelled $m = 1, 2, \ldots, M$, are independent draws from possibly different distributions $p_m(C, s)$ and lead to series of variables $Z_1, Z_2, Z_3, \ldots, Z_M$. The random generation of variables may lead to constant variables, to copies of the same variable, to variables that are each other's complement and to variables that are linearly dependent. In practice, one will often avoid such variables and may reject them. In those cases, the generation of variables is not independent. However, when $C_{\max} = 2$ and $N = 100$, which seems to be a modest population size, then the number of possible variables is already very large and equals 2^{100}. For relatively small numbers of variables, it will happen rarely that two copies are generated or that a constant

is constructed. This seems to appeal to intuition, as it indeed happens rarely that measurements on a population lead to such events in practice.

The population covariance between two realizations of variables, say Z_1 and Z_2, will be denoted by $\Gamma(Z_1, Z_2)$, with

$$\Gamma(Z_1, Z_2) = \frac{1}{N}\sum_{i=1}^{N} s_{1,i}s_{2,i} - \left(\frac{1}{N}\sum_{i=1}^{N} s_{1,i}\right)\frac{1}{N}\sum_{i=1}^{N} s_{2,i}.$$

Essentially, the grouping distributions $p(C, s)$ determine the expected associations that will be found between the variables. For a simple example, let $C_{\max} = 2$ and $N = 3$, let $p_1(C, s)$ be Poisson sampling with equal inclusion probabilities and let $p_2(C, s)$ be Poisson sampling with unequal inclusion probabilities

$$p_1(s_i = 1) = 0.6, \ \forall\, i,$$

and

$$p_2(s_i = 1) = \begin{cases} 0.8 & \text{if } i = 1, \\ 0.6 & \text{if } i = 2, \\ 0.1 & \text{if } i = 3. \end{cases}$$

Suppose that one of the two distributions is selected. Let Z_1 and Z_2 both be randomly generated from this distribution and let $s_{m,i}$ denote the 0–1 indicator for selection of unit i for variable m. The expected probability that any of the variables equals 1 is $P(Z_m = 1) = (1/N)\Sigma_{i=1}^{N} p_k(s_{m,i} = 1)$. The expected probability that they jointly equal 1 is

$$P(Z_1 = 1, Z_2 = 1) = \frac{1}{N}\sum_{i=1}^{N} p_k(s_{1,i} = 1, s_{2,i} = 1) = \frac{1}{N}\sum_{i=1}^{N} p_k^2(s_{1,i} = 1).$$

Hence, the expected covariance between Z_1 and Z_2 equals 0 for p_1 and 0.087 for p_2. It can be shown that, with independent draws from the same distributions $p(C, s)$, the expected covariance of two variables will always be non-negative. Hence, to reach a negative expected covariance we must select different distributions $p(C, s)$. This is in fact what questionnaire designers sometimes do on purpose to identify measurement error; they vary the direction of scales to detect inconsistent answering patterns.

There is a similarity between independent draws from a grouping distribution and exchangeability of sequences of random variables. A sequence is exchangeable when any permutation of the outcome values has the same probability distribution. If a sequence of random variables is generated independently from a grouping distribution, then any permutation of the sequence has the same probability of being generated. Since the numbers of categories per variable are random, a more general definition of exchangeability would be needed to describe independent draws from any grouping distribution. Nonetheless, it holds that exchangeable sequences, like independently drawn variables, have a non-negative covariance; see O'Neill (2009). Note that the exchangeability is in the variables and not in the population units, as was recently discussed by Mealli and Rubin (2015).

Before we move to subclasses of variable-generating distributions, one important observation is made: any combination of multiple variables through a crossing of the categories could be generated directly from one draw of some random-grouping distribution on the population. Consequently, theorems about the properties of a single randomly drawn variable generalize to multiple independently drawn variables.

Lemma 1. Any variable $\tilde{Z}$ that results from the cross-classification of an independently generated set of variables $Z_1, Z_2, Z_3, \ldots, Z_M$ is itself generated from some $\tilde{p}(C, s)$ on the same population.

Proof. Each variable Z_m leads to groups $A_{m,1}, A_{m,2}, \ldots, A_{m,C_m}$. A cross-classification corresponds to repeated intersections of the sets of groups and results in a new number of groups $\tilde{C}$ and a new set of groups $\tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_{\tilde{C}}$. The probability that a specific combination $\tilde{C}$ and $\tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_{\tilde{C}}$ occurs depends on the underlying distributions to the variables $Z_1, Z_2, Z_3, \ldots, Z_M$ and defines $\tilde{p}(C, s)$. Because N is assumed finite, such a distribution always exists.

2.1.2. Uniform grouping and clustered grouping

A natural subclass of grouping distributions is distributions that have equal and independent unit assignment probabilities for all population units. They are termed uniform grouping distributions and are defined as follows.

Definition 1. $p(C, s)$ is a uniform grouping distribution, if conditional on the number of groups C the population units are assigned following a multinomial distribution with sample size parameter N and some cell probabilities, say $\lambda_1^C, \lambda_2^C, \ldots, \lambda_C^C$.

Hence, the family of uniform grouping distributions is a mixture of multinomial distributions where the mixture is defined by the marginal distribution $p(C)$. This family conforms to a quasi-random selection of variables. Note, however, that some groups may not be assigned any units and remain empty. Let the random variable C_A denote the number of non-empty groups. Lemma 2 shows that lemma 1 holds within the family of such distributions.

Lemma 2. Any variable $\tilde{Z}$ that results from the cross-classification of an independently generated set of variables $Z_1, Z_2, Z_3, \ldots, Z_M$ from the family of uniform grouping is itself generated from some uniform grouping $\tilde{p}(C, s)$ on the same population.

Proof. Let $\tilde{p}(C = c) = p\{\min(N, \Pi_{m=1}^{M} C_{m,A}) = c\}$, with $C_{m,A}$ the number of non-empty groups for variable m, and $\tilde{p}(s|C = c)$ follow a multinomial distribution with sample size parameter N and cell probabilities $\tilde{\lambda}_c^{\tilde{C}}$. The $\tilde{\lambda}_c^{\tilde{C}}$s could, in principle, be derived from the cell probabilities $\lambda_1^{C_m}, \lambda_2^{C_m}, \ldots, \lambda_C^{C_m}$ by fixing labels for the intersections of groups of the M variables and then taking products of the cell probabilities. The cell probabilities then need to be adjusted for the condition that the total number of groups is C. This would be a rather cumbersome derivation. More simply we could state, without further specification, that each group in the cross-classification must have some cell probability. $\tilde{p}(C, s)$ is a uniform grouping distribution and is the distribution of the cross-classification of the variables $Z_1, Z_2, Z_3, \ldots, Z_M$. □

Uniform grouping distributions with unequal unit assignment probabilities correspond to targeted selections of variables. However, as long as unit assignment probabilities are unequal to 0 or 1, all variables have a non-zero probability of being selected. This is different when such probabilities are simultaneously equal to 0 or 1 for at least two units in the population. This is termed clustered grouping.

Definition 2. $p(C, s)$ is a clustered grouping distribution if $\exists\, i, j$ for which $p(s_i = s_j) = 1$.

Clustered grouping distributions imply that two units can never be discerned, i.e. the experimenter has no instrument that enables separation of the two units. It should be noted that also for non-clustered grouping distributions it may occur by chance that two units are not

separated by any of the selected measurements and appear in the same category of the resulting variables. Again it holds that a combination of variables generated from (non-)clustered grouping distributions is generated from a (non-)clustered grouping.

Lemma 3. Any variable $\tilde{Z}$ that results from the cross-classification of an independently generated set of variables $Z_1, Z_2, Z_3, \ldots, Z_M$ from non-clustered grouping distributions is itself generated from some non-clustered grouping $\tilde{p}(C, s)$ on the same population.

Proof. For any individual variable every pair of units may end up in different groups with a non-zero probability. It must surely hold for intersections of these groups that two units end up in different intersections with a non-zero probability. □

Lemma 4. Any variable $\tilde{Z}$ that results from the cross-classification of an independently generated set of variables $Z_1, Z_2, Z_3, \ldots, Z_M$ from the same clustered grouping distributions is itself generated from some grouping distribution $\tilde{p}(C, s)$ on the same population with the same clustering.

Proof. For every individual variable, the same clusters of units end up in the groups. It holds for intersections of these groups that clusters still appear in the same intersections. □

The two classes of grouping distributions can be combined and then lead to clustered uniform grouping. Under clustered uniform grouping, clusters instead of units are randomly allocated following a multinomial distribution with C cells and certain cell probabilities.

Lemmas 1–4 together ensure that uniform grouping and clustered uniform grouping provide a rich framework to evaluate associations between variables; first, a set of variables can always be combined in one variable and, second, it is true by definition that a single variable came from a clustered uniform grouping.

