Statistics Netherlands

Discussion Paper

Statistical inference based on randomly generated auxiliary variables

2015 | 15

Barry Schouten

November

Content

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 relations 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 a number of 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 nonresponse in surveys. However, the framework may be more generally used to understand the strength of associations 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.

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 Missing-Completely-at-Random, Missing-at-Random and Not-Missing-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 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, Molenberghs, Beunckens, Sotto and Kenward (2008) for a discussion on Missing-at-Random and Not-Missing-at-Random assumptions and the more general discussion on enriched data through coarsening in Molenberghs, Njeru Nagi, Kenward and Verbeke (2012). The standard approach is a sensitivity analysis in which Not-Missing-at-Random mechanisms are introduced through a number of sensitivity parameters, see for a very elegant example Linero and Daniels (2015). A sensitivity analysis, however, still by-passes the true problem, despite its utility to assess robustness of inference under missing data; it describes a possible path into Not-Missing-at-Random mechanisms, but does not tell us whether it is the right path and how far the path should go. See also Ibrahim, Chen and Lipsitz (2004) and Ibrahim, Chen, Lipsitz and Herring (2005) for a discussion on inference under missing covariates.

Causal inference theory provides structured methodology on how to evaluate causal relations and free the estimation of causal effects from confounding, but it does not tell us why causal relations 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, in order 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 needs to be modelled in order to understand the validity of assumptions underlying to statistical inference. The paper seeks to model the nature with which variables are generated and with which

associations occur between them. In the model, an important role is played by the diversity and diffusion of a population. Diversity is defined as the number of groups in a population with identical scores on all potential variables. Diffusion is defined in terms of the relative sizes of these groups. It is shown that diversity and diffusion are important properties of a population that appear in association measures when variables are treated as (randomly) generated from the universe of variables.

Three main research questions are discussed:

1. Can a sensible framework be constructed that captures random generation of auxiliary variables?
2. What implications follow from the framework about the associations between variables?
3. Can the diversity and diffusion of a population be estimated?

The framework is applied and demonstrated in the setting of missing data. Over the last decades the 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 is hard 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 missing-at-random mechanism. The framework presented here gives conditions to extrapolate the traces of bias found by auxiliary variables to other variables, i.e. to not-missing-at-random mechanisms; when auxiliary variables are randomly drawn from a subset of variables, then any bias found on these variables extends statistically to the subset as a whole.

The original motivation for this paper came from the pursuit to reduce the impact of missing data in surveys through so-called adaptive survey designs (Schouten, Calinescu, Luiten 2013, Wagner et al 2013 and 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 nonresponse is a signal of even larger biases on variables of interest to the survey. Typically, the proportion of explained variation in nonresponse by such variables is rather low, and the designs are often criticized for removing nonresponse 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. When a design is balanced on a random draw from a subset of variables, then the design is expected to show improved balance on other variables in the subset, even after adjustment. Consequently, the results of this paper can be applied more generally to evaluate any data collection or observational study.

An additional, more general, motivation for the framework came from the observation that, regardless of the presence of missing data, auxiliary variables often show little explanatory power for variables of interest. This observation relates to causal effects and the ability to control and manipulate. For the sake of brevity, the

framework is not elaborated for this setting. However, 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 a confounding detected for randomly drawn auxiliary variables from a subset of variables, then the treatment is expected to show a larger effect after controlling for confounding on the subset as a whole. This argument resembles the discussion in Joffe (2000) about confounding by indication. In section 2, the conceptual framework is laid out. In section 3, the estimation of diversity and diffusion is discussed. Section 4 describes the application to missing data in surveys. Section 5 ends with a discussion.

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

This section sets the basic theory and constructs. In the first subsection, some notation is introduced and population diversity and uniformity are defined. In the second subsection, the generation of variables on a population is modelled. In the final subsection, two theorems are presented under the framework where variables are treated as randomly generated from the universe of variables.

2.1 Population diversity and diffusion

Suppose there is a population of interest of size N on which measurements can be made using a set of potential instruments and that the measurements are termed variables once they are stored. Suppose the population consists of G fully homogeneous strata, labelled $g = 1,2,3, \dots, G$, with relative stratum sizes q_g, i.e. $\sum_{g=1}^{G} q_g = 1$. The strata have the same value on all possible variables. Then G is the population diversity and population diffusion is defined as the sum of squared stratum sizes:

Definition: The diversity of a population, G, is the number of strata in which a population can be divided so that all population units are identical, i.e. have the same value on all possible variables. The diffusion of a population, D, is defined as $D = \sum_{g=1}^{G} q_g^2$.

It is straightforward to show that $\frac{1}{G} \leq D < 1$, and the diffusion equals $\frac{1}{G}$ when all strata have an equal size. In section 2.3, it is shown that the diffusion plays an important role in associations between variables.

Clearly, population units are never fully identical; some instruments make continuous measurements and the corresponding variables have continuous measurement levels. So can G be smaller than N? One could clearly argue that truly continuous measurements do not exist and that one always measures on some very fine grid. However, that would just be a diversion and there are two real arguments why for many populations it may hold that $G < N$. The arguments relate to the purpose of measurements. First, there are no continuous measurements that are stable for a meaningful duration of time; measurements will lead to small changes when repeated in short time intervals and one will not view such changes as relevant. Second, and more importantly, there is a limit to what level is relevant to a measurer regardless of time; beyond a certain level there is no control or manipulation. These observations lead to two conclusions: First, diversity and diffusion change in time. They will usually do so very gradually, but sometimes also with shocks due to

immigration/emigration and births/deaths. Second, the actual values that population units have on a variable may be contaminated by noise.

The emphasis on all possible instruments is important as population diversity has no meaning otherwise; the observable diversity of a population depends on the available instruments and may change once new instruments are developed or discovered. Population diversity is about a maximum set of instruments. One may state that there must be a finite number of instruments and that there is, thus, also a limit to the diversity. Such a statement is merely philosophical as, in practice, it will be unknown whether that maximum set of instruments has been developed. The availability of instruments will be incorporated in modelling the sampling procedure of variables in section 2.2.

Diversity and diffusion could also be defined as properties of a superpopulation from which a finite population is drawn. The strata with their stratum sizes may then be seen as the blueprints of the superpopulation, and it's 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.

Two examples are given to illustrate the concepts, one with a designed population and one with an organically grown population.

