Universität Trier

Monte-Carlo Simulation Studies in Survey Statistics – An Appraisal

Jan Pablo Burgard
Patricia Dörr
Ralf Münnich

Research Papers in Economics
No. 4/20

Monte-Carlo simulation studies in survey statistics – an appraisal

Jan Pablo Burgard, Patricia Dörr and Ralf Münnich

Social and Economic Statistic Department — Faculty IV - Economics — Trier University — Universitätsring 15 — 54296 Trier (Germany) e-mail: burgardj@uni-trier.de; doerr@uni-trier.de; muennich@uni-trier.de; *url:* statistik.uni-trier.de.

Abstract: Innovations in statistical methodology is often accompanied by Monte-Carlo studies. In the context of survey statistics two types of inferences have to be considered. First, the classical randomization methods used for developments in statistical modelling. Second, survey data is typically gathered using random sampling schemes from a finite population. In this case, the sampling inference under a finite population model drives statistical conclusions.

For empirical analyses, in general, mainly survey data is available. So the question arises how best to conduct the simulation study accompanying the empirical research. In addition, economists and social scientists often use statistical models on the survey data where the statistical inference is based on the classical randomization approach based on the model assumptions. This confounds classical randomization with sampling inference. The question arises under which circumstances – if any – the sampling design can then be ignored.

In both fields of research – official statistics and (micro-)econometrics – Monte-Carlo studies generally seek to deliver additional information on an estimator's distribution. The two named inferences obviously impact distributional assumptions and, hence, must be distinguished in the Monte-Carlo set-up. Both, the conclusions to be drawn and comparability between research results, therefore, depend on inferential assumptions and the consequently adapted simulation study.

The present paper gives an overview of the different types of inferences and combinations thereof that are possibly applicable on survey data. Additionally, further types of Monte-Carlo methods are elaborated to provide answers in mixed types of randomization in the survey context as well as under statistical modelling using survey data. The aim is to provide a common understanding of Monte-Carlo based studies using survey data including a thorough discussion of advantages and disadvantages of the different types and their appropriate evaluation.

Keywords and phrases: Monte-Carlo simulation, survey sampling, randomization inference, model inference.

1. Introduction

A practical way to study new tools and methodologies in statistics is Monte-Carlo (MC) simulation. Often an estimator's distribution – and its properties derived therefrom – is analytically or computationally too cumbersome for an analysis. Then, resampling techniques relying on the empirical sampling distribution can be as well conclusive when the sample size is efficiently large [Mooney,

1997, pp. 3] and its usage reduces the complexity of the analysis. Halton [1970] defines the Monte-Carlo method as the solution of an estimation problem of a statistical model that has employed *a random sequence of numbers to construct a sample of the population to obtain statistical estimates of the parameter*. The increasing computational power enables thus researchers to apply more and more frequently MC studies to underpin the performance of new methods and to get an idea about the theoretical behavior of an estimator. The frequent use of Monte-Carlo simulation, however, also increases the proneness to conceptual mistakes. Especially in survey statistics, the design of MC studies plays a vital role for the assessment of the studied methodologies. The analysis of survey data has to account for different sources of stochasticity depending on the conceptual set-up. Therefore, the MC evaluation of new statistical methods should account for the probability law under which the proposed estimators should be applied.

Estimators take as input realizations of random variables, which is often also referred to as 'random sample'. In survey statistics, the term random sample has some peculiarities: It does not refer to a random experiment that could theoretically be repeated to infinity (i.e. independent and identical replications of a random variable), but is linked to a subset s of a finite index set $U = \{1, \ldots, N\}$ where s is the realization of a random variable $S \sim P_D$. That is, the first source of stochasticity stems from a survey design $P_D \in \text{Prob}\left(U, 2^U\right)$ that yields some subset selection $s \subset U$. The probability law P_D may depend on design variables $\mathbf{z} := (\mathbf{z}_i : i \in U)$, $\mathbf{z}_i \in \mathbb{R}^{p'}$, $P_D = P_D(\cdot; \mathbf{z})$. For reasons outlined below, we require that P_D is measurable in this second argument. For better distinguishability with the other source of stochasticity which we shall discuss later, we refer thus to s as a survey and call the probability law P_D a survey design. The subset s is then used to draw inference on the finite population U and not on a statistical model [Valliant et al., 2000, p. 1].

Many fields of statistical application are directly or indirectly affected by survey sampling. Official statistics or biological enquiries aim at the estimation of characteristics $\mathbf{y} := (\mathbf{y}_i : i \in U)$, $\mathbf{y}_i \in \mathbb{R}^p$ of a finite population U and functions g_U thereof, for example the total number of animals of a species or the average income in a country. Due to cost and time constraints, surveys $s \subset U$ rather than censuses are realized to gather these informations. For correct inference, s should be realized randomly, i.e. be an outcome of $S \sim P_D$.

Other fields of applications focus on statistical models $\mathcal{M}$, for example in order to discover causal relationships. Estimators $\hat{\vartheta}$ of a model parameter ϑ, where $\mathcal{M} = \{P_{\mathcal{M}_\theta}$ probability measure for all $\theta \in \Theta\}$ should be close to the latter in some statistical sense. The model's parameter space is named Θ that is, any value $\theta \in \Theta$ could theoretically suit the model set-up, but data stem from $P_{\mathcal{M}_\vartheta}$, $\vartheta \in \Theta$. In that case, the characteristics $\mathbf{y}_i$ of units in the finite population U are considered to be realizations of a random variable $Y_i \sim P_{\mathcal{M}_\vartheta}$ and consequently already $\mathbf{y} = (\mathbf{y}_i : i \in U)$ is seen as a classical random sample. Nonetheless, many studies in social sciences use a survey drawn from a finite population $s \subset U$, $(\mathbf{y}_i : i \in s)$, to estimate ϑ for the statistical model $\mathcal{M}_\vartheta$. Thus, survey sampling plays an active role in this case, too. Such fields are for example micro-econometrics or empirical social sciences. Another scientific field

which acknowledges the difference between survey sampling and other statistical modeling approaches is soil science [Brus and de Gruijter, 1997].

In the following, we will focus on likelihood based inference for this second focus in survey statistics. Especially in the context of statistical models $P_{\mathcal{M}}$, Bayesian statistics is often named as a counterpart to likelihood based statistics. The increased computational power also enables to run Monte-Carlo simulations to get posterior distributions in the Bayesian context. However, our focus here lies on simulation in survey sampling and in this case, Bayesian simulations do hardly differ from likelihood based MC studies: Bayesian studies only require an additional generation step of the parameter values ϑ that in this case obey a non-trivial probability law. For this reason, we focus in the following on Fisherian statistics. For Bayesian statistics, we refer to the relevant textbooks such as Gelman et al. [2009]. A brief introductory example is given in Little [1982].

With regard to MC studies, these different aspects of survey sampling imply also different terms of sampling distributions that could be used for simulations. Consequently, a special care has to be laid on MC simulation design when the researcher works with survey data. Therefore, this article seeks to give an overview on the aspects of survey sampling in MC simulation. In the following, we elaborate the different types of inference in survey sampling, the underlying sampling distributions and their implications for Monte-Carlo simulation. Example simulations are described in order to illustrate the stated differences in inference and the resulting variations in an estimator's evaluation. With this work, we hope to illuminate the different MC set-ups found in the literature and to give a guideline for the appropriate choice of the simulation set-up with whom researchers may truely answer the research question that they have posed.

This article is organized as follows. First, some basic mathematical notation is briefly introduced. Second, a general definition of MC simulations and the basic Monte-Carlo algorithm is given. Next, the two main inferences, randomization-based and model-based estimation, are introduced and their implications on Monte-Carlo studies are discussed. Subsequently, combinations of the above inferences in Monte-Carlo studies are discussed and how they are systematized by their types of inference. In the sixth Chapter, examples of applied statistics are described, where the set-up of MC-studies matters. The final section concludes.

2. Mathematical Notation

In the following, probability measures are denoted by $P_{\cdot}$ and the corresponding subscript refers to the context of the probability law. Expectation and variance with respect to $P_{\cdot}$ bear the same subscript. Furthermore, we assume different index sets where one important set, the finite population, is denoted $U = \{1, \ldots, N\}$ with $N \in \mathbb{N}$. Linked to the population U is the random variable S with values in $\mathbb{N}_0^N$. The i-th element of of a realization $S = s$ indicates the number of times that unit $i \in U$ is drawn into a the sample s.

Usually, we denote the probability law of S by P_D, where the D stands for *design*. In the case of without replacement designs, S is a vector in $\{0, 1\}^N$

and we can identify realizations of S with subsets of U. In the case of with replacements designs, S is a multiset with elements from U. It may be that the survey design depends on real-valued *design variables*, $P_D = P_D(\cdot; \mathbf{y}, \mathbf{z})$ where $\mathbf{y} \in \times_{i=1}^{N} \mathcal{Y}$, $\mathcal{Y} \subseteq \mathbb{R}^p$ and $\mathbf{z} \in \times_{i=1}^{N} \mathcal{Z}$, $\mathcal{Z} \subseteq \mathbb{R}^{p'}$ respectively. We assume that P_D is measurable in it's second argument. This means that the survey design belongs to a family of designs with outcomes depending on the points $\mathbf{y}$ and $\mathbf{z}$ in their respective variable spaces $\mathcal{Y}$ and $\mathcal{Z}$. For ease of analysis, we assume that all random variables equal the identity on the underlying probability space. This means that $S = \mathrm{id}_{\mathbb{N}_0^U}$. Obviously, $\mathbf{1}_S$ can only take finitely many values, therefore all moments exist for without replacement designs and the probability law of S is well defined by the ensemble of moments of $\mathbf{1}_S$. We refer to $|S|$, the ℓ_1 norm of S, as the sample size and set $\mathrm{E}_D[|S|] = n$.

In another statistical context, we assume random variables $(Y, Z) = ((Y_i, Z_i) : i = 1, \ldots, N)$. We refer to Y as the variable of interest and to Z as auxiliary or design variables. Again we assume that $(Y, Z) = \mathrm{id}_\Omega$ with $\Omega = \times_{i=1}^{N}(\mathcal{Y} \times \mathcal{Z}) \subset \times_{i=1}^{N} \mathbb{R}^{p+p'}$ where each Y_i takes values in the subset $\mathcal{Y}$ of $\mathbb{R}^p$ and Z_i in the subset $\mathcal{Z}$ of $\mathbb{R}^{p'}$. In order to ease measurability conditions on projections from Ω to $\times_{i=1}^{|S|} \mathcal{Y}$ etc., we assume that the σ-algebra on Ω is the product σ-algebra of the algebras on $\mathcal{Y} \times \mathcal{Z}$ that yields (Y_i, Z_i), $i \in U$, measurable.

