A sampling scheme for estimating the prevalence of a pandemic

Ze Liu, Si-Yu Yi, Jianghu (James) Dong, Min-Qian Liu & Yong-Dao Zhou

View supplementary material

Published online: 20 May 2023.

Submit your article to this journal

View related articles

View Crossmark data

Taylor & Francis
Taylor & Francis Group

Check for updates

A sampling scheme for estimating the prevalence of a pandemic

Ze Liu[a], Si-Yu Yi[a], Jianghu (James) Dong[b], Min-Qian Liu[a], and Yong-Dao Zhou[a]

[a]School of Statistics and Data Science, LPMC & KLMDASR, NanKai University, Tianjin, China; [b]Department of Biostatistics, College of Public Health, University of NE Medical Center, Omaha, Nebraska, USA

ABSTRACT

The spread of COVID-19 makes it essential to investigate its prevalence. In such investigation research, as far as we know, the widely-used sampling methods didn't use the information sufficiently about the numbers of the previously diagnosed cases, which provides a priori information about the true numbers of infections. This motivates us to develop a new, two-stage sampling method in this paper, which utilizes the information about the distributions of both population and diagnosed cases, to investigate the prevalence more efficiently. The global likelihood sampling, a robust and efficient sampler to draw samples from any probability density function, is used in our sampling strategy, and thus, our new method can automatically adapt to the complicated distributions of population and diagnosed cases. Moreover, the corresponding estimating method is simple, which facilitates the practical implementation. Some recommendations for practical implementation are given. Finally, several simulations and a practical example verify its efficiency.

1. Introduction

COVID-19 broke out at the end of 2019 and has become a global pandemic. All countries around the world have been severely affected, such as the United States of America (USA), where the numbers of confirmed cases and deaths increased rapidly from March 2020 to April 2021 (Centers for Disease Control and Prevention 2020). It is essential for the government and health institutions to monitor COVID-19 and control the pandemic by making practical and reasonable plans.

Since some people infected by SARS-CoV-2 are asymptomatic (Sakurai et al. 2020), one main difficulty with the pandemic is that the cumulative number of diagnosed cases cannot represent the number of infections. Many countries have made efforts to investigate the prevalence of COVID-19 (Anand et al. 2020; Stringhini et al. 2020; Pollán et al. 2020; Xu et al. 2020; Havers et al. 2020; Rosenberg et al. 2020; Sood et al. 2020; Ward et al. 2021); however, many of these investigations (Stringhini et al. 2020; Xu et al. 2020; Havers et al. 2020; Sood et al. 2020) only focused on one or several hotspot(s) instead of the whole country.

In addition, when investigating the prevalence nationwide, restricted by the costs, only a small part of the whole population can be investigated, especially when the country has a broad territory area. Therefore, it is important to develop some appropriate sampling strategies. The use of convenience samples (Stringhini et al. 2020; Xu et al. 2020; Havers et al. 2020; Rosenberg et al. 2020; Kissler et al. 2020) is not proper because they are prone to the selection bias, and thus,

---

problematic. Some other literature work (Leong et al. 2021; Parenteau et al. 2021; Knudsen et al. 2021; Tian et al. 2021; Sartorius et al. 2021) drew samples from some representative databases—for example, the medical insurance databases. This strategy can reduce the cost of the survey, but the representativeness of such samples depends on the representativeness of the database. Another popular sampling strategy used in the investigations of prevalence is the multi-stage stratified sampling, or some variants of it. Some recent studies (Jia et al. 2020; Pollán et al. 2020; Ward et al. 2021; Nagashima et al. 2021; Ssentongo et al. 2021; Horton-French et al. 2021; Mulenga et al. 2021; Li, Shan, and Teng 2021) have used these kinds of sampling methods. Compared with the simple random sampling, the variance of the estimator obtained by the stratified sampling is usually smaller. However, this method depends heavily on the construction of the strata in which the inter-homogeneity is required, and sometimes a well-designed stratified sampling strategy may lead to a very complex analysis procedure.

Intuitively, the distribution of the number of cumulative diagnosed cases provides priori information about the true situation of infections, and, therefore, can help with the sampling survey to improve the efficiency. However, none of the above research utilized this priori information sufficiently. In this paper, we propose a new sampling strategy for estimating the total number of infections nationwide. The main feature of this method is that it can flexibly and efficiently utilize the information about the distributions of both population and diagnosed cases. Compared with the stratified multi-stage sampling, our method is more flexible to adapt to various complicated distributions of population and diagnosed cases. The implementation and the corresponding estimating methods of this sampling strategy are also easier. There are two stages in the sampling strategy: first, determining the sampling positions according to some probability density, and then sampling from these positions. The main focus of this paper is on the first stage, in which the probability density may be multimodal and complicated. In this situation, some well-known methods to sample from a general probability density function, including the Markov Chain Monte Carlo (MCMC) method, e.g. Metropolis-Hastings (MH) algorithm (Hastings 1970), and the sampling/importance resampling (SIR) method (Rubin 1987), as well as its variants (Pérez et al. 2005; Ning and Tao 2020), may have a bad performance. The reason is that those methods can't adapt to various kinds of sampling densities (the MCMC method is easy to get stuck at some peak of the density function when the sampling density is multimodal, and the performance of the SIR depends heavily on the choice of the proposal distribution and the quality of the initial samples from the proposal distribution). To overcome these problems, Wang et al. (2015) proposed a new method, called the global likelihood sampling (GLS), and Yi et al. (2022) showed its theoretical properties. We will use GLS in the first stage of our proposed sampling strategy. As for the second stage, any advanced sampling strategy can be used; for simplicity, we only consider the simple random sampling method in this paper.

The rest of this paper is organized as follows. In Sec. 2, we describe the problem and propose the basic approach. The GLS algorithm is also described in this section. In Sec. 3, the optimal settings of our sampling strategy are derived. The complete sampling strategy and the corresponding estimating method are given in Sec. 4, as well as some suggestions for practical implementation. Some numerical simulations are conducted in Sec. 5, in order to find the robust setting of a parameter in our method, and to show the efficiency of our method. To further explain our method, a practical example is presented in Sec. 6. Finally, Sec. 7 concludes this paper. Some additional remarks, details, proofs and simulations are provided in the Supplementary Materials.

2. Preliminaries

2.1. Basic approach

Suppose a pandemic is spreading over a region $\mathscr{R} \subseteq \mathbb{R}^2$, and we want to know its total prevalence over the region $\mathscr{R}$. The total population of this region, the cumulative number of diagnosed cases,

and the cumulative number of real infections are denoted by N_{pop}, N_{diag} and N_{inf}, respectively. The distribution of population is seldom uniform, and this can be described by a non-negative function f_{pop} over $\mathscr{R}$ satisfying $\int_{\mathscr{R}} f_{\text{pop}}(x) \, \mathrm{d}x = N_{\text{pop}}$, that is, $f_{\text{pop}}(x)$ is approximately the number of population per unit area around $x \in \mathscr{R}$. Here after, we call f_{pop} the "density of population". Similarly, there are corresponding densities of diagnosed cases and infections in $\mathscr{R}$, denoted by f_{diag} and f_{inf}, respectively, satisfying $\int_{\mathscr{R}} f_{\text{diag}}(x) \, \mathrm{d}x = N_{\text{diag}}$ and $\int_{\mathscr{R}} f_{\text{inf}}(x) \, \mathrm{d}x = N_{\text{inf}}$, respectively.

The information about the population, N_{pop} and f_{pop}, and that about the diagnosed cases, N_{diag} and f_{diag}, are usually known (see Sec. 4 for more details), but the information about the real infections, N_{inf} and f_{inf}, is hard to get, which is just what we are interested in. Since the total prevalence is $N_{\text{inf}}/N_{\text{pop}}$, estimating the total prevalence is equivalent to estimating the total number of infections N_{inf}. Therefore, for convenience, our goal is to estimate N_{inf} by a sampling survey. We develop a two-stage sampling method in this subsection, and the optimal choices of the adjustable parameters in our method are discussed in Sec. 3.

Denote $\mathscr{Z}_k := \{1, ..., k\}$ for any $k \in \mathbb{N}^+$ where $\mathbb{N}^+$ is the set of all positive integers. Since N_{inf} is the integral of f_{inf} on $\mathscr{R}$, a popular way to approximate N_{inf} is the Monte Carlo method. Here we adopt the importance sampling technique (Lemieux 2009, Section 4.5), which is helpful in reducing the variance. Let φ be a probability density function on $\mathscr{R}$ and $\xi_1, ..., \xi_r$ be r independent and identically distributed (i.i.d.) samples from φ, then the sample mean $r^{-1} \sum_{i=1}^{r} f_{\text{inf}}(\xi_i)/\varphi(\xi_i)$ is an unbiased estimator of N_{inf}. The variance of this estimator depends on the choice of φ (Lemieux 2009, Section 4.5). Hereafter, we call φ the sampling density and call $\xi_1, ..., \xi_r$ the sampling positions. Since f_{inf} is unknown, we have to estimate the values of f_{inf} at $\xi_1, ..., \xi_r$. Therefore, our sampling survey consists of two stages: first, determining the sampling positions $\xi_1, ..., \xi_r$; second, selecting the samples, i.e. the people who will receive the tests, at each sampling position to estimate the values of f_{inf} there. The main focus of this paper is on the first stage, and the simple random sampling is adopted in the second stage. Other advanced sampling methods can also be used in the second stage, which is problem-dependent.