Example - Billiards: Suppose a factory manufactures balls for conventional billiards, i.e. with two white cue balls and one red ball. Manufactured balls thus differ on colour but also on other features. One of the other features is the presence of irregularities in shape. Suppose the population has a total of $G = 36$ *strata, which is 2 (red/white) × 2 (yes/no irregularities) × 9 (other features). Per game, two white balls are manufactured and one red. The other features are equally distributed, except shape. Balls with irregularities are ten times as rare as balls with no irregularities. Hence, white balls without irregularities are 20 times more frequent than red balls with irregularities. It can be seen that* $q_g = \frac{1}{297}$ *if balls in* g *are red with irregularities,* $q_g = \frac{2}{297}$ *if balls in* g *are white with irregularities,* $q_g = \frac{10}{297}$ *if balls in* g *are red without irregularities, and* $q_g = \frac{20}{297}$ *otherwise. The diffusion is* $D = 0.052$*, which is only modestly larger than the lower limit* $1/36 \approx 0.028$*. At some point, the factory decides to remove all balls with irregularities. The diversity then changes to* $G = 18$ *and the diffusion increases to* $D = 0.062$*, which is, however, closer to the lower limit* $\frac{1}{18} \approx 0.056$*.*

Example - Population of a country: In a particular country, the inhabitants can be divided in G *strata with sizes generated from a Dirichlet distribution. Let* $q_1, q_2, \dots, q_G$ *be generated from a* $Dirichlet(G, \alpha)$ *distribution with size parameter G and constant shape parameter* α*. The (expected) diffusion is then approximately equal to* $ED = \frac{\alpha+1}{G\alpha+1}$*.*

A population by itself may be defined as a set of identifiable objects on which measurements can be made. Hence, at least one instrument has already been applied to demarcate the set of objects, e.g. to study humans instead of all mammals. More fundamentally, the identification implies also that measurements have already been made to define objects, e.g. to study households instead of persons. The identification, i.e. the identifier variable, is, therefore, not part of the set of instruments. Furthermore, the strata of any population can be seen as a subset of strata of larger populations, and, vice versa, a subpopulation is a subset of $\mathcal{g} = \{1,2,3,\dots,G\}$.

2.2 The generation of variables on a population

The number of variables that can be formed on a population can be very large, while 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. First, some basic notation is introduced.

An instrument is a random grouping of strata from the set $\mathcal{g}$. Let s_g be the indicator representing to what category stratum g is assigned, and let $s = (s_1, s_2, \dots, s_G)^T$ be the vector of indicators (with T for transpose). Let C be the (random) number of categories of the resulting variable, say Z. Let $p(C, s)$ represent a random grouping probability distribution defined on the set $(\{1\}) \cup (\{2\} \times \{1,2\}^G) \cup (\{3\} \times \{1,2,3\}^G) \cup \dots \cup (\{G\} \times \mathcal{g}^G)$. Here, $\{1\}$ is the constant variable, $\{2\} \times \{1,2\}^G$ are all variables with two categories, $\{3\} \times \{1,2,3\}^G$ are all variables with three categories, etc. Let $A_1, A_2, \dots, A_C$ denote the clusters of strata, i.e. the categories of Z, and let $p_c = \sum_{g \in A_c} q_g$ be the relative size of category c. Let $\delta_{g,c}$ be the 0-1 indicator for the event $\{s_g = c\}$. Finally, let C_{max} be the smallest c with $p[C > c] = 0$. The resulting clusters of population strata represent the categories of the variable Z, with category labels that result from the binding characteristics of the strata and that depend on the instrument measurement level. As a result, each population stratum g has a label z_g, which is constant for all g in the same cluster, i.e. $z_{g_1} = z_{g_2}$ if $\exists c$ with $\delta_{g_1,c} = \delta_{g_2,c} = 1$.

In case of continuous measurement, not Z itself is observed but $\tilde{Z} = Z + \varepsilon^z$, where ε^z reflects irrelevant random noise. Let i be a population unit that falls in group c, then the actual measurement is $Z_i = z_g + \varepsilon_i^z$. The error term ε_i^z varies per unit and is the consequence of instrumental imprecision and/or time instability. It is not the consequence of measurement error, i.e. a malfunction of the instrument or a deficiency in the answering process of a respondent, but measurement error may further reduce the precision or even lead to a different classification of the unit. Hence, for fully continuous measurements, variables may be viewed as being generated from a factor model with G dimensions with weights q_g. This refinement is not further elaborated, however.

Multiple instruments, labelled $m = 1,2, \dots, M$, are independent draws from possibly different distributions $p_m(C, s)$, and lead to series of variables $Z_1, Z_2, Z_3, \dots, Z_M$. 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 $G = 100$, which seems to be a modest population diversity, 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 variable 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 stratum covariance between two realizations of variables, say Z_1 and Z_2, will be denoted by $\Gamma(Z_1, Z_2)$, with

$$\Gamma(Z_1, Z_2) = \frac{1}{G}\sum_{g=1}^{G} z_{1,g} z_{2,g} - \left(\frac{1}{G}\sum_{g=1}^{G} z_{1,g}\right)\left(\frac{1}{G}\sum_{g=1}^{G} z_{2,g}\right).$$

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

$$p_1(s_g = 1) = 0.6, \forall g, \text{ and } p_2(s_g = 1) = \begin{cases} 0.8 & \text{if} \quad g = 1; \\ 0.6 & \text{if} \quad g = 2; \\ 0.1 & \text{if} \quad g = 3. \end{cases}$$

Let Z_1 and Z_2 be randomly generated from one of the two distributions and let $s_{m,g}$ denote the 0-1 indicator for selection of stratum g for variable m. The expected probability that any of the variables equals one is $P[Z_m = 1] = \frac{1}{G}\sum_{g=1}^{G} p_k(s_{m,g} = 1)$. The expected probability that they jointly equal one is $P[Z_1 = 1, Z_2 = 1] = \frac{1}{G}\sum_{g=1}^{G} p_k(s_{1,g} = 1, s_{2,g} = 1) = \frac{1}{G}\sum_{g=1}^{G} p_k^2(s_{1,g} = 1)$. 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, in order to reach a negative expected covariance one has to 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. One important observation is made that will be very helpful in the following: 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 $Z_1, Z_2, Z_3, \dots, Z_M$, is itself generated from some $\tilde{p}(C, s)$ on the same population.

Proof: Each variable Z_m leads to groups $A_{m,1}, A_{m,2}, \dots, A_{m,C_m}$. 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 $\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_{\tilde{C}}$. The probability that a specific combination $\tilde{C}$ and $\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_{\tilde{C}}$ occurs, depends on the underlying distributions to the variables $Z_1, Z_2, Z_3, \dots, Z_M$ and defines $\tilde{p}(C, s)$. Because G is assumed finite, such a distribution always exists. □

A natural subclass of grouping distributions are distributions that have equal assignment probabilities for all strata in $\mathcal{g}$. They are termed uniform grouping and are defined as follows:
Definition: $p(C, s)$ is a uniform grouping distribution if conditional on the number of groups C the strata are assigned following a multinomial distribution with sample size parameter G and some cell probabilities, say $\lambda_1^C, \lambda_2^C, \dots, \lambda_C^C$.

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 from all possible variables for the population. Note, however, that some groups may not be assigned any strata and remain empty. Let the random variable C_A denote the number of non-empty strata. It requires a special construction of $p(C)$ to arrive at a completely random selection from the universe of all variables. This is not derived here. 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 $Z_1, Z_2, Z_3, \dots, Z_M$ from the family of uniform grouping, is itself generated from some uniform grouping $\tilde{p}(C, s)$ on the same population.

Proof: Let $\tilde{p}(C = c) = p(\min(G, \prod_{m=1}^{M} C_{m,R}) = c)$ and $\tilde{p}(s|C = c)$ follow a multinomial distribution with sample size parameter G and cell probabilities $\widetilde{\lambda_c^C}$. The $\widetilde{\lambda_c^C}$'s could, in principle, be derived from the cell probabilities $\lambda_1^{C_m}, \lambda_2^{C_m}, \dots, \lambda_C^{C_m}$ 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 one 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 $Z_1, Z_2, Z_3, \dots, Z_M$. □

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

Definition: $p(C, s)$ is a clustered grouping distribution if $\exists g_1, g_2$ for which $p(s_{g_1} = s_{g_2}) = 1$.