Sampled sub-arrays of $Y = (Y_1, \ldots, Y_N)$ and $Z = (Z_1, \ldots, Z_N)$ are denoted by a subscript, i.e. $Y_I := (Y_i : i \in I)$ and $Z_I = (Z_i : i \in I)$ where I is a combination (possibly with repetition) of U. In the case of without replacement designs, $I \subset U$. From the context it should become clear whether I is a single index (usually indices i and j) or a (multi-)subset of indices. Realizations of Y and Z are denoted by $\mathbf{y}$ and $\mathbf{z}$ respectively and we have the same notation for sub-arrays of realizations, i.e. $\mathbf{y}_I$ and $\mathbf{z}_I$. Furthermore, we assume as data generating process for (Y, Z)

$$(Y, Z) \sim P_{\mathcal{M}_\vartheta}, \tag{2.1a}$$

$$P_{\mathcal{M}_\vartheta} \in \{P_{\mathcal{M}_\theta} \in \mathrm{Prob}(\Omega, \mathcal{A}) : \theta \in \Theta\} =: P_{\mathcal{M}} \quad . \tag{2.1b}$$

For the parameter space, we have $\Theta \subset \tilde{\Theta} \subset \mathbb{R}^q$ to be $P_{\mathcal{M}_\theta}$-measurable for all $\theta \in \Theta$. An estimator for ϑ is a function that is for all $\theta \in \Theta$ $P_{\mathcal{M}_\theta}$-measurable with

$$\hat{\vartheta}_s : \mathcal{X} \to \tilde{\Theta} \tag{2.2}$$

where $s \subseteq U$ and $\mathcal{X} \in \{\times_{i \in s} \mathcal{Y}, \times_{i \in s}(\mathcal{Y} \times \mathcal{Z}), (\times_{i \in s}(\mathcal{Y} \times \mathcal{Z})) \times_{i \in s^c} \mathcal{Z}\}$. This means that an estimator uses the s-sample of the variable of interest and optionally also the auxiliary variable Z or a sub-array thereof. In order to have a general formulation we write

$$\hat{\vartheta} : \mathcal{S} \times \Omega \to \tilde{\Theta} \quad , \tag{2.3}$$

where $\mathcal{S} \subset 2^U$ and $\mathcal{S} = \cup_{\omega \in \Omega} \mathrm{supp}\, P_D(\cdot; Y(\omega), Z(\omega))$. The class of functions that are $P_{\mathcal{M}_\theta}$-measurable for all $\theta \in \Theta$ is denoted by $\mathcal{F}$ and the class of estimators

is denoted by $\hat{\Theta} \subset \mathcal{F}$. Often, estimators are (higher order) integrable for all $\theta \in \Theta$, then we have even $\hat{\Theta} \subset \cap_{\theta \in \Theta} \mathcal{L}^m (\Omega, \mathcal{A}, P_{m_\vartheta})$ where $m \in \mathbb{N}$.

3. Monte-Carlo Simulations

The term Monte-Carlo simulation is made up of two parts, and it is quite illuminating to give here the view of Kalos and Whitlock [1986, pp. 2]: *"A distinction is sometimes made between simulation and Monte Carlo. In this view, simulation is a rather direct transcription into computing terms of a natural stochastic process [...]. Monte Carlo, by contrast is the solution by probabilistic methods of nonprobabilistic problems. This distinction is somewhat useful, but often impossible to maintain."* Indeed, in the context of survey sampling, this differentiation is possible only in a very limited manner. As already stated above, often, the aim is to learn about an estimator's distribution by an empirical approximation thereof. In that sense, what was given as a short definition of Monte-Carlo agrees with the view of Kalos and Whitlock [1986]. In the case of survey sampling, though, this typically implies to draw realizations of $S \sim P_D$, where P_D follows deterministic rules and S is a random variable. This corresponds to a 'direct transcription into computing terms of a natural stochastic process', i.e. simulation. Hence, in the following, Monte-Carlo and simulation are thought as an ensemble and often, we use MC simulation and MC study interchangeably.

Assume that interest lies on the distribution of an estimator $g : \mathcal{P} \to \mathbb{R}^q$, where $\mathcal{P}$ is the set of outcomes of an distribution estimator $\hat{P}$ for P and $g(P)$ is the estimand, a statistic's value that depends on the probability law P. Having an estimator $\hat{P}$ for P, it is thus possible to express the estimator of $g(P)$ as $g\left(\hat{P}\right)$. The distribution of the estimator is thus expressed by $P\left(g(\hat{P}) \in \cdot\right)$. We fix as a general MC procedure the following: Given a transcription of P into computing terms, it is possible to generate independently random realizations of a variable $X \sim P$ indexed by $b = 1, \ldots, B$ leading to realizations of the estimator $\hat{P}$, say $\hat{P}_1, \ldots, \hat{P}_B$. Then these realizations of empirical distributions can be used to construct an empirical probability measure that assigns mass $\frac{1}{B}$ to each realization $g(\hat{P}_b)$, $b = 1, \ldots, B$. This gives an empirical distribution for the general estimator $g(\hat{P})$. A summary of these steps is given in Algorithm 1.

Algorithm 1 General Monte-Carlo Simulation Algorithm

Require: A probability measure P and an estimator of a statistic $g(P)$, i.e. $g(\hat{P})$, $B \in \mathbb{N}$
Ensure: A Monte-Carlo distribution for the estimator $g(\hat{P})$
  for $b = 1, \ldots, B$ **do**
    Generate a realization of $X \sim P$
    Derive the empirical distribution of P, namely $\hat{P}_b$
    Evaluate the estimator g at $\hat{P}_b$
  end for
  Set the empirical probability function of $g(\hat{P})$ to $\frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\{g(\hat{P}_b)\}}(\cdot)$

MC studies rely on properties coming with large sample sizes: $P\left(g(\hat{P}) \in \cdot\right)$

is proxied by its empirical counterpart. *"The principle behind Monte-Carlo simulation is that the behavior of a statistic in random samples can be assessed by the empirical process of actually drawing lots of random samples and observing this behavior."* [Mooney, 1997, p. 3]. At least for independent realizations of random variables, relative frequencies are known to converge uniformly to their probabilities [Vapnik and Chervonenkis, 1971]. Applying this theorem to empirical frequencies stemming from MC-simulation, this then implies that the expectation of bounded estimators converge to the expectation under the underlying distribution. For survey sampling under the design-based set-up with a maximum sample size, this holds even for every statistic and therefore calculus of properties of $g(\hat{P})$ is only a combinatorial problem. Furthermore, also for the model-based simulation set-up, laws of large numbers are often applicable when the distribution is such that second order moments exist. Note, however, that counterexamples can be built where the law of large numbers does not hold. Hence, many MC runs ($B \gg 0$) cannot replace prior theoretical considerations.

4. Types of Inference

4.1. Classical Inference in Empirical Research

4.1.1. Model-based Inference

Under model-based estimation, we understand the classical estimation theory e.g. often used in (micro-)econometrics: The estimand is a parameter $\vartheta \in \Theta$ or a statistic thereof, $g(\vartheta)$. For simplicity we set $g(\vartheta) = \vartheta$. It is supposed that we have realizations $(\mathbf{y}, \mathbf{z})$ of the random variables $(Y, Z) \sim P_{\mathcal{M}_\vartheta}$ where it is only known that $P_{\mathcal{M}_\vartheta} \in P_{\mathcal{M}}$.

As the estimand is ϑ, we denote estimators by $\hat{\vartheta}$. instead of g used in the previous section. In general, the choice of an estimator $\hat{\vartheta} \in \mathcal{F}$ depends on the particular statistical problem and the statistician's requirements on statistical properties of $\hat{\vartheta}$ ('closeness to ϑ in some statistical sense'). One possible example for required statistical properties may be unbiasedness, i.e. $\mathrm{E}_{\mathcal{M}_\theta}\left[\hat{\vartheta}\,(Y, Z)\right] = \theta$ for all $\theta \in \Theta$. An estimator might then be considered to be 'close' when it is uniformly optimal in squared deviation

$$\mathrm{E}_{\mathcal{M}_\theta}\left[\left(\hat{\vartheta} - \theta\right)^2\right] = \inf_{\hat{\theta} \in \mathcal{F}} \mathrm{E}_{\mathcal{M}_\theta}\left[\left(\hat{\theta} - \theta\right)^2\right]$$

within the subclass of unbiased estimators in $\mathcal{F}$, that is minimality subject to

$$\mathrm{E}_{\mathcal{M}_\theta}\left[\hat{\vartheta}\,(Y, Z)\right] = \theta \quad .$$

Other types of 'closeness' and statistical properties are possible, though.

In the framework of survey sampling, model-based estimation thus means that the N-dimensional array of characteristics $\mathbf{y}$, and design variables $\mathbf{z}$, are

considered to be realizations of $(Y, Z) \sim P_{m_\vartheta}$. Hence, ideally, estimators $\hat{\vartheta}_U$ using the complete information $(\mathbf{y}, \mathbf{z})$ would be employed to estimate ϑ. However, not all elements of the characteristic $\mathbf{y}$ are observable in survey sampling but only the subset $s \subset U$ due to the sampling process described above.

Then, obviously, another estimator $\hat{\vartheta}_s$ is required, mapping first from Ω to a projective space $\mathcal{X}$ and then to $\tilde{\Theta}$. In general, the employed estimator $\hat{\vartheta}_U$ that would be used for parameter estimation if all elements in $\mathbf{y}$ and $\mathbf{z}$ were observed, belongs to a sequence of estimators $\{\hat{\vartheta}_{U_k}\}_{k\in\mathbb{N}}$ with $\ldots \subset U_k \subset U_{k+1} \subset \ldots$ and $|U_k| \to_{k\to\infty} \infty$. Researchers ignoring the sampling design P_D tend to choose $\hat{\vartheta}_s \in \{\hat{\vartheta}_{U_k}\}_{k\in\mathbb{N}}$ as $s \subset U$, hoping that the statistical properties of $\{\hat{\vartheta}_{U_k}\}_{k\in\mathbb{N}}$ also hold for $\hat{\vartheta}_s$ though s is a random sample realization.

This means that such researchers assume that inference on P_{m_ϑ} is equal to inference based on the probability measure

$$P_{m_\vartheta,D}(s \times A) := \int_A P_D\left(s; Y(\omega), Z(\omega)\right) \mathrm{d}P_{m_\vartheta}(\omega) \quad , \tag{4.1}$$

where $A \in \mathcal{A}$. Rubin-Bleuer and Kratina [2005] and Boistard et al. [2015] leave out the random variable Y in the definition of $P_{m_\vartheta,D}$ because they assume that the survey design is non-informative [Pfeffermann, 1993, Definition 3] for the estimator $\hat{\vartheta}_s$. Implicitely, they assume with the short cut $Z = \pi_U^z((Y, Z))$ that

$$P_{m_\vartheta,D}(s \times A) = \int_A P_D\left(s; Z(\omega)\right) \mathrm{d}P_{m_\vartheta}(\omega) \tag{4.2}$$

because in that case, for a measurable $C \subset \tilde{\Theta}$, we have for $P_{m_\vartheta,D}\left(\hat{\vartheta}_s \in C\right)$ and $F_Y^C(\omega) = \{y \in \times_{i=1}^N \mathcal{Y} : y \in Z^{-1}(\omega)\}$

$$\begin{aligned}
P_{m_\vartheta,D}\left(\hat{\vartheta}_s \in C\right) &= \sum_{s\in\mathcal{S}} \int_\Omega P_D\left(s; Z(\omega)\right) \cdot \mathbb{1}_C\left(\hat{\vartheta}_s(s, (Y, Z)(\omega))\right) \mathrm{d}\, P_{m_\vartheta}(\omega) \\
&= \sum_{s\in\mathcal{S}} \int_\Omega P_D(s; Z(\omega)) \cdot P_{m_\vartheta}\left(\hat{\vartheta}_s(s, Y, Z) \in C | Z(\omega)\right) \mathrm{d}\, P_{m_\vartheta}(\omega) \\
&= \int_\Omega \underbrace{\sum_{s\in\mathcal{S}} P_D(s; Z(\omega)) \cdot P_{m_\vartheta}\left(\hat{\vartheta}_s(s, Y, Z) \in C | Z(\omega)\right)}_{= P_{m_\vartheta}\left(\hat{\vartheta}_s(S,Y,Z)\in C | Z(\omega)\right)} \mathrm{d}\, P_{m_\vartheta}(\omega) \\
&= P_{m_\vartheta}\left(\hat{\vartheta}(S, Y, C) \in C\right)
\end{aligned} \tag{4.3}$$

Under the assumption of survey non-informativity, it is thus theoretically justified to take $\hat{\vartheta}_s \in \{\hat{\vartheta}_N\}_{N\in\mathbb{N}}$ and researcher can limit their focus on desirable properties of the sequence $\{\hat{\vartheta}_{U_k}\}_{k\in\mathbb{N}}$ and the elements therein.

