Statistics Surveys
Vol. 14 (2020) 1–31
ISSN: 1935-7516
https://doi.org/10.1214/19-SS127

Estimating the size of a hidden finite set: Large-sample behavior of estimators

Si Cheng

Department of Biostatistics, University of Washington
e-mail: chengsi@uw.edu

Daniel J. Eck

Department of Statistics, University of Illinois Urbana-Champaign
e-mail: dje13@illinois.edu

and

Forrest W. Crawford

Department of Biostatistics, Yale School of Public Health
Department of Statistics & Data Science, Yale University
Department of Ecology & Evolutionary Biology, Yale University
Yale School of Management
e-mail: forrest.crawford@yale.edu

Abstract: A finite set is "hidden" if its elements are not directly enumerable or if its size cannot be ascertained via a deterministic query. In public health, epidemiology, demography, ecology and intelligence analysis, researchers have developed a wide variety of indirect statistical approaches, under different models for sampling and observation, for estimating the size of a hidden set. Some methods make use of random sampling with known or estimable sampling probabilities, and others make structural assumptions about relationships (e.g. ordering or network information) between the elements that comprise the hidden set. In this review, we describe models and methods for learning about the size of a hidden finite set, with special attention to asymptotic properties of estimators. We study the properties of these methods under two asymptotic regimes, "infill" in which the number of fixed-size samples increases, but the population size remains constant, and "outfill" in which the sample size and population size grow together. Statistical properties under these two regimes can be dramatically different.

MSC 2010 subject classifications: 62D05, 62F12, 62P25.
Keywords and phrases: Capture-recapture, German tank problem, multiplier method, network scale-up method.

Received February 2019.

Contents

1 Introduction 2
2 Setting and notation 4
2.1 Hidden sets 4

2.2 Asymptotic regimes 5
2.3 Statistical properties of estimators 6
3 Ordered sets: The German tank problem 7
4 Bernoulli trials 9
4.1 Binomial N parameter 9
4.2 Waiting times 11
4.3 The network scale-up method 12
4.3.1 Sampling from the general population 13
4.3.2 Sampling from the hidden population 14
4.4 Estimating a total with unequal sampling probabilities 15
5 Other unordered sets 16
5.1 Capture-recapture experiments 16
5.1.1 Two-sample estimation 16
5.1.2 k-sample estimation 18
6 Discussion 20
Acknowledgements 22
Supplementary Material 22
References 22

1. Introduction

Estimating the size of a hidden finite set is an important problem in a variety of scientific fields. Often practical constraints limit researchers' access to elements of the hidden set, and direct enumeration of elements may be impractical or impossible. In demographic, public health, and epidemiological research, researchers often seek to estimate the number of people within a given geographic region who are members of a stigmatized, criminalized, or otherwise hidden group [3, 53, 108, 1]. For example, researchers have developed methods for estimating the number of homeless people [76, 35], human trafficking victims [105, 111], sex workers [62, 74, 61, 70, 114], men who have sex with men [39, 88, 74, 115, 116, 92, 61, 97, 93], transgender people [85, 97], drug users [69, 55, 64, 54, 100, 87, 97, 61, 52], and people affected by disease [121, 120, 56, 94, 72, 15]. In ecology, the number of animals of a certain type within a geographic region is often of interest [104, 25, 51, 71]. Effective wildlife protection, ecosystem preservation, and pest control require knowledge about the size of free-ranging animal populations [102, 46, 60]. In intelligence analysis, military science, disaster response, and criminal justice applications, estimates of the size of hidden sets can give insight into the size of a threat or guide policy responses. Analysts may seek information about the number of combatants in a conflict, military vehicles [96, 49], extremists [36], terrorist plots [67, 68], war casualties [98], people affected by a disaster [6], and the extent of counterfeiting [118].

Statistical approaches to estimating the size of a hidden set fall into a few general categories. Some approaches are based on traditional notions of random sampling from a finite population [57, 8]. Others leverage information about the ordering of units [96, 49], or relational information about "network" links between units [76, 126, 7, 84, 40, 100]. Single- or multi-step sampling procedures that involve record collection or "marking" of sampled units – called capture-recapture experiments – are common when random sampling is possible [19, 34, 42, 104, 91, 55]. Sometimes exogenous, or population-level data can help: when the proportion of units in the hidden set with a particular attribute is known *a priori*, then the proportion with that attribute in a random sample can be used to estimate the total size of the set [124, 125, 78, 54, 92, 99]. Still other methods use features of a dynamic process, such as the arrival times of events in a queueing process, to estimate the number of units in a hidden set [67, 68].

Alongside these practical approaches, corresponding theoretical results provide justification for particular study designs and estimators, based on large-sample (asymptotic) arguments. Guidance for prospective study planning often depends on asymptotic approximation. For example, sample size calculation may be based on asymptotic approximation if the finite-sample distribution of an estimator is not identified or hard to analyze [24, 33, 81]. In retrospective analysis of data and the comparison of statistical approaches, researchers may choose estimators based on large-sample properties like asymptotic unbiasedness, efficiency and consistency if closed-form expressions for finite-sample biases and variances are hard to derive [119, 38]. Claims about the large-sample performance of estimators depend on specification of a suitable asymptotic regime, and it is well known that estimators can perform differently under different asymptotic regimes. Asymptotic theory in spatial statistics provides some perspective on what it means to obtain more data from the same source: informally, an "infill" asymptotic regime assumes a bounded spatial domain, with the distance between data points within this domain going to zero. An "increasing domain" or "outfill" asymptotic regime assumes that the minimum distance between any pair of points is bounded away from zero, while the size of the domain increases as the sample size increases. The latter is usually the default asymptotic setting considered by researchers studying the properties of spatial smoothing estimators [79, 82, 31]. However, under infill asymptotics, these desirable asymptotic properties of smoothing estimators often do not hold: even when consistency is guaranteed, the rate of convergence may be different [30, 106, 79, 123, 22]. When the size of the population from which the sample is drawn is the estimand of interest, intuition about large-sample properties of estimators can break down, but a similar asymptotic perspective is useful in studying the properties of estimators for the size of a hidden set: an infill asymptotic regime takes the total population size to be fixed, while the number of samples from this population increases; the outfill regime permits the sample size and population size to grow to infinity together.

In this paper, we review models and methods for estimating the size of a hidden finite set in a variety of practical settings. First we present a unified characterization of set size estimation problems, formalizing notions of size, sampling, relational structures, and observation. We then introduce the non-asymptotic regime in which sample size tends to the population size, and define the "infill" and "outfill" asymptotic regimes in which the sample size and population size may increase. We investigate a range of problems, query models, and estimators, including the German tank problem, failure time models, the network scale-up estimator, the Horvitz-Thompson estimator, the multiplier method, and capture-recapture methods. We characterize consistency and rates of estimation errors for these estimators under different asymptotic regimes. We conclude with discussion of the role of substantive and theoretical considerations in guiding claims about statistical performance of estimators for the size of a hidden set.

2. Setting and notation

2.1. Hidden sets

Let U be a set consisting of all elements from a specified target population. In general, U can be discrete or continuous. Let $\mu(\cdot)$ be a measure defined on U such that $\mu(U) < \infty$. The *size* of U is $\mu(U)$. We call U a *hidden* set if the members of U are not directly enumerable, or if its size $\mu(U)$ cannot be ascertained from a deterministic query. When U is a finite set of discrete elements, $\mu(U) = |U| := N$ is the cardinality of U.