2.2. Population diversity and diffusion

Here, two general properties of a population are defined relative to a variable-generating or random-grouping distribution: diversity and diffusion.

Definition 3. The diversity of a population with respect to a variable-generating distribution p, denoted as $G(p)$, is the number of clusters of units for distribution p.

Definition 4. The diffusion of a population with respect to a variable-generating distribution p, denoted as $D(p)$, is the sum of the squared relative cluster sizes for distribution p. More specifically, when $G(p)$ clusters exist, labelled $g = 1, 2, 3, \ldots, G(p)$, with relative cluster sizes q_g, i.e. $\Sigma_{g=1}^{G(p)} q_g = 1$, then $D(p) = \Sigma_{g=1}^{G(p)} q_g^2$.

In other words, the diversity is the maximal grid or resolution of the variable-generating distribution, whereas the diffusion measures the amount of clustering via the second moment of the relative grid sizes. It is straightforward to show that $1/G(p) \leqslant D(p) < 1$, and the diffusion equals $1/G(p)$ when all clusters have an equal size. In Section 2.3, it is shown that the diffusion plays an important role in associations between variables. The definitions imply that for a uniform grouping distribution $G(p) = N$ and $D(p) = 1/N$. In what follows, the dependence on p is suppressed in the notation.

So far, the terms instruments and measurements have been used loosely, but a more fundamental discussion is needed. What is meant by measurements? First, to talk of a population, the units that are contained in it must be identifiable and separable in place and time. Hence,

identifiers of population units or locations of units in time are not part of the measurements; they serve to demarcate the units. Second, some measurements are already employed to separate a population from larger populations, e.g. humans from animals; such measurements will lead to constants. Third, measurements need to apply to the unit itself and not to its environment, e.g. questions like who is the current President or what is the age of your mother? Finally, one may conjecture that measurers, apart from identifiers and location variables, never have a sufficient range of measurements to separate all population units, i.e. they always apply clustered grouping. The observable diversity G is smaller than the population size N; it depends on the available instruments and may change once new instruments have been developed or discovered.

Diversity and diffusion could also be defined as properties of a superpopulation from which a finite population is drawn. The clusters with their cluster sizes may then be seen as the blueprints of the superpopulation, and its diversity is unrelated to the actual size of the finite population, N. However, the actual size of the population, obviously, censors and masks the real diversity and diffusion of the underlying superpopulation. For this reason, diversity and diffusion are defined as properties of a finite population.

2.3. Associations between randomly generated variables

Suppose that an analysis is directed at explaining a variable of interest Y by using auxiliary variables $(X_1, X_2, \ldots, X_M)^{\mathrm{T}}$. For the sake of demonstration, let Y be quantitative. A researcher may then be interested in the variance $S^2(Y_X)$, where Y_X is the projection of Y on the space that is formed by the variables $(X_1, X_2, \ldots, X_M)^{\mathrm{T}}$.

Let the projection for unit i, $Y_{X,i}$, be defined as

$$Y_{X,i} = \sum_{c=1}^{C} \Delta_{i,c} \frac{\sum_{h=1}^{N} \Delta_{h,c} y_h}{\sum_{h=1}^{N} \Delta_{h,c}}, \tag{1}$$

with y_i the value of Y for unit i. Next, let $\bar{Y}_X$ be the average of the projected values and let $S^2(y)$ be the variance of the measurement values of Y. It is easy to show that $\bar{Y}_X = \bar{y}$ always holds, regardless of the grouping distribution.

First, suppose that there is one auxiliary variable X; then the following theorem applies (with C_A the number of non-empty cells).

Theorem 1. If X is generated from a uniform grouping distribution, then by Taylor series approximation

$$E\{S^2(Y_X)\} = \frac{E(C_A) - 1}{N - 1} \sum_{i=1}^{N} (y_i - \bar{y})^2. \tag{2}$$

For a proof of theorem 1, see Appendix A.

Since, because of lemma 2, a combination of a series of independently generated variables from uniform grouping distributions is itself generated from a uniform grouping distribution, theorem 1 also applies to series of variables. The sizes of the cell probabilities $\lambda_1^C, \lambda_2^C, \ldots, \lambda_C^C$ in the uniform grouping determine the expected number of non-empty cells, $E(C_A)$.

A researcher may be interested in the proportion of unexplained variance, $R^2(Y)$:

$$R^2(Y, X) = 1 - \frac{S^2(Y_X)}{S^2(y)}. \tag{3}$$

Under the conditions of theorem 1, the expected value of equation (3) reduces to

$$E\{R^2(Y,X)\} = 1 - \frac{E(C_A) - 1}{N - 1}.$$

For clustered uniform grouping, a similar result can be derived. Let the clusters be labelled $g = 1, 2, \ldots, G$, q_g be the relative cluster size and y_g be the average value within cluster g.

Theorem 2. If X is generated from a clustered, uniform grouping distribution, then by Taylor series approximation

$$E\{S^2(Y_X)\} = \frac{G\{E(C_A) - 1\}}{G - 1} \sum_{g=1}^{G} q_g^2 (y_g - \bar{y})^2, \tag{4}$$

where higher order terms are cubic in the q_g (and y_g). If, additionally, $\Gamma\{q_g^2, (y_g - \bar{y})^2\} = 0$ and $\Gamma\{q_g, (y_g - \bar{y})^2\} = 0$, then

$$E\{S^2(Y_X)\} = \frac{G\{E(C_A) - 1\} D(p)}{G - 1} S_B^2(y), \tag{5}$$

where $S_B^2(y) = (1/G)\Sigma_{g=1}^{G}(y_g - \bar{y})^2$ is the between variance for the clusters.

For a proof of theorem 2, see Appendix A.

The two conditions $\Gamma\{q_g^2, (y_g - \bar{y})^2\} = 0$ and $\Gamma\{q_g, (y_g - \bar{y})^2\} = 0$ are very similar in nature and assume a lack of relationship between cluster sizes and deviances between the cluster y_g and the overall mean. When the cluster sizes are equal, i.e. $q_g = 1/G$, then the conditions hold.

Note that theorem 2 is a generalization of theorem 1. Because of lemmas 2 and 4, theorem 2 can be extended again to series of variables generated from clustered, uniform grouping distributions with the same clustering of population units, i.e. sampled from the same subset of variables.

2.4. Expected numbers of categories

Next to the diffusion parameter D, the expected number of groups or variable categories, $E(C_A)$, determines the expected proportion of variance that is explained. The literature on sample occupancy numbers and occupancy probabilities (see Song *et al.* (2007)) can be employed to derive expressions for $E(C_A)$ under clustered uniform grouping (and the special case of uniform grouping).

When one variable is drawn from a clustered uniform grouping distribution with cell probabilities $\lambda_1^c, \lambda_2^c, \ldots, \lambda_c^c$ and $p_c = P(C = c)$, then the expected number of non-empty cells is

$$E(C_A) = \sum_{c=1}^{C} p_c \left\{ c - \sum_{m=1}^{c} (1 - \lambda_m^c)^G \right\}. \tag{6}$$

For equal cell probabilities, equation (6) reduces to

$$E(C_A) = \sum_{c=1}^{C} p_c c \left\{ 1 - \left(1 - \frac{1}{c} \right)^G \right\}, \tag{7}$$

and, when $p_L = 1$, for some L, then

$$E(C_A) = L \left\{ 1 - \left(1 - \frac{1}{L} \right)^G \right\}. \tag{8}$$

From equation (6), the number of categories over M independently generated variables from a clustered uniform grouping can be derived by noting that C cells are formed with probability $\Sigma_{\{(c_1,\ldots,c_M):\Pi_{k=1}^{M} c_k = C\}} \Pi_{m=1}^{M} p_{C_m}$ and that cell $(c_1, c_2, \ldots, c_M)$ has cell probability $\Pi_{m=1}^{M} \lambda_{c_m}^{C_m}$. For