4.1.2. Monte-Carlo Inference for Empirical Research

The general principles of MC studies also apply when these are used in survey sampling under model-based inference. Actually, the description given in the

previous section has only to be broken down to the probability laws introduced in this Section, P_{m_ϑ} and $P_{m_\vartheta,D}$. Due to the derivation (4.3) that underlies model-based inference, the focus shall lie here on the probability measure P_{m_ϑ}. In the case that model violations (i.e. Assumption 4.2) shall be studied using MC, however, there is no possibility to avoid the joint measure $P_{m,D}$. We refer to Section 5 for a discussion of corresponding MC studies.

Adjusting Algorithm 1 to the notation of this section on model-based inference, we write here for an estimator $\hat{\vartheta}$ instead of g and for the distribution underlying the estimator $P_{m_\vartheta}\left(\hat{\vartheta} \in \cdot\right)$ instead of $P(g \in \cdot)$. The estimator $\hat{\vartheta}$ directly depends on the empirical distribution of (Y, Z) because it was assumed that $P_{m_\vartheta}\left(\hat{\vartheta} \in \cdot\right) \in \mathcal{A}$ and (Y, Z) is the source of stochasticity. Therefore, the estimator's empirical distribution results from the standardized count measure of $\hat{\vartheta}$ evaluated at the realizations $(\mathbf{y}, \mathbf{z})$.

In mathematical terms, the description of Mooney [1997] can be summarized as follows: Assume that a distributional property ι of an estimator $\hat{\vartheta}$ for ϑ,

$$\begin{aligned} \hat{\vartheta} &: \mathcal{X} \to \tilde{\Theta} \\ \iota &: P_m \times \hat{\Theta} \to \mathcal{E} \end{aligned} \tag{4.4}$$

is of interest, and $\mathcal{E}$ is an appropriate normed space, usually $\mathcal{E} \in \{\mathbb{R}^k, P_m\}$. As the sample $s \subset U$ is considered to be a realization of the law P_{m_ϑ} without any subsampling procedure, we have here usually $\mathcal{X} = \times_{i \in s}(\mathcal{Y} \times \mathcal{Z})$. Examples for ι are $\hat{\vartheta}$'s distribution ($\iota = P_{m_\vartheta}\left(\hat{\vartheta} \in \cdot\right)$) or expectation ($\iota(P_{m_\vartheta}, \hat{\vartheta}) = \int \hat{\vartheta} \,\mathrm{d}\, P_{m_\vartheta}$). We write interchangeably $\iota\left(P_{m_\vartheta}, \hat{\vartheta}\right) = \iota\left(P_{m_\vartheta}(\hat{\vartheta} \in \cdot)\right)$.

The Monte-Carlo estimator for $\iota\left(P_{m_\vartheta}, \hat{\vartheta}\right)$, $\hat{\iota}^{\mathrm{MC},B}$ based on $b = 1, \ldots, B$ MC runs is based on the empirical measures

$$\hat{P}^B_{m_\vartheta} := \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left((Y_b, Z_b) \in \cdot\right) \text{ and } \hat{P}^B_{m_\vartheta}(\hat{\vartheta} \in \cdot) := \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left(\hat{\vartheta}_b \in \cdot\right), \tag{4.5}$$

where

$$\hat{\vartheta}_b := \hat{\vartheta}(Y_b, Z_b), \quad (Y_b, Z_b) \sim P_{m_\vartheta} \quad \forall b = 1, \ldots, B \quad . \tag{4.6}$$

The MC distribution of the estimator is thus (4.5) and the MC expectation for $\iota = \mathrm{E}_{m_\vartheta}\left[\hat{\vartheta}\right]$ is

$$\hat{\iota}^{\mathrm{MC},B}\left(P_{m_\vartheta}, \hat{\vartheta}\right) = \frac{1}{B}\sum_{b=1}^{B} \hat{\vartheta}_b \quad .$$

The procedure for general ι is summarized in Algorithm 2. The motivation for Algorithm 2 is the strong law of large numbers: If $\int_\Omega g \,\mathrm{d}\, P_{m_\vartheta}$ exists, then the

mean of the independent realizations in Algorithm 2, converges $P_{\mathcal{M}_\vartheta}$-almost surely to $\int_\Omega g \,\mathrm{d}\, P_{\mathcal{M}_\vartheta}$ because the empirical probability measure $\hat{P}^B_{\mathcal{M}_\theta}$ converges uniformly to the true probability law $P_{\mathcal{M}_\theta}$. However, note that delimiting the Monte-Carlo study to a fixed B does not reveal that $\hat{g}_B$ might not converge when the expectation does not exist.

Algorithm 2 Basic Idea of Model-based Monte-Carlo

Require: $0 \ll B \in \mathbb{N}$, probability space $\left(\Omega, \mathcal{A}, P_{\mathcal{M}_\vartheta}\right)$, estimator $\hat{\vartheta}$, property $\iota(P_{\mathcal{M}_\vartheta}, \hat{\vartheta})$
Ensure: A MC estimate $\iota^{\mathrm{MC},B}$
  for $b = 1, \ldots, B$ **do**
    Realize $(Y, Z) \sim P_{\mathcal{M}_\vartheta}$. Denote the realization $(\mathbf{y}_b, \mathbf{z}_b)$
    Evaluate $\hat{\vartheta}(\mathbf{y}_b, \mathbf{z}_b) =: \hat{\vartheta}_b$
  end for
  Set $\hat{P}^B_{\mathcal{M}_\vartheta}(\hat{\vartheta} = x) \triangleq \frac{1}{B}\sum_{b=1}^B \mathbb{1}_{\{x\}}(\hat{\vartheta}_b)$
  Calculate distributional properties of $\hat{\vartheta}$ using $\hat{P}^B_{\mathcal{M}_\vartheta}(\hat{\vartheta} = \cdot)$, $\hat{\iota}^{\mathrm{MC},B} = \iota\left(\hat{P}^B_{\mathcal{M}_\vartheta}(\hat{\vartheta} \in \cdot)\right)$

Note again that Algorithm 2 ignores the sample design P_D and sample realization s. This is due to the fact that for model-based estimation in survey sampling, the design is assumed ignorable and inference is thus based on $P_{\mathcal{M}_\vartheta}$ only. As discussed in Section 5.1, this is motivated by the asymptotic independency of P_D and $P_{\mathcal{M}_{v}artheta}$ [Rubin-Bleuer and Kratina, 2005]. In that case, simulation from $P_{\mathcal{M}_\vartheta}$ is much easier and sufficient for the research question, except for the study of model violations, cf. Section 5.

Under the model, subsampling from U is not required or alternatively, we set $s := U$ for the estimator $\hat{\vartheta}_s$ from the previous section. Note that if interest lies on $Y|Z = \mathbf{z}$, for example in regression analysis, it is an alternative to sample Z from its marginal distribution once and simulate Y conditional on Z that is kept fixed during the simulation. In that case, only Y must be B times realized in the study, using its conditional distribution on $Z = \mathbf{z}$. Confer Section 5.1 for that case. This, however, reduces the variability of the estimator and therefore impacts ι. In fact, it is then $\iota\left(P_{\mathcal{M}_\vartheta}(\hat{\vartheta} \in \cdot|Z)\right)$ that is studied and estimated.

From a computational point of view, it might happen that already the sampling $(Y, Z) \sim P_{\mathcal{M}_\vartheta}$ is cumbersome or not implemented in standard statistical programs. In these cases, the generation of random variables may be subject to Monte-Carlo methods, too. This is a sub-field of (pseudo) random number generation. An overview about MC-methods such as acceptance-rejection methods, importance sampling, simulations of Markov chains and the Metropolis-Hastings algorithm is given in Gentle [2006]. The main idea of these MC methods is (rather than integral approximation that we outlined in the paragraph before) to sample random numbers from a *proposed* distribution Q that is easier to handle than $P_{\mathcal{M}_\vartheta}$. Those random numbers are then put into relation to those that would result from the desired distribution $P_{\mathcal{M}_\vartheta}$. This 'putting into relation' might include a reweighting scheme, rejection of too improbable realizations or convergence to a stable state related to $P_{\mathcal{M}_\vartheta}$. Though the accuracy and/or speed of convergence of these algorithms are of interest, too, they are not generally

part of the application of MC-methods in the field of survey sampling. However, note that nonetheless some estimation algorithms used in our context employ MC-sampling (cf. Booth and Hobert [1999] in a model-based context).

4.2. Inference for Sampling from Finite Populations

4.2.1. Design-based Inference

In the following, we return to the finite index set $U = \{1, \dots, N\}$ and statistics g of characteristics $\mathbf{y}$ as estimands. Survey sampling deals with randomly drawn subsets s of U where s is thus a realization of a random variable S under a survey design P_D. As the survey design might depend on design variables $\mathbf{z}$ (and possibly also $\mathbf{y}$ under informativity), we have $P_D = P_D(\cdot; \mathbf{y}, \mathbf{z})$. We require that P_D is a probability measure in the first argument and for each $s \in \mathcal{S}$ measurable in the second argument [Rubin-Bleuer and Kratina, 2005, Boistard et al., 2015], P_D is thus a probability kernel. We define furthermore $\mathcal{S}$ to be the set of all realizable samples under the design P_D, i.e. $\mathcal{S} \subseteq \mathbb{N}_0^N$. We differentiate between $\mathcal{S}$ and $\mathbb{N}_0^N$ as sample space because some estimators g_S might only be defined for samples of fixed size, i.e. when P_D is such that $|S| \equiv n$.

In survey sampling, interest then lies in the study of estimators of the type

$$g_S : \mathcal{S} \times \Omega \to \mathbb{R}^q \tag{4.7}$$

where we remind that $(\mathbf{y}, \mathbf{z}) \in \Omega$ and the notation $g_S : \mathcal{X} \to \mathbb{R}^q$ is not admissible because $\mathcal{X}$ includes $\times_{i \in s} \mathcal{Y}$ and as s is a random variable, the projection space $\mathcal{X}$ is random, which does not suit the definition of a fixed domain. g_S shall return estimates for a finite population statistic $g_U(\mathbf{y})$ where

$$g_U : \times_{i=1}^N \mathcal{Y} \to \mathbb{R}^q \quad . \tag{4.8}$$

Note that the domains of g_U and g_S might not only differ because g_S employs random realizations $s \in \mathcal{S}$, but also because the estimator g_S can employ (a subarray of) auxiliary information $\mathbf{z}$. Note that for the extension of the type (4.1) that we aim at in the next section, both g_S and g_U should also be measurable.

