*Paper for the 26th International Workshop on Household Survey Nonresponse*
*September 2-4, 2015, Leuven, Belgium*

# Nonresponse adjustment by design?

Barry Schouten – Statistics Netherlands

## 1 – Motivation

Adaptive survey designs have been studied for some years. The designs essentially attempt to reduce the impact of nonresponse error given an available budget. They employ (more) explicit quality and cost functions, and, most, importantly, identify different relevant strata to which design features are adapted or tailored.

The reduction of nonresponse error impact comes through reduced adjustment weight variation, i.e. increasing precision, and through improved balance or representativeness, i.e. decreasing bias. As such adaptive survey design adjusts nonresponse by design, rather than just by estimation alone.

To date, implementation of adaptive survey designs is still very modest and mostly experimental. The reasons are twofold. First, in general, any manipulation of actual data collection designs is very hard and has to pass organizational and IT barriers. Second, the ability to reduce bias is debated because the same auxiliary variables can also used in the estimation afterwards, and the explanatory power of these variables is often relatively weak.

In the paper, I will discuss theoretical conditions for the efficacy of adaptive survey designs in reducing bias, even after adjustment afterwards in the estimation. I will show that auxiliary variables do not necessarily need to have a strong explanatory power in order to detect whether a design needs to be favoured to another.

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. 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. 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. 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 uniformity of a population. The framework is applied and demonstrated in the setting of survey nonresponse. Without a complete theory about the causes for nonresponse, 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. The motivation for this paper comes from the pursuit to reduce the impact of nonresponse in surveys through so-called adaptive survey designs (Schouten, Calinescu, Luiten 2013, Wagner et al 2013 and Särndal and Lundquist 2014). 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.

## 2 – A framework for the generation of variables on a population

### 2.1 – Population diversity and uniformity

Suppose there is a population of interest 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 arose, either by construction or by evolution, as a random draw from $G$ strata, labelled $g = 1,2,3, \ldots, G$, 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 population diversity and population uniformity are defined as the number of strata and the variation in 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 uniformity of a population, U, is defined as* $U = 1 - \sum_{g=1}^{G}(q_g - \frac{1}{G})^2$.

Clearly, population units are never fully identical; some instruments make continuous measurements and the corresponding variables have continuous measurement levels. So can $G$ be finite or even countable? 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 it is natural to have a finite number of strata. 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 uniformity 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 non-relevant noise.

The diversity is bounded by the actual size of the population, say *N*. The set of available instruments at a given time, obviously, limit the estimation of the diversity and uniformity of a population. 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.

### 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. 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 $\mathscr{g}$. Let $s_g$ be the indicator representing to what group stratum $g$ is assigned, and let $s = (s_1, s_2, \ldots, s_G)^T$ be the vector of indicators. Let $C$ be the (random) number of groupings or categories of the resulting variable. Let $p(C,s)$ represent a random grouping probability distribution defined on $\{2\} \times \{1,2\}^G \cup \{3\} \times \{1,2,3\}^G \cup \ldots \cup \{G\} \times \mathcal{g}^G$. 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 a variable, say $Z$, 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 $\delta_{g_1,c} = \delta_{g_2,c} = 1$.

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 population stratum covariance between two realizations of variables, say $Z_1$ and $Z_2$, will be denoted by $\Gamma(Z_1, Z_2)$, with $\Gamma(Z_1,Z_2) = \frac{1}{G}\sum_{g=1}^{G} z_{1,g} z_{2,g} - \left(\frac{1}{G}\sum_{g=1}^{G} z_{1,g}\right)\left(\frac{1}{G}\sum_{g=1}^{G} z_{2,g}\right)$.

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.

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 $\lambda_1^C, \lambda_2^C, \ldots, \lambda_C^C$.*

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 strata and remain empty. Let the random variable $C_A$ denote the number of non-empty strata.

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* $p\left(s_{g_1} = s_{g_2}\right) = 1$.

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.

## 2.3 – Associations between randomly generated variables

Suppose that an analysis is directed at explaining a variable of interest $Y$ using auxiliary variables $(X_1, X_2, \ldots, X_M)^T$. 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 $(X_1, X_2, \ldots, X_M)^T$.

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

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

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.

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 it holds approximately that*

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

*If, additionally,* $\Gamma\left(q_g^2, \left(y_g - \bar{y}\right)^2\right) = 0$ *and* $\Gamma\left(q_g, \left(y_g - \bar{y}\right)^2\right) = 0$, *then*

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

It can be shown that 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 $\Gamma\left(q_g^2, \left(y_g - \bar{y}\right)^2\right) = 0$ and $\Gamma\left(q_g, \left(y_g - \bar{y}\right)^2\right) = 0$ 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.