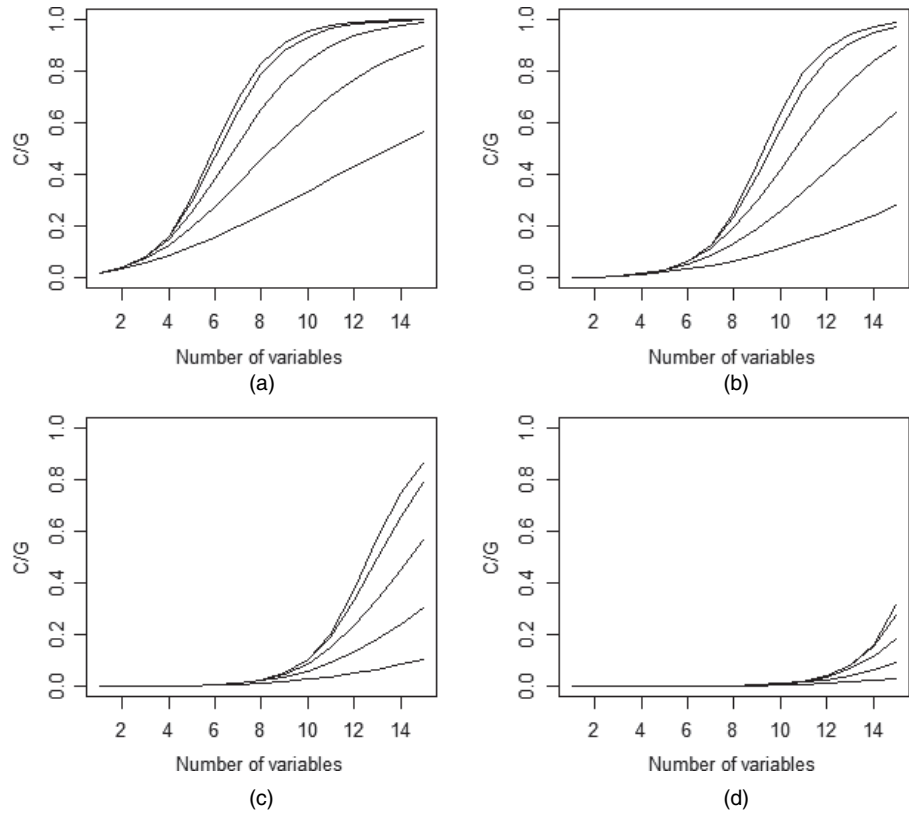

Fig. 1. Expected number of categories, $E(C_A)$, relative to the population diversity G, over M variables for various values of the population diversity (a) $G = 100$, (b) $G = 1000$, (c) $G = 10000$ and (d) $G = 100000$: per panel, the number of independently generated variables from uniform grouping is varied from $M = 1$ to $M = 15$; $P(C = 2) = 1$ in the clustered uniform grouping distribution and five values of λ_1^2 are shown, $\lambda_1^2 \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$

general probabilities p_c and $\lambda_1^c, \lambda_2^c, \ldots, \lambda_c^c$, the expression is cumbersome to write down. Instead, it is illustrated how fast the expected number of categories converges for $p_2 = 1$, i.e. two cells are formed per variable, for different values of the population diversity G and for different values of cell probability λ_1^2. Note that $\lambda_2^2 = 1 - \lambda_1^2$.

Fig. 1 depicts the expected number of categories relative to the population diversity G for increasing numbers of variables. Four populations are considered with an increasing diversity, $G = 100, 1000, 10\,000, 100\,000$. The cell probability λ_1^2 is varied from 0.5 to 0.9. As expected, convergence depends strongly on the population diversity and on the cell probability λ_1^2. For $G = 100$, all expected numbers of categories are close to G, except for $\lambda_1^2 = 0.9$, after $M = 15$ variables have been drawn. For $G = 100\,000$, 31 219 and 2987 non-empty cells are expected to be formed for cell probability $\lambda_1^2 = 0.5$ and $\lambda_1^2 = 0.9$ respectively.

Generally, as the length of the series increases, the expected number of groups, $E(C_A)$, will increase with it and the proportion of unexplained variance will decrease. However, although asymptotically $E(C_A) \to G$, under uniform grouping, the increases in the number of groups become smaller with every new variable and convergence is very slow. In practice, this finding is encountered in settings where auxiliary variables are used because they happen to be available.

2.5. Drawing samples from a population

In practice, often a population is observed through (random) samples. The theory in the previous sections can readily be applied to samples when variable-generating distributions are groupings of the resulting sampled subpopulation. However, when variables are generated on the full population first and next are applied to the sample, then the theory no longer holds. In this section, we briefly discuss the consequences of applying such variables to samples.

For convenience, we assume a simple random sample without replacement of size n from the population (of size N). The relative cluster sizes q_g are then replaced by random variables $q_{g,n}$ that vary from one sample to another. The absolute cluster size $nq_{g,n}$ has a binomial distribution, $\mathrm{Bin}(Nq_g, n/N)$. Now, theorem 2 can be extended by first conditioning on the $q_{g,n}$ and next taking expectations over the $q_{g,n}$. This gives

$$E_s[E_p\{S^2(Y_X)\}] = \frac{G\{E_p(C_A)-1\}}{G-1} \sum_{g=1}^{G} E_s(q_{g,n}^2)(y_g-\bar{y})^2,$$

where E_p denotes expectation over the variable-generating distribution and E_s over the sample. From the binomial distribution it follows that

$$E_s(q_{g,n}^2) = q_g^2 + \frac{1}{n}\left(1-\frac{n}{N}\right)q_g,$$

and we can conclude that simple random samples lead to a sample-size-dependent overestimation of the total variation.

The sample diffusion, $D_n = \Sigma_{g=1}^{G} q_{g,n}^2$, has expected value $E_s(D_n) = D + (1/n)(1-n/N)$. For relatively small samples, the sample diffusion will tend to $1/n$ and the diversity will be close to the sample size.

3. Estimation of population diversity and diffusion

Can the population diversity G be identified for clustered grouping distributions? The answer to this question is no, unless the grouping distribution is modelled. For clustered uniform grouping, it would be possible, but a large number of randomly drawn variables is needed, unless it is known or assumed that cluster sizes are equal.

However, the diffusion parameter D, that showed its importance in theorem 2, can be estimated on a relatively small series of variables. D is the inverse of the diversity G, when all clusters have an equal size, $q_g = 1/G$. Hence, when clusters have an equal, or nearly equal, size, estimating the diffusion parameter implies estimating the diversity.

Suppose that $(X_1, X_2, \ldots, X_M)^{\mathrm{T}}$ is generated independently from a clustered uniform grouping distribution. An obvious statistic to consider is Pearson's χ^2-statistic between pairs of variables. It is shown that this statistic is a simple function of D. Consider two variables, say X_1 and X_2. Let C_m be the number of non-empty groups for variable m, and let $\delta_{g,c}^m$ be the 0–1 indicator for cluster g in group c for variable m. The population χ^2-test statistic, when variables X_1 and X_2 are used to form a contingency table, is denoted as $\chi_N^2(1,2)$ and is defined in terms of observed and expected frequencies under independence

$$\chi_N^2(1,2) := \sum_{k=1}^{C_1}\sum_{l=1}^{C_2} \frac{\left(N\sum_{g=1}^{G} q_g\delta_{g,k}^1\delta_{g,l}^2 - N\sum_{g=1}^{G} q_g\delta_{g,k}^1\sum_{g=1}^{G} q_g\delta_{g,l}^2\right)^2}{N\sum_{g=1}^{G} q_g\delta_{g,k}^1\sum_{g=1}^{G} q_g\delta_{g,l}^2}, \tag{9}$$

which, because clustered uniform grouping is independent of the cluster sizes q_g, can be simplified to

$$\chi^2_N(1,2) = ND \sum_{k=1}^{C_1} \sum_{l=1}^{C_2} \frac{(G_{k,l} - G_{k.}G_{.l}/G)^2}{G_{k.}G_{.l}/G}, \tag{10}$$

with $G_{k,l}$ the count of clusters in cell (k,l) of the contingency table and $G_{k.}$ and $G_{.l}$ the marginal counts of clusters for X_1 and X_2 respectively. The expectation of equation (10) can be derived by conditioning on the numbers of groups C_1 and C_2. Following standard theory, the conditional expectation equals the degrees of freedom multiplied by D, i.e.

$$E\{\chi^2_N(1,2)|C_1, C_2\} = ND(C_1 - 1)(C_2 - 1), \tag{11}$$

and, hence,

$$E\{\chi^2_N(1,2)\} = ND\{E(C_A) - 1\}^2. \tag{12}$$

For equal cluster sizes, $q_g = 1/G$, equation (11) equals the degrees of freedom times the average absolute cluster size N/G.

Based on a series of variables $(X_1, X_2, \ldots, X_M)^{\mathrm{T}}$ the population diffusion can be estimated from realizations of Pearson's χ^2-statistics on all pairs of variables in a data set. Let $\chi^2_N(m_1, m_2)$ represent the statistic for variables X_{m_1} and X_{m_2} based on the full population. In practice, the statistic will often be estimated on the basis of a sample, say of size n. Let $\chi^2_n(m_1, m_2)$ be the sample-based statistic. To provide an asymptotically unbiased estimator for the population χ^2-statistic (10), $\chi^2_n(m_1, m_2)$ needs to be multiplied by N/n. Treating parameters that may drive the distribution of the groups sizes C_m as nuisance parameters, an estimator for the population diffusion may be derived by maximizing the conditional likelihood given the group sizes. To derive the estimator, it is assumed that $\{1/(ND)\}\chi^2_N(m_1, m_2)$ follows a χ^2-distribution with $(C_{m_1} - 1)(C_{m_2} - 1)$ degrees of freedom. It can be shown that the maximum conditional likelihood is obtained for

$$\hat{D} = \frac{\sum_{m_1=1}^{M-1} \sum_{m_2=m_1+1}^{M} \chi^2_n(m_1, m_2)/n}{\sum_{m_1=1}^{M-1} \sum_{m_2=m_1+1}^{M} (C_{m_1} - 1)(C_{m_2} - 1)}. \tag{13}$$

In deriving equation (13), it is ignored that the $\chi^2_n(m_1, m_2)$ are based on only M variables, which introduces some dependence between them.