An example where no auxiliary information $\mathbf{z}$ is used in g_S is the Horvitz-Thompson (HT) estimator [Horvitz and Thompson, 1952]

$$g_S(s, \mathbf{y}, \mathbf{z}) \triangleq \sum_{i \in U} d_i \cdot \mathbb{1}_s(i) \cdot \mathbf{y}_i \tag{4.9}$$

where

$$d_i \triangleq \frac{1}{\mathrm{E}_S[\mathbb{1}_S(i)]} \quad . \tag{4.10}$$

The Generalized Regression Estimator (GREG) [Särndal et al., 1992] on the other hand uses auxiliary information

$$g_S(s, \mathbf{y}, \mathbf{z}) \triangleq \sum_{i \in U} d_i \cdot \mathbb{1}_s(i) \cdot \mathbf{y}_i + \mathbf{B}_s \left(\sum_{i \in U} \mathbf{z}_i - \sum_{i \in U} d_i \cdot \mathbb{1}_s(i) \cdot \mathbf{z}_i \right) \tag{4.11}$$

with

$$\mathbf{B}_s \triangleq \left(\sum_{i\in U} d_i \cdot \mathbb{1}_s(i) \cdot \mathbf{z}_i^T \mathbf{z}_i\right)^{-1} \left(\sum_{i\in U} d_i \cdot \mathbb{1}_s(i) \cdot \mathbf{z}_i \mathbf{y}_i^T\right) \quad . \tag{4.12}$$

It is thus assumed that the auxiliary information $\mathbf{z}_i$ is known for all $i \in U$. $\mathbf{B}_s$ is hence a realization of B_S which in turn is an estimator for the least squares regression matrix at the finite population level $\mathbf{B}_U$

$$\mathbf{B}_U \triangleq \left(\sum_{i\in U} \mathbf{z}_i^T \mathbf{z}_i\right)^{-1} \left(\sum_{i\in U} \mathbf{z}_i \mathbf{y}_i^T\right) \quad .$$

Note that in a pure survey sampling context, no causal relation between $\mathbf{z}$ and $\mathbf{y}$ has to be assumed. B_U is purely descriptive and being a least squares solution, it holds that the residuals $\mathbf{y} - \mathbf{z}B_U$ always sum up to zero. These two estimators (4.9) and (4.11) make obvious, however, that basic design estimation (here expressed by $\{d_i\}_{i\in S}$) needs to be available to ensure unbiased estimation. *Sample informativity* thus results when this is not the case.

As $|\mathcal{S}| < \infty$ for most designs, the statistical properties ι (moments, quantiles etc.) of g_S are calculable. However, for complex designs or $N \gg 0$, this is practically infeasible and thus Monte-Carlo methods represent an alternative [Halton, 1970]. From a theoretical point of view, asymptotics for g_S are not applicable because both $|U|$ is fixed and the sample size $|\mathcal{S}|$ is bounded by N. Consequently, asymptotic studies of g_S when $g_S \in \{g_{S,N}\}_{N\in\mathbb{N}}$ require monotonically growing finite populations $\{U_N\}_{N\in\mathbb{N}}$, $\ldots \subset U_{N-1} \subset U_N \subset U_{N+1} \subset \ldots$, and an adjustment of the sample design P_{D_N} to the respective population is required. If the population size is large with respect to $\mathrm{E}_D[|S|]$, however, it can be sufficient to increase the expected sample size within the analysis or simulations in order to study asymptotic behavior of a sequence of estimators g_S.

If, on the other hand, real asymptotics are to be studied, i.e. growing finite populations are necessary, a data generating process (DGP) for both, the variable of interest Y and the design variables Z must be assumed as in (2.1a) which then again leads to the analysis of $P_{\mathcal{M}_\vartheta,D}$ [Boistard et al., 2015].

4.2.2. Monte-Carlo Inference in Survey Statistics

In consistency with the statistical framework described above, simulation studies under the sampling randomization framework require a fixed finite population U and a vector with characteristics $(\mathbf{y}, \mathbf{z}) \in \Omega$. In order to exclude any 'confounding' with a statistical model like the ones introduced in Section 4.1, these data stem ideally from a real-world population, for example from register data [Burgard, 2013]. Then the statistic of interest, g_U could directly be derived from the register and neither estimator g_S nor MC studies are required. Consequently, simulation studies on register data are usually run when the register contains a

characteristic $\tilde{\mathbf{y}}$ that is similar to the one of interest, $\mathbf{y}$. The simulation study is then conducted assuming that the estimator g_S behaves similary for $\mathbf{y}$ and $\tilde{\mathbf{y}}$.

Nonetheless, register data might not be either available or accessible to the researcher. Then, a close-to-reality synthetic data set as basis for the simulation study is an alternative. Whilst we refer to simulation studies based on register data as *(purely) design-based*, those relying on synthetic data are referred to as *realistic design-based*. An example for such synthetic but realistic data for social sciences is AMELIA [Burgard et al., 2017], stemming from the AMELI project [Alfons et al., 2011b]. An introduction to (quasi) design-based simulation studies is given in Alfons et al. [2011a]. In the DACSEIS project, realized samples from real world data were used as populations from which samples were drawn [Münnich et al., 2003, work package 3]. When the finite population is generated differently, e.g. by a statistical model, we come to the context of mixed MC studies and inferences to be discussed in the next section.

From these data $\mathbf{y}$, $\mathbf{z}$, sub-arrays are drawn by generating index samples s of $S \sim P_D$. The sampling estimator g_S then can be evaluated using $\mathbf{y}_s$ (and possibly $\mathbf{z}$ or $\mathbf{z}_s$). The distribution of $P_D(g_S \in \cdot)$ can then be numerically approximated after $b = 1, \ldots, B$ repetitions. This empirical distribution in return allows to derive empirical statistics $\iota^{\mathrm{MC},B}$ based on the distribution of g_S such as bias, mean squared error (MSE) etc.

Putting purely design-based and realistic design-based MC simulations together (ignoring the difference between $\tilde{\mathbf{y}}$ and $\mathbf{y}$), they are summarized in Algorithm 3. If the study is quasi design-based, the data either originates from a synthetic data set or is generated by a pre-defined DGP before the MC-loop.

Algorithm 3 Basic Idea of Design-Based Monte-Carlo

Require: $0 \ll B \in \mathbb{N}$, finite population U with characteristics $\mathbf{y}$ and $\mathbf{z}$, probability space $\left(\mathcal{S}, 2^{\mathcal{S}}, P_D(\cdot; \mathbf{y}, \mathbf{z})\right)$, estimator g_S
Ensure: Estimate $\iota^{\mathrm{MC},B}\left(P_D\left(g_S \in \cdot\right)\right)$
  for $b = 1, \ldots, B$ **do**
    Realize $S \sim P_D$. Denote the realization s_b
    Evaluate $g_S\left(s_b, \mathbf{y}, \mathbf{z}\right) =: g_{s_b}$
  end for
  Set $\hat{P}_D^B(g_S = x) \triangleq \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\{x\}}(g_{s_b})$ (where $x \in \mathbb{R}^q$)
  Calculate distributional properties of g_S, $\iota^{\mathrm{MC,B}}\left(P_D(g_S \in \cdot)\right) = \iota\left(\hat{P}_D^B(g_S \in \cdot)\right)$

It is remarkable – though not very surprising given the definition of Monte-Carlo studies – that the basic principles in design-based and model-based simulation studies are quite similar: In both cases, the distribution of an estimator and consequently additional properties ι are unknown (be it either $P_{\mathcal{M}_\vartheta}(\hat{\vartheta} \in \cdot)$ or $P_D(g_S \in \cdot)$) and is approximated by its empircal counterpart. This is underdone in order to derive information on the behavior of the estimator under the given probability law ($P_{\mathcal{M}_\vartheta}$ or P_D), i.e. it is assumed that B is large enough to have a good approximation due to convergence properties. The differentiation between the statistical models $P_{\mathcal{M}_\vartheta}$ and P_D that either assume infinitely many identically and independently distributed random variables (Y_b, Z_b), $b \in \mathbb{N}$ or a

fixed population with infinitely many identically and independently distributed random variables S_b, $b \in \mathbb{N}$ is nonetheless essential for the interpretation of simulation results.

5. Combinations of Simulation Types

In this section, we discuss mixtures of the previously named simulation types which constitute the majority of settings in current research in survey statistics. Besides the realistic design-based framework which we already named in the previous section, we differentiate in addition between MC studies that are *synthetic design-based*, *quasi model-based*, *conditional model-based* or *model-based under finite populations*.

5.1. Model-based Simulation under Finite Populations

The integration of a randomization based probability law P_D and the data generating model $P_{\mathcal{M}_\vartheta}$ into $P_{\mathcal{M}_\vartheta,D}$ that was already mentioned in Section 4.1, is of major interest in social sciences: This is the framework in which most analyses in empirical social science and micro-econometrics are conducted (though some researcher assume erroneously the model-based framework). Model-based inference is sought, but the data from a surveys is used. That is, an integrated framework like in Rubin-Bleuer and Kratina [2005] is required that also leads to the probability law $P_{\mathcal{M}_\vartheta,D}$ in Equation (4.1). In that scenario, the DGP for the finite population $P_{\mathcal{M}_\vartheta}$ is often referred to as *superpopulation model* [Särndal et al., 1978, for example]. An overview on the differences between design-based and model-based estimation is given in Brus and de Gruijter [1997].

Many results concerning estimators $\hat{\vartheta}$ (and also the asymptotics of sequences $\{g_{S_N}\}_{N\in\mathbb{N}}$) under non-informative design P_D and thus $P_{\mathcal{M}_\vartheta,D}$ are already known [Pfeffermann, 1993, Rubin-Bleuer and Kratina, 2005, Boistard et al., 2015, Kott, 2018]. For example, Rubin-Bleuer and Kratina [2005, Theorem 5.1] state that a uniformly in $\mathbf{z}$ design-consistent estimator $\hat{\vartheta}_s(\cdot;\mathbf{z})$ for an estimate $\hat{\vartheta}_U(\mathbf{y};\mathbf{z})$ (i.e. an estimator using survey data that converges for almost each $\mathbf{z}$ (and $P_D(\cdot;\mathbf{u})$) to its finite population estimate), converges in probability under $P_{\mathcal{M}_\vartheta,D}$, too. Boistard et al. [2015] extend the functional central limit theorem to the pseudo empirical cumulative distribution function (the HT estimator of the empirical cumulative distribution function)

$$\hat{F}^{\mathrm{HT}}(t) \triangleq \frac{1}{N}\sum_{i=1}^{N} d_i \cdot \mathbb{1}_S(i) \cdot \mathbb{1}_{(-\infty,t]}(Y_i)$$

when Y_i is 1-dimensional and P_D meets certain requirements. This means, that $\{\hat{F}(t) - F(t)\}_{t\in\mathbb{R}}$ converges weakly to a zero-mean Gaussian process for $N, n \to \infty$, which is a weaker result than the Glivenko-Cantelli theorem which would be applicable for $\sum_{i=1}^{N} \mathbb{1}_{(-\infty,t]}(Y_i)$.