Suppose the desired total sample size is n, and the sample size at ξ_i is $\varsigma(\xi_i)$ for $i \in \mathscr{Z}_r$, where ς is a positive function on $\mathscr{R}$ such that

$$\int_{\mathscr{R}} \varsigma(x)\varphi(x) \, \mathrm{d}x = \frac{n}{r}. \tag{1}$$

The constraint (1) provides that the expectation of the total sample size $\sum_{i=1}^{r} \varsigma(\xi_i)$ is n. It is known that n and r affect ς, and we consider n and r as the hidden parameters of ς. Apparently, n/r serves as a multiplicative factor in ς. Hereafter, we call ς the allocation function of the sample sizes. Then, for each $i \in \mathscr{Z}_r$, let's derive an estimator of $f_{\text{inf}}(\xi_i)$. For simplicity, assume $\varsigma(\xi_i)$, the sample size at ξ_i, is an integer; otherwise, some rounding procedure is needed, but the corresponding results still hold approximately, as long as $\varsigma(\xi_i)$ is not too small. Given the sampling position ξ_i, in the second stage, $\varsigma(\xi_i)$ sample from ξ_i are tested, and denote the number of infections in the $\varsigma(\xi_i)$ samples by τ_i. Intuitively, one person's possibility of infection is associated with the infection status of others in that person's household/neighborhood/community. This correlation can be characterized by the local prevalence at ξ_i, i.e. $\rho_{\text{inf}}(\xi_i) := f_{\text{inf}}(\xi_i)/f_{\text{pop}}(\xi_i)$. Moreover, $\varsigma(\xi_i)$ is usually very small compared to the population around ξ_i. Hence, we can consider, approximately, that $\tau_i \mid \xi_i \sim \text{Bi}(\varsigma(\xi_i), \rho_{\text{inf}}(\xi_i))$, where 'Bi' represents the binomial distribution. An unbiased estimator of $f_{\text{inf}}(\xi_i)$ (conditional on ξ_i) is

$$\hat{f}_{\text{inf}}(\xi_i) := f_{\text{pop}}(\xi_i) \cdot \frac{\tau_i}{\varsigma(\xi_i)}, \tag{2}$$

with its variance being

$$\mathrm{Var}\left[\hat{f}_{\mathrm{inf}}(\boldsymbol{\xi}_i) \mid \boldsymbol{\xi}_i\right] = \frac{f_{\mathrm{inf}}(\boldsymbol{\xi}_i)\left[f_{\mathrm{pop}}(\boldsymbol{\xi}_i) - f_{\mathrm{inf}}(\boldsymbol{\xi}_i)\right]}{\varsigma(\boldsymbol{\xi}_i)}. \tag{3}$$

Combining the results of the two stages, we obtain an unbiased estimator of the total number of infections N_{inf},

$$\hat{N}_{\mathrm{inf}} := \frac{1}{r}\sum_{i=1}^{r}\frac{\hat{f}_{\mathrm{inf}}(\boldsymbol{\xi}_i)}{\varphi(\boldsymbol{\xi}_i)}. \tag{4}$$

The unbiasedness of $\hat{N}_{\mathrm{inf}}$ can be easily verified using the law of total expectation as follows:

$$\mathrm{E}(\hat{N}_{\mathrm{inf}}) = \mathrm{E}\left[\frac{\hat{f}_{\mathrm{inf}}(\boldsymbol{\xi}_1)}{\varphi(\boldsymbol{\xi}_1)}\right] = \mathrm{E}\left\{\mathrm{E}\left[\frac{\hat{f}_{\mathrm{inf}}(\boldsymbol{\xi}_1)}{\varphi(\boldsymbol{\xi}_1)}\bigg|\boldsymbol{\xi}_1\right]\right\} = \mathrm{E}\left[\frac{f_{\mathrm{inf}}(\boldsymbol{\xi}_1)}{\varphi(\boldsymbol{\xi}_1)}\right] = \int_{\mathscr{R}}\frac{f_{\mathrm{inf}}(x)}{\varphi(x)}\cdot\varphi(x)\,\mathrm{d}x = N_{\mathrm{inf}}. \tag{5}$$

Similarly, we can obtain its variance

$$\mathrm{Var}(\hat{N}_{\mathrm{inf}}) = \frac{1}{r}\left\{\mathrm{Var}\left[\frac{f_{\mathrm{inf}}(\boldsymbol{\xi}_1)}{\varphi(\boldsymbol{\xi}_1)}\right] + \mathrm{E}\left[\frac{f_{\mathrm{inf}}(\boldsymbol{\xi}_1)\left[f_{\mathrm{pop}}(\boldsymbol{\xi}_1) - f_{\mathrm{inf}}(\boldsymbol{\xi}_1)\right]}{\varsigma(\boldsymbol{\xi}_1)[\varphi(\boldsymbol{\xi}_1)]^2}\right]\right\} = \frac{1}{r}\cdot\left[v_0(\hat{N}_{\mathrm{inf}}) + v_1(\hat{N}_{\mathrm{inf}})\right]. \tag{6}$$

By the central limit theorem,

$$\frac{\hat{N}_{\mathrm{inf}} - N_{\mathrm{inf}}}{\sqrt{\mathrm{Var}(\hat{N}_{\mathrm{inf}})}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1) \quad \text{as} \quad r \to +\infty, \tag{7}$$

where '$\xrightarrow{\mathrm{d}}$' means the convergence in distribution, and $\mathrm{N}(0,1)$ is the standard normal distribution. Based on (7), we can construct the approximate confidence intervals for N_{inf} when r is large enough. It is hinted by Supplementary Material S4.3 that the least r required for well approximating this asymptotic distribution can be very small, which is completely achievable in practice. The estimator $\hat{N}_{\mathrm{inf}}$ in (4) can also be considered based on a kernel estimator of f_{inf}; see Supplementary Material S1.

2.2. Global Likelihood sampling

The sampling strategy introduced in Sec. 2.1 involves sampling from a bi-variate probability density function φ. Since the form of φ can be various, multimodal and complicated, the extraction of $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_r$ is not easy, and some well-known methods, e.g. MCMC and SIR, may be not suitable. Instead, we adopt the GLS algorithm, which applies to the multimodal and complicated cases, to generate the r sampling positions. Detailed discussion can refer to Zhou et al. (2021). Algorithm 1 describes the GLS method for generating i.i.d. samples $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_r$ from φ. The GLS used here is simplified compared to the original one in Wang et al. (2015).

In the inputs of Algorithm 1, the kernel $\tilde{\varphi}$ is an arbitrary positive constant multiple of the desired sampling density φ. The uniform design D is a set of M points scattered evenly in $\overline{\mathscr{R}} = [0,1]^2$; refer to Fang et al. (2018) for more about uniform designs. Intuitively, the better the uniformity of D, the better $\mathscr{P}^{(i)}$ approximates φ for each $i \in \mathscr{Z}_r$, and so the better the quality of each sample $\boldsymbol{\xi}_i$. Since the generated $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_r$ are i.i.d., the setting of M is not related to the setting of r. Therefore, we should take a large enough M, say 100 or more, to assure that the empirical distribution of $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_r$ approximates φ well.

There are some additional remarks about the region $\overline{\mathscr{R}}$ and the uniform design D in Algorithm 1 in Supplementary Material S2.

Algorithm 1: GLS algorithm

Let the smallest rectangle containing $\mathscr{R}$ be $\overline{\mathscr{R}} \subseteq \mathbb{R}^2$; without loss of generality, suppose $\overline{\mathscr{R}} = [0,1]^2$.

Input: (i) r: the number of samples from the sampling density; (ii) $\tilde{\varphi}$: the kernel of the desired sampling density φ; (iii) $D = \left\{ \boldsymbol{q}_j : j \in \mathscr{Z}_M \right\}$: a uniform design on $\overline{\mathscr{R}} = [0,1]^2$.

Step 1. Loop. Repeat Steps 2 to 4 for $i = 1, ..., r$.

Step 2. Random shift. Generate $\boldsymbol{\delta}^{(i)} \sim \mathrm{U}(\overline{\mathscr{R}})$ and let $D^{(i)} = \left\{ \boldsymbol{q}_j \oplus \boldsymbol{\delta}^{(i)} : j \in \mathscr{Z}_M \right\} = D \oplus \boldsymbol{\delta}^{(i)}$, where the operator $\oplus$ means the addition modulo 1, i.e. if $\boldsymbol{\delta}^{(i)} = (\delta_1^{(i)}, \delta_2^{(i)})$, $\boldsymbol{q}_j = (q_{j1}, q_{j2})$ and $\boldsymbol{q}_j \oplus \boldsymbol{\delta}^{(i)} = (q_{j1}^{(i)}, q_{j2}^{(i)})$, then for $k \in \{1,2\}$, $q_{jk}^{(i)} = (q_{jk} + \delta_1^{(i)}) - \mathrm{floor}(q_{jk} + \delta_1^{(i)})$, where $\mathrm{floor}(t)$ is the greatest integer no more than t.

Step 3. Likelihood. For each $x \in D^{(i)}$, let $w^{(i)}(x) = \tilde{\varphi}(x) / \sum_{y \in D^{(i)}} \tilde{\varphi}(y)$, where $\tilde{\varphi}(x) = 0$ for $x \in \overline{\mathscr{R}} \setminus \mathscr{R}$. Then $\mathscr{P}^{(i)} = \left\{ (x, w^{(i)}(x)) : x \in D^{(i)} \right\}$ is a discrete distribution on $D^{(i)}$.