We seek to learn about the size of U by sampling its elements. Define a probability space $(U, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is a σ-field, and $\mathbb{P}$ is a probability measure on $(U, \mathcal{F})$. The measure $\mathbb{P}$ represents a probabilistic query mechanism by which we may draw subsets of the elements of U. For each possible sample $s \in \mathcal{F}$, defining $\mathbb{P}(s)$ gives a notion of *random sampling.* Sequential sampling designs can be specified by defining the sequential sampling probabilities $\mathbb{P}(S_i = s_i | s_1, \ldots, s_{i-1})$. Sequential samples are denoted as $\boldsymbol{s} = (s_1, \ldots, s_k)$, and the sample size is defined as $|s_1| + \cdots + |s_k|$, the sum of the cardinality of each sample, which can be larger than $\mu(U)$ under with-replacement sampling. An estimator $\delta(\boldsymbol{s})$ of $\mu(U) = N$ is a functional of $\mathcal{F}$ onto $\mathbb{R}^+$ or $\mathbb{N}$.

Elements of the hidden set U, or of a sample s from U, may have attributes, labels, or relational structures that permit estimation of $\mu(U)$ from a subset. An element $i \in U$ may be labeled or have attributes X_i, which may be continuous, discrete, unordered, or ordered. The elements of U may be connected via a relational structure, such as a graph $G = (U, E)$, where the vertex set is U, and edges $\{i, j\} \in E$ represent relationships between elements. Alternatively, the sampling mechanism may impose a structure on the elements of a sample: if $s_1 \subseteq U$ and $s_2 \subseteq U$ are samples from U, then the intersection $M = s_1 \cap s_2$ is the set of elements in both samples. An *observation* on the sample $\boldsymbol{s}$ consists of statistics that reflect these attributes, labels or structures of the units in $\boldsymbol{s}$, such as the value of attributes $\{X_i\}$, network degrees in a graph or size of the intersection of samples $|M|$.

An example serves to make this setting and notation more concrete. Consider the problem of estimating the number of injection drug users in a city [e.g. 69, 54, 100]. This is an important task in public health research and drug use epidemiology because injection drug use may contribute to transmission of infectious diseases such as hepatitis C virus (HCV) and human immunodeficiency virus (HIV). Policymakers considering educational and intervention programs to mitigate the harms of injection drug use require accurate estimates of the size of the target population. In this context, U is the set of injection drug users in the city, and we wish to estimate the size of this set, $\mu(U) = |U| = N$. The probability space is $(U, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is a σ-field consisting of subsets of U, and $\mathbb{P}$ is a probabilistic query distribution assigning probabilities to each set in $\mathcal{F}$. For example, if $s \in \mathcal{F}$ is a subset of U, then $\mathbb{P}(s)$ represents a mechanism for randomly sampling a subset of $|s|$ members of U. An individual injection drug user $i \in U$ may have an attribute X_i representing, for example, the number of times i has experienced an overdose and been taken to the local hospital. In addition, relational information may be available in the form of a graph or network $G = (U, E)$, where E is the set of pairs $\{i, j\}$ that are "connected" via syringe sharing or social relationships.

2.2. Asymptotic regimes

We now formalize asymptotic regimes relevant for hidden set size estimation.

Definition 1 (Asymptotic regime)**.** Let $(U_t, \mathcal{F}_t, \mathbb{P}_t)$ be a probability space defined for each $t = 1, 2, \ldots$, and let $\boldsymbol{s}_t = \{s_1^{(t)}, \ldots, s_{k_t}^{(t)}\}$ be the set of k_t samples from U, with $|\boldsymbol{s}_t| = \sum_{i=1}^{k_t} |s_i|$. An asymptotic regime is a sequence $\{\boldsymbol{s}_t, U_t, \mathbb{P}_t\}_{t=1}^{\infty}$ such that the limits $\lim_{t\to\infty} |\boldsymbol{s}_t|$ and $\lim_{t\to\infty} \mu(U_t)$ exist (infinity included).

We first define the trivial finite-population regime, in which the sampled set approaches the fixed population U.

Definition 2 (Finite-population regime)**.** Let U be a hidden discrete set of fixed size. The finite-population (non-asymptotic) regime is $U_t = U$ for all t and $\boldsymbol{s}_t = U$ for all $t > t_0$, where $t_0 < \infty$ is a positive integer.

Next, we define the "infill" asymptotic regime that arises when sampling repeatedly (with replacement between different samples) from a set of fixed finite size. This regime is an example of a superpopulation model [58, 14] which reproduces the original population $U_t = U$ for each t.

Definition 3 (Infill asymptotic regime)**.** Let $(U_t = U, \mathcal{F}_t = \mathcal{F}, \mathbb{P}_t)$ be a sequence of probability spaces, where $\mathbb{P}_t$ assigns probability $\mathbb{P}(s_i^{(t)} | s_1^{(t)}, \ldots, s_{i-1}^{(t)})$ to sequential samples $s_1^{(t)}, \ldots, s_{k_t}^{(t)} \in \mathcal{F}$ for any t. The infill asymptotic regime is a sequence $\{\boldsymbol{s}_t, U_t = U, \mathbb{P}_t\}_{t=1}^{\infty}$, where $|s_j^{(t)}|$ (any $j \in [k_t]$) and $\mu(U_t)$ are both fixed and bounded, and the number of samples $k_t \to \infty$ as $t \to \infty$.

Sometimes it can be difficult to conceptualize sampling infinitely many times from U, or the sampling design may be subject to practical constraints, so that sampling only a single or fixed number of samples, or a fixed proportion of the total population, is allowed. It is therefore also reasonable to study the performance of estimators under an asymptotic regime in which a *single* sample is obtained from the hidden set, where the size of the sample and hidden set may tend to infinity together.

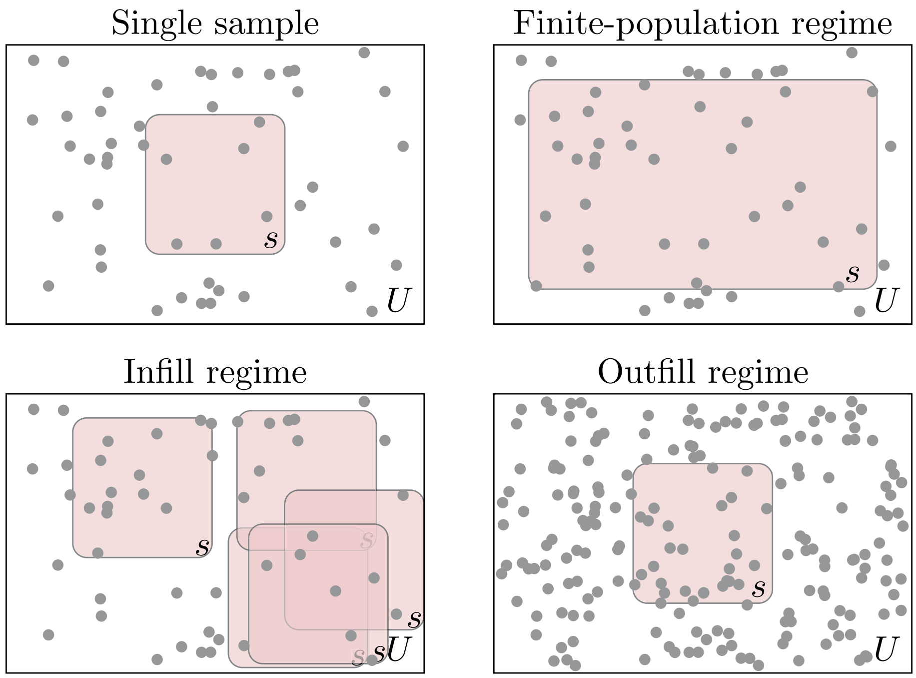

Fig 1. *Illustration of different regimes for discrete sets. Units are indicated by circles. The sample s "expands" to U under the finite-population regime. Infinitely repeated samples of a fixed size are drawn from a fixed population under infill asymptotics. Under outfill, s and U grow simultaneously with s approaching a fixed proportion of U.*