Clustered grouping distributions imply that two strata can never be discerned, i.e. the experimenter has no instrument that enables separation of the two sets of elements. It should be noted that also for non-clustered grouping distributions it may occur by chance that two strata 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 $Z_1, Z_2, Z_3, \dots, Z_M$ 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 strata may end up in different groups with a non-zero probability. It must surely hold for intersections of these groups that two strata 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 $Z_1, Z_2, Z_3, \dots, Z_M$ 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 strata end up in the groups. It holds for intersections of these groups that clusters still appear in the same intersections. □

The lemmas 1 to 4 together ensure that uniform grouping and clustered 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 grouping distribution that is uniform and clustered. The two examples of section 2.1 are elaborated.

Example – Billiards: Obviously, balls are designed to show predictable and reproducible behaviour, such that games effectively select the best players in a tournament. The manufacturer periodically performs test on the balls. There is an intensive manual test based on inspection and a superficial simple ballistics test. The manual test can be seen as uniform grouping, the testing employee is free to choose measurement. The ballistics test is a clustered uniform grouping, as, for example, colour and irregularities in shape are not recorded.

Example – Population of a country: A Crime Victimisation survey is conducted in the country. The questionnaire contains a range of survey items related to victimisation, perceptions of safety and judgment about the performance of authorities and police to avoid and reduce crime. The authorities have registered information on the population from other data collections and use these to understand the sentiments and mechanisms behind victimisation. Both the questionnaire and the administrative data may be seen as (different) clustered uniform groupings.

2.3 Associations between randomly generated variables

Suppose that an analysis is directed at explaining a variable of interest Y using auxiliary variables $(X_1, X_2, \dots, X_M)^T$. For the sake of demonstration, let Y be quantitative, i.e. its category labels y_g correspond to measurement values. A researcher may then be interested in the variance $S^2(Y_X)$, where Y_X is the projection of Y on the space formed by the variables $(X_1, X_2, \dots, X_M)^T$.

Let the projection for stratum g, $Y_{X,g}$, be defined as

$$Y_{X,g} = \sum_{c=1}^{C} \delta_{g,c} \frac{\sum_{h=1}^{G} \delta_{h,c} q_h y_h}{\sum_{h=1}^{G} \delta_{h,c} q_h}, \quad (1)$$

with y_g the value of Y on stratum g. Next, let $\bar{Y}_X$ be the average of the projected values and let $S^2(y)$ be 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.
Here, the variation in the individual error terms ε^y over population units is ignored. This variation adds noise and tends to decrease the amount of variation that can be explained. Hence, it is assumed that the number of observed units is sufficiently large.

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

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

$$ES^2(Y_X) = \frac{G(EC_A - 1)}{G-1} \sum_{g=1}^{G} q_g^2 (y_g - \bar{y})^2, \quad (2)$$

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

$$ES^2(Y_X) = \frac{G(EC_A - 1)D}{G-1} S^2(y), \quad (3)$$

Proof: 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 size of the cell probabilities $\lambda_1^C, \lambda_2^C, \dots, \lambda_C^C$ in the uniform grouping is irrelevant. Hence, it does not matter whether cells are formed at very different sizes or nearly equal sizes. 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 relation between stratum sizes and deviances between the y_g and their mean. When the stratum sizes are equal, i.e. $q_g = \frac{1}{G}$, then the conditions hold.

A researcher may be interested in the proportion of unexplained variance, $R^2(Y)$, defined as one minus the coefficient of determination

$$R^2(Y, X) = 1 - \frac{S^2(Y_X)}{S^2(y)}. \quad (4)$$

Under the conditions of theorem 1, the expected value of (4) reduces to $ER^2(Y,X) = 1 - \frac{G(EC_A-1)D}{G-1}$. Generally, as the length of the series increases, the expected number of groups, EC_A, will increase with it and the proportion of explained variance will decrease. However, although asymptotically $EC_A \to G$, 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 were auxiliary variables are used because they happen to be available.

For clustered uniform grouping, a similar result can be derived. Let the grouping distribution have K clusters of strata, labelled $k = 1,2,\dots,K$, let Q_k be the size of cluster k, i.e. the sum of the q_g in cluster k, and $\bar{y}_k$ be the average of Y in cluster k weighted by the q_g. Theorem 2 is the analogue of theorem 1.

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

$$ES^2(Y_X) = \frac{K(EC_A-1)}{K-1}\sum_{k=1}^{K} Q_k^2(\bar{y}_k - \bar{y})^2 \qquad (5)$$

If, additionally, $\Gamma(Q_k^2, (\bar{y}_k - \bar{y})^2) = 0$ and $\Gamma(Q_k, (\bar{y}_k - \bar{y})^2) = 0$, then

$$ES^2(Y_X) = \frac{K(EC_A-1)}{K-1} D_Q S_B^2(y). \qquad (6)$$

with $S_B^2(y)$ the between variance based on the clusters and $D_Q = \sum_{k=1}^{K} Q_k^2$.

Proof: The proof follows directly from theorem 1 by replacing G by K, and the strata $g = 1,2,\dots,G$ by the clusters $k = 1,2,\dots,K$. □

Because of lemma's 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 strata, i.e. sampled from the same subset of variables.
The two examples are further elaborated to illustrate the implications of the two theorems.

Example – Billiards: The interest of the manufacturer lies in the behaviour of the balls in game situations: the proportion of shots that leads to a detectable deviation. Let this be Y. The simple ballistics test and intensive manual test both deliver auxiliary variables that may explain Y, but have different costs associated with them. Suppose that colour nor irregularities relate to Y. This implies that (3) holds. The diffusion equals $D = 0.052$ in case balls with irregularities are kept in the population, while it is $D = 0.062$ when they are removed. For the manual test, (4) is then equal to $ER^2(Y,X) = 1 - \frac{G-1}{G}D(EC_A - 1) = 1 - 0.053(EC_A - 1)$ and $ER^2(Y,X) = 1 - 0.065(EC_A - 1)$, with and without removal of balls with irregularities, respectively. The simple test clusters colour and irregularities, but let it satisfy uniform grouping on the other measurements. Suppose, $K = 9$, equally sized clusters remain when colour and irregularities are clustered in the simple test on the $G = 36$ strata. Hence, $D_Q = \frac{1}{9}$ and (6) reduces to $ES^2(Y_X) = \frac{1}{8}(EC_A - 1)S_B^2(y)$. Since Y does not relate to colour

and irregularities, the between variance is equal to the overall variance, $S_B^2(y) = S^2(y)$*, and (4) reduces to* $ER^2(Y,X) = 1 - 0.125(EC_A - 1)$*. Note that* C_A *follows different distributions for the different settings.*

Example – Population of a country: Suppose that the auxiliary variables arose from uniform grouping and are used to explain one of the variables of interest from the Crime Victimisation survey. The stratum sizes were randomly constructed from a symmetric Dirichlet distribution and the squared stratum sizes can be assumed to be unrelated in any way to the key survey variable. The expected diffusion equals $ED = \frac{1}{G} + \frac{G-1}{G(G\alpha+1)} = \frac{\alpha+1}{G\alpha+1}$*. Again (3) holds and* $ER^2(Y,X) = 1 - \frac{G(EC_A-1)ED}{G-1} \cong 1 - (EC_A - 1)\frac{\alpha+1}{G\alpha+1}$*, when* G *is large.*

3. Estimation of population diversity and diffusion

Can the population diversity G be identified? The answer to this question is clearly no, because, at any given point in time, it will be unknown whether all potential instruments are available to be applied to the population. It will, hence, only be possible to estimate the diversity over the clusters formed by all available instruments. However, simultaneously, it must be noted that the observable population diversity will in most populations, especially those that are organically grown, change very gradually. It should, thus, be possible to estimate this parameter using a large range of independent sets of measurements on the population. In the following, the limitation to all available instruments is ignored, but should be kept in mind.