In model-based simulation studies under finite populations, there can be two aspects of interest: Either the performance of estimators g_S for varying population statistics (due to the choice and realizations following $P_{\mathcal{M}_\theta}$) or the impact of survey designs on estimates of model parameters $\theta \in \Theta$.

First assume that an estimator g_S of a statistic $g_U(\mathbf{y})$ is of interest. The model-based framework under finite populations allows to exclude – unlike a design-based simulation study – that the performance of g_S depends only on the specific characteristics $\mathbf{y}$ and $\mathbf{z}$ found in the finite population. This is especially important when the estimator g_S is either model-assisted like the GREG [Särndal et al., 1992] or even model-based like it is common in small area estimation [Rao, 2005]. Alternatively, the impact of different designs P_D can be studied and under which properties of Z conditioning P_D on Z might help to increase efficiency of g_S. Also, the violation of the non-informativity assumption, $P_D = P_D(\cdot; (\mathbf{y}, \mathbf{z}))$ instead of $P_D = P_D(\cdot; \mathbf{z})$ may be studied. Note that this is also possible (with a different inference) in design-based studies.

Second, for causal inference in empirical social science and microeconometrics, it is usually an estimator for $\vartheta \in \Theta$ that is going to be studied. In contrast to purely model-based studies, where only the impact of the choice $P_{\mathcal{M}_\vartheta}$ can be studied, model-based simulations under finite populations offer also the possibility to study how important asymptotic independence assumptions of P_D and $P_{\mathcal{M}_\vartheta}$ are.

As $P_{\mathcal{M}_\vartheta, D}$ consists of two steps chaining P_D and $P_{\mathcal{M}_\vartheta}$, it is the easiest way to design the MC study also in two steps. The law generating the finite population precedes the sampling process, which must also be taken into account in the simulation set-up.

A summary of model-based MC studies under finite populations is given in Algorithm 4. We refer to the estimator here as $g \in \{g_S, \hat{\vartheta}_s\}$. Note again that in this setting, the projection from Ω to $\mathcal{X}$ is a random variable itself, which does not allow to set $\mathcal{X}$ as domain. Therefore, we write like in Section 4.2 for the domain of an estimator $g : \mathcal{S} \times \Omega \to \mathbb{R}^q$.

Algorithm 4 Basic Idea of Model-based Monte-Carlo Studies under Finite Populations

Require: $B \in \mathbb{N}$, probability space $(\Omega, \mathcal{A}, P_{\mathcal{M}_\vartheta})$, probability space $(\mathcal{S}, 2^{\mathcal{S}}, P_D)$, estimator g

Ensure: Estimate $\hat{\iota}^{\mathrm{MC},B}\left(P_{\mathcal{M}_\vartheta, D}, g\right) = \iota\left(\hat{P}^B_{\mathcal{M}_\vartheta, D}(g \in \cdot)\right)$

for $b = 1, \ldots, B$ **do**
- Generate $(Y, Z) \sim P_{\mathcal{M}_\vartheta}$. Denote the realization $(\mathbf{y}_b, \mathbf{z}_b)$
- Define the survey design $P_D^b := P_D(\cdot; \mathbf{y}_b, \mathbf{z}_b)$
- Realize $S \sim P_D^b$. Denote the realization s_b
- Evaluate $g(s_b, \mathbf{y}_b, \mathbf{z}_b) =: g_{s_b}$

end for

Set $\hat{P}^B_{\mathcal{M}_\vartheta, D}(g = x) \triangleq \frac{1}{B} \sum_{b=1}^B \mathbb{1}_{\{x\}}(g_{s_b})$

Approximate the distributional properties ι of g by $\iota^{\mathrm{MC},B} \hat{P}_{\mathcal{M}_\vartheta, D}(g \in \cdot) = \iota\left(\hat{P}^B_{\mathcal{M}_\vartheta, D}(g \in \cdot)\right)$

Often, simulation set-up 4 can be simplified because only the conditional distribution $Y|Z = \mathbf{z}$ is of interest. For example in linear regression, we have

usually the assumption $Y_i \sim_{\text{ind}} N(\mathbf{z}_i^T\boldsymbol{\beta}, \sigma^2)$. To reflect this assumption, it is sufficient to generate once the auxiliary data Z and to consider for each $A \in \mathcal{A}$ the conditional probability $P_{m_{\vartheta},D}\left(A \cap \pi^y\left((\pi^z)^{-1}(\mathbf{z})\right)\right) =: P_{m_\vartheta}(Y \in A|Z = \mathbf{z})$ for measurable $A \subset \mathcal{Y}$.

5.2. Quasi Model-based Simulation

Quasi model-based simulation differs only sometimes from model-based simulation under finite populations. Depending on the survey design P_D, it is not always necessary to generate a finite population to sample from $P_{m_\vartheta,D}$ previous to survey sampling in order to generate (Y_S, Z_S).

Consider for example $Z_i \sim \text{Bern}(p)$ and $Y_i|Z_i = k \sim_{\text{ind}} Q_k$ where $k = 1, 2$. Concerning the survey design P_D, units $i \in U$ with $Z_i = 1$ get sampled with probability q_1 and $q_0 < q_1$ otherwise, i.e. P_D corresponds to Bernoulli sampling. The sampling rates are thus pq_1 and $(1-p)q_0$. Then it is sufficient to draw n_1 units $Y_i|Z_i = 1 \sim Q_1$ and n_0 units $Y_j|Z_i = 0 \sim Q_0$ where $N_1 \sim \text{Bin}(N, p)$ and $n_1 \sim \text{Bin}(N_1, q_1)$ and $n_0 \sim \text{Bin}(N - N_1, q_2)$. This means, S is directly generated without the help of a finite population.

If the asymptotic behavior of g when $N, n \to \infty$ is studied, it is even possible to leave out the bounding by N in the example and implied by $P_{m_\vartheta,D}\left(\pi_s((Y, S)) \in \cdot\right)$. It is simply taken S as the finite population, not under the law P_{m_ϑ} but $P_{m_\vartheta}\left(\pi_s \in \cdot\right)$. Note, that this method is especially feasible when $P_D(i \in S \wedge j \in S|Z) = P_D(i \in S|Z) \cdot P_D(j \in S|Z)$ but can get complicated otherwise. If implementable, however, one gains possibly computational efficiency.

Leaving out the two-step procedure from Algorithm 4, one needs to formulate from $P_{m_\vartheta,D}\left((Y, Z) \in \cdot\right)$ to $P_{m_\vartheta,D}\left(\pi_s((Y, S)) \in \cdot\right)$ where π_s is the coordinate projection from U to S. If this is simple, one can reduce the overall computational effort. Note, though, that in this setting, the estimator g cannot be contrasted with a finite population statistic, as units $U \setminus S$ are not generated. Estimators thus can only be compared to statistics of model parameters ϑ, which means that only estimators with estimands stemming from ϑ can be studied.

5.3. Conditional Model-based Simulation

Algorithm 2 summarizes what Burgard [2013] calls 'purely model-based': Both, auxiliary variables Z and the variable of interest Y are generated in each MC run. It can, however, often be observed that only the variable of interest Y is generated within the simulation loop [Burgard and Dörr, 2018, e.g]. This observation is due to the fact that often, we have only distributional assumptions on Y given Z. If sampling is non-informative, this analysis also allows to keep one survey-design fixed, even though it may depend on auxiliaries. In that case, the probability law under study is $P_{m_\vartheta,D}\left(\cdot \times (Y, \mathbf{z})\right)$. The advantage is less computational effort and that also close-to-reality auxiliary variables may be used. Less assumptions on a marginal law of Z have to be made. On the other hand,

the drawback is that conclusions are only (under constraints) transferable to other auxiliary data different from the realization $\mathbf{z}$. Because the generation of the dependent variable in each simulation run then depends on the realization $\mathbf{z}$, we refer to those simulation studies as conditional model-based MC studies. Those may occur in a model-based (under finite populations) or quasi-model based setting.

5.4. Synthetic Design-based Simulation

There are different possibilities to deal with the joint probability law $P_{\mathcal{M}_\vartheta,D}$ as pointed out in Burgard [2013]: Alternatively to Section 5.1, one could study $P_{\mathcal{M}_\vartheta,D}(S = \cdot \times (Y,Z) = (\mathbf{y},\mathbf{z}))$, which Burgard [2013] denotes 'smooth design-based' and we refer to synthetic design-based: In fact, this probability law is simply a design-based law, i.e. survey design, as we required P_D to be a probability kernel.

In fact, this does not vary very much from a quasi-design based simulation study besides the fact that there is due to the impact of $P_{\mathcal{M}_\vartheta}$ possibly more structure in the data and the DGP of the finite population's characteristic is better known to the researcher. This can help to evaluate estimators g_S that assume models. Note, that conclusions on the behaviour of g_S on any other finite population than the generated one is not admissibile and similar behaviour of g_S on similar realizations of the finite population characteristics $(\mathbf{y},\mathbf{z})$ is suggestive.

For this reason, we do not further consider this case here and refer to Section 4.2.2. A brief summary of such MC studies is given in Algorithm 5.

Algorithm 5 Basic Idea of Synthetic Design-Based Monte-Carlo

Require: $0 \ll B \in \mathbb{N}$, index set U, probability space $(\Omega, \mathcal{A}, P_{\mathcal{M}_\vartheta})$ with dimension N, probability kernel P_D, estimator g_S
Ensure: Estimate $\iota^{\mathrm{MC},B}\left(P_D\left(g_S \in \cdot; (Y,Z) = (\mathbf{y},\mathbf{z})\right)\right)$
  Realize $(Y,Z) \sim P_{\mathcal{M}_\vartheta}$
  Determine $P_D\left(\cdot; (\mathbf{y},\mathbf{z})\right) = P_{\mathcal{M}_\vartheta,D}\left(\cdot \times \{(\mathbf{y},\mathbf{z})\}\right)$ and the corresponding probability space on 2^U
  for $b = 1, \ldots, B$ **do**
    Realize $S \sim P_D$. Denote the realization s_b
    Evaluate $g_S\left(s_b, \mathbf{y}, \mathbf{z}\right) =: g_{s_b}$
  end for
  Set $\hat{P}_D^B\left(g_S = x; (\mathbf{y},\mathbf{z})\right) \triangleq \frac{1}{B}\sum_{b=1}^B \mathbb{1}_{\{x\}}(g_{s_b})$ (where $x \in \mathbb{R}^q$)
  Calculate distributional properties of g_S, $\iota^{\mathrm{MC,B}}\left(P_D(g_S \in \cdot; (\mathbf{y},\mathbf{z}))\right) = \iota\left(\hat{P}_D^B(g_S \in \cdot; (\mathbf{y},\mathbf{z})\right)$

To finish this section, an overview on the terms statistic of interest, estimand, estimator and others in the context of different simulation set-ups is given in Table 1.