It may again be assumed that variables are generated on the full population before they are applied to the sample. If so, the relative cluster sizes need to be replaced by sample-based versions and the expectation of the sample-based χ^2-statistic deviates from equation (11) because of the sampling variation. As was shown in Section 2.5, the expectation of the sample diffusion shows a sample-size-dependent overestimation of the true population diffusion. As a consequence, equation (13) also has a non-negative bias.

To this point, the cell probabilities λ_c^C have remained unspecified, but they obviously determine the size of groups and, therefore, the appearance of variables. Without further specification or model it is infeasible to estimate them, unless the number of variables is unrealistically large. The modelling and estimation of these parameters is left to future papers; a straightforward option might be to model them as drawn from Dirichlet prior distributions.

4. Application of framework to missing data in surveys

In this section, the general theory of Section 2 is applied to missing data in surveys. For missing data, the aim is to understand the magnitude of selection effects based on a set of auxiliary variables. The auxiliary variables are assumed to be drawn from a variable-generating distribution and bias in these variables translates to other non-observed variables.

In the application, the focus is on bias of population means. Let Y be a variable of interest in the survey that is observed only for respondents and let $\hat{\mu}_Y$ be an estimator for its population mean μ_Y, employing a vector of auxiliary variables, $X = (X_1, X_2, \ldots, X_M)^{\mathrm{T}}$. Let ρ_i represent the response probability of population unit i. The response probability is defined as the variation in response that is the result of changing individual and data collection circumstances that the survey cannot control or is not willing to control. The definition of a response probability has been subject to debate over recent decades. The postulation of response probabilities conforms to a so-called random-response model, as opposed to a fixed response model in which population units either respond or not, e.g. Schouten *et al.* (2009) and Särndal (2011). Here, we adopt a random-response model. Individual circumstances that cannot be fully controlled are, for example, the mood of the sample person or the family circumstances as a whole. Data collection circumstances that cannot be fully controlled are, for example, weather conditions, big (inter)national events like sports championships, the mood of interviewers or the malfunction of Internet connections. Let $\rho_X(x_i)$ be the response propensity for unit i given the auxiliary variables, e.g. Rosenbaum and Rubin (1983).

Schouten *et al.* (2016) showed that for the doubly robust estimator the maximal absolute remaining bias after adjustment, $|B(\hat{\mu}_Y)| = |E(\hat{\mu}_Y) - \mu_Y|$, under missingness not at random non-response, is proportional to

$$|B(\hat{\mu}_Y)| \propto \sqrt{[\{\mathrm{CV}^2(\rho) - \mathrm{CV}^2(\rho_X)\} R^2(Y, X)]}, \tag{14}$$

where CV is the coefficient of variation, $\mathrm{CV}(\rho) = S(\rho)/\bar{\rho}$, and the proportion of unexplained variance, $R^2(Y, X)$, is defined by equation (3). Either when the proportion of unexplained variance is 0 or when the variance of response propensities equals that of the true response probabilities, then equation (14) equals 0 and bias is removed. Schouten *et al.* (2015) showed that similar upper limits can be constructed for the inverse propensity weighting estimator and the generalized regression estimator.

Schouten *et al.* (2009) defined response to be strongly representative when response probabilities are constant for all population units and weakly representative with respect to X when the response propensities $\rho_X(x)$ are a constant function. $\mathrm{CV}(\rho)$ and $\mathrm{CV}(\rho_X)$ can be seen as the distance to strongly and weakly representative responses respectively.

In Section 4.1, first the setting is discussed where the range of survey target variables is wide and the objective may be to obtain a general representativeness of response. Next, in Section 4.2, the setting is discussed where there are a few key variables for which representativeness is needed.

4.1. Detection of general bias due to missing data

Suppose that the objective is to detect bias due to non-response in a survey, and that the variables of interest are diffuse and large in number, e.g. an omnibus survey or panel. In this setting, the interest is not in bias in a specific variable of interest. So the unexplained variance in equation (14) is of no direct interest and attention may be restricted to the first term

$$B_1 = \sqrt{\{\mathrm{CV}^2(\rho) - \mathrm{CV}^2(\rho_X)\}}. \tag{15}$$

Obviously, the ρ_i cannot be measured directly as they are largely unknown to population units themselves. The response to a survey is, however, one realization of this variable for the sample, and the response propensity ρ_X can be estimated.

Now, in the theorems of Section 2.2, let $y_i = \rho_i$ be the variable that needs to be explained. Since $\bar{\rho}_X = \bar{\rho}$ for any grouping distribution, $\mathrm{CV}(\rho_X)$ is $S(\rho_X)$ divided by the constant $\bar{\rho}$.

When X is constructed by uniform grouping, then by theorem 1

$$E\{\mathrm{CV}^2(\rho_X)\} = \frac{E(C_A) - 1}{N - 1}\mathrm{CV}^2(\rho), \tag{16}$$

and, when X is constructed by clustered uniform grouping, then by theorem 2

$$E\{\mathrm{CV}^2(\rho_X)\} = \frac{G\{E(C_A) - 1\}D}{G - 1}\mathrm{CV}_B^2(\rho), \tag{17}$$

when $\Gamma\{q_g^2, (\rho_g - \bar{\rho})^2\} = 0$ and $\Gamma\{q_g, (\rho_g - \bar{\rho})^2\} = 0$, and $\mathrm{CV}_B^2(\rho) = S_B^2(\rho)/\bar{\rho}^2$. It seems reasonable to assume that the covariances are negligibly small, as the diversity of most survey target populations may be expected to be relatively large and the q_g to be relatively small and close in size.

Ignoring higher order terms, it holds for the expected value of equation (15) that

$$E(B_1) \cong \sqrt{[\mathrm{CV}^2(\rho) - E\{\mathrm{CV}^2(\rho_X)\}]} = \sqrt{\left[E\{\mathrm{CV}^2(\rho_X)\}\frac{N - E(C_A)}{E(C_A) - 1}\right]}, \tag{18}$$

$$E(B_1) \cong \sqrt{[\mathrm{CV}^2(\rho) - E\{\mathrm{CV}^2(\rho_X)\}]} = \sqrt{\left[E\{\mathrm{CV}_B^2(\rho_X)\}\left[\frac{G - 1}{G\{E(C_A) - 1\}D} - 1\right]\right]}, \tag{19}$$

under uniform and clustered uniform grouping respectively.

Consequently, when two different survey designs lead to different $\mathrm{CV}(\rho_X)$ for variables $(X_1, X_2, \ldots, X_M)^{\mathrm{T}}$ from a uniform grouping distribution, then the design with the lowest value is to be preferred; a lower value implies that the expected remaining bias after adjustment with X by using a range of estimators is also smaller for an arbitrary other variable. When the grouping distribution is clustered uniform, then the same holds for an arbitrary other variable with the same clustering.

A natural follow-up question is whether it is sensible to pursue actively a survey response with a smaller $\mathrm{CV}(\rho_X)$ in the data collection stage. It can be shown that this is true under (clustered) uniform grouping. In adaptive survey designs, e.g. Schouten *et al.* (2013) and Wagner *et al.* (2013), resources are reallocated in between waves of a survey or during data collection to reduce the risk of non-response bias. Different groups, identified by using auxiliary variables, receive different treatments. Schouten *et al.* (2013) suggested formulating the allocation problem as a mathematical optimization problem with $\mathrm{CV}(\rho_X)$ as objective function, subject to cost, precision and logistical constraints. Within the range of designs that satisfy the constraints, the optimization prefers a design that has smallest $\mathrm{CV}(\rho_X)$. Theorem 3 in Appendix B shows that the optimized design is at least as good as the best strategy.

In the application of Section 5, it is demonstrated how a series of auxiliary variables can be employed to evaluate attrition in a panel, i.e. where the focus is on general representativeness.

4.2. Detection of non-response bias in a variable of interest

Very often surveys have a restricted set of topics and variables of interest. Say that Y is one such variable, generated from some grouping distribution $p_Y(C, s)$. Because a survey sample unit is usually informed about the topics and purpose of the survey, it cannot be assumed that Y is generated independently from the response probability variable ρ. However, when

a set of auxiliary variables $X = (X_1, X_2, \ldots, X_M)^{\mathrm{T}}$ is generated independently from both the survey outcome variable Y and the response probability ρ, then still they can be employed to make statements about the non-response bias for Y, and, subsequently, also about expected differences in bias for different survey designs.