Step 4. Sampling. Generate ξ_i in $D^{(i)}$ from the discrete distribution $\mathscr{P}^{(i)}$.

Output: r i.i.d. samples from φ : $\xi_1, ..., \xi_r$.

---

3. Optimal settings of the parameters

In the two-stage sampling strategy introduced in Sec. 2.1, there are three adjustable parameters: the sampling density φ, the allocation function of the sample sizes ς, and the number of sampling positions r. From (5), (6) and (7), these three parameters do not affect the unbiasedness of $\hat{N}_{\mathrm{inf}}$, but do affect its variance and distribution. In this section, we discuss the optimal settings of these parameters, where 'optimal' means to minimize the variance of $\hat{N}_{\mathrm{inf}}$.

First, we derive the theoretical minimum variance of $\hat{N}_{\mathrm{inf}}$ as follows.

Proposition 3.1. *For* $\mathrm{Var}(\hat{N}_{\mathrm{inf}})$ *in (6), we have:*

i. *The term* $v_1(\hat{N}_{\mathrm{inf}})/r$ *does not depend on* r.

ii. *Let* $\tilde{f} := \sqrt{f_{\mathrm{inf}}(f_{\mathrm{pop}} - f_{\mathrm{inf}})}$, *and* $\kappa(\varphi^2/\tilde{f}) := \inf\left\{ [\varphi(x)]^2/\tilde{f}(x) : x \in \mathrm{supp}(\varphi) \right\}$, *where* $\mathrm{supp}(\varphi)$, *the support set of* φ, *is the closure of* $\{ x : x \in \mathscr{R} \text{ and } \varphi(x) > 0 \}$. *For any given* $r \in \mathbb{N}^+$ *and* φ *satisfying* $\kappa(\varphi^2/\tilde{f}) > 0$, $v_1(\hat{N}_{\mathrm{inf}})/r$ *is minimized subject to the constraint (1), when for any* $x \in \mathscr{R}$,

$$\varsigma(x) = \frac{\tilde{f}(x)/\varphi(x)}{r \int_{\mathscr{R}} \tilde{f}(y) \, \mathrm{d}y} \cdot n. \tag{8}$$

The minimum of $v_1(\hat{N}_{\mathrm{inf}})/r$ *is* $\left[\int_{\mathscr{R}} \tilde{f}(x) \, \mathrm{d}x \right]^2/n$.

iii. *When* $\varphi \propto f_{\mathrm{inf}}$, $v_0(\hat{N}_{\mathrm{inf}}) = 0$.

iv. *Assume* $\kappa(f_{\mathrm{inf}}^2/\tilde{f}) > 0$, *then* $\mathrm{Var}(\hat{N}_{\mathrm{inf}})$ *is minimized subject to the constraints (1) and* $\kappa(\varphi^2/\tilde{f}) > 0$, *when* $r \in \mathbb{N}^+$ *is arbitrary,* $\varphi \propto f_{\mathrm{inf}}$, *and* ς *is set as (8). The minimum of* $\mathrm{Var}(\hat{N}_{\mathrm{inf}})$ *is*

$$\frac{1}{n} \left[\int_{\mathscr{R}} \tilde{f}(x) \, \mathrm{d}x \right]^2. \tag{9}$$

The proof of Proposition 3.1 is in Supplementary Material S3. Proposition 3.1 derives the optimal settings of (r, φ, ς), as well as the theoretical minimum value of $\mathrm{Var}(\hat{N}_{\mathrm{inf}})$. In (ii) and (iv), the constraint $\kappa(\varphi^2/\tilde{f}) > 0$ is needed to guarantee the optimality of the solutions. Such constraint is mild and can be easily satisfied, e.g. when is bounded and $\kappa(\varphi) > 0$. Even without this constraint, the solutions in (ii) and (iv) are still stationary points of the respective objective functions, and they are definitely not maximizers. Also note that in the optimal settings that minimize $\mathrm{Var}(\hat{N}_{\mathrm{inf}})$, r can be arbitrary, and it only affects ς through a multiplicative factor $1/r$ in (8). That means $\mathrm{Var}(\hat{N}_{\mathrm{inf}})$ is not affected by the tradeoff between r (affecting the thoroughness of the exploration of the region $\mathscr{R}$) and the sample sizes at the sampling positions (affecting the local sampling variances), provided $\varphi \propto f_{\mathrm{inf}}$. This result is verified by additional numerical simulations in Supplementary Material S4.3.

However, the above exact optimal settings of φ and ς cannot be achieved in practice because those optimal settings depend on f_{inf}, which is unknown, and is just what we want to estimate. Instead, we need to find the nearly optimal settings of φ and ς. We first consider the following mechanism to determine an initial rough estimate of f_{inf}. If a pilot sampling survey is implemented or some historical data are available, then we can take a kernel estimator like that in Supplementary Material S1, based on the available data, to roughly estimate f_{inf}. Otherwise, without any extra knowledge, the only information about f_{inf} is $0 \leqslant f_{\mathrm{diag}} \leqslant f_{\mathrm{inf}} \leqslant f_{\mathrm{pop}}$, which implies that there exists a function γ from $\mathscr{R}$ to $[0, 1]$ such that $f_{\mathrm{inf}} = \gamma \cdot f_{\mathrm{pop}} + (1 - \gamma) \cdot f_{\mathrm{diag}}$. It is impossible to know the form of γ in this case, since finding γ is equivalent to finding f_{inf}. A simple but reasonable way is to roughly estimate γ by a constant function, i.e. to choose a proper constant $\check{\gamma} \in [0, 1]$, and take

$$\check{f}_{\mathrm{inf}} := \check{\gamma} \cdot f_{\mathrm{pop}} + (1 - \check{\gamma}) \cdot f_{\mathrm{diag}} \tag{10}$$

as an initial rough estimate of f_{inf}. This estimation combines the information of both f_{pop} and f_{diag}, which can make our method efficient. With the estimator $\check{f}_{\mathrm{inf}}$, the nearly optimal setting of φ is $\check{f}_{\mathrm{inf}}/[\check{\gamma} \cdot N_{\mathrm{pop}} + (1 - \check{\gamma}) \cdot N_{\mathrm{diag}}]$. As for ς, we can use $\check{\tilde{f}} := \sqrt{\check{f}_{\mathrm{inf}}(f_{\mathrm{pop}} - \check{f}_{\mathrm{inf}})}$ to estimate $\tilde{f}$ in (8), but there remains an integral $\int_{\mathscr{R}} \tilde{f}(y) \, \mathrm{d}y$ to approximate. Since the form of $\tilde{f}$ is generally complicated, it is appropriate to approximate this integral using the Monte Carlo method. In fact, since the r positions $\xi_1, ..., \xi_r$ are i.i.d. samples from φ, the sample mean

$$\frac{1}{r} \cdot \sum_{i=1}^{r} \frac{\check{\tilde{f}}(\xi_i)}{\varphi(\xi_i)}$$

is an unbiased estimator of the integral $\int_{\mathscr{R}} \tilde{f}(y) \, \mathrm{d}y$. Therefore, for each $i \in \mathscr{Z}_r$, the nearly optimal sample size at the sampling position ξ_i, which is an approximation of the exact optimal $\varsigma(\xi_i)$, is

$$n_{\xi_i} := \frac{\check{\tilde{f}}(\xi_i)/\varphi(\xi_i)}{\sum\limits_{j=1}^{r} \check{\tilde{f}}(\xi_j)/\varphi(\xi_j)} \cdot n.$$

With the nearly optimal setting of φ, it is simplified to

$$n_{\xi_i} = \frac{\sqrt{\left[f_{\mathrm{pop}}(\xi_i) - \check{f}_{\mathrm{inf}}(\xi_i)\right]/\check{f}_{\mathrm{inf}}(\xi_i)}}{\sum\limits_{j=1}^{r} \sqrt{\left[f_{\mathrm{pop}}(\xi_j) - \check{f}_{\mathrm{inf}}(\xi_j)\right]/\check{f}_{\mathrm{inf}}(\xi_j)}} \cdot n. \tag{11}$$

The integral $\int_{\mathscr{R}} \tilde{f}(y) \, \mathrm{d}y$ can be estimated more accurately, for example, by using a larger set of Monte Carlo samples rather than $\{\xi_1, ..., \xi_r\}$. However, we still recommend the aforementioned approach for the following four reasons:

i.   From the discussion in the next paragraph, r should be large, so this estimator is plausible;
ii. Even if the integral $\int_{\mathscr{R}} \tilde{f}(y) \, \mathrm{d}y$ in (8) is replaced by a biased one, say $t \cdot \int_{\mathscr{R}} \tilde{f}(y) \, \mathrm{d}y$ where $t > 0$, it is easy to check that one, say ausible;approach for the following four reasons:n/t instead of n, i.e., under the constraint (1) in the right-hand side of which n is replaced by n/t. Therefore, this integral only controls the expectation of the total sample size, and does not affect the optimality of e that in Supplementary Material S1, based on the
iii. Most importantly, this approach guarantees $\sum_{i=1}^{r} n_{\xi_i} = n$, i.e., the total sample size is exactly the desired value n, before rounding the sample sizes at those sampling positions to integers. A deterministic total sample size is usually more desirable in practice, than a random total sample size with expectation being n which is indicated by the constraint (1);
iv. It is simple.