TABLE 1
Overview of Statistical Terms under Different Inferences

Term	Mathematical Notation
(Quasi) design-based: Finite population is available or (partially synthetically) generated. Samples are drawn repeatedly from the universe according to a given sampling design to generate the MC distribution of the estimator of interest.	
Statistic of interest	$g_U : \times_{i=1}^{N} \mathcal{Y} \to \mathbb{R}^q$
Estimand	$g_U(\mathbf{y}) \in \mathbb{R}^q$
Estimator	$g : \mathcal{S} \times \Omega \to \mathbb{R}^q$
Estimator's distribution	$P_D\left(g \in \cdot\right)$
Monte-Carlo probability	$\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\{\cdot\}}\left(g(s_b, \mathbf{y}, \mathbf{z})\right)$ where $S_b \sim P_D$
Model-based: Variables of interest and explanatories are considered to originate from a statistical model. They are realized repeatedly according to the model to generate the MC distribution of the estimator of interest.	
Statistic of interest	$g : \Theta \to \mathbb{R}^q$ (usually and henceforth $g = \mathrm{id}$)
Estimand	$\vartheta \in \Theta$
Estimator	$\hat{\vartheta} : \Omega \to \tilde{\Theta} \supset \Theta$
Estimator's distribution	$P_{\mathcal{M}_\vartheta}\left(\hat{\vartheta} \in \cdot\right)$
Monte-Carlo probability	$\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\{\cdot\}}\left(\hat{\vartheta}(\mathbf{y}_b, \mathbf{z}_b)\right)$ where $(Y_b, Z_b) \sim P_{\mathcal{M}_\vartheta}$
Model-based under Finite Populations: Characteristics in the finite population are considered to be realized by a statistical model. Samples are drawn from the finite population. Both steps are repeated to generate the MC distribution of the estimator of interest.	
Statistic of interest	$g : \Theta \to \mathbb{R}^q$ (usually and henceforth $g = \mathrm{id}$)
Estimand	$\vartheta \in \Theta$
Estimator	$\hat{\vartheta} : \mathcal{S} \times \Omega \to \tilde{\Theta} \supset \Theta$
Estimator's distribution	$P_{\mathcal{M}_\vartheta, D}\left(\hat{\vartheta} \in \cdot\right)$
Monte-Carlo probability	$\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\{\cdot\}}\left(\hat{\vartheta}(\mathbf{y}_b, \mathbf{z}_b)\right)$ where $(Y_b, Z_b) \sim P_{\mathcal{M}_\vartheta}$ and $s_b \sim P_D\left(\cdot; \mathbf{y}_b, \mathbf{z}_b\right)$
Conditional Model-based (under Finite Populations): The auxiliary variables of the statistical model that generates the variable of interest are considered to be fixed. Except for this peculiarity, the statistical model (and sampling process) and therefore the generation of the estimator's MC distribution are realized as before.	
Statistic of interest	$g : \Theta \to \mathbb{R}^q$ (usually and henceforth $g = \mathrm{id}$)
Estimand	$\vartheta \in \Theta$
Estimator	$\hat{\vartheta} : (\mathcal{S}\times)\Omega \to \tilde{\Theta} \supset \Theta$
Estimator's distribution	$P_{\mathcal{M}_{\vartheta}(\cdot, D)}\left(\hat{\vartheta} \in \cdot \vert Z = \mathbf{z}\right)$
Monte-Carlo probability	$\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\{\cdot\}}\left(\hat{\vartheta}(s_b, \mathbf{y}_b, \mathbf{z})\right)$ where $Y_b \sim P_{\mathcal{M}_\vartheta, D}(\cdot \vert Z = \mathbf{z})$ (and $s_b \sim P_D(\cdot; \mathbf{y}_b, \mathbf{z})$)

6. Examples

6.1. Small Area Estimation

An illustration of the different types of inference and their implication on MC studies is easily demonstrated using a basic set-up in the field of small area estimation (SAE): In SAE, the finite population is the ensemble of $d = 1, \dots, D$ partitions, $U = \cup_{d=1}^{D} U_d$. Hence, it is possible to formulate estimands and estimators for any projection π_d from U to U_d, $d = 1, \dots, D$. The basic idea is to define the estimators as

$$g_{s_d} \triangleq \mathrm{E}_{m_\vartheta, D}\left[g_{U_d} | Y, Z, S\right] \quad . \tag{6.1}$$

It is thus obvious that small area estimators can at most be unbiased for the estimand $\mathrm{E}_{m_\vartheta}\left[g_{U_d}(Y)\right]$ and, for non-trivial cases, never for $g_{U_d}(\mathbf{y})$. Estimators that rely exclusively on observations $S \cap U_d$ are in that setting inconvenient because the efficiency of estimators is usually an increasing function in the sample size. On the other hand, if P_{m_ϑ} is now such that variables $\pi_{U_d}(Y)$ and $\pi_{U_{d'}}(Y)$ share some joint parameters in ϑ even for $d \neq d'$, units from $U_{d'} \cap S$ can be used for the prediction in $U_d \cap S^c$, which is denoted as 'borrowing strength'.

Usually, it is assumed that Y is generated by mixed models: Additional to an identically and independently distributed, zero-centered idiosyncratic error ε_i for each $h(Y_i)$, i.e. $h(Y_i) = \mu_d + \varepsilon_i$ where $\mu_d = \mathrm{E}[h(Y_i) | i \in U_d]$, a zero-centered random area effect ν_d, $\mathrm{E}_{m_\vartheta}\left[\nu_d\right] = 0$, $d = 1, \dots, D$, is shared for all $\{h(Y_i)\}_{i \in U_d}$. h is an appropriate transformation to restablish the assumptions on the errors; usually the identity. However, h can also denote, for example, the Box-Cox or logarithmic transformation [Rojas-Perilla et al., 2017, Zimmermann, 2018].

It is now interesting how the estimators $\{g_{s_d}\}_{d=1,\dots,D}$ behave under different simulation scenarios, especially as they are usually contrasted to design-based estimators. Those are usually defined under fixed populations, for example the HT [Horvitz and Thompson, 1952]. Keeping thus the realizations of Y and Z fixed, thus can be justified by the comparison with such estimators (and the application of SAE in practice, where the population is fixed, too). This yields (synthetic) design-based simulation studies like in Burgard [2013]. On the other hand, repeated realizations of the finite population can be argued for by the definition of the SAE, which makes the model-based MC studies enter the scene [Verret et al., 2015, Wagner et al., 2017, Zimmermann, 2018].

Furthermore, the 'true values' with whom the MC outcomes are contrasted with, need to be documented as well: One could either contrast with the finite population statistic $g_{U_d}(\mathbf{y}_b)$ in MC run b. Under the model-based set-up, this is of course unfavorable for SAEs in contrast to design-based statistics. On the other hand, one could argue that this is closer to reality. Or, one could contrast all estimators with $\mathrm{E}_{m_\vartheta, D}\left[g_{U_D}(Y)\right]$ as this value is consistent with the framework of $P_{m_\vartheta, D}$ and is also defined for design-based estimators such as the

HT. Remember that

$$\begin{aligned}\mathrm{E}_{m_\vartheta,D}\left[\left(g_{s_d}-g_{U_d}\right)^2\right] &= \mathrm{E}_{m_\vartheta,D}\left[\left(g_{s_d}-\mathrm{E}_{m_\vartheta,D}\left[g_{U_d}\right]\right)^2\right]+\mathrm{Var}\left[g_{U_d}\right]\\ &\quad+2\cdot\mathrm{E}\left[g_{s_d}-\mathrm{E}_{m_\vartheta,D}\left[g_{U_d}\right],g_{U_d}-\mathrm{E}_{m_\vartheta,D}\left[g_{U_d}\right]\right] \quad ,\end{aligned}$$

where the last term is usually non-negative. Thus, choosing the first option yields a larger MC variation than contrasting the estimators with their model-design expectation.

Depending on the simulation set-up, different conclusions on the performance of estimators $\{g_{S_d}\}_{d=1,\ldots,D}$ can be made. Depending on the probability law under study in the MC study, the global statistic against which the estimator g_{S_d} is contrasted, must be chosen. In the example with mixed models, as the random effect is under the model and unconditionally zero-centered, it may be asked whether a small area estimator should be contrasted against $\mathrm{E}_{m_\vartheta}\left[g_{U_d}|\nu_d\right]$ or $\mathrm{E}_{m_\vartheta}\left[g_{U_d}\right]$ in a model-based simulation study [Zimmermann, 2018]. This is an even more important question when model-based SAE is compared to design-based estimators. When $S \perp Y$ given the auxiliary information Z, one could study the marginal distribution of the estimator given Y and Z, integrating out the sample. This is then possible even when conditioning on ν_d. When $Y \not\perp S|Z$, it may become impossible to integrate out S without keeping conditioning on ν.

We demonstrate these differences in inference with a small simulation study. We generate once $X_i \sim_{iid} \mathrm{Pois}(1)$ where $i = 1,\ldots,1000$ and assign randomly to each outcome a domain $d = 1,\ldots,50$ where $|U_d| \equiv 20$. In each of the $B = 1000$ simulation runs, we generate for $i \in U_d$

$$Y_i = 1 + x_i + \nu_d + \varepsilon_i \tag{6.2a}$$

$$\nu_d \sim_{iid} N(0,3^2) \tag{6.2b}$$

$$\varepsilon \sim_{iid} N(0,1). \tag{6.2c}$$

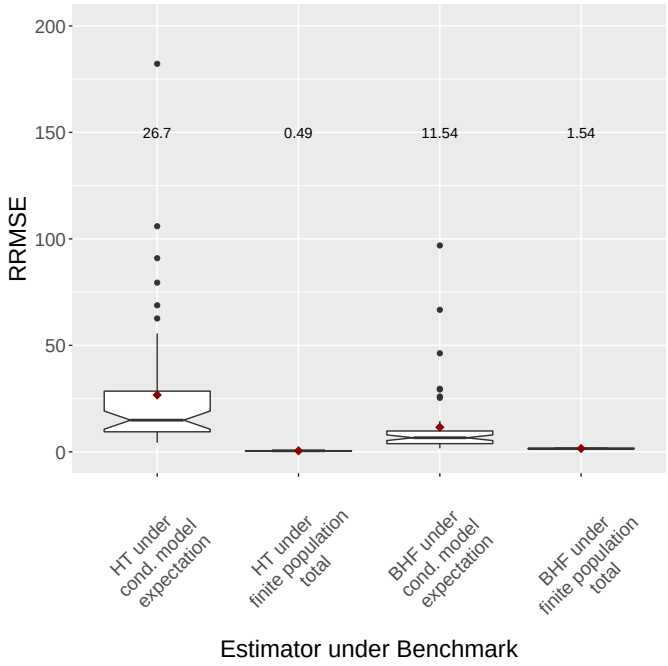
From each domain, $B = 1000$ times a simple random sample is drawn with $|U_d \cap S| \equiv 5$ without replacement and we evaluate for each domain the HT estimator [Horvitz and Thompson, 1952] $\hat{\tau}_d^{HT}$ and the Battese-Harter-Fuller estimator [Battese et al., 1988] $\hat{\tau}_d^{BHF}$ for the domain total of $\mathbf{y}$ (or Y). This means, that we are in the context of conditional model-based simulation under finite populations.