One option to employ auxiliary variables is to derive the expectation of equation (14) under clustered uniform grouping. This is, however, not straightforward, as such an evaluation will involve the covariance between the response probabilities and the values of Y. It is, therefore, left to future research.

Another option to employ auxiliary variables is to construct relevant subsets of auxiliary variables based on the observed associations of auxiliary variables to either the survey outcome variable or the response probability. Under the first approach, auxiliary variables are included in the analysis whenever they have a minimal association with the survey outcome variable and are otherwise discarded. The bias component (15) can then again be evaluated and compared as in Section 4.1, but only for the selected auxiliary variables. Under the second approach, auxiliary variables are included in the analysis whenever they have a minimal association with response and are otherwise discarded. Now the other bias component, the coefficient of determination $R^2(Y, X)$, can be evaluated, but again for the selected variables only. The subset depends on the approach that is followed, i.e. a minimal association with the survey outcome variable or to the response probability, and to the original clustering of the auxiliary variable grouping itself and the height of the selection threshold. A minimal association may be operationalized by setting a lower threshold to the coefficient of determination or to other association measures like Cramèr's V.

Obviously, none of the auxiliary variables may meet the requirement, in which case no statements are possible. This may occur either by the small number of variables or by the clustering to which they are subject themselves in their generation. In fact, for clustered uniform grouping, the probability of generating a variable that satisfies the threshold may be 0. However, when the probability of accepting a variable is positive, then application of theorem 2 to the resulting grouping distribution may still be deemed useful and informative. When the auxiliary variables are themselves generated by uniform grouping, then obviously the survey outcome variables (under approach 1) or the response probabilities (under approach 2) are themselves elements of the subset.

In the application of Section 5, the effect of attrition on a number of survey outcome variables of general interest in a panel is evaluated. This is done by accepting auxiliary variables only when they show a minimal association with these outcome variables.

5. A case-study: the Dutch 'Longitudinal Internet studies for the social sciences' panel

To show the utility of the framework in checking validity of inference under missing data, the theory of Sections 2–4 is applied to the Dutch on-line 'Longitudinal Internet studies for the social sciences' (LISS) panel of the Center for Economic Research and Data at Tilburg University. The panel was established in November 2007 and panel members were recruited from simple random samples of the general population. First, the diversity of the panel population is discussed. Next, the attrition in the panel over the years 2007–2014 is evaluated from a general viewpoint and from the focus on four survey variables of interest. Finally, utility of the framework is discussed. The data set and R code can be found at `www.risq-project.eu/tools.html`.

5.1. The diversity of the panel members

People who register for the LISS panel receive monthly invitations to participate in up to three

surveys. The panel is set up for academic purposes; academic researchers can apply for a survey to be fielded in the panel. As a service to these panel users and to avoid the repetition of similar questions in different surveys, a series of 10 core studies is conducted for each panel member. The 10 core studies are repeated annually or biannually. The themes of the core studies are assets, family and household, income, health, housing, personality, politics and values, religion and ethnicity, social integration and leisure, and work and schooling. From the LISS panel, 2205 panel members were selected who participated in all LISS panel core studies that were fielded between December 2007 and December 2009. These 2205 people are followed in the subsequent five years up to December 2014 and their answers to the core study surveys are treated as auxiliary variables.

A total of 1306 variables was selected from the 2480 survey questions in the 10 core studies. The survey questions that were not selected either did not apply to the person or could not be used for various reasons. The following survey items were omitted: questions about other people than the panel member or about general knowledge, open-ended questions, questions with continuous measurement level and no obvious classification, questions with a high rate (greater than 5%) of refusal answers, and factual questions with a high rate (greater than 5%) of 'do not know' answers. Part of the questions with a continuous measurement level were, however, available as recoded categorical variables and were used to replace the original non-coded questions. 'Do not know' answers to non-factual questions were coded as separate substantive categories. Given the range of topics and the sheer size, the core study variables are viewed as the universe from which variables can be selected at random. The average number of variable categories for the 1306 selected questions is 4.95 but is skewed to the right. The median number of categories is 4. The variables include 10 demographic and economic variables that are asked or linked in many cross-sectional surveys: gender, age, marital status, ethnicity, income, educational level, type of household, degree of urbanization of area of residence, type of dwelling and socio-economic status. In the following sections, these 10 variables will be compared with random draws of variables. They are denoted as X_D.

Can the diversity of the 2205 selected panel members be estimated? First, it should be noted that core study survey questions were taken from existing general population surveys. Hence, it must be assumed that they were first formed for the population as a whole and then applied to the panel. Given the reasoning of Section 2.5 and given that the sample of 2205 people is small relative to the full population, the diversity must tend to 2205. Indeed, no two panel members were found who answered identically to all 1306 questions. Hence, $G = 2205$ is taken as the diversity parameter. As a check, the diversity was also estimated by using equation (13). In doing so, it was assumed that the clusters are equal in size, so that the diffusion is the inverse of the diversity G. The resulting estimate for G is $\hat{G} = 431$, which is substantially lower than 2205. Hence, there is some clustering in the 1306 survey questions, even after mixing the 10 core study surveys. This is most likely to be due to the repetition of similar survey questions to form psychometric and sociometric latent constructs. This clustering is ignored in what follows when drawing variables but may have a small effect.

5.2. Evaluating the general effect of attrition in the panel

The attrition patterns in the years 2010–2014 of the 2205 panel members are evaluated by using their answers to the 1306 questions from 2007 to 2009. The panel has an infinite horizon; panel members drop out only when they become inactive for a longer period. Five time points are considered, January 1st of 2010, 2011, 2012, 2013 and 2014. The first row of Table 1 contains the rates of attrition at each time point. Since the panel has a wide range of possible survey

Table 1. Estimated coefficients of variation averaged over all single variables, for random draws of 10 variables and for the 10 demographic variables

		Results for the following years:				
		2010	*2011*	*2012*	*2013*	*2014*
Attrition rate (%)		12	21	27	31	35
Single X	$\overline{\text{CV}}(\rho_{X_1})$	0.006	0.010	0.014	0.017	0.019
	$\widehat{\text{CV}}(\rho)$	0.141	0.239	0.336	0.393	0.451
Random $X_1, \ldots, X_{10}$	$\overline{\text{CV}}(\rho_{X_{10}})$	0.018	0.033	0.048	0.057	0.066
	$\frac{\overline{\text{CV}}(\rho_{X_{10}})}{\widehat{\text{CV}}(\rho)} \times 100$ (%)	13	14	14	15	15
Demographic	$\text{CV}(\rho_{X_D})$	0.033	0.069	0.089	0.106	0.127
	$\frac{\text{CV}(\rho_{X_D})}{\widehat{\text{CV}}(\rho)} \times 100$ (%)	23	29	26	27	28

topics, the general representativeness of the panel is of interest. Following Section 4.1, the representativeness is assessed by coefficients of variation CV of attrition propensities estimated at each time point. Four questions are asked.

(a) How does the expected total CV of attrition probabilities change over time?
(b) Do the demographic variables explain more variation than variables drawn from a uniform grouping?
(c) How much of the maximal variation comes from substantive variables?
(d) Is it sensible to adapt panel data collection?

The first two questions are answered by estimating three different CVs:

(a) the CV of attrition propensities in models with only one variable at a time averaged over all 1306 variables,
(b) the CV of attrition propensities in models with 10 randomly drawn variables, averaged over 200 independent draws and
(c) the CV of attrition propensities in a model with the 10 demographic variables.

The three are denoted as $\overline{\text{CV}}(\rho_{X_1})$, $\overline{\text{CV}}(\rho_{X_{10}})$ and $\text{CV}(\rho_{X_D})$ and are given in the second, fourth and last rows of Table 1.

Consider, first, the average single variable $\overline{\text{CV}}(\rho_{X_1})$ (second row). The values are relatively small. Since $G = 2205$ and $\widehat{E(C_A)} = 4.95$, theorem 2 tells us that the expected proportion of variance that is explained by one randomly drawn variable should indeed only be $(4.95 - 1)/2204 \times 100\% = 0.18\%$ of the CV of the response probabilities over the 2205 panel members. By application of equation (17), the third row of Table 1 estimates the true coefficient of variation as

$$\widehat{\text{CV}}(\rho) = \sqrt{\left\{ \frac{G-1}{\widehat{E(C_A)} - 1} \right\}} \overline{\text{CV}}(\rho_X), \tag{20}$$

which is roughly 24 times the average single variable CV. From 2010 to 2014, the values increase steadily and the representativeness becomes weaker when the panel members grow older.