While, in general, large numbers of variables are needed to estimate the observable population diversity and uniformity, the diffusion parameter D, which showed its importance in theorem 1, can be estimated on a relatively small series of variables. It should be noted that D is the inverse of the diversity G when all strata have an equal size, $q_g = \frac{1}{G}$. Hence, when strata have an equal, or nearly equal, size, estimating the diffusion parameter implies estimating the diversity.

Suppose $(X_1, X_2, \dots, X_M)^T$ is generated independently from a uniform grouping distribution. An obvious statistic to consider is the chi-square 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 stratum g in group c for variable m. The chi-square test statistic is denoted as $\chi^2_{m_1,m_2}$, when variables m_1 and m_2 are used to form a contingency table. Taking $m_1 = 1$ and $m_2 = 2$, it is defined in terms of observed and expected frequencies under independence

$$\chi^2_{1,2} := \sum_{k=1}^{C_1}\sum_{l=1}^{C_2} \frac{\left(\frac{\sum_{g=1}^{G} q_g \delta^1_{g,k}\delta^2_{g,l}}{G} - \frac{\sum_{g=1}^{G} q_g \delta^1_{g,k}}{G}\frac{\sum_{g=1}^{G} q_g \delta^2_{g,l}}{G}\right)^2}{\frac{\sum_{g=1}^{G} q_g \delta^1_{g,k}}{G}\frac{\sum_{g=1}^{G} q_g \delta^2_{g,l}}{G}}, \tag{7}$$

which, because uniform grouping is independent of the stratum sizes q_g, be simplified to

$$\chi^2_{1,2} = D\sum_{k=1}^{C_1}\sum_{l=1}^{C_2} \frac{\left(\frac{G_{k,l}}{G} - \frac{G_{k.}}{G}\frac{G_{.l}}{G}\right)^2}{\frac{G_{k.}}{G}\frac{G_{.l}}{G}}, \tag{8}$$

with $G_{k.}$ and $G_{.l}$ the marginal counts of strata on X_1 and X_2, respectively. The expectation of (7) 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.

$$E(\chi^2_{1,2}|C_1, C_2) = D(C_1-1)(C_2 - 1), \quad (9)$$

and, hence,

$$E(\chi^2_{1,2}) = D(EC_A - 1)^2. \quad (10)$$

Note that for equal stratum sizes, $q_g = \frac{1}{G}$, (9) equals the degrees of freedom divided by G.

Based on a series of variables $(X_1, X_2, \dots, X_M)^T$ the population diffusion can be estimated from realizations of Pearson's chi-square test statistics on all pairs of variables in a sample data set. The Pearson's chi-square test statistic needs to be divided by the sample size to obtain an asymptotically unbiased estimator for the population chi-square statistic (8). Treating parameters that may drive the distribution of the groups sizes C_m as nuisance parameters, an estimator for the population diversity may be derived by maximizing the conditional likelihood given the group sizes. In order to derive the estimator it is assumed that $G\chi^2_{m_1,m_2}$ follows a chi-square 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

$$\widehat{D} = \frac{\sum_{m_1=1}^{M-1}\sum_{m_2=m_1+1}^{M}\hat{\chi}^2_P(m_1,m_2)/n}{\sum_{m_1=1}^{M-1}\sum_{m_2=m_1+1}^{M}(C_{m_1}-1)(C_{m_2}-1)}, \quad (11)$$

with n the size of the sample and $\hat{\chi}^2_P(m_1, m_2)$ the Pearson's chi-square test statistic between X_{m_1} and X_{m_2}. In deriving (11), it is ignored that the $\hat{\chi}^2_P(m_1, m_2)$ are based on only M variables, which introduces some dependence between them.

To this point, the cell probabilities λ_c^C 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 using Dirichlet prior distributions, as is done in the example on a population of a country.

Example – population of a country: The diffusion parameter is estimated based on a real survey, the 2011 Dutch Crime Victimisation Survey (CVS). The survey is conducted based on probability samples from the Dutch municipality registers. The target population consists of all inhabitants of The Netherlands aged 15 years and older. To the sample a series of auxiliary variables can be linked from various administrative data sources available at Statistics Netherlands. In the application, ten variables are considered: age (15-25 years, 25-35 years, 35-45 years, 45-55 years, 55-65 years, 65-75 years, 75 years or older), ethnicity (native, western non-native, non-western non-native), gender (male, female), individual annual income (0-3k Euro, 3-10k Euro, 10-15k Euro, 15-20k Euro, 20-30k Euro, 30k Euro or more), province of residence (twelve provinces of The Netherlands), registered phone number (yes, no), subscription to an unemployment office (yes, no), type of household (single, single parent, not married couple, not married couple with children, married couple, married couple with children, other type), type of income (job, allowance, other), and urbanization level of the area of residence (very strong, strong, moderate, little, not). The variables are

taken as they are defined and used by the social statistics department for publication purposes. The size of the CVS sample is $n = 8766$.
Table 1 contains the numbers of categories for the auxiliary variables and the standardized Pearson's chi-squares. The variables have an average of 4.9 categories. From the chi-square statistics, an estimate for the population diffusion is derived; estimator $\widehat{D}$ *in (11) equals* $\widehat{D} = 0.0055$. *Based on a naïve bootstrap of the chi-square statistics, the 95%-confidence interval is* $(0.0025, 0.0094)$, *which is still rather wide for* $M = 10$ *variables. If indeed the stratum sizes were generated using a Dirichlet distribution, then* $G = \frac{\alpha+1-D}{\alpha D}$, *and* G *may be estimated as well for any choice of* α *as* $\widehat{G} = \frac{\alpha+1-\widehat{D}}{\alpha\widehat{D}}$. *If* $\alpha \to \infty$, *then the relative stratum sizes become equal and* $\widehat{G} = 182$.