Definition 4 (Outfill asymptotic regime)**.** Let $(U_t, \mathcal{F}_t, \mathbb{P}_t)$ be a sequence of probability spaces, where $\mathbb{P}_t$ assigns probability $\mathbb{P}(s_i^{(t)}|s_1^{(t)}, \ldots, s_{i-1}^{(t)})$ to $s_1^{(t)}, \ldots, s_{k_t}^{(t)} \in \mathcal{F}_t$ for any t. The outfill asymptotic regime is a sequence $\{\boldsymbol{s}_t, U_t, \mathbb{P}_t\}$ such that $\mu(U_t) \to \infty$ and $n_i^{(t)} := |s_i^{(t)}| \to \infty$ with $n_i^{(t)}/\mu(U_t) \to c_i \in [0, \infty)$ for each $i \in [k_t]$ as $t \to \infty$, where $\lim_{t\to\infty} k_t$ may be finite or infinite.

The ratio c_i can be greater than one when sampling is with replacement. The sample sizes mentioned above can be deterministic or random. In the latter case, all regimes can be defined in a similar way, e.g. $\mathbb{E}|\boldsymbol{s}_t|/\mu(U_t) \to c_i$. We are primarily interested in the outfill asymptotic regime with $k_t = 1$ for all t. The binomial model as well as the multiplier and capture-recapture methods, described below, are special cases where k_t may be greater than one. Figure 1 illustrates different regimes in general discrete settings.

2.3. Statistical properties of estimators

Let $\delta(\boldsymbol{s}_t)$ be an estimator of $\mu(U_t)$, defined for each t. We are interested in the statistical properties of $\delta(\boldsymbol{s}_t)$ under the asymptotic regimes described above. An estimator is called *unbiased* if $\mathbb{E}_t[\delta(\boldsymbol{s}_t)] = \mu(U_t)$ for all t, where $\mathbb{E}_t(\cdot)$ denotes expectation with respect to $\mathbb{P}_t$. Under an asymptotic regime $\{\boldsymbol{s}_t, U_t, \mathbb{P}_t\}_{t=1}^{\infty}$, an estimator $\delta(\boldsymbol{s}_t)$ is *asymptotically unbiased* if $\lim_{t\to\infty} \mathbb{E}_t[\delta(\boldsymbol{s}_t)] - \mu(U_t) = 0$. There may be some slightly biased estimators whose variance is smaller than that of every unbiased estimator. A common way to balance the trade-off between the bias and variance is to evaluate the *mean squared error* (MSE), defined as $MSE[\mu(U_t), \delta(\boldsymbol{s}_t)] = \mathbb{E}\left[(\delta(\boldsymbol{s}_t) - \mu(U_t))^2\right] = (\mathbb{E}[\delta(\boldsymbol{s}_t)] - \mu(U_t))^2 + \text{Var}[\delta(\boldsymbol{s}_t)]$. The asymptotic MSE under a given regime is defined as $\lim_{t\to\infty} MSE(\mu(U_t), \delta(\boldsymbol{s}_t))$.

An estimator $\delta(\boldsymbol{s}_t)$ that satisfies $\lim_{t\to\infty} \mathbb{P}_t(|\delta(\boldsymbol{s}_t) - \mu(U_t)| > \varepsilon) = 0$ for any $\varepsilon > 0$ under a particular asymptotic regime $\{\boldsymbol{s}_t, U_t, \mathbb{P}_t\}$ is called *consistent* for $\mu(U_t)$. An estimator $\delta(\boldsymbol{s}_t)$ is called *MSE consistent* for $\mu(U_t)$ under a certain asymptotic regime if $MSE[\delta(\boldsymbol{s}_t), \mu(U_t)] \to 0$ as $t \to \infty$ under that asymptotic setting. MSE consistency implies consistency. Under a particular asymptotic regime, we call a sequence of estimates $\delta(\boldsymbol{s}_t)$ *asymptotically normal* with mean ξ, variance σ^2/t^r and rate t^r if the cumulative distribution function (CDF) of $t^r\,(\delta(\boldsymbol{s}_t) - \xi)$ converges to the CDF of a $N(0, \sigma^2)$ random variable, denoted by $t^r\,(\delta(\boldsymbol{s}_t) - \xi) \xrightarrow{L} N(0, \sigma^2)$.

3. Ordered sets: The German tank problem

Suppose each unit in the hidden set $i \in U$ has a distinct label $X_i \in \mathbb{R}$, so that the labels give a natural ordering of the elements in U: we can define units $i < j$ if $X_i < X_j$. One common scenario for discrete U is that the X_i's are consecutive integers. Another common situation when U is equivalent to an interval in $\mathbb{R}$ is that $\cup_{i\in U} X_i$ equals that interval. An observation of samples from an ordered set U consists of sampled units s and their labels $\{x_i :\ i \in s\}$.

In 1943, the Economic Warfare Division of the American Embassy in London initiated a project to learn about the capacity of the German military using serial numbers found on German equipment [96, 50]. In a simple conceptualization of the problem, let $U = \{1, \dots, N\}$ and consider sampling $n = |s|$ units without replacement from U with probability $\mathbb{P}(s) = 1/\binom{N}{n}$. With k_t i.i.d. repeated samples, an estimator $\delta(\boldsymbol{s})$ for N is a functional of the observations, including the sample sizes and observed labels $X_{1,1}, \dots, X_{1,n}, \dots, X_{k_t,1}, \dots, X_{k_t,n}$. For example, to estimate the total number of participants in a marathon, if all N participants are numbered by the consecutive integers $1, \dots, N$, one could randomly record the first n numbers they saw in the race, and estimate the total based on the observed numbers.

For the kth sample $X_{k,1}, \dots, X_{k,n}$, we let $X_{k(n)}$ be the nth order statistic in the sample. With one sample, the maximum likelihood estimator (MLE) for N is $\widehat{N}_{MLE} = X_{(n)}$, which is negatively biased. Goodman [49] proposed an unbiased estimator

$$\widehat{N}_G = \frac{n+1}{n} X_{(n)} - 1, \tag{1}$$

which is a uniformly minimum-variance unbiased estimator (UMVUE), with $\text{Var}(\widehat{N}_G) = (N-n)(N+1)/n(n+2)$. An alternative estimator of N takes into