However, rounding the sample sizes at those sampling positions to integers can make the total sample size different from n, which also happens in the allocation of sample sizes in traditional sampling techniques. If this matters, one can apply some more sophisticated rounding methods to maintain the total sample size equaling n: for example, rounding n_{ξ_i} to $\mathrm{floor}(n_{\xi_i}) + 1$ if the fractional part $n_{\xi_i} - \mathrm{floor}(n_{\xi_i})$ is in the top $\left[n - \sum_{j=1}^{r} \mathrm{floor}(n_{\xi_j}) \right]$ ones among the r fractional parts, and rounding n_{ξ_i} to $\mathrm{floor}(n_{\xi_i})$ otherwise.

Next, we consider the number of sampling positions r. Although Proposition 3.1(iv) says that r can be arbitrary in the exact optimal settings, r indeed affects $\mathrm{Var}(\hat{N}_{\inf})$ when the settings of φ and ς are nearly optimal. By Proposition 3.1(i), r affects $\mathrm{Var}(\hat{N}_{\inf})$ mainly through the term $v_0(\hat{N}_{\inf})/r$. Since $v_0(\hat{N}_{\inf})$ does not depend on r, r should be large in order to reduce the variance when φ is not exactly proportional to $f_{\inf}$, as long as n_{ξ_i} is not too small to estimate the value of $f_{\inf}$ at ξ_i for each $i \in \mathscr{L}_r$. By (7), a large r also helps to obtain a good approximate distribution of $\hat{N}_{\inf}$. The numerical simulations in Supplementary Material S4.3 show that larger r can notably reduce the variance and improve the coverage rate of the confidence interval under the nearly optimal settings. Note that in practice, larger r may also increase the cost and the difficulty of the sampling survey, since more positions have to be sampled. These results are consistent with the classical sampling theory (Cochran 1977).

In addition, the costs at different sampling positions may be different in practice, which can be quantified by a cost function c over $\mathscr{R}$. Then our goal is to find the appropriate parameters which can minimize both of $\mathrm{Var}(\hat{N}_{\inf})$ and the total cost. For example, the weighted sum of $\mathrm{Var}(\hat{N}_{\inf})$ and the total cost can be used as an optimality criterion, i.e. $\mathrm{Var}(\hat{N}_{\inf}) + \gamma_c \int_{\mathscr{R}} c(x) \varsigma(x) \varphi(x) \, \mathrm{d}x$ where $\gamma_c \in [0, +\infty)$ reflects the importance of the total cost. For such cases, an analysis similar to the above can be performed, but it may be difficult to derive the explicit expressions. Instead, the numerical optimization algorithms can be considered to solve this problem, which is beyond the scope of this paper.

4. Sampling and estimating

Based on the discussion about the nearly optimal settings of the parameters in Sec. 3, we show the complete sampling strategy and the estimating method in this section. Some suggestions for the practical implementation are also given.

The details of the two-stage sampling strategy are described in Algorithm 2, which combines the basic approach in Sec. 2.1 with the GLS algorithm in Sec. 2.2.

Algorithm 2: Two-stage sampling strategy

Input: (i) $\check{\gamma}$: the constant in $[0, 1]$ to give a rough estimator of f_{inf} by (10); (ii) D: the uniform design on $[0, 1]^2$ used in Algorithm 1; (iii) n: the total sample size; (iv) r: the number of sampling positions.

Step 1. Rough estimate. Obtain $\check{f}_{\text{inf}}$, an initial rough estimator of f_{inf}, by (10). The kernel of the sampling density φ is set to be $\check{f}_{\text{inf}}$.

Step 2. Sampling positions. Generate r i.i.d. sampling positions $\xi_1, ..., \xi_r$ in $\mathscr{R}$ from the sampling density φ by the GLS in Algorithm 1 with inputs $(r, \check{f}_{\text{inf}}, D)$.

Step 3. Allocation of sample sizes. For each $i \in \mathscr{Z}_r$, calculate round(n_{ξ_i}), the sample size at the sampling position ξ_i, where n_{ξ_i} is calculated by (11) and 'round' is the function rounding a real number into the nearest integer.

Step 4. Test. For each $i \in \mathscr{Z}_r$, select round(n_{ξ_i}) people at position ξ_i by the simple random sampling and then implement tests on them.

Next, we give some remarks about Algorithm 2 as follows.

i.   About f_{pop} and f_{diag}. These two density functions can usually be obtained or approximated in practice. For example, we usually have the number of population in each administrative district, and we can approximate f_{pop} by a piecewise constant function whose value on each administrative district is a constant equaling the ratio of the number of population in this district and the area of this district. We can similarly obtain f_{diag}. In order to gather more information, we should use as fine administrative division as possible. The kernel methods like that in Supplementary Material S1 can also be applied to smooth the densities.

ii. About $\check{\gamma}$. It is difficult to theoretically optimize the setting of $\check{\gamma}$, but the numerical studies will give some recommendations in Sec. 5.1.

iii. About n. In practice, the n is usually determined by both the requirement of precision and the restriction of costs, thus $\text{Var}(\hat{N}_{\text{inf}})$ should be roughly estimated before the implementation of the sampling strategy. This can be done by using (9), which can be approximated by

$$\frac{1}{n} \left[\frac{\mathbf{m}(\mathscr{R})}{M} \sum_{x \in D} \check{f}(x) \right]^2$$

where $\mathbf{m}(\mathscr{R})$ is the area of $\mathscr{R}$, and D is the uniform design with M points used in Algorithm 2. The n should be set such that the above estimator of $\text{Var}(\hat{N}_{\text{inf}})$ is smaller than some pre-defined threshold about the precision.

iv. About n_{ξ_i}. As mentioned in Sec. 3, r should be as large as possible, provided that the cost will not exceed the budget, and no n_{ξ_i} is too small to estimate $f_{\text{inf}}(\xi_i)$. An additional way to avoid small n_{ξ_i} is to modify the calculation method of the sample sizes at those sampling positions, i.e., the n_{ξ_i} in Step 3 of Algorithm 2 as

$$n_{\xi_i} := \frac{\sqrt{\left[f_{\text{pop}}(\xi_i) - \check{f}_{\text{inf}}(\xi_i) \right] / \check{f}_{\text{inf}}(\xi_i)}}{\sum\limits_{j=1}^{r} \sqrt{\left[f_{\text{pop}}(\xi_j) - \check{f}_{\text{inf}}(\xi_j) \right] / \check{f}_{\text{inf}}(\xi_j)}} \cdot (1 - \eta)n + \frac{1}{r} \cdot \eta n, \qquad (12)$$

where $\eta \in [0, 1]$ is a properly chosen constant. Since the allocation method of sample sizes

in (11) is nearly optimal, η in (12) should not be too large. The proper setting of η should make the minimum sample size among those sampling positions just achieve some pre-defined threshold. Further, if n_{ξ_i} is not very small compared to the population around the sampling position ξ_i, finite population corrections (Cochran 1977; Lohr 2019) should be applied. For such case, expression (3) becomes

$$\mathrm{Var}\Big[\hat{f}_{\mathrm{inf}}(\xi_i) \mid \xi_i\Big] = [1 - \rho_S(\xi_i)] \cdot \frac{f_{\mathrm{inf}}(\xi_i)\big[f_{\mathrm{pop}}(\xi_i) - f_{\mathrm{inf}}(\xi_i)\big]}{\varsigma(\xi_i)},$$

where $\rho_S(\xi_i)$ refers to the sampling fraction at ξ_i and $i \in \mathscr{Z}_r$, and $\nu_1(\hat{N}_{\mathrm{inf}})$ in (6) becomes

$$\mathrm{E}\left\{[1 - \rho_S(\xi_1)] \cdot \frac{f_{\mathrm{inf}}(\xi_1)\big[f_{\mathrm{pop}}(\xi_1) - f_{\mathrm{inf}}(\xi_1)\big]}{\varsigma(\xi_1)[\varphi(\xi_1)]^2}\right\}.$$

v. About the sampling survey at each sampling position. For each $i \in \mathscr{Z}_r$, in practice, the n_{ξ_i} samples come from not exactly ξ_i, but a neighborhood of ξ_i. For convenience, this neighborhood can be chosen as some small district, like a city, village, or community, containing ξ_i or just close to ξ_i. It won't affect the property of this sampling strategy as long as the diameter of the neighborhood is negligible compared to the whole region $\mathscr{R}$. For example, the samples at each ξ_i can be drawn from the people whose current residences are within the circle centered at ξ_i with radius 10km. Some other popular sampling methods, such as the stratified sampling, cluster sampling, multi-stage sampling and other techniques (Lohr 2019), can be used to estimate $f_{\mathrm{inf}}(\xi_i)$ more efficiently. The corresponding results can be obtained similarly with suitable modifications.

Our main purpose is to estimate the cumulative total number of infections N_{inf}. With the testing results obtained by the above sampling strategy, we can estimate N_{inf} by using (2) and (4). The variance of the estimator can also be estimated by (6), and then an approximate confidence interval (CI) can be constructed according to (7). This estimating procedure is discribed in detail as follows.