Table 1: The numbers of groups and the standardized Pearson's chi-squares $\hat{\chi}_P^2(m_1, m_2)/n$ *for the auxiliary variables in the CVS example. Chi squares are given times 1000.*

	1	2	3	4	5	6	7	8	9	10
C_m	7	3	2	6	12	2	2	7	3	5
$\hat{\chi}_P^2(1, m_2)/n$	-	29	4	336	9	20	15	494	540	9
$\hat{\chi}_P^2(2, m_2)/n$	-	-	0	19	31	41	21	50	8	79
$\hat{\chi}_P^2(3, m_2)/n$	-	-	-	122	1	0	0	6	5	1
$\hat{\chi}_P^2(4, m_2)/n$	-	-	-	-	10	5	13	103	764	4
$\hat{\chi}_P^2(5, m_2)/n$	-	-	-	-	-	9	1	16	5	411
$\hat{\chi}_P^2(6, m_2)/n$	-	-	-	-	-	-	4	33	0	29
$\hat{\chi}_P^2(7, m_2)/n$	-	-	-	-	-	-	-	8	8	1
$\hat{\chi}_P^2(8, m_2)/n$	-	-	-	-	-	-	-	-	199	46
$\hat{\chi}_P^2(9, m_2)/n$	-	-	-	-	-	-	-	-	-	2

To end this section, two remarks are in place. First, this paper is conceptual and estimator (11) is merely meant to initiate discussion. There are alternative, and perhaps more natural, statistics that can be employed, like the order statistics of the absolute covariances found between the $(X_1, X_2, \dots, X_M)^T$. Second, the number of variables must be relatively large to precisely estimate the population diffusion. Consequently, in many settings, the actual standard error may still be large.

4. Application to missing data in surveys

In this section, the general theory of section 2 is applied to missing data in surveys. 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 nonresponse in a survey, and that the variables of interest are diffuse and large in number. In this setting, the interest is not in bias on a specific variable of interest. A vector of auxiliary variables, $(X_1, X_2, \dots, X_M)^T$, is available and ρ represents the response probability of a population element. The focus may then be on the coefficient of variation of the response probabilities, $CV(\rho) = S(\rho)/\bar{\rho}$, as a general measure of risk of nonresponse bias. It is easy to show that $CV(\rho)$ bounds the standardized absolute bias of any arbitrary variable, say Y. The bias of the mean of Y due to nonresponse, $B(Y)$, divided by its standard deviation, is approximately equal to

$$\frac{|B(Y)|}{S(Y)} = \frac{|\mathrm{cov}(Y,\rho)|}{S(Y)\bar{\rho}}, \tag{12}$$

and

$$\frac{|B(Y)|}{S(Y)} \leq \frac{S(\rho)}{\bar{\rho}}. \tag{13}$$

For a more elaborate discussion see Schouten, Cobben and Bethlehem (2009) and Särndal (2011).

Schouten, Cobben, Lundquist and Wagner (2014) show that the square root of the difference between the squared coefficients of variation of the true response probabilities and the response propensities,

$$\sqrt{CV^2(\rho) - CV^2(\rho_X)} \tag{14}$$

appears as a general term in the maximal absolute remaining nonresponse bias for most commonly used adjustment estimators. They show that the maximal absolute remaining bias of the expansion estimator, the generalized regression estimator, the inverse propensity weighting estimator, and the doubly robust estimator all are proportional to (14). It follows from Taylor linearization that

$$E\sqrt{CV^2(\rho) - CV^2(\rho_X)} \cong \sqrt{CV^2(\rho) - ECV^2(\rho_X)} + \frac{1}{2}\frac{var(S^2(\rho_X))}{\bar{\rho}^4(CV^2(\rho) - ECV^2(\rho_X))^{3/2}}, \tag{15}$$

so that $ECV^2(\rho_X)$ is also a crucial term in nonresponse adjusted estimates. The second term in (15) contains the variance of the response propensity variance, $\mathrm{var}(S^2(\rho_X))$. This variance is an interesting quantity for further research, as it determines the uncertainty about conclusions based on response propensity variation for a selected set of variables.

If one would be able to measure the ρ and when they are all strictly positive, then nonresponse bias on any variable can be removed. The actual realization of a survey may be seen as an instrument that measures this variable. However, it is a far from perfect instrument as per element only one realization is available, and it is, hence, contaminated by random, circumstantial influences. For this reason, researchers usually move towards the response propensity ρ_X, i.e. the projection of ρ on the space spanned by the auxiliary variables (Rosenbaum and Rubin 1983).

Now, let in the theorems of section 2.2, $y_g = \rho_g$ be the variable that needs to be explained. Since $\bar{\rho}_X = \bar{\rho}$ for any grouping distribution, $CV(\rho_X)$ is $S^2(\rho_X)$ divided by the constant $\bar{\rho}$.

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

$$ECV^2(\rho_X) = \frac{G(EC_A - 1)D}{G-1} CV^2(\rho), \tag{16}$$

when $\Gamma(q_g^2, (\rho_g - \bar{\rho})^2) = 0$ and $\Gamma(q_g, (\rho_g - \bar{\rho})^2) = 0$. In the following, it is assumed that the covariances are negligibly small. This is reasonable 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. Given that $\frac{G}{G-1} \approx 1$, the first term in (13) can be rewritten to

$$\sqrt{CV^2(\rho) - ECV^2(\rho_X)} = \sqrt{1 - (EC_A - 1)D}\, CV(\rho) = \sqrt{1 - \frac{1}{(EC_A - 1)D}}\, ECV(\rho_X). \tag{17}$$

This allows for an important conclusion: When two different survey or data collection designs lead to different $CV(\rho_X)$ and when the variables $(X_1, X_2, \dots, X_M)^T$ follow a uniform grouping distribution, then the design with the lowest value of the CV is to be preferred; a lower value implies that the expected remaining bias after adjustment with X using a range of estimators is also smaller for an arbitrary other, but not observed, variable. In other words, the design with the lower value of the CV is to be favoured when it comes to the risk of nonresponse bias.

A natural follow-up question is whether it is sensible to pursue a survey response with a smaller $CV(\rho_X)$ in the data collection stage. It is shown that, again under uniform grouping, this is true.

In adaptive survey designs, e.g. Schouten et al (2013) and Wagner et al 2013), resources are re-allocated in between waves of a survey or during data collection in order to reduce the risk of nonresponse bias. Different strata, identified using auxiliary variables, get different treatments. Schouten et al (2013) suggest to

formulate the allocation problem as a mathematical optimization problem with $CV(\rho_X)$ 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 $CV(\rho_X)$. As an alternative, they suggest to minimize $S^2(\rho_X)$ subject to an additional constraint on the response rate $\bar{\rho}$. Say, for example, T strategies are available, labelled $d = 1,2, \dots, T$, where design d has response probabilities ρ_d. The optimization creates a mix of these strategies based on the observed response propensities $\rho_{X,d}$, $d = 1,2, \dots, T$, 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 Theorem 3 shows that the optimized design is at least as good as the best strategy.