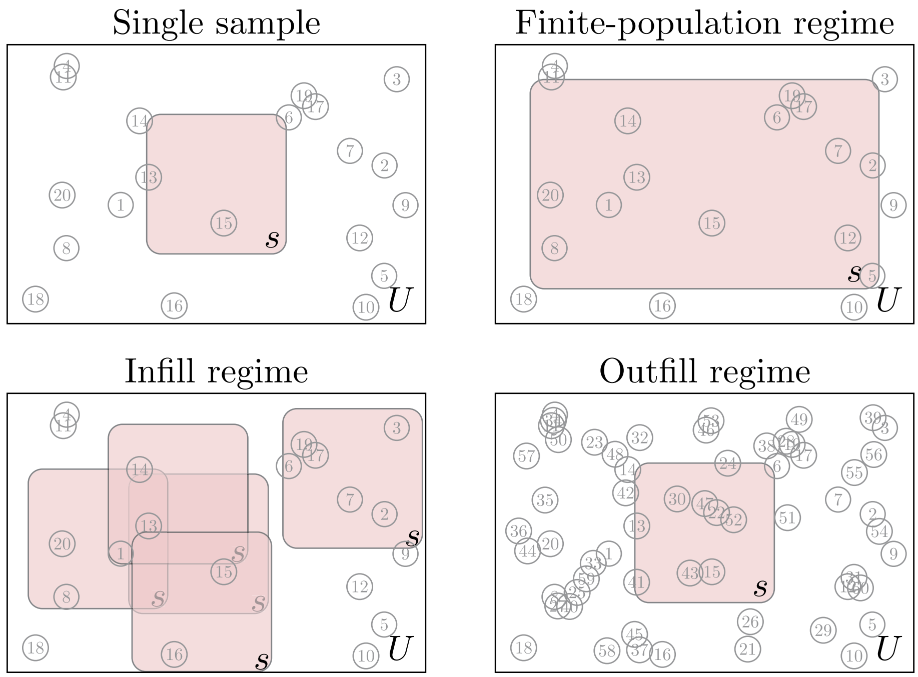

FIG 2. *Illustration of a single sample, and the finite-population, infill, and outfill regimes for the German tank problem. Units with their labels are represented by circles with numbers inside.*

account the gap between $X_{(n)}$ and N, and adjusts for the bias with the average gap between order statistics [49]. The estimator

$$\widehat{N}_2 = X_{(n)} + \frac{X_{(n)} - X_{(1)}}{n-1} - 1, \tag{2}$$

is also unbiased, with $\mathrm{Var}(\widehat{N}_2) = n(N-n)(N+1)/(n-1)(n+1)(n+2)$. The estimator N_2 can also be modified to estimate N when the labels do not start with 1. In particular,

$$\widehat{N}_3 = \frac{(n+1)\left(X_{(n)} - X_{(1)}\right)}{n-1} - 1$$

is the UMVUE of N when the initial label is unknown [49], with $\mathrm{Var}(\widehat{N}_3) = 2(N-n)(N+1)/(n-1)(n+2)$.

When there is more than one sample, we take the MLE as the maximizer of the joint sampling probability $\mathbb{P}_t(s_1, \ldots, s_{k_t})$, which is $\max_{i\in[k_t]} X_{i(n)}$, the largest observed value across all k_t samples. For estimators with closed forms like $\widehat{N}_G, \widehat{N}_2, \widehat{N}_3$, we derive k_t estimates $\delta(s_i^{(t)})$, $i = 1, \ldots, k_t$ based on each sample, and take their average as the estimator. In remaining sections, we average the estimators under infill by default, except for the models where infinite without-replacement sampling is feasible (e.g. Section 4.1). We consider the infill asymptotic regime where $n_t = n, N_t = N$ and $k_t \to \infty$, and the outfill regime where $n_t, N_t \to \infty, k_t = 1$ with $n_t/N_t \to c \in (0,1)$. Figure 2 illustrates different regimes for the German tank problem. We have the following asymptotic results (proof provided in the supplementary materials [23]):

Theorem 3.1. *Under the finite-population and infill regimes, $\widehat{N}_{MLE}, \widehat{N}_G, \widehat{N}_2$, $\widehat{N}_3$ are consistent. Under the outfill regime, all estimators above are asymptotically unbiased with asymptotic MSE $O(1)$ and inconsistent. Whether the initial label is known or not does not change the rate of MSE of the UMVUE.*

4. Bernoulli trials

Consider a discrete hidden set U consisting of N unlabeled, indistinguishable units. A sample s from U arises by associating a binary indicator $Y_i \sim$ Bernoulli(p) to each $i \in U$, for fixed $0 < p < 1$, where different realizations of the Y_i's can be generated in different draws. The probability p may be known or unknown. A single sample consists of the subset of units with positive indicators, $s = \{i \in U : \ Y_i = 1\}$. This is a frequently encountered situation in computer science, ecology, business, epidemiology, and many other fields [44, 107, 15, 72].

4.1. Binomial N parameter

We first assume that p is known. A single sample s from U gives a statistic $Q := n = |s| = \sum_{i \in U} Y_i$ which has Binomial(N, p) distribution. When there are k independent samples, we assume they are generated by the same mechanism, so $\mathbb{P}(Q_1 = q_1, \ldots, Q_k = q_k) = \prod_{i=1}^{k} \binom{N}{q_i} p^{q_i}(1-p)^{N-q_i}$. The method of moments estimator (MME) $\widehat{N}_{MME} = \bar{Q}/p$ is an unbiased estimator of N. There are two versions of the MLE, derived from continuous and discrete likelihood equations respectively. The continuous MLE, $\widehat{N}'_{MLE}$ is the solution of $\partial L/\partial N = 0$ (take $Q_{(k)}$ if it is larger than the solution), and the discrete MLE $\widehat{N}_{MLE}$ is the largest N such that $L(N) - L(N-1) \geq 0$.

The finite-population regime arises when $k = 1$ and $p \to 1$, i.e. when all units are associated with indicator 1 and observed in a single sample. We consider the infill asymptotic regime with $N_t = N$ and $k_t \to \infty$. The outfill regime is $k_t, N_t \to \infty$ with $k_t/N_t \to c > 0$. Figure 3 shows how the sampling mechanism varies under different regimes for the binomial N model. The following theorem combines results in [9, 41] and states the consistency of estimates under the infill asymptotic regime, along with error rates under the outfill regime. In particular, the estimation error of $Q_{(k)}$ increases with N under the outfill regime.

Theorem 4.1. *Under the finite-population regime, $\widehat{N}_{MME}$, $Q_{(k)}$ and $\widehat{N}_{MLE}$ are consistent. Under infill asymptotics, $\widehat{N}_{MLE}, \widehat{N}_{MME}$, $Q_{(k)}$, and $\widehat{N}'_{MLE}$ after rounding to the nearest integer, are consistent [9]. Under outfill asymptotics, $\widehat{N}_{MME}$ and $\widehat{N}'_{MLE}$ are both asymptotically unbiased and normal with variance*

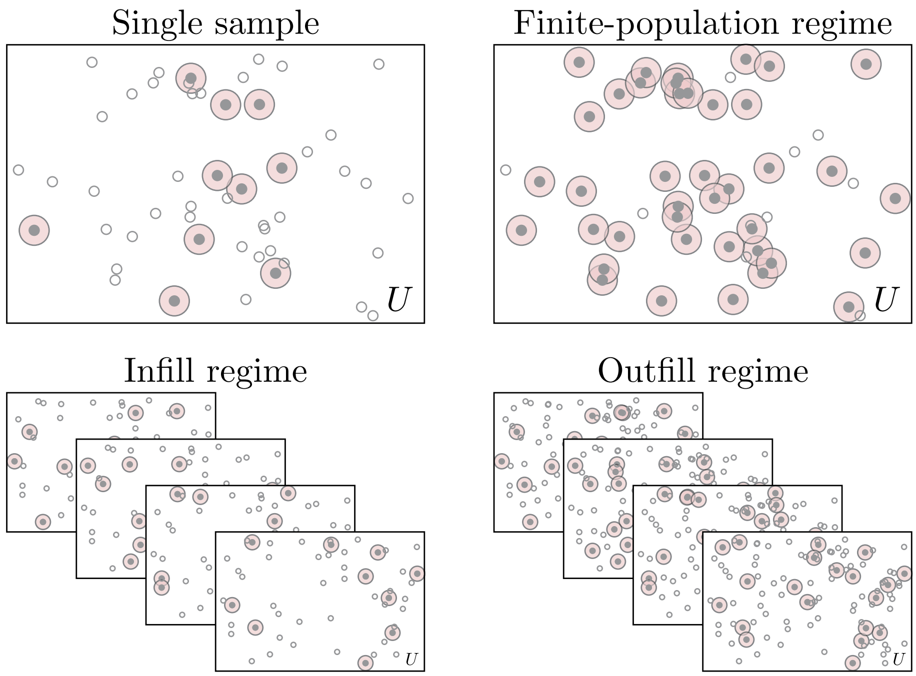

Fig 3. *Illustration of the sampling mechanism for the binomial model, and the finite-population, infill and outfill asymptotic regimes. The solid points with red circles are units with indicator 1 (which are therefore in the sample), and the rest are unobserved.*