Consider next the $\overline{\text{CV}}(\rho_{X_{10}})$ and $\text{CV}(\rho_{X_D})$ (fourth and sixth rows). The $\overline{\text{CV}}(\rho_{X_{10}})$ are approximately three times larger than the $\overline{\text{CV}}(\rho_{X_1})$, whereas the $\text{CV}(\rho_{X_D})$ are approximately six times

Table 2. Estimated total coefficients of variation divided by maximal coefficients of variation

Year	2010	2011	2012	2013	2014
$\frac{\widehat{\text{CV}}(\rho)}{\text{CV}_{\max}} \times 100$ (%)	39	47	55	59	62

Table 3. Preferred time points over 200 random draws of 10 auxiliary variables

Year	2010	2011	2012	2013	2014
Fraction preferred CV (%)	69	18	9	3	1

larger. The fifth and last rows in Table 1 show the estimated relative size of the coefficients relative to the estimated total coefficient, $\widehat{\text{CV}}(\rho)$. For the randomly drawn variables, the relative sizes fluctuate between 13% and 15%, and for the demographic variables they fluctuate between 23% and 29%. It must be concluded that the demographic variables explain much more variation than do average auxiliary variables.

Note that the $E(C_A) - 1$ for 10 randomly drawn variables is on average only nine times larger than for a single variable, confirming the slow convergence of explained variation for randomly drawn variables.

To answer the third question, the maximal coefficients are estimated. The maximal CV can be derived by assuming that all response probabilities are either 0 or 1, i.e.

$$\text{CV}_{\max} = \frac{\sqrt{\{\bar{\rho}(1-\bar{\rho})\}}}{\bar{\rho}}. \tag{21}$$

Table 2 presents the estimated coefficients $\widehat{\text{CV}}(\rho)$ divided over the $\text{CV}_{\max}$. The fractions grow steadily from 2010 to 2014 but still make up only 62% at the end. Hence, a considerable proportion of the maximal variation is not accounted for by the 1306 variables. Given that these variables are viewed as the universe of variables, a substantial part of the variation seems to come from panel data collection circumstances that cannot, or are not, controlled for.

The last question is about the utility of adaptation of panel data collection to known values of the auxiliary variables. Table 1 shows that the growing attrition from 2010 to 2014 coincides with growing CVs. However, the 1306 variables are not constant in their preferred time point of attrition. Table 3 presents the fractions of the 200 draws in which a time point was preferred in terms of the estimated coefficients of variation by the 10 randomly drawn variables. In around 70% of the draws, January 2010 was preferred, and for the remaining draws other years were preferred. This result implies that there is a potential for adaptive survey designs that vary the horizon of panel membership and/or refreshment strategies over auxiliary variables.

5.3. Evaluating the effect of attrition on key survey variables

Whereas in Section 5.2 the focus was on general representativeness, the focus, here, is on specific survey variables. From the core studies, four variables were selected that are viewed as key survey variables within the corresponding studies. They are usually asked as one of the first questions in the questionnaire or in one of the main questionnaire modules. The variables are 'How would you describe your general health?', 'Do you practice sports?', 'Do you agree or disagree with the

Table 4. Numbers of selected variables and selected demographic variables

	Number of selected X	
	All variables	*Demographic X*
Health	187	4
Sports	235	3
Surveys	25	0
Dwelling	411	10

statement "Surveys are important for society"?' and 'How satisfied are you with your current dwelling?'. The four questions are repeated in each wave of the corresponding core study. The answers to these four questions in the 2007–2009 core studies are used to follow the panel in the years 2010–2014.

The first approach of Section 4.2 is followed: auxiliary variables are selected on the basis of a minimal association with each of the four survey outcome variables and the coefficients of variation of the attrition propensities for the variables selected are evaluated and compared. Variables were maintained in the models for attrition whenever their value of Cramèr's V exceeded 0.05. This value seems rather small. However, the expected value of Cramèr's V is equal to the square root of the diffusion for a uniform grouping and is only about 0.02. For higher threshold values, the number of variables selected decreases very quickly. Two questions are asked:

(a) 'What representativeness patterns are found by auxiliary variables that satisfy the threshold?' and
(b) 'Are the same patterns found by demographic variables that satisfy the threshold?'.

To answer the first question, the selection of auxiliary variables is evaluated. The second and third columns in Table 4 show the overall number of variables and the number of demographic variables respectively that satisfy the threshold for each of the survey variables. The numbers are largest for the 'Dwelling' variable and all demographic variables reach the threshold. For the 'Survey' attitude variable, the numbers are smallest and none of the demographic variables reaches the threshold.

To answer the first question, 200 independent draws of 20 variables from the 1306 variables were generated, for each set, variables that did not reach the threshold were deleted and the average coefficient of variation over the 200 draws was computed; to avoid large numbers of empty models for attrition, 20 variables were drawn. Table 5 shows the average coefficients of variation for the subsets of selected auxiliary variables (the second to fifth rows). For comparison, the overall (estimated) coefficient of variation $\widehat{\mathrm{CV}}(\rho)$ of Section 5.2 is given in the first row. The results show that representativeness grows weaker for all four subsets when time progresses, but with different rates and with slightly different patterns. For the 'Dwelling' subset the representativeness is weakest throughout all years, whereas for the 'Surveys' subset it is strongest. For the 'Surveys' subset the coefficient stabilizes after 2012, whereas for the other variables there is a constant increase.

To answer the second question, the coefficients of variation for the demographic variables selected were estimated and are given in Table 5 (the sixth to ninth rows). The eighth row is

Table 5. Estimated total coefficients of variation, estimated coefficients of variation for the survey variables and estimated coefficients of variation for demographic variables after filtering on a minimal association to key survey variables per time point

		Results for the following years:				
		2010	*2011*	*2012*	*2013*	*2014*
$\widehat{\mathrm{CV}}(\rho)$		0.141	0.239	0.336	0.393	0.451
Random $X_1, \ldots, X_{20}$	Health	0.008	0.017	0.024	0.030	0.037
	Sports	0.006	0.016	0.026	0.031	0.038
	Surveys	0.003	0.005	0.009	0.008	0.008
	Dwelling	0.012	0.022	0.033	0.042	0.050
Demographic X	Health	0.030	0.065	0.075	0.095	0.102
	Sports	0.017	0.046	0.065	0.081	0.096
	Surveys	—	—	—	—	—
	Dwelling	0.032	0.069	0.090	0.106	0.127

empty, because none of the demographic variables is included in the 'Surveys' subset. Now, the coefficients are much larger, which may be expected from the findings of Section 5.2. Like for the random draws, the demographic variables indicate that the 'Dwelling' subset shows the weakest representativeness in all years. However, for the 'Health' and 'Sports' subsets, patterns are slightly different and do not fully confirm the findings for the random draws.

5.4. Checking validity of statistical inference

In the application, does the framework assist in checking the validity of inference? Two subsets of auxiliary variables were distinguished: a random selection and standard demographic and socio-economic variables. In a panel, random subsets may be obtained in a panel recruitment survey or a baseline core study. The demographic subset may also be obtained in a recruitment survey or by linkage from administrative data. The standard subset is available and used in analyses and adjustment of non-response bias in many surveys. In inference about the variables of interest in the LISS panel, researchers will normally attempt to account for panel attrition by weighting response or imputing non-response given the auxiliary variables. To remove the selection bias, the attrition must be missing at random for a survey variable of interest using such subsets of auxiliary variables.

Under the framework, it becomes clear that attrition is not missing at random for variables of interest conditionally on both subsets of auxiliary variables and that remaining bias exists. It does show that the standard subset of auxiliary variables is relatively powerful. However, through the coefficients of variation of estimated propensities, the random subset does allow for estimating a maximal bias over all variables and over subsets of variables. Furthermore, it allows for ranking different designs to deal with attrition.

6. Discussion

This paper is based on the rationale that the nature of variables that are used for statistical inference is often discarded but is crucial in evaluating assumptions under which inference is valid. Such variables may be assumed to be picked in some random fashion from the universe of potential variables. Depending on the diversity of the population, the size of this 'universe' is

larger or smaller. Little diversity implies that the set of potential variables is small and independent draws of variables show more association. Four research questions were posed about the construction of a framework for the generation of variables, the consequences for associations between variables, the estimation of population parameters that are important in these associations and the utility of the framework for statistical inference. They are briefly discussed, before limitations and future research are discussed.

It is shown that a framework can be constructed for random generation of auxiliary variables. The straightforward approach is to enumerate and label all possible variables and to draw variables at random. This approach is, however, not useful as it does not model collinearity, which is the driving force in associations between variables. For this reason, an approach was taken where the population units are randomly grouped to form variables. Obviously, the framework needs more discussion and evaluation. For example, continuous variables have been ignored completely.

Associations between variables explored by using two classes of variable-generating distributions are considered: uniform grouping and clustered grouping. These two classes seem sufficiently wide to model a wide range of settings. The first, uniform grouping, amounts to a fully random selection of variables and leads to powerful conclusions about associations. When auxiliary variables are indeed selected at random, then they detect traces of missing data bias and associations and enable conclusions beyond the mere associations that they themselves show. The second, clustered grouping, corresponds to a random selection from subsets of the universe of variables. Clustering essentially bounds the potential to extrapolate observed associations and limits conclusions.