Step 1. For $i = 1, ..., r$, calculate $\hat{f}_{\mathrm{inf}}(\xi_i)$ by (2) using the testing result τ_i at ξ_i and replacing $\varsigma(\xi_i)$ by the actual sample size at ξ_i, and calculate $\varphi(\xi_i)$ where $\varphi = \check{f}_{\mathrm{inf}} / \big[\check{\gamma} \cdot N_{\mathrm{pop}} + (1 - \check{\gamma}) \cdot N_{\mathrm{diag}}\big]$.

Step 2. Obtain the point estimator $\hat{N}_{\mathrm{inf}}$ of the total number of infections N_{inf} according to (4).

Step 3. According to (6), let

$$\begin{aligned}
\hat{v}_0(\hat{N}_{\mathrm{inf}}) &:= \frac{1}{r-1} \sum_{i=1}^{r} \left[\frac{\hat{f}_{\mathrm{inf}}(\xi_i)}{\varphi(\xi_i)} - \hat{N}_{\mathrm{inf}}\right]^2, \\
\hat{v}_1(\hat{N}_{\mathrm{inf}}) &:= \frac{1}{r} \sum_{i=1}^{r} [1 - \rho_S(\xi_i)] \cdot \frac{\hat{f}_{\mathrm{inf}}(\xi_i)\big[f_{\mathrm{pop}}(\xi_i) - \hat{f}_{\mathrm{inf}}(\xi_i)\big]}{n_{\xi_i}[\varphi(\xi_i)]^2}, \\
\hat{v}(\hat{N}_{\mathrm{inf}}) &:= \frac{1}{r} \cdot \big[\hat{v}_0(\hat{N}_{\mathrm{inf}}) + \hat{v}_1(\hat{N}_{\mathrm{inf}})\big],
\end{aligned} \tag{13}$$

then $\hat{v}(\hat{N}_{\mathrm{inf}})$ is an estimator of $\mathrm{Var}(\hat{N}_{\mathrm{inf}})$.

Step 4. According to (7), an approximate $1 - \alpha$ CI of N_{inf} is $\Big[\hat{N}_{\mathrm{inf}} - z_{\alpha/2}\sqrt{\hat{v}(\hat{N}_{\mathrm{inf}})},\ \hat{N}_{\mathrm{inf}} + z_{\alpha/2}\sqrt{\hat{v}(\hat{N}_{\mathrm{inf}})}\Big]$, where $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of $\mathrm{N}(0,1)$.

Hence, by the two-stage sampling strategy and the estimating method above, we can obtain the estimator of N_{inf}, its estimated variance and approximate CI. Note that these estimators cannot be improved by generating extra sampling positions from φ other than $\xi_1, ..., \xi_r$, since we can only get the estimator $\hat{f}_{\text{inf}}$ at $\xi_1, ..., \xi_r$ based on the testing results obtained in Step 4 of Algorithm 2. In practice, the settings of ς and φ are only nearly optimal, so $\text{Var}(\hat{N}_{\text{inf}})$ is different from its theoretic minimal value (9), and we don't have to estimate (9) based on the sampling data. The complete procedure is summarized in Figure 1. In the next two sections, we will show some numerical simulations and a practical example to verify the validity of our proposed sampling strategy.

5. Numerical simulation

In Sec. 3 and Sec. 4, we discussed how to set the parameters in the two-stage sampling strategy, except for the coefficient $\check{\gamma}$ in (10). In this section, we first give a robust setting for $\check{\gamma}$ in the mini-max sense through some numerical simulations, when there is little information about the underlying true f_{inf}. Then we compare our proposed sampling strategy with some other popular methods to verify the efficiency of our method.

5.1. Robust setting of $\check{\gamma}$

In the oracle situation, when the exact optimal settings of ς and φ discussed in Sec. 3 can be achieved, the variance of $\hat{N}_{\text{inf}}$ is minimized, denoted by $\text{Var}(\hat{N}_{\text{inf}} \mid \varphi \propto f_{\text{inf}})$. On the other hand, when we do not know the true f_{inf} and have to use the initial rough estimator (10) to determine the sample sizes and the sampling density, the variance of $\hat{N}_{\text{inf}}$, denoted by $\text{Var}(\hat{N}_{\text{inf}} \mid \varphi \propto \check{f}_{\text{inf}})$, depends on the quality of $\check{f}_{\text{inf}}$, and thus depends on $\check{\gamma}$. Therefore, in order to measure the performance of different settings of $\check{f}_{\text{inf}}$ or $\check{\gamma}$, we define the standardized standard deviation (SSD) as

$$\text{SSD}(\check{\gamma}) = \text{SSD}(\check{f}_{\text{inf}}) := \sqrt{\frac{\text{Var}(\hat{N}_{\text{inf}} \mid \varphi \propto \check{f}_{\text{inf}})}{\text{Var}(\hat{N}_{\text{inf}} \mid \varphi \propto f_{\text{inf}})}}.$$

Our goal is to find the setting of $\check{\gamma}$ such that the maximum of SSD over all possible true f_{inf}'s is minimized at that $\check{\gamma}$.

In the simulations, the region $\mathcal{R} = [0, 1]^2$ is the unit square and it is divided into four equal-sized sub-squares. Let the populations in them be 20, 40, 60 and 80, multiplied by 1×10^4
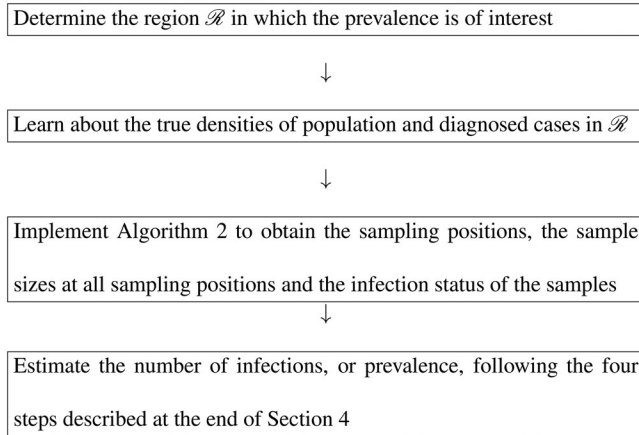
Determine the region $\mathcal{R}$ in which the prevalence is of interest

↓

Learn about the true densities of population and diagnosed cases in $\mathcal{R}$

↓

Implement Algorithm 2 to obtain the sampling positions, the sample sizes at all sampling positions and the infection status of the samples

↓

Estimate the number of infections, or prevalence, following the four steps described at the end of Section 4

Figure 1. The complete procedure for implementing the proposed sampling method.

Table 1. Settings of f_{inf} and $\check{\gamma}$ in the simulations in Series E.

	Group E1	Group E2	Group E3
γ	$\{0.05, 0.1, ..., 0.9, 0.95\}$	$\{0.05, 0.1, ..., 0.45, 0.5\}$	$\{0.5, 0.55, ..., 0.9, 0.95\}$
$\check{\gamma}$	$\{0.05, 0.1, ..., 0.9, 0.95\}$	$\{0.05, 0.1, ..., 0.45, 0.5\}$	$\{0.5, 0.55, ..., 0.9, 0.95\}$

respectively, and the numbers of diagnosed cases in them be 6, 8, 4 and 2, multiplied by 1×10^4 respectively. Dividing those numbers by 1/4 will obtain the corresponding densities in the sub-squares. Following is the corresponding graph.

$$\text{Populations } (\times 10^4): \begin{array}{|c|c|} \hline 20 & 40 \\ \hline 60 & 80 \\ \hline \end{array}, \quad \text{Diagnosed cases } (\times 10^4): \begin{array}{|c|c|} \hline 6 & 8 \\ \hline 4 & 2 \\ \hline \end{array}.$$

The total population $N_{\text{pop}} = 200 \times 10^4$ and the total number of diagnosed cases $N_{\text{diag}} = 20 \times 10^4$. We set the size of the uniform design used in the GLS algorithm $M = 210$, the total sample size $n = 1 \times 10^4$ and the number of sampling positions $r = 50$. In different groups of simulations, the settings of f_{inf} will be different. For each setting of f_{inf} and $\check{\gamma}$, $\text{Var}(\hat{N}_{\text{inf}})$ is calculated using the sample variance of $\hat{N}_{\text{inf}}$ over 200 independent simulations.

In the first series of simulations, Series E, assume the true density of the infections f_{inf} is a convex combination of f_{pop} and f_{diag}, i.e. $f_{\text{inf}} = \gamma f_{\text{pop}} + (1 - \gamma) f_{\text{diag}}$, where $\gamma \in [0, 1]$ is a constant. Series E contains three groups of simulations, E1 to E3, whose settings are shown in Table 1, and the corresponding results are given in Figure 2. For each setting of f_{inf} (or γ) in Group E1, the coefficient $\check{\gamma}$ takes 19 different values, and the corresponding 19 values of $\text{SSD}(\check{\gamma})$ form a black curve in Figure 2a. The robust setting of $\check{\gamma}$ is the one that minimizes the maximum SSD among the 19 settings of f_{inf}, i.e. the one that minimizes the red bold curve in Figure 2a, which is marked out by a red circle. The other two subfigures are obtained similarly. The three subfigures in Figure 2 present a similar phenomenon that in the domain of γ, the maximum SSD is large when $\check{\gamma}$ is close to the bounds of the domain, while the maximum SSD is minimized when $\check{\gamma}$ is neither too large nor too small. Therefore, Figure 2 indicates that when a possible range of the true value of γ is available, the mid-point of that range may be a robust setting of $\check{\gamma}$.