$O(1)$. *The "relative error" of the discrete MLE,* $(\widehat{N}_{MLE} - N)/N^{\alpha} \xrightarrow{P} 0$ *for any* $\alpha > 1/2$. *The "relative error" of* $Q_{(k)}$ *with* $\alpha = 1$ *goes to* $p - 1$ *in probability.*

When p is unknown, the situation does not improve: negative or unstable estimates may occur, and Bayesian approaches are usually adopted to avoid these issues. Blumenthal and Dahiya [9] adopted a conjugate prior $\text{Beta}(a, b)$ for p and an improper uniform prior $p(N) \propto 1$ for N; the posterior is proper if and only if $a > 1$ [65]. Blumenthal and Dahiya [9] showed that the posterior mode $\widehat{N}_m$ is consistent under infill asymptotics, and satisfies

$$\frac{\sqrt{n}}{N}\left(\widehat{N}_m - N\right) \xrightarrow{L} N\left(0, \frac{2(1-p)^2}{p^2}\right)$$

under the outfill regime. In particular, the MSE rate is slower compared to $O(1)$ as in Theorem 4.1 when p is known.

A special case of the Binomial scenario arises for zero-truncated counts. For example, a registry may record the number of times each unit has been observed, but zero counts are not recorded. Distributional assumptions can be used to estimate the proportion of unobserved zero counts, leading to estimates of the set size. Zero-truncated counting models have been used to estimate size of hard-to-reach populations, including drug users [32, 11], undocumented immigrants

[110, 10], criminal population [109, 13], the number of infected households in an epidemic [103], and species richness in ecology [117, 26]. To illustrate, associate to each unit $i \in U$ a realization of the attribute $Y_i \sim \text{Poisson}(\lambda)$. A sample from U is $s = \{i \in U : \; Y_i > 0\}$ and an observation on s is $\{Y_i : \; i \in s\}$, the set of all positive counts. For one sample, the sampling mechanism is given by $\mathbb{P}(y_1, \ldots, y_{|s|}|s) = \prod_{i \in s} \lambda^{y_i}/(e^{\lambda} - 1) y_i!$. Estimating λ under this model reveals the proportion of zero counts, $p = 1 - e^{-\lambda}$, and estimation of N proceeds as in the Binomial(N, p) case outlined above. The asymptotic results in Theorem 4.1 follow.

4.2. Waiting times

Sometimes the state of a hidden unit may change, thereby making it known to an observer. For example, terrorist plots may change state from "hidden" to "executed", making them observable by intelligence agents [67]. The temporal pattern of such state changes may give insight into the number of hidden units. Properties of waiting times to an event have been exploited to estimate the number of units in studies of terrorism, crime, and estimation of epidemiological risk population sizes [67, 43, 27, 28].

Suppose U is a set of N hidden units in existence at time 0, each of which is at risk of "failure" at some future time. To each $i \in U$, associate a failure time $T_i \sim \text{Exponential}(\lambda)$, and suppose failure times are observed up to some finite observation time $T > 0$. A sample is the set of units that have failed by the end of study, $s = \{i \in U : \; T_i < T\}$ with $|s| = n$, and an observation on s is $\{T_i : i \in s\}$. With repeated sampling, a new observation is independent of all previous observations, taken after all units are set to be "at risk" over again. We consider the finite-population regime in which $T \to \infty$ so that all failures are observed, the infill regime in which T and N are fixed with the number of repeated observations $k_t \to \infty$, and the outfill regime in which $T_t, N_t \to \infty$ with $T_t/N_t \to c > 0$. For example, if U is the set of hidden terrorist plots [e.g. 67, 68], the finite-population regime keeps $|U| = N$ constant, while letting the maximum observation time $T \to \infty$, so that eventually every plot in U is executed and thereby revealed to the observer. The infill regime consists of keeping N and T constant, while obtaining (hypothetical) repeated realizations of the same N plots over $[0, T]$. The outfill regime lets both the observation time T and number of plots N go to infinity together, so that more plots are added, while the observation time increases. Figure 4 illustrates each regime under the waiting time model.

Let $\Delta_i := T_i - T_{i-1}$ be the waiting time between the $(i-1)$th and ith failure. The sampling mechanism is given by

$$\begin{aligned} \mathbb{P}(t_1, \ldots, t_n | N, \lambda) = \lambda^n \cdot \prod_{i=1}^{n} (N - i + 1) \cdot \exp\left[-\lambda \sum_{i=1}^{n} (N - i + 1) \Delta_i \right] \\ \cdot \exp[-\lambda (N - n)(T - t_n)], \end{aligned}$$

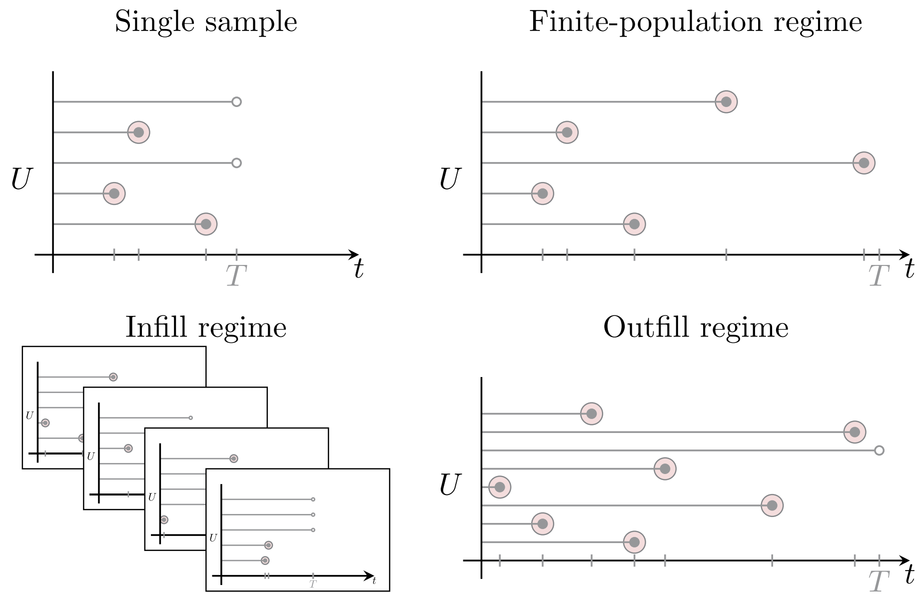

Fig 4. *Illustration of the waiting time model. The observed event times are subject to right censoring at $t = T$, that is, events that occur before T are observed. Solid dots with red shades indicate observed event times. The finite-population regime is that $T \to \infty$ so that all events are observed. Infill asymptotics amounts to generating different realizations of the failure times. Under the outfill regime, T and the total number of units N both increase toward infinity.*

which gives rise to the likelihood $L(t_1, \dots, t_n; N)$. Alternatively, if we ignore the timing of events, the observed number of events can be characterized by a binomial model $\mathbb{P}(n|N, \lambda) = \binom{N}{n}(1 - e^{-\lambda T})^n e^{-\lambda T(N-n)}$, which yields $L_2(n; N)$. Maximizing L and L_2 lead to two estimates, $\widehat{N}_{MLE}$ and $\widehat{N}'_{MLE}$ of N. It is easy to verify that $\partial \log L/\partial N = \partial \log L_2/\partial N$, so $\widehat{N}_{MLE}$ and $\widehat{N}'_{MLE}$ are identical. The timing of events does not contain more information about N than the total number of events. The asymptotic behavior of $\widehat{N}_{MLE}$ follows from the discussion in Section 4.1: when λ is known, $\widehat{N}_{MLE}$ is consistent under finite-population and infill regimes. Under the outfill regime, it is unbiased and asymptotically normal with variance $O(1)$.