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 approximation*

$$ES^2(Y_X) = \frac{K(EC_A - 1)}{K-1} \sum_{k=1}^{K} Q_k^2 (\bar{y}_k - \bar{y})^2 \tag{4}$$

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

$$ES^2(Y_X) = \frac{K(EC_A - 1)}{K-1} D_Q S_B^2(y). \tag{5}$$

Theorem 2 can be extended to series of variables generated from clustered, uniform grouping distributions, i.e. sampled from the same subset of variables.

## 3 – Application to nonresponse in surveys

### 3.1 – Detection of general bias due to nonresponse

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 $\rho$ 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

$$\frac{|B(Y)|}{S(Y)} = \frac{|\mathrm{cov}(Y,\rho)|}{\mathrm{S}(Y)\bar{\rho}}, \tag{6}$$

and

$$\frac{|B(Y)|}{S(Y)} \leq \frac{S(\rho)}{\bar{\rho}}. \tag{7}$$

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,

$$\sqrt{CV^2(\rho) - ECV^2(\rho_X)} \tag{8}$$

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 (8) so that $ECV^2(\rho_X)$ is also a crucial term in nonresponse adjusted estimates.

If one would be able to measure the $\rho$ 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 $\rho_X$, i.e. the projection of $\rho$ on the space spanned by the auxiliary variables (Rosenbaum and Rubin 1983). Now, let in the theorems, $y_g = \rho_g$ be the variable that needs to be explained.

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

$$ECV^2(\rho_X) = \frac{G(EC_A-1)D}{G-1}CV^2(\rho),. \tag{9}$$

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 $\frac{G}{G-1} \approx 1$, (8) can be rewritten to

$$\sqrt{CV^2(\rho) - ECV^2(\rho_X)} = \sqrt{1-(EC_A-1)D}\,CV(\rho) = \sqrt{1 - \frac{1}{(EC_A-1)D}}\,ECV(\rho_X). \tag{10}$$

This allows for an important conclusion: When two different survey or data collection designs lead to different $CV(\rho_X)$ and when the variables $(X_1, X_2, \ldots, X_M)^T$ follow 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 $X$ using a range of estimators is also smaller for an arbitrary other variable.

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, 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

$CV(\rho_X)$. Say, for example, $T$ strategies are available, labelled $d = 1,2,\dots,T$, where design $d$ has response probabilities $\rho_d$. 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. It can be shown that the optimized design is at least as good as the best strategy.

*Example*: Suppose the interest is in general representativeness of the survey response. Five data collection designs are considered for the Dutch Crime Victimisation survey (CVS): 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. Ten auxiliary variables are available: age, ethnicity, gender, individual annual income, province of residence, registered phone number, subscription to an unemployment office, type of household, type of income, and urbanization level. 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 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, which slightly outperforms the Web → face-to-face. The single mode design Web is by far the least favourite. If the ten auxiliary variables are believed to be generated randomly, then mail → face-to-face is expected to have the least bias on any other randomly drawn variable.

*Table 1: 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 first but last 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 |

### 3.2 – 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. In these settings, it is more useful to consider specific bias rather than general bias. Consider one variable of interest, say $Y$. 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

$$\sqrt{(CV^2(\rho) - ECV^2(\rho_X))ER^2(Y,X)}, \tag{11}$$

where $ER^2(Y,X)$ is the expected proportion of unexplained variance.

Theorem 2 can now be used. As clustered, uniform grouping distributions conform to random draws of variables from a subset of variables, this theorem 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.

Theorem 2 allows for acceptance-rejection schemes on generated variables in $(X_1, X_2, \ldots, X_M)^T$ to create random subsets of variables with useful features. Suppose variable $X_m$ is accepted for the derivation of response propensities whenever the proportion of unexplained variance is lower than a specified threshold $\theta$, 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, 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.

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. 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 (5) 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 $\rho$ in its support, then the between variance in (5) approximates the overall variance of response probabilities.

*Table 3: 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 |

*Example –continued*: The topics of the CVS 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 $C_V > 0.15$. Under the first threshold, respectively, two (gender and urbanization), zero and three variables (age, type of household and urbanization) are selected for the three survey variables. Under the second threshold, these numbers are one (gender), zero and one (age). Hence, no statement is made about variable police performance ($\mathrm{Y}_2$). Table 3 shows the coefficients of variation under the two thresholds. Design preferences do not change 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, but the other designs have shifted roles and the single mode design face-to-face is now least favourite. However, when $C_V > 0.15$, then face-to-face is favourite, although the difference with the two sequential designs is small.

## 4 - Discussion

This paper is based on the rationale that the nature of variables, that are used for inference under nonresponse, 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. 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.

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.

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.