From the MC simulation, we get the empirical distribution of the estimators, and study the relative mean squared error (RRMSE) across domains as a quality measure: The empirical version for B simulation runs is

$$\mathrm{RMSE}^{\mathrm{MC}}\left[\hat{\tau}_d\right] := \sqrt{\frac{1}{B}\cdot\sum_{b=1}^{B}\left(\hat{\tau}_d(s_b)-\mathrm{E}\left[\tau_d|\mathbf{x}\right]\right)^2}$$

$$\mathrm{RRMSE}^{\mathrm{MC}}\left[\hat{\tau}_d\right] := \frac{\mathrm{RMSE}^{\mathrm{MC}}\left[\hat{\tau}_d\right]}{\mathrm{E}\left[\tau_d|\mathbf{x}\right]}$$

FIG 1. *Results of Model-based Simulation Study under Different Benchmarks*



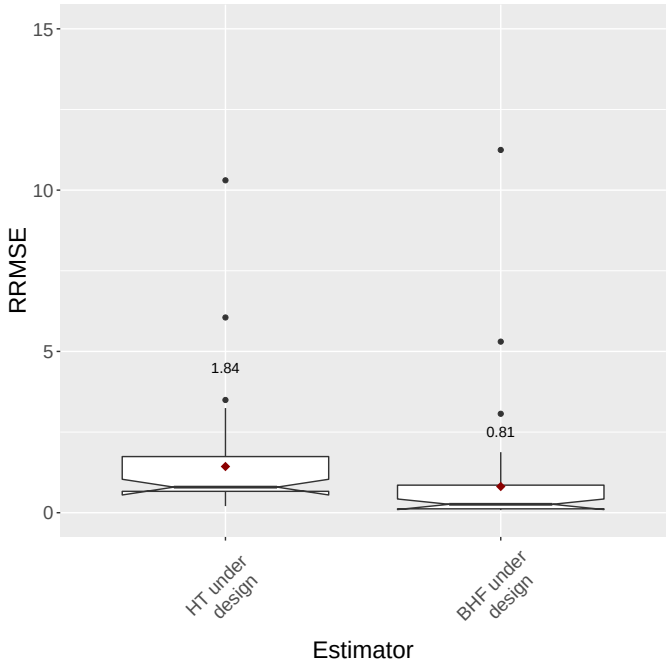
where either

$$\mathrm{E}\left[\tau_d|\mathbf{x}\right] = 20 + \sum_{i \in U_d} x_i \quad \text{or} \quad \mathrm{E}\left[\tau_d|\mathbf{x}\right] = \sum_{i \in U_d} y_i^{(b)} \quad ,$$

where $y_i^{(b)}$ is the b-th realization of the random variable Y_i. The first definition of the expectation $\mathrm{E}\left[\tau_d|\mathbf{x}\right]$ is the conditional model-based one whilst the second is the design-based expectation in each simulation run. Obviously, the results will depend on the choice of expectation as already outlined.

The RRMSEs for both the BHF and the HT under the different types of inference are illustrated in Figure 1. The difference using different benchmarks is striking: There is not only a quantitative difference, but also the qualitative result on which estimator to prefer based on RRMSE is inconclusive: Under the model (conditional on $\mathbf{x}$), the BHF has a clear advantage. On the other hand, the RRMSE of the BHF is larger then of the HT in the finite population total contrast. This is due to the fact that in this setting, conditional on $\mathbf{y}^{(b)} = (y_1^{(b)}, \ldots, y_N^{(b)})^T$, the BHF is biased. This bias seems to outweigh possible efficiency gains. Alternatively to the model-based simulation under finite populations, a (quasi) model-based simulation study could be run, what Burgard [2013] calls 'smooth design-based'. For this purpose, we realize *once* Y according to the Model (6.2). Then, $B = 1000$ times a simple random sample within all

FIG 2. *Results of Quasi Design-based Simulation Study*

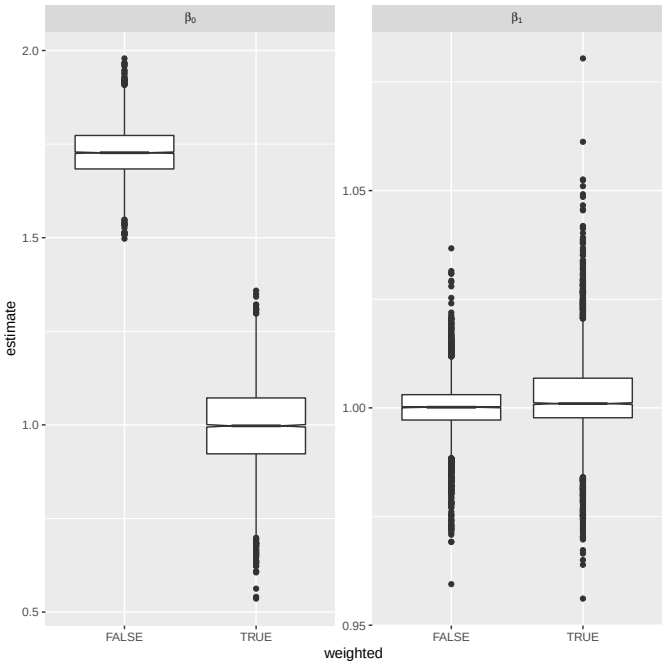


domains with $|U_d \cap S| \equiv 5$ is drawn, that is, we employ a stratified sampling design. The MC-RRMSE is plotted in Figure 2. As the Model (6.2) underlies the finite population, the BHF performs well although it is not unbiased under the design, meaning that the efficiency gain of the (correctly) assumed model outweighs the design bias. Note furthermore, that the scaling of the RRMSE is again different and that the HT does not perform too bad compared with the BHF under this set-up. To conclude, it is thus essential to determine in advance the desired probability law to be studied using the MC-simulation. Due to the highly diverging conclusions that may result from the varying set-ups, the chosen scenario must be communicated in great detail.

6.2. Regression Analysis

Whilst the first example concerns summary statistics for finite population inference, the second example deals with regression analysis, common in econometrics or social science for example. We demonstrate the problems with model-based estimation in a set-up of finite populations: It is common in social science to infer on estimators like under a model-based setting although a survey sample has generated the data. This is obviously problematic when the sampling mechanism P_D depends on the variable of interest, because in that case, we have $P_{m_\vartheta,D}(Y,S|Z) \neq P_{m_\vartheta}(Y|Z) \cdot P_D(S|Z)$ and thus integrating out the sample is

FIG 3. *Monte-Carlo Distribution of Regression Estimators*



not easy.

To illustrate this, we consider the following data generating process in each simulation run $b = 1, \ldots, 5000$

$$Y_i^{(b)} = 1 + X_i + \varepsilon_i^{(b)}, \qquad i = 1, \ldots, N \tag{6.3a}$$

$$\varepsilon_i \sim_{iid} N(0, 1) \tag{6.3b}$$

$$\log X_i \sim_{iid} N(0, 4) \quad . \tag{6.3c}$$

and we sample $n = 100$ units across two strata: 75 units are drawn under those with $\varepsilon_i > 0.5$ and the remaining units are drawn amongst those with $\varepsilon_i \leq 0.5$. The explanatory variables X_i, $i = 1, \ldots, N$, are generated once. We are thus in a conditional model-based framework. As illustrated in Boxplot 3, the informative subsetting process, $Y \not\perp S|X$, impacts the point estimation when the regression is unweighted. Of course, this also affects inference: The coverage rate for the intercept for the 95% confidence interval is 0, when the regression is unweighted, whilst it is about 94% for the weighted regression.

In summary, this exemplary simulation study illustrates model-based simulations under finite populations. Though in principle it would have been possible to generate 25 and 75 units from the respective truncated normal distribution, it is conceptually easier to take the workaround using a finite population to sample from. Furthermore, this allows to calculate survey weights, which would

not have been possible in quasi model-based simulation.

7. Further Notes

7.1. Nonresponse

In the intersection of model- and design based inference occurs also the phenomenon and treatment of survey nonresponse. Units sampled into the survey S sometimes do not answer either to the complete survey questionnaire or specific questions, leading to completely missing observations or at least missing variables.

When simulations are conducted in order to study the performance of nonresponse treatments, their conception and the inferential context of nonresponse is consequently of importance, too.

Nonresponse can either be understood as another subsampling process from S, that might again depend on auxiliaries Z or variables of interest Y. Or it can be understood as an additional binary variable that is attributed to each unit in the finite population, that is generated from a probability model possibly employing Z and Y. Define the variable $R_i \sim \text{Bern}(p_i)$ equal to one if unit i responds and zero else. In the first setting, R_i is conditional on $\mathbb{1}_S(i)$ and only defined for $i \in S$. In the second, R_i is defined for all $i \in U$. In both settings, it is possible to differentiate between *missing completely at random* (MCAR), *missing at random* (MAR), and not missing at random (NMAR) [Rubin, 1996]. In the design-based setting these may be described as follows: There exists no function f such that $p_i = f(\mathbf{y}, \mathbf{z})$ (MCAR), $p_i = f(\mathbf{y})$ (MAR) and there exists a function h such that $p_i = h(\mathbf{y}, \mathbf{z})$ (NMAR) for all $i \in U$ respectively. In the model-based setting MCAR, MAR and NMAR refer to $R \perp (Y, Z)$, $R \perp Y$ and $R \not\perp Y$ respectively.

The different concepts of nonresponse impact the design of MC studies, though. In the first framework, units in one and the same finite population might once respond to the survey and another time not. Consequently, the finite population total exists and could theoretically be determined if all units in U were observed.

A non-Bayesian application of this first framework is found in Bjørnstad [2007]. There, nonresponse can be considered to be another sampling process from the finite population, and the final survey is the intersection of the sampling process $R_i \sim \text{Ber}(p_i)$ and the realization of $S \sim P_D$. This approach is in so far interesting as it needs not assume that $\mathbf{y}$ is a random realization. Also this implies that nonresponse should be modelled prior to survey sampling within a MC study.

In the second, nonresponse is part of the model-based, data generating process and for a given finite population, the outcome of R_i is fixed regardless whether $i \in S$. In contrast to the first setting, the finite population total would even be unknown if all units $i \in U$ were observed. This setting is especially interesting when *multiple imputation* (MI) [Rubin, 1996] is considered to remedy

nonresponse. MI is a model-driven methodology, seeking to explain missingness in the variable of interest Y by the observed answers Z, i.e. MAR is assumed. As is learned from Sections 2.3 to 2.5 in Rubin [1996], MI is put into the second framework, implying that MC simulations studying MI are correctly conducted simulating nonresponse once at the finite population level and repeating the generation of finite populations. Note that even under MAR assumptions, MI becomes critical when $S \not\perp Y|Z$ because in that case, $(S, R) \not\perp Y|Z$ and the estimation of the correct imputation model becomes difficult.

The study of other violations of model assumptions such as the generation process of R, though, also allows to use MI in the firstly named context, i.e. in generating nonresponse in each sample realization $S = s$. Benchmarks in this model-based setting are ambiguous like in the SAE case and need to be communicated clearly because the conclusions can depend on the choice of benchmarks like in Section 6.1.