4.3. The network scale-up method

Estimating the number of vertices in a hidden network or graph is an important problem in sociology, epidemiology, computer science, and intelligence applications [76, 6, 126, 84, 40, 77, 73]. A subgraph of a larger graph may contain information about the size of the larger graph [40, 5, 83]. The network scale-up method (NSUM) [76] provides an estimate for the size of a hidden population by making use of network information from a sub-sample of individuals.

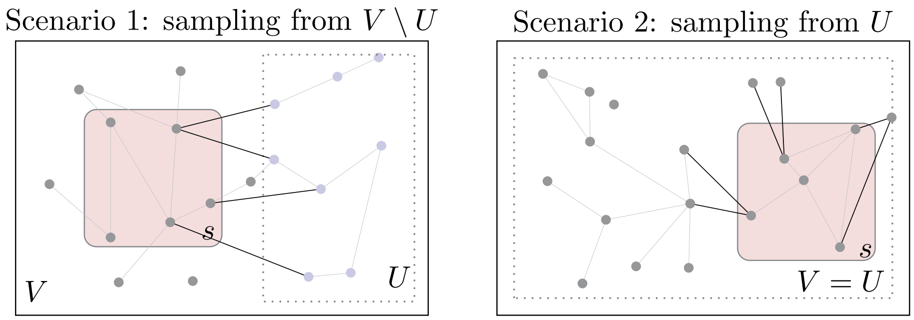

FIG 5. *Illustration of the two common scenarios for the network scale-up method. In scenario 1, $U \subset V$ and we observe a randomly chosen subset s of $V \setminus U$ and number of edges from each unit in s to U (thick lines) and to $V \setminus U$ (thin lines) respectively. In scenario 2, $U = V$ and we observe the induced subgraph (edges represented by thin lines) from a randomly chosen subset s of U as well as the pendant edges (thick lines) between s and $U \setminus s$.*

Consider a graph $G_V = (V, E)$, where V is a set of M units and $\{i, j\} \in E$ means that $i, j \in V$ are connected. V is called the *total population*, and a subset $U \subseteq V$ of size N is the *hidden population*. Assume G_V is *simple*, and has no parallel edges or self-loops. The network of U is $G_U = (U, E_U)$, where $E_U = \{\{i, j\} : i \in U, j \in U, \{i, j\} \in E\}$. We call $V \setminus U$ the *general population*. A sample from a subset of V, along with network degrees of the sampled units within and outside of that subset provides information for learning about the size of U. Suppose the total population network G_V is generated from the Erdős-Rényi random graph model [37] in which each pair of distinct vertices is connected independently by an edge with probability $\Pr(\{i, j\} \in E) = \pi$. The likelihood of a random graph $G_V = (V, E)$ from the Erdős-Rényi model with $|V| = N$ and connection probability π is

$$\Pr(G_V) = \pi^{|E|}(1-\pi)^{\binom{N}{2}-|E|}$$

where $|E|$ is the number of edges and $\binom{N}{2}$ is the number of unordered distinct pairs of vertices. Two common sampling scenarios – sampling from $V \setminus U$ and directly from U – are illustrated in Figure 5.

4.3.1. Sampling from the general population

We consider sampling uniformly at random from the general population $V \setminus U$ with a fixed sample size $|s| = n$. The sampling mechanism is $\mathbb{P}(s \mid |s| = n) = \binom{M-N}{n}^{-1}$. For a sample $s \in V$, we observe network degrees $d_i^V := \sum_{j \in V} \mathbb{1}\{E_{ij} = 1\}$ and $d_i^U := \sum_{j \in U} \mathbb{1}\{E_{ij} = 1\}$ for each $i \in s$. As an empirical example, suppose we wish to estimate the number of people who died in an earthquake [e.g. 5]. We cannot survey the dead (members of U) but we can survey living people ($V \setminus U$) to determine how many people they know (d_i^V), and how many they know who died as a result of the earthquake (d_i^U).

Under the Erdős-Rényi model, $\mathbb{E}d_i^V = (M-1)\pi \approx M\pi$ and $\mathbb{E}d_i^U = N\pi$, $\forall i \in V \setminus U$. Taking the ratio and canceling out π yields the MME

$$\widehat{N}_{NS} = M \cdot \frac{\sum_{i=1}^n d_i^U}{\sum_{i=1}^n d_i^V}. \tag{3}$$

Conditional on d_i^V, d_i^U follows hypergeometric distribution for each i. The same estimator can also be derived under a different model assumption. Killworth et al. [76] considered a model where d_i^U is Binomial$(d_i^V, N/M)$ given d_i^V, and (3) is then the MLE under this binomial model, which is unbiased with variance $(M-N)N/\sum_i d_i^V$.

We consider the finite-population regime in which $n \to (M-N)$, i.e. $s \to V \setminus U$. Under the infill regime, M, N, n are fixed and the number k_t of repeated samples $s \subseteq V \setminus U$ goes to infinity. The outfill regime is that $M_t, N_t, n_t \to \infty$ such that $N_t/M_t \to c_1 \in (0,1)$, $n_t/(M_t - N_t) \to c_2 \in (0,1)$, and $k_t = 1$.

Sometimes an intermediate step in deriving $\widehat{N}_{NS}$ is the estimation of personal network sizes d_i^V. If unbiased estimates $\hat{d_i^V}$ are plugged in, $\widehat{N}_{NS}$ would have a positive bias by Jensen's inequality since $1/x$ is a convex function. Let us assume for now that the d_i^V's are observed true values. Theorem 4.2 states the asymptotic properties of $\widehat{N}_{NS}$ under the Erdős-Rényi assumption (proof provided in the supplementary materials [23]).

Theorem 4.2. *$\widehat{N}_{NS}$ has a positive bias $N/(M-1)$. It is not necessarily consistent under the finite-population regime, and converges to a positively biased quantity under infill. It is asymptotically normal with bias c_1 and variance $O(1)$ under the outfill regime.*

4.3.2. Sampling from the hidden population

When possible, a random sample from the hidden population U can also lead to a valid estimate. Consider a random sample $s \subseteq U$ where G_U follows the Erdős-Rényi model with edge probability π. We observe the nodes $i \in s$, as well as network degrees $d_i^s := \sum_{j\in s} \mathbb{1}\{E_{ij} = 1\}$ and $d_i^U := \sum_{j \in U} \mathbb{1}\{E_{ij}=1\}$, for each individual $i \in s$. Then, $\mathbb{E}\left(\sum_{i\in s} d_i^U\right) = 2\binom{n}{2}\pi$ and $\mathbb{E}\left(\sum_{i\in s} d_i^s\right) = \pi n(N-1)$. Canceling out π yields the MME, which is often simplified to

$$\widehat{N} = \frac{n\sum_{i=1}^n d_i^U}{\sum_{i=1}^n d_i^s}. \tag{4}$$

Chen, Karbasi and Crawford [21] investigated the behavior of $\widehat{N}$ with finite-sample as well as with large n, but did not specify the relationship between N and n under the asymptotic setting. In our setting, the finite-population regime is $n \to N$ with N fixed. The infill regime is that n, N are fixed and the sampling procedure is infinitely repeated. The outfill asymptotic regime is that $n_t, N_t \to \infty$ with $n_t/N_t \to c \in (0,1)$. Then we have the following theorem for the asymptotic properties of $\widehat{N}$ (proof provided in the supplementary materials [23]).