A first approach is presented to estimate the diversity and diffusion of a population; key parameters in associations between variables. Importantly, such an endeavour to construct estimators will always be limited to observable diversity and diffusion, i.e. given the subset of variables for which we have instruments. However, in organically grown populations, these features of a population should be relatively stable in time and change only gradually.

In the setting of missing data, the framework turned out to be a very useful way of looking at auxiliary variables. Assumptions about missing data mechanisms are substituted by assumptions about the generation of auxiliary variables. This substitution, obviously, does not solve the missing data problem. Nevertheless, assumptions about the generation of auxiliary variables may be more intuitive and, therefore, easier to check. With a range of data sets it may be possible to estimate the population diversity. From there, it may be possible to construct a basis of variables for a population, i.e. to select a set of variables that are not correlated and describe the full diversity of a population. Furthermore, it may be possible to judge a set of variables, e.g. from a survey, on their cohesion relative to a fully random set of variables. Such features would help checking validity of variable generation.

The framework could be applied in a relatively straightforward way to explore confounding of selection and treatment effects in causal inference, by viewing available confounders as randomly picked. For causal effects, the aim would be to understand the amount of confounding with selection in observational studies or quasi-randomized trials. Covariates, either baseline or observed during treatment, may be assumed to follow a variable-generating distribution, so that again confounding effects can be translated to other, non-observed variables. The same argument as given for inference under missing data can be applied: when one treatment shows a larger effect than another, after controlling for confounding detected for randomly drawn auxiliary variables, then the treatment is expected to show a larger effect after controlling for confounding on all variables. This argument resembles the discussion in Joffe (2000) about confounding by indication. An application to causal inference is left to future papers.

Given that one accepts the framework, there are still some challenges. First, the number of auxiliary variables must be large to draw conclusions. Essentially, the variables are just draws and, as usual, quite a few are needed to obtain a precise picture of the parameters of interest, i.e. population diversity and specific diversity of variables of interest. For numbers of auxiliary variables that are common in practice, precision may often remain too low. Second, it is assumed that variables are measured without error and are intrinsic to the population units. If an instrument shows faulty measurements or if a person provides answers with some measurement error, then the variables become obscured by the noise that is added. As a result, the diversity of the population is judged to be much higher than it really is, as all associations become attenuated. Third, and most importantly, it seems most reasonable that variables are generated from subsets of possible variables, i.e. by clustered random grouping. Consequently, conclusions apply to subsets as well and may underestimate the full diversity. It is plausible that the auxiliary variables that are used most frequently have actually proved themselves in time to be relevant in a broad sense. Probably the archetype variables are gender and age. The case-study hints at this conclusion.

These challenges may be picked up in future research. It would be worthwhile to attempt to replicate the estimation of population diversity as is done in this paper. One would have to deal with the complication of measurement error, but, as mentioned above, many populations may be expected to have stable features in time.

Acknowledgements

I thank Joep Burger, Mark van der Loo and Geert Molenberghs for their comments on earlier versions of the paper. Furthermore, I thank the Associate Editor and especially the referee who have spent considerable effort in reviewing the manuscript and really helped to improve the content and methodology. Finally, I express my gratitude to the Center for Economic Research and Data, Tilburg University, for being able to use the LISS panel data set.

Appendix A: Proof of theorems 1 and 2

Since theorem 1 is a special case of theorem 2, i.e. by taking clusters identical to units, it suffices to prove theorem 2.

Let $G_c = \Sigma_{g=1}^{G} \delta_{g,c}$ be the number of clusters that is assigned to cell c and $p_c = \Sigma_{g=1}^{G} \delta_{g,c} q_g$ be the relative size of cell c. In deriving the expectation $E\{S^2(Y_X)\}$, first the conditional expectations $E\{S^2(Y_X)|C_A = M, G_1 = n_1, \ldots, G_M = n_M\}$ are evaluated. The variance can be expressed as

$$S^2(Y_X) = \sum_{c=1}^{C} p_c \left(\frac{\sum_{h=1}^{G} \delta_{h,c} q_h y_h}{p_c} - \bar{y} \right)^2 = \sum_{c=1}^{C} G_c \frac{p_c}{G_c} \left(\frac{\sum_{h=1}^{G} \delta_{h,c} q_h y_h}{G_c} \frac{G_c}{p_c} - \bar{y} \right)^2, \tag{22}$$

so that

$$E\{S^2(Y_X)|C_A = M, G_1 = n_1, \ldots, G_M = n_M\} = \sum_{c=1}^{M} n_c E \left\{ \frac{p_c}{n_c} \left(\frac{\sum_{h=1}^{G} \delta_{h,c} q_h y_h}{n_c} \frac{n_c}{p_c} - \bar{y} \right)^2 \,\middle|\, G_c = n_c \right\}. \tag{23}$$

In equation (23) the p_cs and $\delta_{h,c}$s are the only random variables as n_c is fixed. Setting $A = \Sigma_{h=1}^{G} \delta_{h,c} q_h y_h / n_c$ and $B = p_c / n_c$, the conditional expectation on the right-hand side of equation (23) can be written as

$$E\left\{B\left(\frac{A}{B}-\bar{y}\right)^2\middle|G_c=n_c\right\}=E\left(\frac{A^2}{B}\middle|G_c=n_c\right)-2\bar{y}E(A|G_c=n_c)+\bar{y}^2E(B|G_c=n_c). \tag{24}$$

A second-order Taylor series approximation of the first term of equation (24) around $(E(A), E(B))$ leads to

$$E\left(\frac{A^2}{B}\right)=\frac{E(A)^2}{E(B)}+\frac{\mathrm{var}(A)}{E(B)}-2\frac{\mathrm{cov}(A,B)E(A)}{E(B)^2}+\frac{\mathrm{var}(B)E(A)^2}{E(B)^3}$$
$$+O[E\{B-E(B)\}^3]+O[E\{B-E(B)\}\{A-E(A)\}^2]+O[E\{B-E(B)\}^2\{A-E(A)\}], \tag{25}$$

where the condition $G_c=n_c$ is omitted from the notation for convenience.

The third- and higher order terms of equation (25) are ignored in what follows. The other terms in equation (25) can be evaluated:

$$E(B|G_c=n_c)=\frac{1}{G},$$
$$\mathrm{var}(B|G_c=n_c)=\frac{1}{n_c}\left(1-\frac{n_c}{G}\right)\frac{1}{G-1}\sum_{g=1}^{G}\left(q_g-\frac{1}{G}\right)^2, \tag{26}$$

$$E(A|G_c=n_c)=\frac{\bar{y}}{G},$$
$$\mathrm{var}(a|G_c=n_c)=\frac{1}{n_c}\left(1-\frac{n_c}{G}\right)\frac{1}{G-1}\sum_{g=1}^{G}\left(q_gy_g-\frac{\bar{y}}{G}\right)^2, \tag{27}$$

$$\mathrm{cov}(A,B|G_c=n_c)=\frac{1}{n_c}\left(1-\frac{n_c}{G}\right)\frac{1}{G-1}\sum_{g=1}^{G}q_g^2y_g-\frac{1}{n_c}\left(1-\frac{n_c}{G}\right)\frac{1}{G-1}\frac{\bar{y}}{G}. \tag{28}$$

Combining equations (24)–(28) gives

$$\begin{aligned}E\left\{B\left(\frac{A}{B}-\bar{y}\right)^2\middle|G_c=n_c\right\}&=\frac{1}{n_c}\left(1-\frac{n_c}{G}\right)\frac{G}{G-1}\bar{y}^2\sum_{g=1}^{G}\left(q_g-\frac{1}{G}\right)^2\\&\quad+\frac{1}{n_c}\left(1-\frac{n_c}{G}\right)\frac{G}{G-1}\sum_{g=1}^{G}\left(q_gy_g-\frac{\bar{y}}{G}\right)^2\\&\quad-2\frac{1}{n_c}\left(1-\frac{n_c}{G}\right)\frac{G}{G-1}\bar{y}\sum_{g=1}^{G}q_g^2y_g+2\frac{1}{n_c}\left(1-\frac{n_c}{G}\right)\frac{1}{G-1}\bar{y}^2\\&=\frac{1}{n_c}\left(1-\frac{n_c}{G}\right)\frac{G}{G-1}\left(\sum_{g=1}^{G}q_g^2y_g^2+\bar{y}^2\sum_{g=1}^{G}q_g^2-2\bar{y}\sum_{g=1}^{G}q_g^2y_g\right)\\&=\frac{1}{n_c}\left(1-\frac{n_c}{G}\right)\frac{G}{G-1}\sum_{g=1}^{G}q_g^2(y_g-\bar{y})^2.\end{aligned} \tag{29}$$