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

$$E\, min_{\tilde{\rho}\in\tilde{P}}\, S^2(\tilde{\rho}_X) \leq \frac{EG(EC_A-1)D}{G-1} min_d\, S^2(\rho_d), \qquad (18)$$

$$E\, min_{\tilde{\rho}\in\tilde{P}}\, CV(\tilde{\rho}_X) \leq \frac{G(EC_A-1)D}{G-1} min_d\, CV(\rho_d), \qquad (19)$$

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

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.
From (16), it is straightforward to derive the expected remaining within variance of response probabilities in any stratum c formed by a uniform grouping variable X, say $S_w^2(\rho|X = c)$. Since under uniform grouping all strata have the same distributional properties, it must hold that

$$ES_w^2(\rho|X = c) = E\left(\frac{G(EC_A-1)D}{C(G-1)}\right) S^2(\rho). \qquad (20)$$

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 $\rho_{X11}, \rho_{X12}, \rho_{X21}$ and ρ_{X22}. So $\rho_{Xkl} = (\rho_{X,k}(1), \rho_{X,l}(2))^T$. Obviously, $\rho_{Xkk} = \rho_{X,k}$, as both strata always get design k.
Now the left hand terms of (18) and (19) reduce to

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

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

and, by standard probability theory, it holds that $E\min_{k,l}(S^2(\rho_{Xkl})) \leq \min_{k,l}(ES^2(\rho_{Xkl}))$ and $E\min_{k,l}(CV(\rho_{Xkl})) \leq \min_{k,l}(ECV(\rho_{Xkl}))$. Since it is true that $ES^2(\rho_{Xkk}) = ES^2(\rho_{X,k}) = \frac{G(EC_A-1)D}{G-1}S^2(\rho_k)$ and $ECV(\rho_{Xkk}) = ECV(\rho_{X,k}) = \frac{G(EC_A-1)D}{G-1}CV(\rho_k)$,

the theorem holds for $T = 2$ and $C = 2$. □

Example – population of a country - continued: Suppose the interest is in general representativeness of the survey response. Five data collection designs are considered: Web only, mail only, face-to-face only, Web → face-to-face and mail → face-to-face. The last two designs are sequential; face-to-face is only offered to nonrespondents in Web and mail, respectively. Table 2 shows the coefficients of variation for the ten variables in the five designs. The coefficients are shown per variable, averaged over the ten variables and for a model in which all variables are included. The ten variables together show a preference for the sequential design mail → face-to-face ($CV = 0.16$), which slightly outperforms the Web → face-to-face ($CV = 0.18$). The single mode design Web is by far the least favourite ($CV = 0.36$). Given the estimated population diffusion and estimated expected number of groups, the average coefficients of variation per design over the ten variables capture an estimated $100\% \times \sqrt{(4.9-1) \times 0.0055}$ = 15% of the coefficient of variation of the response probabilities. Table 2 shows that the ten variables together have roughly a three times higher value for the coefficient of variation than the average per variable, and, hence, capture roughly 45% of the true coefficient. These estimates should, of course, be treated very carefully as they are subject to assumptions and a large imprecision in the population diversity estimate.

Table 2: Coefficients of variation (CV) for the ten auxiliary variables for five survey designs (Web only, mail only, face-to-face only, Web → face-to-face and mail → face-to-face). The last but one column gives the average value over the ten variables. The last column gives the value when all variables are selected simultaneously.

Design	1	2	3	4	5	6	7	8	9	10	Av	All
W	0.21	0.14	0.07	0.28	0.07	0.07	0.01	0.18	0.15	0.05	0.12	0.36
M	0.18	0.16	0.05	0.14	0.06	0.06	0.04	0.19	0.05	0.04	0.10	0.29
F	0.09	0.13	0.00	0.00	0.14	0.05	0.01	0.11	0.04	0.13	0.07	0.23
W→F	0.06	0.08	0.01	0.08	0.10	0.08	0.01	0.11	0.06	0.10	0.07	0.18
M→F	0.08	0.09	0.02	0.09	0.05	0.03	0.04	0.10	0.05	0.04	0.06	0.16

4.2 Detection of nonresponse bias on a variable of interest

Very often surveys have a restricted set of topics and a small set of variables of interest $(Y_1, Y_2, \dots, Y_L)^T$. In these settings, it is more useful to consider specific bias rather than general bias as is done in section 4.1. It is assumed that $(Y_1, Y_2, \dots, Y_L)^T$ are independently generated from the same grouping distribution $p_Y(C, s)$. Again a vector of auxiliary variables, $(X_1, X_2, \dots, X_M)^T$, is available, generated from $p_X(C, s)$, and ρ represents the response probability of a population element.

Consider one variable of interest, say Y. The standardized nonresponse bias on the location is given by (10), but can also be expressed in terms of the response propensities ρ_Y and bounded by

$$\frac{|B(Y)|}{S(Y)} = \frac{|\text{cov}(Y,\rho_Y)|}{\bar{\rho}} \leq \frac{S(\rho_Y)}{\bar{\rho}}. \tag{21}$$

Schouten, Cobben, Lundquist and Wagner (2014) show that the maximal absolute remaining bias after adjustment using the expansion, generalized regression, inverse propensity weighting or doubly robust estimators is proportional to

$$\sqrt{(CV^2(\rho) - CV^2(\rho_X))R^2(Y,X)}, \tag{22}$$

where the proportion of unexplained variance $R^2(Y,X)$ is defined by (4).

Section 4.1 tells us that the coefficient of variation points at data collection designs that have smaller expected nonresponse bias, even after adjustment, for arbitrary variables. However, this implication does not hold for every variable, and it may not hold for a variable of interest. Clearly, it is true that, when $p_X(C,s) = p_Y(C,s)$, strong statements are possible and $ECV^2(\rho_X) = ECV^2(\rho_Y)$. In this scenario, auxiliary variables are generated in the same way as survey questions or observations are generated and obviously provide direct evidence of nonresponse bias. This scenario is very unrealistic.
However, theorem 2 is much more powerful than it appears at first sight. Clustered, uniform grouping distributions conform to random draws from a subset of the universe of variables, i.e. the subset of variables in which two or more strata are always assigned to the same category. As a consequence, theorem 2 is very helpful in translating response propensity variation on X to Y in two different ways: It can be used to set up acceptance-rejection schemes for auxiliary variables and it can be used to evaluate a targeted selection of variables. The two options are briefly discussed.

Theorem 2 allows for acceptance-rejection schemes on generated variables in $(X_1, X_2, \dots, X_M)^T$ to create random subsets of variables with useful features. Let us return to the variable of interest Y. Suppose variable X_m is accepted for the derivation of response propensities whenever the proportion of unexplained variance is lower than a specified threshold θ, e.g. $R^2(Y,X_m) < \theta$. If the proportion is larger, then the variable is discarded. It is straightforward to show that the resulting subseries of auxiliary variables, $\tilde{X}$, is generated from a clustered, uniform grouping distribution. The series may be empty, i.e. there may not be X_m that satisfy the criterion, in which case no statements can be made. However, if the series $\tilde{X}$ exists, then it represents a random draw from the subset of variables to which also the variable of interest belongs. A smaller $CV(\rho_{\tilde{X}})$ for one design than another design implies that in expectation the CV for any arbitrary other variable from the same subset is also smaller. Still this result does not mean that the bias for the variable of interest is really smaller, but evidence is growing as from (23) there is less room for the remaining bias to move around.
Clearly, the proportion of unexplained variance can only be derived using the observed data and is, therefore, in general biased itself by the missing data. This is a

consequence that cannot be avoided, but using a threshold θ implies some robustness to any bias in R^2.