Theorem 4.3. *Under the finite-population regime, $\widehat{N}$ converges to N. Under infill asymptotics, $\widehat{N}$ is always positively biased conditioning on $|E_s| > 0$ [21], and is hence inconsistent. Under outfill asymptotics, $\widehat{N}$ is asymptotically normal with bias $(1-c)/c$ and variance $O(1)$.*

4.4. Estimating a total with unequal sampling probabilities

A generalization of binomial models allows for heterogeneity in the inclusion, or "success" probabilities p, that is, when the sampling is not uniformly at random. Horvitz and Thompson [57] proposed unbiased estimators for population means and totals under the setting of sampling without replacement from finite population, where the selection probabilities can be unequal. The Horvitz-Thompson (HT) estimator for the population total is $\widehat{N} = \sum_{i \in s} 1/p_i$, where $p_i = \mathbb{E}(\mathbb{1}\{i \in s\})$ is the probability that unit $i \in U$ is sampled in s. The estimator $\widehat{N}$ is unbiased for the total population size N. This estimator and its variants have been applied to the estimation of animal abundance [12] and other fields. We consider a deterministic sample size n. Then the variance of $\widehat{N}$ is [57]

$$\mathrm{Var}(\widehat{N}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(p_{ij} - p_i p_j)\left(\frac{1}{p_i} - \frac{1}{p_j}\right)^2, \tag{5}$$

where p_{ij} is the joint probability that units i and j are both in the sampled set s, and $p_{ii} = p_i$. The finite-population regime amounts to letting $p_i \to 1$ for any i. Under the infill regime, p_i, p_{ij}, N are fixed and the number of repeated samples $k_t \to \infty$. Under the outfill regime, N and n both increase to infinity such that $n/N \to c \in (0,1)$. Figure 6 shows the non-uniform sampling mechanism under each regime.

Specifically, we consider the following setting to illustrate the asymptotic behavior of the HT estimator. Suppose U consists of H clusters, where the hth cluster has N_h units. We assume that H is known in advance, while N_h is observed only if a unit from cluster h is sampled. In each sample, a total of n units are sampled from U by the following procedure: first a cluster h is drawn uniformly at random each with probability $1/H$. Then one unit is drawn from the N_h units in that cluster, also uniformly at random, without replacement. We assume that $\min_{h\in[H]} N_h > n$. An observation on sample s consists of the units in s, their cluster membership, and the sizes of clusters that they belong to.

When there are repeated observations, we assume they follow the same design and are mutually independent. In this setting, the outfill regime is defined such that each cluster in the original population is replicated and appears t times in U_t. The cluster sizes are fixed at $N_h^{(t)} = N_h$ and the number of clusters increases as $H_t = tH$. $N = \sum_{h=1}^{H}$ is fixed and the estimand is $N_t = Nt$. The sample size satisfies $n_t/N_t \to c \in (0,1)$. We then have the following theorem about the consistency of $\widehat{N}$ under each regime (proof provided in the supplementary materials [23]). In particular, the variance of $\widehat{N}$ grows with N under the outfill regime.

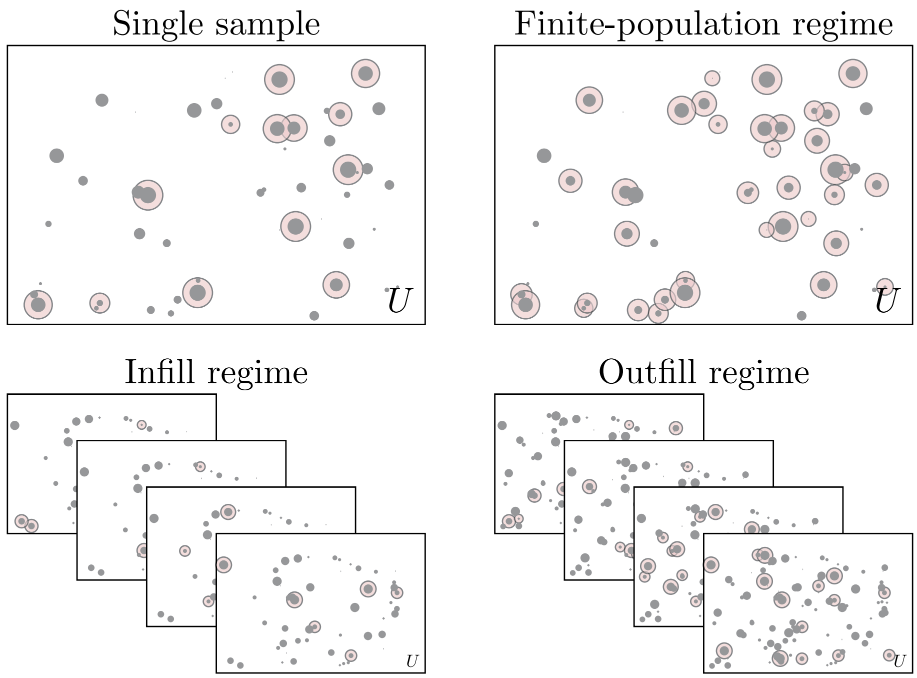

Fig 6. *Illustration of the single sample, finite-population, infill and outfill regimes for the general HT estimator. The probability of being sampled for each point here is visualized as its size.*

Theorem 4.4. *$\widehat{N}$ is consistent under the finite-population regime, and MSE consistent under infill asymptotics. $\widehat{N}$ is unbiased and asymptotically normal with variance $O(N)$ under the outfill regime.*

5. Other unordered sets

5.1. Capture-recapture experiments

Capture-recapture (CRC) refers to a broad class of methods to estimate the size of hidden populations for which random sampling is possible [104, 42, 34, 90, 18, 63]. Estimation of the population size is based on the overlap between two or more random samples [56, 94, 111, 88]. While a wide variety of CRC estimators have been developed [101, 122, 90, 18, 75], we focus here on the two- and k-sample CRC estimators with homogeneity within a closed population.

5.1.1. Two-sample estimation

We first consider the common case of two-sample CRC. Let U be a hidden finite set of size N, where each unit $i \in U$ has binary attributes (X_i^1, X_i^2), which are all $(0,0)$ in the beginning. We draw a sample $s_1 \subseteq U$ with size n_1 from U, and set $X_i^1 = 1$ for all $i \in s_1$. Then a second sample s_2 with size n_2 is drawn, independent from s_1 and uniformly at random, and we set $X_i^2 = 1$ for all $i \in s_2$. We observe $(X_i^1, X_i^2)_{i \in s_1 \cup s_2}$, and let $m = \sum_{i \in U} \mathbb{1}\{(X_i^1, X_i^2) = (1,1)\}$. In ecology, to estimate the abundance of an animal species, researchers could first capture n_1 animals from that species, mark them and then release them. After the captured animals have mixed well with the remaining ones, researchers could capture n_2 animals again, uniformly at random, and record the number m of animals captured in the first step. Then m follows a hypergeometric distribution conditioning on N, n_1 and n_2, i.e. the mechanism of generating the observations can be defined as $\mathbb{P}(m|s_1, s_2) = \binom{n_1}{m}\binom{N-n_1}{n_2-m}/\binom{N}{n_2}$. The MME, $\widehat{N}_L = n_1 n_2/m$, is also known as the Lincoln-Petersen estimator [80, 89].