In order to verify this conclusion, another series of simulations, Series R, is conducted. In this series, the true density of the infections is more complex than that in Series E. For each $i \in \mathscr{Z}_4$, the density of infections in the i-th sub-square is $f_{\text{inf},i} = \gamma_i f_{\text{pop},i} + (1 - \gamma_i) f_{\text{diag},i}$, where $\gamma_i \in [0, 1]$, and $f_{\text{pop},i}$ and $f_{\text{diag},i}$ are the densities of population and diagnosed cases in the i-th sub-square, respectively. Series R also contains three groups of simulations, R1 to R3. The settings are shown in Table 2, in which there are 20 independent settings for $\gamma_1, ..., \gamma_4$ in each group, and the corresponding results are given in Figure 3. For example, in Group R1, each setting of f_{inf} is obtained by generating $\gamma_1, ..., \gamma_4 \overset{\text{i.i.d.}}{\sim} U[0.02, 0.98]$. Under each setting of f_{inf}, $\check{\gamma}$ takes the 19 values in $\{0.05, 0.1, 0.15, ..., 0.9, 0.95\}$ one by one, and the corresponding values of $\text{SSD}(\check{\gamma})$ are calculated, and they form a black curve in Figure 3a. The robust setting of $\check{\gamma}$ is the one that minimizes the maximum SSD, i.e. the one that minimizes the red bold curve in Figure 3a, which is marked out by a red circle. The three subfigures in Figure 3 seem more tanglesome than those in Figure 2, which is caused by the complex setting of f_{inf} in Series R. However, the phenomenon presented in Figure 3 is similar to that in Figure 2, i.e. in each subfigure, the red bold curve becomes high when $\check{\gamma}$ is near the end of the domain of the γ_i's, but becomes low when $\check{\gamma}$ is around the mid-point of that domain. It indicates again that the mid-point of the possible range of γ_i's is a robust setting of $\check{\gamma}$. Some additional simulations in Supplementary Material S4.1 also present the similar phenomenon clearly. Therefore, according to the above simulations under different settings of γ, when a possible range of the value of γ is available, we should set $\check{\gamma}$ around the mid-point of that range, which is robust in the sense that the maximum value of SSD over all possible settings of f_{inf} would not be very large.
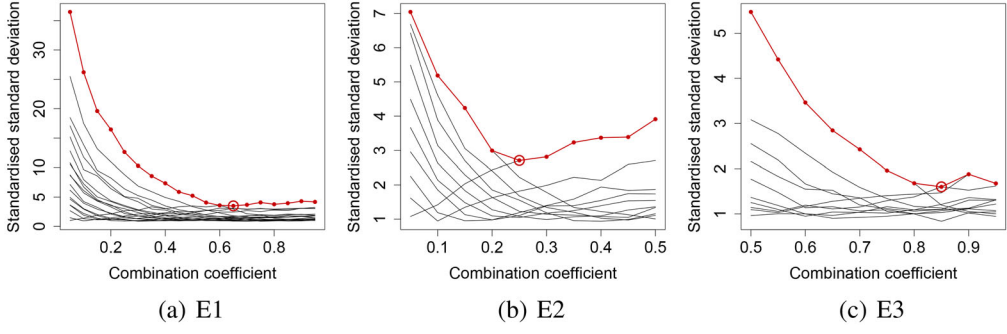
Figure 2. Results of the simulations in Series E: curves of SSD against $\check{\gamma}$.

Table 2. Settings of f_{inf} and $\check{\gamma}$ in the simulations in Series R.

	Group R1	Group R2	Group R3
γ_i's	$[0.02, 0.98]$	$[0.02, 0.53]$	$[0.47, 0.98]$
$\check{\gamma}$	$\{0.05, 0.1, ..., 0.9, 0.95\}$	$\{0.05, 0.1, ..., 0.45, 0.5\}$	$\{0.5, 0.55, ..., 0.9, 0.95\}$
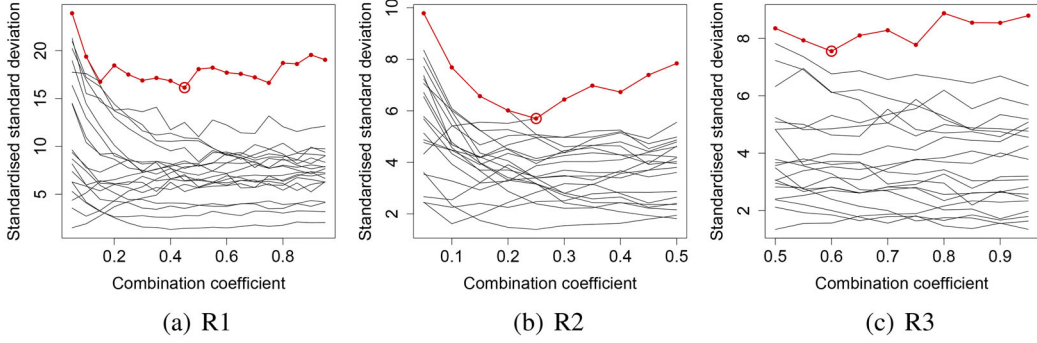
Figure 3. Results of the simulations in Series R: curves of SSD against $\check{\gamma}$.

5.2. Comparisons

In this subsection, we show some comparisons for the GLS against other popular methods, including the SIR, MH and the stratified sampling. In this simulation, the region $\mathscr{R} = [0, 1]^2$ is divided into 16 equal-sized sub-squares (which is similar to the previous subsection) within each of which the density of the population is uniform, while the densities of the diagnosed cases and infections are proportional to a normal distribution. In each of the 16 sub-squares, the population is fixed, but the numbers of diagnosed cases and infections are generated randomly. Please refer to Supplementary Material S4.2 for more details about the settings. As for the sampling methods, we use the same sampling strategy in Algorithm 2, except for the sampler in Step 2. In our proposed method, the size of the uniform design used in the GLS $M = 210$. The proposal distributions in the SIR and MH samplers are set as the uniform distribution on $\mathscr{R} = [0, 1]^2$, and the normal distribution with covariance matrix equal to $1/16 \cdot \mathbf{I}_2$, respectively. In the three methods GLS, SIR and MH, the number of sampling positions $r = 16$, and the value of $\check{\gamma}$ is set 0 so that $\varphi \propto f_{\text{diag}}$. In the stratified sampling, the strata are the 16 sub-squares and the Neyman allocation is used to determine the sample sizes in each stratum. The total population $N_{\text{pop}} = 8000 \times 10^4$ and the total sample size $n = 1 \times 10^4$.

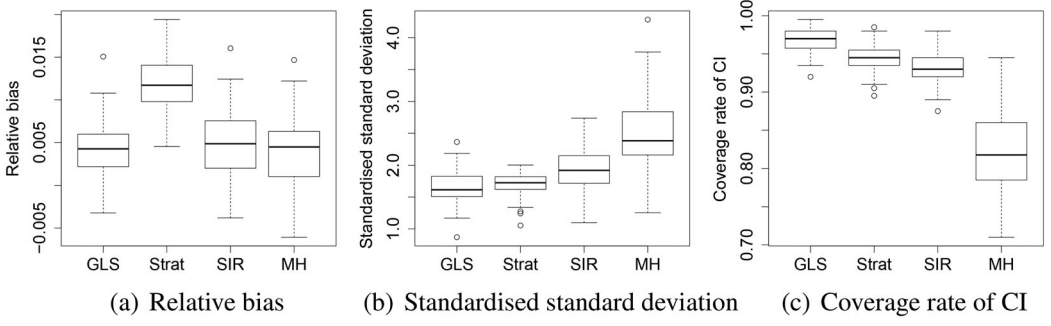We use the following three criteria to compare their performances:

Figure 4. Comparisons of different sampling methods.

i.   Relative bias: the ratio of the bias of $\hat{N}_{\text{inf}}$ to the true value of N_{inf};
ii. Standardized standard deviation: the ratio of the standard deviation of the estimator of N_{inf}, obtained by a particular method, to that obtained by our method with the exact optimal settings, which is similar to that in Sec. 5.1;
iii. Coverage rate of CI: the coverage rate of the approximate 95% CI.

We use 100 different random settings of f_{inf} and f_{diag}. Under each of the 100 settings, for each method, the values of these three criteria are calculated through 200 independent simulations. The results are presented in Figure 4. The unbiasedness of the estimators of N_{inf} obtained by these four methods is verified through Figure 4a. From Figure 4b, we can find that our method has the smallest SSD, i.e. the minimum variance of the estimator of N_{inf}. The main reasons are that the performance of the SIR sampler depends heavily on the quality of the initial samples from the proposal distribution, that the MH sampler is easy to get stuck at some peak of the density function, and that the stratified sampling method does not utilize the information about the population and the diagnosed cases sufficiently. From Figure 4c, we can also find that our method has the highest coverage rate of the approximate 95% CI of N_{inf}. Therefore, we can conclude that our method is efficient, in the sense that its estimator of N_{inf} is unbiased, with a small variance and a high coverage rate of the CI.