7.2. Resampling Methods

Independently from MC studies, basic resampling techniques such as the bootstrap encounter problems in survey sampling: When inference aims at $P_{m_\vartheta, D}$, there is no infinitely growing sample of realizations $\mathbf{y}$. Furthermore, when inference aims at P_D, there is only one sample $S = s$ available in practice which makes resampling difficult. Hence, the bootstrap procedure must be adapted to reflect the sampling process P_D within one sample realization, for example this is done in Sitter [1992b]. Competitors of the bootstrap such as the jackknife and the Balanced Repeated Replication (BRR), can also be seen in this context. For example, the BRR [McCarthy, 1966] may be considered to mimic the sample design with simply half of the scheduled sample size. Adjusted jackknife methods for complex samples, on the other hand, account for the impact of one included observation $\mathbf{y}_i$ in the sample s under design P_D. An analysis of the properties in stratified samples is given in Krewski and Rao [1981].

It is, however, an alternative to go the other way around, confer for example Gross [1980] or Booth et al. [1994] or Shao [2003]: Taking the population statistic g_U as a function of the discrete probability law $(Y, Z) \sim \text{Unif}_{(\mathbf{y},\mathbf{z})} =: P_U$, the authors state the objective of finite population resampling to plug-in realizations from $Y \sim \hat{P}_U$ into g_U where the estimator $\hat{P}_U$ is an empirical estimate for P_U employing the sample realization $S = s$. Note, that though the qualities of the bootstrap variants based on the second motivation are proven, the motivation is contrary to what we established as a sampling randomization framework.

As both motivations lead to non-parametric bootstrap variants that account for P_D, these methods are conceptually easily applicable to design-based MC studies. Nonetheless, resampling methods within simulation studies are highly computer intensive as any resampling size K scales up to $B \times K$ because the resampling is redone in each simulation run $b = 1, \ldots, B$ [Shao, 2003, Münnich et al., 2015]. Sitter [1992a] considers four different finite populations (based on real world sample data) from which random samples were drawn. We consider

this an extended design-based simulation study. Münnich et al. [2015] even run resampling methods in combination with SAE and nonresponse scenarios.

The parametric bootstrap [Hinkley, 1988], on the other hand, assumes that $Y \sim P_{\mathcal{M}_\vartheta}$ and makes inference with respect to $Y \sim P_{\mathcal{M}_{\hat{\vartheta}}}$. In that case, the estimator $\hat{\vartheta}$ should account for the survey design P_D. When inference on $P_{\mathcal{M}_\vartheta,D}$ is aimed at, though, a design-adapted estimator is not sufficient to yield accurate resampling inference, at least when the design is informative or the asymptotic independence of P_D and $P_{\mathcal{M}_\vartheta}$ is not applicable. For model-based MC studies in survey sampling, though, the parametric bootstrap is a useful method to yield estimators for higher order statistics in each simulation run whose distribution can then be evaluated using the MC probability $\hat{P}^B_{\mathcal{M}_\vartheta}$.

8. Discussion

In this paper, an overview of distinct statistical methodologies in survey sampling is given and the implications on MC-studies are described. Depending on the objective of a researcher's analysis, there are different probability laws that can be studied on survey data and the applicable probability law determines the set-up of the simulation study. We differentiate between a randomization approach returning a (synthetic) design-based Monte-Carlo simulation set-up, and model-based approaches that assume that a parametric statistical model has generated the data. The latter requires a model-based simulation set-up. Third, there are hybrid cases that account for both sources of stochasticity and are named (quasi or conditional) model-based scenarios, partially under finite population. We cite formulae on how to combine both probability laws and describe an adequate Monte-Carlo set-up. The theoretical results are demonstrated in exemplary Monte-Carlo studies.

The differentiation between the statistical frameworks is not always clear-cut, especially when estimators from different points of view shall be compared in the simulation study. Such examples are small area estimation and nonresponse. Especially in these hybrid cases, the communication of the Monte-Carlo framework has to be detailed and not all (empirical) papers on these fields are comparable to each other due to varying underlying assumptions.

Of course, within an overview paper hardly all details can be discussed extensively considering all hybrid versions of MC-studies in survey sampling. Nonetheless, we hope to have sharpened the view on important details of the overall statistical model that affect the set-up of simulation studies. Especially the quasi and synthetic design-based variants are highly subject to the availability of (close-to-reality) populations. Thus, the generation of such pseudo-populations will be a major task in future statistical research. In addition, the definition of the underlying population and response mechanisms extend straight to the debate of Big Data in statistical analysis because in Big Data, it might not be clear whether any unit in the population of interest could be contacted with positive probability and how the response patterns look like. Furthermore, machine learning algorithms (that are often implicit statistical models) that are not run

on image or online data but on populations of human beings encounter the problem of sub-sampling from a finite population, too. Design-based Monte-Carlo studies can help to learn about the behavior of machine learning algorithms in this (realistic) context.

The other way around, repeated Monte-Carlo studies are very computer intensive and it is not always a priori clear, what an adequate number of Monte-Carlo replications is or what could be chosen as a good proposal distributions. Feeding a machine learning algorithm with pairs of MC replications or proposal distributions respectively and the resulting MC error for supervised or affirmative learning, could help to give answers to these questions or to allow simulations with reduced Monte-Carlo variance.

Acknowledgements

The paper was finally written within the RIFOSS project (Research Innovation for Official and Survey Statistics), funded by the German Federal Statistical Office. However, the research was developed over longer period including several projects, i.e. the EU projects DACSEIS, KEI, AMELI, BLUE-ETS, MAKSWELL, and InGRID as well as the German Census 2011 sampling and estimation project to name the main projects.

References

Andreas Alfons, Jan Pablo Burgard, Peter Filzmoser, Beat Hulliger, Jan-Philipp Kolb, Stefan Kraft, Ralf Münnich, Tobias Schoch, and Matthias Templ. The ameli simulation study. Research Project Report WP6 – D6.1, FP7-SSH-2007-217322 AMELI, 2011a. URL http://ameli.surveystatistics.net.

Andreas Alfons, Peter Filzmoser, Beat Hulliger, Jan-Philipp Kolb, Stefan Kraft, Ralf Münnich, and Matthias Templ. Synthetic data generation of silc data. Research Project Report WP6 – D6.2, FP7-SSH-2007-217322 AMELI, 2011b. URL http://ameli.surveystatistics.net.

George E. Battese, Rachel M. Harter, and Wayne A. Fuller. An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36, 1988.

Jan F Bjørnstad. Non-Bayesian multiple imputation. *Journal of Official Statistics*, 23(4):433–452, 2007.

Hélène Boistard, Hendrik P Lopuhaä, and Anne Ruiz-Gazen. Functional central limit theorems in survey sampling. *ArXiv e-prints*, 1509, 2015.

James G. Booth and James P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Methodology)*, 61(1):265–285, 1999.

James G. Booth, Ronald W. Butler, and Peter Hall. Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89(428): 1282–1289, 1994.

D.J. Brus and J.J. de Gruijter. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, 80(1):1–44, 1997. ISSN 0016-7061.

Jan Pablo Burgard. *Evaluation of small area techniques for applications in official statistics*. PhD thesis, Trier University, 2013.

Jan Pablo Burgard and Patricia Dörr. Survey-weighted generalized linear mixed models. Research Papers in Economics 1, Trier University, 2018.

Jan Pablo Burgard, Jan-Philipp Kolb, Hariolf Merkle, and Ralf Münnich. Synthetic data for open and reproducible methodological research in social sciences and official statistics. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 11(3):233–244, Dec 2017. ISSN 1863-8163.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2 edition, 2009. ISBN 0203491289.

James E. Gentle. *Random Number Generation and Monte Carlo Methods*. Springer Science & Business Media, 2 edition, 2006.

S. Gross. Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods*, volume 1814184. American Statistical Association Alexandria, VA, 1980.

John H. Halton. A retrospective and prospective survey of the monte carlo method. *Siam Review*, 12(1):1–63, 1970.

David V. Hinkley. Bootstrap methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(3):321–337, 1988.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Malvin H. Kalos and Paula A. Whitlock. *Monte Carlo Methods*, volume 1. John Wiley & Sons, 1986. ISBN 0471898392.

Phillip S. Kott. A design-sensitive approach to fitting regression models with complex survey data. *Statistics Surveys*, 12:1–17, 2018. ISSN 1935-7516. .

D. Krewski and Jon N.K. Rao. Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, pages 1010–1019, 1981.

Roderick J. A. Little. Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77(378):237–250, 1982.

Philip J. McCarthy. Replication – an approach to the analysis of data from complex surveys. DHEW Publication 79-1269, US Department of Health, Education and Welfare, 1966. Vital and Health Statistics Series 2(14).

Christopher Z. Mooney. *Monte Carlo Simulation*. Sage Publ., 1997. ISBN 0803959435.

Ralf Münnich, Josef Schürle, Wolf Bihler, Harm-Jan Boonstra, Paul Knottnerus, Nico Nieuwenbroek, Alois Haslinger, Seppo Laaksonen, Doris Eckmair, Andreas Quatember, Helga Wagner, Jean-Pierre Renfer, Ueli Oetliker, and Rolf Wiegert. Monte-carlo simulation study on european surveys. Research Project Report WP3 – D3.1/3.2, IST-2000-26057 DACEIS, 2003. URL http://www.dacseis.de.

Ralf Münnich, Siegfried Gabler, Christian Bruch, Jan Pablo Burgard, Tobias

Enderle, Jan-Philipp Kolb, and Thomas Zimmermann. Tabellenauswertungen im zensus unter berücksichtigung fehlender werte. *AStA Wirtschafts-und Sozialstatistisches Archiv*, 9(3-4):269–304, 2015.

Danny Pfeffermann. The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, pages 317–337, 1993.

Jon N.K. Rao. *Small Area Estimation*. John Wiley & Sons, 2005.

Natalia Rojas-Perilla, Sören Pannier, Timo Schmid, and Nikos Tzavidis. Data-driven transformations in small area estimation. Discussion paper, School of Business & Economics: Economics, 2017.

Donald B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996.

Susana Rubin-Bleuer and Ioana Schiopu Kratina. On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33(6): 2789–2810, 2005.

Carl-Erik Särndal, Ib Thomsen, Jan M. Hoem, D. V. Lindley, O. Barndorff-Nielsen, and Tore Dalenius. Design-based and model-based inference in survey sampling [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 27–52, 1978.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer, 1992.

Jun Shao. Impact of the bootstrap on sample surveys. *Statistical Science*, 18 (2):191–198, 2003.

Randy Rudolf Sitter. Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20(2):135–154, 1992a.

Randy Rudolf Sitter. A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87(419):755–765, 1992b.

Richard Valliant, Alan H. Dorfman, and Richard M. Royall. *Finite Population Sampling and Inference - A Prediction Approach*. Probability and Statistics. Wiley, 2000.

V. N. Vapnik and A. Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

François Verret, Jon N.K. Rao, and Michael A. Hidiroglou. Model-based small area estimation under informative sampling. *Survey Methodology*, 41(2):333–347, 2015.

Julian Wagner, Ralf Münnich, Joachim Hill, Johannes Stoffels, and Thomas Udelhoven. Non-parametric small area models using shape-constrained penalized b-splines. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4):1089–1109, 2017.

Thomas Zimmermann. *The interplay between sampling design and statistical modelling in small area estimation*. PhD thesis, Trier University, 2018.