The acceptance-rejection scheme might also be extended to the full set of variables of interest, $(Y_1, Y_2, \dots, Y_L)^T$. It is, however, not so straightforward how to make such a simultaneous choice for all variables of interest. An option is to set a threshold to the average proportion of unexplained variance, $\frac{1}{L}\sum_{l=1}^{L} R^2(Y_l, X_m) < \theta$. It is imperative that the threshold is not too low in order to keep all variables of interest in the subset of variables themselves; if the variables of interested are very diverse, then one should take a very low threshold and one would be in the setting of section 4.1. Theorem 2 can also be used to consider settings where the auxiliary variables are explicitly designed to relate to the variables of interest or to the missing-data-mechanism itself. These settings may occur when auxiliary variables are taken from so-called paradata measurements (e.g. Kreuter 2013), i.e. observations and recordings made during survey data collection, like interviewer observations about the dwelling and household or call record data. If auxiliary variables are generated from a clustered grouping distribution that also has the variables of interest in its support, then the between variance in (6) approximates that of the clustered grouping distribution corresponding to the variables of interest. If the auxiliary variables are generated from a clustered grouping distribution that has the response probability ρ in its support, then the between variance in (6) approximates the overall variance of response probabilities.

Example – population of a country - continued: The topics of the survey consist of neighbourhood cohesion, neighbourhood problems, safety on the streets and in general, victimisation, safety measures taken, contact with the local police, performance of the police and performance of the municipality. Three variables of interest are considered: a 0-1 indicator for feeling unsafe at times (Y_1), a 0-1 indicator for being satisfied with police performance (Y_2), and the number of victimisations in the past year (Y_3).Two thresholds are set to select variables based on Cramer's V: $C_V > 0.10$ and $C_V > 0.15$. Table 3 presents the C_V values between the ten auxiliary variables and the three target variables. It follows that under the first threshold, respectively, two (X_3 = gender and X_{10} = urbanization), zero and three variables (X_1 = age, X_8 = type of household and X_{10} = urbanization) are selected for the three survey variables. Under the second threshold, these numbers are one (X_3 = gender), zero and one (X_1 = age). Hence, no statement is made about variable police performance (Y_2). Table 4 shows the coefficients of variation under the two thresholds for the five designs of table 2. Design preferences do not change compared with table 2 when selecting auxiliary variables for target variable past victimisation (Y_3). For target variable feeling unsafe (Y_1) the picture is somewhat unclear. When variables are selected based on the criterion $C_V > 0.10$, then the sequential design mail → face-to-face still scores best ($CV = 0.05$), but the other designs have shifted roles and the single mode design face-to-face is now least favourite ($CV = 0.14$). However, when $C_V > 0.15$, then face-to-face is favourite ($CV = 0.00$), although the difference with the two sequential designs is small.

Table 3: Cramer's V (C_V) between the ten auxiliary variables and the target variables.

	1	2	3	4	5	6	7	8	9	10
$C_V(Y_1,X)$	0.08	0.05	0.20	0.06	0.09	0.02	0.00	0.07	0.03	0.11
$C_V(Y_2,X)$	0.05	0.02	0.00	0.04	0.08	0.03	0.00	0.03	0.04	0.03
$C_V(Y_3,X)$	0.18	0.04	0.00	0.08	0.09	0.03	0.00	0.12	0.09	0.12

Table 4: Coefficients of variation (CV) per design for the four auxiliary variables that relate to the variable of interest. The second to seventh column give average and combined values for auxiliary variables that have $C_V > 0.10$ and the last three columns give values for auxiliary variables that have $C_V > 0.15$. For Y_2 no auxiliary variables satisfy these criteria.

Design	$C_V > 0.10$						$C_V > 0.15$		
	Y_1		Y_2		Y_3		Y_1	Y_2	Y_3
	Av	All	Av	All	Av	All	Gender	NA	Age
W	0.06	0.10	-	-	0.15	0.29	0.07	-	0.21
M	0.05	0.08	-	-	0.14	0.24	0.05	-	0.18
F	0.07	0.14	-	-	0.11	0.19	0.00	-	0.09
W→F	0.06	0.10	-	-	0.09	0.16	0.01	-	0.06
M→F	0.03	0.05	-	-	0.07	0.13	0.02	-	0.08

5. Discussion

This paper is based on the rationale that the nature of variables, that are used for inference under missing data and causal 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.

Three research questions were posed about the construction of a framework for the generation of variables, the consequences for associations between variables, and the estimation of population parameters that are important in these associations. They are briefly 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 were the population is made of a countable number of strata that are randomly grouped to form variables. A countable population diversity seems natural from the point of view of relevance and time-stability; beyond a certain precision, measurement values reflect irrelevant, circumstantial noise. In the framework, variables with a continuous measurement level are truncated and the error term is assumed to be random and to vary over population units. As a consequence, for continuous variables, the model resembles a factor model with a countable number of dimensions. This model could be elaborated, which is not done here. In general, the framework needs more discussion and evaluation. In the setting of missing data, however, it turned out to be a very useful way of looking at auxiliary variables.

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 allow for conclusions beyond the mere associations 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, uniformity 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, uniformity 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. Hence, with a range of data sets it may be possible to actually estimate them. 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.

Given that one accepts the framework, there are still a number of challenges. First, the number of auxiliary variables must be large in order to draw conclusions. Essentially, the variables are just draws and, as usual, quite a few are needed to get 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 get 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 is hard to believe that variables are generated by random grouping. It seems more 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 imaginable 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.
These challenges may be picked up in future research. It would be worthwhile to attempt to estimate the population diversity on a large panel data set containing many variables about the same population. 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.

References

Heckman, J.J. (2008), Econometric causality, International Statistical Review, 76, 1 – 27.

Ibrahim, J.G., Chen, M.H., Lipsitz, S.R. (2004), Monte Carlo EM for missing covariates in parametric regression models, Biometrics, 55 (2), 591 – 596.

Ibrahim, J.G., Chen, M.H., Lipsitz, S.R., Herring, A.H. (2005), Missing data methods for generalized linear models. A comparative review, Journal of the American Statistical Association, 100, 332 – 346.

Joffe, M.M. (2000), Confounding by indication: The case of calcium channel blockers, Pharmacoepidemiology and Drug Safety, 9, 37 – 41.

Kreuter, F. (2013), Improving surveys with paradata. Analytic use of process information, Edited book, Wiley Series in Survey Methodology, John Wiley & Sons.

Linero, A.R., Daniels, M.J. (2015), A flexible Bayesian approach to monotone missing data in longitudinal studies with nonignorable missingness with application to an acute schizophrenia clinical trial, Journal of the American Statistical Association, 110 (509), 45 – 55.

Little, R.J.A., Rubin, D.B. (2002), Statistical Analysis with Missing Data, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, NY, USA.

Molenberghs, G., Beunckens, C., Sotto, C. Kenward, M.G. (2008), Every missingness not at random model has a missingness at random counterpart with equal fit, Journal of the Royal Statistical Society B, 70 (2), 371 – 388.

Molenberghs, G., Njeru Njagi, E., Kenward, M.G., Verbeke, G. (2012), Enriched-data problems and essential non-identifiability. Int Journal of Statistics in Medical Research, 1, 16 – 44.

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

Robins, J.M., Hernán, M.A. (2009), Estimation of the causal effects of time-varying exposures, Chapter 23 in Longitudinal Analysis, Handbook of Modern Statistical Methods, Eds Fitzmaurice, G, Davidian, M., Verbeke, G., Molenberghs, G., Chapman & Hall/CRC, Bacon Raton, USA.

Rosenbaum, P. R. & 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, Journal of the American Statistical Association, 100 (469), 322 – 331.