6. A Practical example

To further illustrate the two-stage sampling strategy, an example based on the situation of COVID-19 in the USA is presented in this section. The 'USA' mentioned in this section refers to the 50 states of the USA, as well as the District of Columbia, excluding other territories of the USA. The cumulative numbers of diagnosed cases of the 51 administrative districts in the USA can be obtained from the Centers for Disease Control and Prevention (2020). Assume that the true densities of population f_{pop}, diagnosed cases f_{diag} and infections f_{inf} are the densities of population, diagnosed cases up to December 27th, 2020, and diagnosed cases up to April 22nd, 2021 in the USA, respectively. The corresponding densities are depicted in Figure 5a–c, with the unit being km^{-2}, and the specific data are shown in Supplementary Material S5. The total population and the numbers of diagnosed cases and infections are $N_{\text{pop}} = 331.319 \times 10^6$, $N_{\text{diag}} = 29.701 \times 10^6$ and $N_{\text{inf}} = 31.467 \times 10^6$, respectively. The total sample size n is set as 10000, which is quite small compared to the total population of the USA. For comparison, we use both our method and the classical stratified sampling to estimate the total number of infections N_{inf}. Based on the settings of N_{pop} and N_{diag}, it is reasonable to assume that the values of the true γ are in the range $[0, 0.1]$, and we set $\check{\gamma} = 0.05$ to obtain the initial rough estimation of f_{inf}. In our method, we choose $r = 250$ and $M = 210$, as recommended in the previous sections, and the settings of the
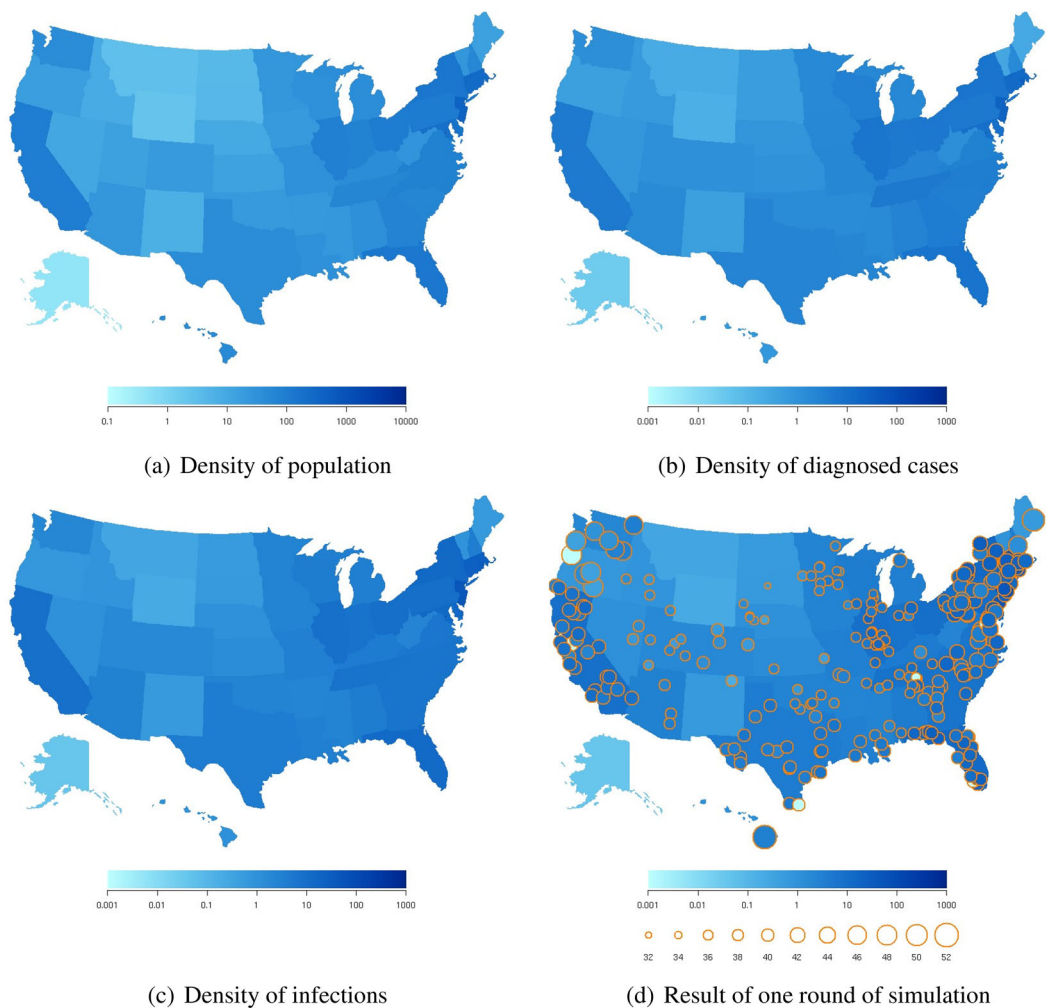
Figure 5. (a)–(c): The densities of population, diagnosed cases and infections; (d): the result of one round of simulation.

Table 3. Results of the comparison between our method and the stratified sampling.

Method	Sample mean of $\hat{N}_{inf}$	Sample standard deviation of $\hat{N}_{inf}$	Coverage rate of CI
Our method	31.612×10^6	0.995×10^6	99.5%
Stratified sampling	31.691×10^6	1.000×10^6	95.0%

sampling density φ and the allocation method of the sample sizes are nearly optimal. As for the stratified sampling, the strata are defined as the 51 administrative districts, which is consistent with the area partition when recoding the epidemic-related data; the Neyman allocation (Lohr 2019) was used. The details about the stratified sampling are described in Supplementary Material S5.

The results of 200 rounds of independent simulations are shown in Table 3. Though the standard deviation of the estimator obtained by our proposed method is slightly smaller than that of the stratified sampling, our method has higher coverage rate of CI. Note that in practice, the samples must be drawn from several selected sampling positions in each stratum. Due to this, the stratified multi-stage sampling is often used (Pollán et al. 2020), which can further increase the variance of the estimator. For the stratified multi-stage sampling method, big efforts have to be

made in order to adapt to a complicated distribution of population, and the variance may become difficult to estimate. However, our method can automatically utilize the complicated information about the distributions of population and diagnosed cases, while also maintaining the simplicity in estimating N_{inf} and its variance.

Further, we show the result of one of the 200 rounds of simulations to explain our method. In this round of simulation, the estimated number of infections $\hat{N}_{\text{inf}} = 29.70 \times 10^6$, with the estimated standard deviation $\sqrt{\hat{v}(\hat{N}_{\text{inf}})} = 1.36 \times 10^6$. The approximate 95% CI does cover the true N_{inf}. The sampling positions are presented in Figure 5d, where the background is the true f_{inf}, and the circles represent the sampling positions, with their colors showing the estimated values of f_{inf} in km^{-2}, and their diameters showing the sample sizes there. The final total sample size in this round is 9999, smaller by 1 than the expected $n = 10000$, due to the round-off error at each sampling position. The sample sizes at these sampling positions are nearly the same, which is determined by the characters of f_{pop} and f_{diag}, as well as $\check{\gamma}$.

7. Conclusions

In this paper, we propose a novel, two-stage sampling strategy to estimate the number of infections of a pandemic. Our method can sufficiently utilize the information about the distributions of both the population and the diagnosed cases, which can gather information more flexibly than the existing methods, and hence, more efficient. Moreover, our two-stage sampling strategy does not involve any discrete structures like strata or clusters, therefore, it can easily and automatically adapt to the complicated distributions of population and diagnosed cases, and the corresponding estimating method keeps simple. The GLS algorithm used in our method is also easy to implement, since it does not need a proposal distribution. Its performance is robust against the complexity and multimodality of the sampling density, which overcomes the drawbacks of the other popular samplers for general probability densities, such as the SIR or MH algorithm.

Since the true density of the infections is not known in practice, obtaining the exact optimal settings of the sampling density and the allocation function in this sampling strategy is unrealistic. Instead, we discuss the nearly optimal ones for practical implementation, which is based on an initial rough estimate of the density of infections. The total sample size can be determined by the pre-defined estimation precision. The small sample sizes at the sampling positions can be avoided by modifying the allocation of sample sizes based on a convex combination. In addition, in the second stage of our method, we just consider the simple random sampling method for simplicity. To further improve the efficiency and eliminate the selection bias, other sampling methods like the stratified sampling can be taken into account, with some slight modifications for the corresponding formulae in the second stage. Further, by the numerical simulations, we discuss the robust setting of the combination coefficient for the rough estimator of the density of infections in the minimax sense. We also compare the GLS algorithm with the SIR, MH and stratified sampling methods in terms of the relative bias, standard deviation and coverage rate of CI. It shows that the GLS algorithm has the smallest standard deviation and the highest coverage rate of the approximate 95% CI. Moreover, we apply our method to the investigation of COVID-19 in the USA. Our method shows good performance and has higher coverage rate of CI compared with the stratified sampling. Hence, these simulations and the practical example verify the efficiency of our proposed two-stage sampling method, whatever the sampling density is.

Acknowledgements

Funding

References

Anand, S., M. Montez-Rath, J. Han, J. Bozeman, R. Kerschmann, P. Beyer, J. Parsonnet, and G. M. Chertow. 2020. 'Prevalence of SARS-CoV-2 antibodies in a large nationwide sample of patients on dialysis in the USA: A cross-sectional study. *Lancet (London, England)* 396 (10259):1335–44. doi:10.1016/S0140-6736(20)32009-2.

Centers for Disease Control and Prevention. 2020. 'COVID data tracker', https://covid.cdc.gov/covid-data-tracker/#cases_totalcases.