Now, inserting equation (29) in equation (23) gives

$$\begin{aligned}E\{S^2(Y_X)|C_A=M,G_1=n_1,\ldots,G_M=n_M\}&=\frac{G}{G-1}\sum_{g=1}^{G}q_g^2(y_g-\bar{y})^2\sum_{c=1}^{M}\left(1-\frac{n_c}{G}\right)\\&=\frac{G}{G-1}(M-1)\sum_{g=1}^{G}q_g^2(y_g-\bar{y})^2.\end{aligned} \tag{30}$$

Equation (30) does not depend on the G_c, so the conditioning on $G_1=n_1,\ldots,G_M=n_M$ can be removed:

$$E\{S^2(Y_X)|C_A=M\}=\frac{G}{G-1}(M-1)\sum_{g=1}^{G}q_g^2(y_g-\bar{y})^2. \tag{31}$$

Now, weighting equation (31) by the probabilities $P(C_A=M)$ leads to equation (5):

$$E\{S^2(Y_X)\}=\frac{G}{G-1}\{E(C_A)-1\}\sum_{g=1}^{G}q_g^2(y_g-\bar{y})^2. \tag{32}$$

Equation (31) can be rewritten as

$$E\{S^2(Y_X)\} = \frac{G}{G-1}\{E(C_A)-1\}[\Gamma\{q_g^2, (y_g-\bar{y})^2\} + DS^2(y) - D\Gamma\{q_g, (y_g-\bar{y})^2\}], \tag{33}$$

so the additional conditions, $\Gamma\{q_g^2, (y_g-\bar{y})^2\}=0$ and $\Gamma\{q_g, (y_g-\bar{y})^2\}=0$ lead to equation (6).

Appendix B: Proof of expected gain of adaptive survey design

Say that T strategies are available, labelled $d=1,2,\ldots,T$, where design d has response probabilities ρ_d. The adaptive survey design optimization creates a mix of these strategies based on the observed response propensities $\rho_{X,d}, d=1,2,\ldots,T$, which leads to a design with response probabilities $\tilde{\rho}$ and response propensities $\tilde{\rho}_X$. In general, $\tilde{\rho}_X \neq \rho_{X,d}$ but is a mix of the $\rho_{X,d}(c)$ over groups and strategies, unless one of the strategies is superior to all possible mixes. The following theorem holds for uniform grouping and can be extended to clustered uniform grouping following a similar reasoning.

Theorem 3. If X is generated from a uniform grouping distribution, then

$$E\{\min_{\tilde{\rho}\in\tilde{P}} S^2(\tilde{\rho}_X)\} \leqslant \frac{E(C_A)-1}{N-1}\min_d S^2(\rho_d), \tag{34}$$

$$E\{\min_{\tilde{\rho}\in\tilde{P}} \mathrm{CV}^2(\tilde{\rho}_X)\} \leqslant \frac{E(C_A)-1}{N-1}\min_d \mathrm{CV}^2(\rho_d), \tag{35}$$

where minimization is over $\tilde{P}=\{\rho=(\rho_1,\rho_2,\ldots,\rho_C)^{\mathrm{T}} | \rho_c \in \{\rho_{X,1}(c),\ldots,\rho_{X,D}(c)\}, c=1,2,\ldots,C\}$.

Proof. A proof is given for the case where $T=2$ and a single X is generated with $P(C=2)=1$. Generalizations to arbitrary T and to general $p(C)$ are straightforward but cumbersome in notation. Lemma 2 again helps to generalize to series of variables.

It is straightforward to derive the expected remaining within variance of response probabilities in any group c formed by a uniform grouping variable X, say $S_{\mathrm{w}}^2(\rho|X=c)$. Since under uniform grouping all clusters have the same distributional properties, it must hold that

$$E\{S_{\mathrm{w}}^2(\rho|X=c)\} = E\left\{\frac{E(C_A)-1}{C(N-1)}\right\}S^2(\rho). \tag{36}$$

Hence, a smaller observed variance of response propensities for a particular survey design gives a smaller expected remaining within variance for that design.

For $D=2$ and $C=2$, there are four possible designs: $d=1$ is assigned to both $X=0$ and $X=1$, $d=1$ is assigned to $X=0$ and $d=2$ is assigned to $X=1$, $d=2$ is assigned to $X=0$ and $d=1$ is assigned to $X=1$, and $d=2$ is assigned to both $X=0$ and $X=1$. The resulting response propensities are denoted by ρ_{X11}, ρ_{X12}, ρ_{X21} and ρ_{X22}. So $\rho_{Xkl}=(\rho_{X,k}(1),\rho_{X,l}(2))^{\mathrm{T}}$. Obviously, $\rho_{Xkk}=\rho_{X,k}$, as both groups always have design k.

Now, the left-hand terms of equations (34) and (35) reduce to

$$E\{\min_{\tilde{\rho}\in\tilde{P}} S^2(\tilde{\rho}_X)\} = E\{\min_{k,l} S^2(\rho_{Xkl})\},$$

$$E\{\min_{\tilde{\rho}\in\tilde{P}} \mathrm{CV}^2(\tilde{\rho}_X)\} = E\{\min_{k,l} \mathrm{CV}^2(\rho_{Xkl})\},$$

and, by standard probability theory, it holds that

$$E\{\min_{k,l} S^2(\rho_{Xkl})\} \leqslant \min_{k,l} E\{S^2(\rho_{Xkl})\}$$

and

$$E\{\min_{k,l} \mathrm{CV}(\rho_{Xkl})\} \leqslant \min_{k,l} E\{\mathrm{CV}(\rho_{Xkl})\}.$$

Since it is true that

$$E\{S^2(\rho_{Xkk})\} = E\{S^2(\rho_{X,k})\} = \frac{E(C_A)-1}{N-1}S^2(\rho_k)$$

and

$$E\{\mathrm{CV}^2(\rho_{Xkk})\} = E\{\mathrm{CV}^2(\rho_{X,k})\} = \frac{E(C_A)-1}{N-1}\mathrm{CV}^2(\rho_k),$$

theorem 3 holds for $T = 2$ and $C = 2$.

References

Heckman, J. J. (2008) Econometric causality. *Int. Statist. Rev.*, **76**, 1–27.

Joffe, M. M. (2000) Confounding by indication: the case of calcium channel blockers. *Pharmepidem. Drug Safty*, **9**, 37–41.

Kreuter, F. (2013) *Improving Surveys with Paradata: Analytic Use of Process Information*. New York: Wiley.

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

Mealli, F. and Rubin, D. B. (2015) Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, **102**, 995–1000.

Molenberghs, G., Beunckens, C., Sotto, C. and Kenward, M. G. (2008) Every missingness not at random model has a missingness at random counterpart with equal fit. *J. R. Statist. Soc.* B, **70**, 371–388.

Molenberghs, G., Njeru Njagi, E., Kenward, M. G. and Verbeke, G. (2012) Enriched-data problems and essential non-identifiability. *Int. J. Statist. Med. Res.*, **1**, 16–44.

O'Neill, B. (2009) Exchangeability, correlation and the Bayes' effect. *Int. Statist. Rev.*, **77**, 241–250.

Pearl, J. (2009) *Causality: Models, Reasoning and Inference*, 2nd edn. New York: Cambridge University Press.

Robins, J. M. and Hernán, M. A. (2009) Estimation of the causal effects of time-varying exposures. In *Longitudinal Analysis: Handbook of Modern Statistical Methods* (eds G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs), ch. 23. Boca Raton: Chapman and Hall–CRC.

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.

Rubin, D. B. (2005) Causal inference using potential outcomes. *J. Am. Statist. Ass.*, **100**, 322–331.

Särndal, C. E. (2011) Dealing with survey nonresponse in data collection, in estimation. *J. Off. Statist.*, **27**, 1–21.

Särndal, C. E. and Lundquist, P. (2014) Accuracy in estimation with nonresponse: a function of degree of imbalance and degree of explanation. *J. Surv. Statist. Methodol.*, **2**, 361–387.

Schouten, B., Calinescu, M. and Luiten, A. (2013) Optimizing quality of response through adaptive survey designs. *Surv. Methodol.*, **39**, 29–58.

Schouten, B., Cobben, F. and Bethlehem, J. (2009) Indicators for the representativeness of survey response. *Surv. Methodol.*, **35**, 101–113.

Schouten, B., Cobben, F., Lundquist, P. and Wagner, J. (2016) Does balancing survey response imply less nonresponse bias? *J. R. Statist. Soc.* A, **179**, 727–748.

Seaman, S., Galati, J., Jackson, D. and Carlin, J. (2013) What is meant by "Missing at random"? *Statist. Sci.*, **28**, 257–268.

Song, R., Green, T., McKenna, M. and Glynn, K. (2007) Using occupancy models to estimate the number of duplicate cases in a data system without unique identifiers. *J. Data Sci.*, **5**, 53–66.

Wagner, J., West, B. T., Kirgis, N., Lepkowski, J. M., Axinn, W. G. and Kruger Ndiaye, S. (2013) Use of paradata in a responsive design framework to manage a field data collection. *J. Off. Statist.*, **28**, 477–499.