Särndal. C.E. (2011), The 2010 Morris Hansen Lecture: Dealing with survey nonresponse in data collection, in estimation, Journal of Official Statistics, 27 (1), 1 – 21.

Särndal, C.E. and P. Lundquist (2014), Accuracy in estimation with nonresponse: A function of degree of imbalance and degree of explanation, Journal of Survey Statistics and Methodology, 2 (4), 361 – 387.

Schouten, B., Cobben, F., Bethlehem, J. (2009), Indicators for the representativeness of survey response, Survey Methodology, 35 (1), 101 – 113.

Schouten, B., Cobben, F., Lundquist, P., Wagner, J. (2014), Theoretical and empirical evidence for balancing of survey response by design, Discussion paper 201415, Statistics Netherlands, The Hague, available at www.cbs.nl.
Schouten, B., Calinescu, M., Luiten, A. (2013), Optimizing quality of response through adaptive survey designs, Survey Methodology, 39 (1), 29 – 58.
Seaman, S., Galati, J., Jackson, D., Carlin, J. (2013), What is meant by "Missing at random"?, Statistical Science, 28 (2), 257 – 268.
Wagner, J., West, B.T., Kirgis, N., Lepkowski, J.M., Axinn, W.G., Kruger Ndiaye, S. (2013), Use of paradata in a responsive design framework to manage a field data collection, Journal of Official Statistics, 28 (4), 477 – 499.

Appendix A: Proof of Theorem 1

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

$$S^2(Y_X) = \sum_{c=1}^{C} p_c \left(\frac{\sum_{h=1}^{G} \delta_{h,c} q_h y_h}{p_c} - \bar{y} \right)^2 = \sum_{c=1}^{C} G_c \frac{p_c}{G_c} \left(\frac{\sum_{h=1}^{G} \delta_{h,c} q_h y_h}{G_c} \frac{G_c}{p_c} - \bar{y} \right)^2. \quad \text{(A1)}$$

So that

$$E(S^2(Y_X)|C_A = M, G_1 = n_1, \dots, G_M = n_M) = \sum_{c=1}^{M} n_c E\left(\frac{p_c}{n_c} \left(\frac{\sum_{h=1}^{G} \delta_{h,c} q_h y_h}{n_c} \frac{n_c}{p_c} - \bar{y} \right)^2 | G_c = n_c \right). \quad \text{(A2)}$$

The expectation term in (A2) can be written as (omitting the condition $G_c = n_c$)

$$EX\left(\frac{Y}{X} - \bar{y}\right)^2 = E\frac{Y^2}{X} - 2\bar{y}EY + \bar{y}^2 EX, \quad \text{(A3)}$$

and a second order Taylor approximation of the first term of (A3) around (EX, EY) leads to

$$E\frac{Y^2}{X} = \frac{(EY)^2}{EX} + \frac{\text{var}(Y)}{EX} - 2\frac{\text{cov}(X,Y)EY}{(EX)^2} + \frac{\text{var}(X)(EY)^2}{(EX)^3} + O(E(X-EX)^3) + O(E(X-EX)(Y-EY)^2) + O(E(X-EX)^2(Y-EY)). \quad \text{(A4)}$$

The third and higher order terms of (A4) are ignored. It holds that

$$E(X|G_c = n_c) = \frac{1}{G}, \quad \text{var}(X|G_c = n_c) = \frac{1}{n_c}\left(1 - \frac{n_c}{G}\right)\frac{1}{G-1}\sum_{g=1}^{G}\left(q_g - \frac{1}{G}\right)^2. \quad \text{(A5)}$$

$$E(Y|G_c = n_c) = \frac{\bar{y}}{G}, \quad \text{var}(Y|G_c = n_c) = \frac{1}{n_c}\left(1 - \frac{n_c}{G}\right)\frac{1}{G-1}\sum_{g=1}^{G}\left(q_g y_g - \frac{\bar{y}}{G}\right)^2. \quad \text{(A6)}$$

$$\text{cov}(X,Y|G_c = n_c) = \frac{1}{n_c}\left(1 - \frac{n_c}{G}\right)\frac{1}{G-1}\sum_{g=1}^{G} q_g^2 y_g - \frac{1}{n_c}\left(1 - \frac{n_c}{G}\right)\frac{1}{G-1}\frac{\bar{y}}{G}. \quad \text{(A7)}$$

Combining (A3) to (A7) gives

$$E(S^2(Y_X)|C_A = M, G_1 = n_1, \dots, G_M = n_M) = \frac{1}{n_c}\left(1 - \frac{n_c}{G}\right)\frac{G}{G-1}\bar{y}^2\sum_{g=1}^{G}\left(q_g - \frac{1}{G}\right)^2 + \frac{1}{n_c}\left(1 - \frac{n_c}{G}\right)\frac{G}{G-1}\sum_{g=1}^{G}\left(q_g y_g - \frac{\bar{y}}{G}\right)^2 - 2\frac{1}{n_c}\left(1 - \frac{n_c}{G}\right)\frac{G}{G-1}\bar{y}\sum_{g=1}^{G} q_g^2 y_g + 2\frac{1}{n_c}\left(1 - \frac{n_c}{G}\right)\frac{1}{G-1}\bar{y}^2 =$$

$$= \frac{1}{n_c}\left(1 - \frac{n_c}{G}\right)\frac{G}{G-1}\left(\sum_{g=1}^{G} q_g^2 y_g^2 + \bar{y}^2\sum_{g=1}^{G} q_g^2 - 2\bar{y}\sum_{g=1}^{G} q_g^2 y_g\right)$$

$$=\frac{1}{n_c}\left(1-\frac{n_c}{G}\right)\frac{G}{G-1}\sum_{g=1}^{G} q_g^2\left(y_g-\bar{y}\right)^2.$$

(A8)

Now, filling in (A8) in (A2) gives

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

(A9)

(A9) does not depend on the G_c, so that the conditioning on $G_1=n_1,\dots,G_M=n_M$ can be removed

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

(A10)

Now, weighting (A10) by the probabilities $P[C_A=M]$ leads to (2)

$$ES^2(Y_X)=\frac{G}{G-1}(EC_A-1)\sum_{g=1}^{G} q_g^2\left(y_g-\bar{y}\right)^2. \quad \text{(A11)}$$

(A11) can be rewritten to

$$ES^2(Y_X)=\frac{G}{G-1}(EC_A-1)\left[\Gamma(\mathrm{q_g^2},(\mathrm{y_g}-\bar{\mathrm{y}})^2)+DS^2(y)-D\Gamma(\mathrm{q_g},(\mathrm{y_g}-\bar{\mathrm{y}})^2)\right],$$

(A12)

so that the additional conditions, $\Gamma(\mathrm{q_g^2},(\mathrm{y_g}-\bar{\mathrm{y}})^2)=0$ and $\Gamma(\mathrm{q_g},(\mathrm{y_g}-\bar{\mathrm{y}})^2)=0$, lead to (3).

Explanation of symbols

.	Data not available
*	Provisional figure
**	Revised provisional figure (but not definite)
x	Publication prohibited (confidential figure)
–	Nil
–	(Between two figures) inclusive
0 (0.0)	Less than half of unit concerned
Empty cel	Not applicable
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/'15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/'13–2014/'15	Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive

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

Colofon

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

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information