Cochran, W. G. 1977. *Sampling techniques*. 3rd ed. New York: John Wiley & Sons.

Fang, K.-T., M.-Q. Liu, H. Qin, and Y.-D. Zhou. 2018. *Theory and application of uniform experimental designs*. 1st ed. Singapore and Beijing: Springer and Science Press.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57 (1):97–109. doi:10.1093/biomet/57.1.97.

Havers, F. P., C. Reed, T. Lim, J. M. Montgomery, J. D. Klena, A. J. Hall, A. M. Fry, D. L. Cannon, C.-F. Chiang, A. Gibbons, et al. 2020. Seroprevalence of antibodies to SARS-CoV-2 in six sites in the United States, March 23–May 3, 2020. *JAMA Internal Medicine* 180 (12):1576–86. doi:10.1001/jamainternmed.2020.4130.

Horton-French, K., E. Dunlop, R. M. Lucas, G. Pereira, and L. J. Black. 2021. Prevalence and predictors of vitamin D deficiency in a nationally representative sample of Australian adolescents and young adults. *European Journal of Clinical Nutrition* 75 (11):1627–36. doi:10.1038/s41430-021-00880-y.

Jia, L., Y. Du, L. Chu, Z. Zhang, F. Li, D. Lyu, Y. Li, Y. Li, M. Zhu, H. Jiao, et al. 2020. Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China: A cross-sectional study. *The Lancet. Public Health* 5 (12):e661–e671. doi:10.1016/S2468-2667(20)30185-7.

Kissler, S. M., N. Kishore, M. Prabhu, D. Goffman, Y. Beilin, R. Landau, C. Gyamfi-Bannerman, B. T. Bateman, J. Snyder, A. S. Razavi, et al. 2020. Reductions in commuting mobility correlate with geographic differences in SARS-CoV-2 prevalence in New York City. *Nature Communications* 11 (1):4674. doi:10.1038/s41467-020-18271-5.

Knudsen, A. K. S., K. Stene-Larsen, K. Gustavson, M. Hotopf, R. C. Kessler, S. Krokstad, J. C. Skogen, S. Øverland, and A. Reneflot. 2021. 'Prevalence of mental disorders, suicidal ideation and suicides in the general population before and during the COVID-19 pandemic in Norway: A population-based repeated cross-sectional analysis. *The Lancet Regional Health. Europe* 4 (100071):100071. doi:10.1016/j.lanepe.2021.100071.

Lemieux, C. 2009. *Monte Carlo and Quasi-Monte Carlo sampling*. 1st ed. New York: Springer.

Leong, P.-Y., J.-Y. Huang, J.-Y. Chiou, Y.-C. Bai, and J. C.-C. Wei. 2021. The prevalence and incidence of systemic lupus erythematosus in Taiwan: A nationwide population-based study. *Scientific Reports* 11 (1):5631. doi:10.1038/s41598-021-84957-5.

Li, Y., Z. Shan, and W. Teng. 2021. Estimated change in prevalence of abnormal thyroid-stimulating hormone levels in China according to the application of the kit-recommended or NACB standard reference interval. *EClinicalMedicine* 32 (100723):100723. doi:10.1016/j.eclinm.2021.100723.

Lohr, S. L. 2019. *Sampling: Design and analysis*. 2nd ed. Boca Raton: CRC Press.

Mulenga, L. B., J. Z. Hines, S. Fwoloshi, L. Chirwa, M. Siwingwa, S. Yingst, A. Wolkon, D. T. Barradas, J. Favaloro, J. E. Zulu, et al. 2021. Prevalence of SARS-CoV-2 in six districts in Zambia in July, 2020: A cross-sectional cluster sample survey. *The Lancet. Global Health* 9 (6):e773–e781. doi:10.1016/S2214-109X(21)00053-X.

Nagashima, S., K. Ko, C. Yamamoto, E. Bunthen, S. Ouoba, C. Chuon, M. Ohisa, A. Sugiyama, T. Akita, M. S. Hossain, et al. 2021. Prevalence of total hepatitis A antibody among 5 to 7 years old children and their mothers in Cambodia. *Scientific Reports* 11 (1):4778. doi:10.1038/s41598-021-83710-2.

Ning, J., and H. Tao. 2020. Randomized quasi-random sampling/importance resampling. *Communications in Statistics - Simulation and Computation* 49 (12):3367–79. doi:10.1080/03610918.2018.1547398.

Parenteau, C. S., E. C. Lau, I. C. Campbell, and A. Courtney. 2021. Prevalence of spine degeneration diagnosis by type, age, gender, and obesity using Medicare data. *Scientific Reports* 11 (1):5389. doi:10.1038/s41598-021-84724-6.

Pérez, C. J., J. Martín, M. J. Rufo, and C. Rojano. 2005. Quasi-random sampling importance resampling. *Communications in Statistics - Simulation and Computation* 34 (1):97–112. doi:10.1081/SAC-200047112.

Pollán, M., B. Pérez-Gómez, R. Pastor-Barriuso, J. Oteo, M. A. Hernán, et al. 2020. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): A nationwide, population-based seroepidemiological study. *The Lancet* 397 (10250):535–44.

Rosenberg, E. S., J. M. Tesoriero, E. M. Rosenthal, R. Chung, M. A. Barranco, L. M. Styer, M. M. Parker, S.-Y. John Leung, J. E. Morne, D. Greene, et al. 2020. Cumulative incidence and diagnosis of SARS-CoV-2 infection in New York. *Annals of Epidemiology* 48:23–9.e4. doi:10.1016/j.annepidem.2020.06.004.

Rubin, D. B. 1987. A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association* 82 (398):543–6. doi:10.2307/2289460.

Sakurai, A., T. Sasaki, S. Kato, M. Hayashi, S.-I. Tsuzuki, T. Ishihara, M. Iwata, Z. Morise, and Y. Doi. 2020. Natural history of asymptomatic SARS-CoV-2 infection. *The New England Journal of Medicine* 383 (9):885–6. doi:10.1056/NEJMc2013020.

Sartorius, B., J. Cano, H. Simpson, L. S. Tusting, L. B. Marczak, M. K. Miller-Petrie, B. Kinvi, H. Zoure, P. Mwinzi, S. I. Hay, et al. 2021. Prevalence and intensity of soil-transmitted helminth infections of children in sub-Saharan Africa, 2000-18: A geospatial analysis. *The Lancet. Global Health* 9 (1):e52–e60. doi:10.1016/S2214-109X(20)30398-3.

Sood, N., P. Simon, P. Ebner, D. Eichner, J. Reynolds, E. Bendavid, and J. Bhattacharya. 2020. Seroprevalence of SARS-CoV-2specific antibodies among adults in Los Angeles County, California, on April 10–11, 2020. *Jama* 323 (23):2425–7. doi:10.1001/jama.2020.8279.

Ssentongo, P., A. E. Ssentongo, D. M. Ba, J. E. Ericson, M. Na, X. Gao, C. Fronterre, V. M. Chinchilli, and S. J. Schiff. 2021. Global, regional and national epidemiology and prevalence of child stunting, wasting and underweight in low- and middle-income countries, 2006. *Scientific Reports* 11 (1):5204. doi:10.1038/s41598-021-84302-w.

Stringhini, S., A. Wisniak, G. Piumatti, A. S. Azman, S. A. Lauer, H. Baysson, D. De Ridder, D. Petrovic, S. Schrempft, K. Marcus, et al. 2020. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): A population-based study. *Lancet (London, England)* 396 (10247):313–9. doi:10.1016/S0140-6736(20)31304-0.

Tian, H., Y. Hu, Q. Li, L. Lei, Z. Liu, M. Liu, C. Guo, F. Liu, Y. Liu, Y. Pan, et al. 2021. Estimating cancer survival and prevalence with the Medical-Insurance-System-based Cancer Surveillance System (MIS-CASS): An empirical study in China. *EClinicalMedicine* 33 (100756):100756. doi:10.1016/j.eclinm.2021.100756.

Wang, Y. J., J.-H. Ning, Y.-D. Zhou, and K.-T. Fang. 2015. A new sampler: Randomized likelihood sampling, *in* 'Souvenir Booklet of the 24th International Workshop on Matrices and Statistics', 255–61.

Ward, H., C. Atchison, M. Whitaker, K. E. C. Ainslie, J. Elliott, L. Okell, R. Redd, D. Ashby, C. A. Donnelly, W. Barclay, et al. 2021. SARS-CoV-2 antibody prevalence in England following the first peak of the pandemic. *Nature Communications* 12 (1):905. doi:10.1038/s41467-021-21237-w.

Xu, X., J. Sun, S. Nie, H. Li, Y. Kong, M. Liang, J. Hou, X. Huang, D. Li, T. Ma, et al. 2020. Seroprevalence of immunoglobulin M and G antibodies against SARS-CoV-2 in China. *Nature Medicine* 26 (8):1193–5. doi:10.1038/s41591-020-0949-6.

Yi, S.-Y., Z. Liu, M.-Q. Liu, and Y.-D. Zhou. 2022. Global likelihood sampler for multimodal distributions. *Journal of Computational and Graphical Statistics*, online. doi:10.1080/10618600.2023.2165499.

Zhou, Y.-D., P. He, J.-H. Ning, and K.-T. Fang. 2021. *Methods and Applications of Random Simulations.* Beijing: Higher Education Press (in Chinese).