We consider the finite-population regime with $n_2 \to N$. The infill regime is that N, n_1, n_2 are fixed and repeated sample pairs $\{s_1^{(t)}, s_2^{(t)}\}$ are drawn with $t \to \infty$. The outfill regime is given by $N^{(t)}, n_1^{(t)}, n_2^{(t)} \to \infty$ with $n_i^{(t)}/N^{(t)} \to c_i \in (0,1)$ for $i = 1, 2$.

Previous results exist on the bounds or estimates of biases and variances. These were implicitly based on asymptotic approximations: Chapman [19] showed a lower bound for the bias

$$\mathbb{E}\left(\widehat{N}_L\right) - N \geq N\left[\frac{N}{n_1 n_2} + 2\left(\frac{N}{n_1 n_2}\right)^2\right]$$

under outfill, and bounded the variance as

$$\mathrm{Var}(\widehat{N}_L) > N^2\left[\left(\frac{N}{n_1 n_2}\right) + \left(\frac{N}{n_1 n_2}\right)^2\right]$$

under asymptotic approximation that was satisfied by the outfill regime. Though these no longer hold under finite-sample setting, calculations in [19] showed that $\widehat{N}_L$ has a considerable bias under a range of settings. A less biased estimator

$$\widehat{N}_C = \frac{(n_1+1)(n_2+1)}{m+1} - 1, \tag{6}$$

was proposed [19], with bias

$$\mathbb{E}(\widehat{N}_C) - N = -\frac{(N-n_1)!(N-n_2)!}{N!(N-n_1-n_2-1)!} \tag{7}$$

for any n_1, n_2, N, and variance

$$\mathrm{Var}(\widehat{N}_C) \sim N^2\left[\frac{N}{n_1 n_2} + 2\left(\frac{N}{n_1 n_2}\right)^2 + 6\left(\frac{N}{n_1 n_2}\right)^3\right] \tag{8}$$

under outfill [19], where $\sim$ means the difference between two quantities decay to 0. We have the following asymptotic result of $\widehat{N}_L$ and $\widehat{N}_C$ (proof provided in the supplementary materials [23]). Specifically, both estimators have infinitely increasing estimation error under the outfill asymptotic setting.

Theorem 5.1. *Under the finite-population regime, $\widehat{N}_L$ and $\widehat{N}_C$ are consistent. Under infill asymptotics, $\widehat{N}_L$ is positively biased and has MSE $O(1)$ for at least a range of values of n_1, n_2, N. $\widehat{N}_C$ is negatively biased, but the bias is within 1 if $n_1 + n_2 + 1 < N/2$ and $n_1 n_2/N > \log N$ [19]. Under the outfill regime, $\widehat{N}_L$ has bias at least $O(1)$ and variance at least $O(N)$. $\widehat{N}_C$ is asymptotically unbiased with variance $O(N)$. Furthermore, $\widehat{N}_C$ and $\widehat{N}_L$ are inconsistent with $\mathbb{P}(|\widehat{N}_C - N| < \varepsilon) \to 0$ and $\mathbb{P}(|\widehat{N}_L - N| < \varepsilon) \to 0$ for some $\varepsilon > 0$ when $n_1 = c_1 N, n_2 = c_2 N$.*

Further, Chapman [19] showed that no estimator can be unbiased for all possible values of N, n_1 and n_2.

A similar but slightly different sampling mechanism gives rise to the multiplier method, also called the method of benchmark multiplier (MBM). In practice, researchers may know the number of hidden units with a certain trait. The overall prevalence of that trait in the hidden population, if available from estimation, would provide an estimate for the size of the hidden population. Often the prevalence is estimated through expert opinion, historical data, or from a separate sample [48, 45, 55].

We consider the last approach. The idea of MBM can be expressed with a sampling mechanism similar to CRC, except that the first sample s_1 is fixed under infill asymptotics. That is, the known sub-population of hidden units with a certain trait is fixed. The size n_1 of s_1 is called the *benchmark*. The proportion m/n_2 gives the *multiplier*, which is an estimate of the prevalence p. Again, m follows a hypergeometric distribution, so the MME for N is $\widehat{N}_{MBM} = n_1 n_2/m$, which is often called the multiplier estimator. $\widehat{N}_{MBM}$ takes the same form as the Lincoln-Petersen CRC estimator. Asymptotic behaviors of $\widehat{N}_{MBM}$, as summarized in Theorem 5.2, are essentially the same as that of $\widehat{N}_L$ for CRC.

Theorem 5.2. *$\widehat{N}_{MBM}$ is consistent under the finite-population regime. Under infill asymptotics, $\widehat{N}_{MBM}$ is inconsistent with MSE $O(1)$. Under the outfill regime, when $n_1 = c_1 N$, and $n_2 = c_2 N$, $\widehat{N}_{MBM}$ is inconsistent with MSE at least $O(N)$. $\mathbb{P}(|\widehat{N}_{MBM} - N| < \varepsilon) \to 0$ for some $\varepsilon > 0$.*

5.1.2. k-sample estimation

We now consider the generalized setting of k samples. In this scenario, we draw k samples $s_1, \dots, s_k \subseteq U$ with deterministic sizes $n_1, \dots, n_k$ respectively. We assume the probability $p_j := n_j/N$ of being observed in the jth sample is the same for each unit for $j = 1, \dots, k$. In each sample (say s_j), we give the observed units a label that is different for different j's, and record the capture history $\mathcal{H}_{j,i} = (I_1^{(i)}, \dots, I_j^{(i)})$ of each unit $i \in s_j$, where $I_l = 1$ if $i \in s_l$ and 0 otherwise ($l \leq j$). Then an observation on a sequence of samples $\boldsymbol{s} = \{s_1, \dots, s_k\}$ is a 2^k contingency table $T = \{T_{I_1 \dots I_k}\}_{I_1, \dots, I_k \in \{0,1\}^k}$ [42], where the entry corresponding to $I_1, \dots, I_k$ is $\sum_{i \in U} \mathbb{1}(I_1^{(i)} = I_1, \dots, I_k^{(i)} = I_k)$, the number of units with capture history $\mathcal{H}_k = (I_1, \ldots, I_k)$. Let r be the sum of known entries in the contingency table – only the entry $T_{0\ldots0}$ is unobserved. In plain words, following the animal abundance example, researchers could instead draw k random samples. In the first $k-1$ samples, animals that are captured will be given a mark that is unique for each sample. The contingency table summarizes the capture history for all observed animals – how many animal(s) are observed in, or absent from, which sample(s). From the contingency table we have m_i, the number of already marked individuals in s_i, and M_i, the total number of marked individuals in U before s_i is drawn. The sampling scheme then follows a generalized hypergeometric distribution:

$$\mathbb{P}(T|s_1, \ldots, s_k) = \frac{N!}{\prod_{I_1,\ldots,I_k \in \{0,1\}^k} T_{I_1\ldots I_k}!(N-r)!} \prod_{i=1}^{k} \binom{N}{n_i}^{-1}. \tag{9}$$

Maximizing the likelihood (9) gives the MLE of N as the solution of

$$\left(1 - \frac{r}{N}\right) = \prod_{i=1}^{k} \left(1 - \frac{n_i}{N}\right), \tag{10}$$

which is unique, finite and greater than r if $s_1 \cap \ldots \cap s_k$ is non-empty and $|s_i| < r$ for all $i \leq k$ [34]. We restrict our interest to this case only. Setting $k = 2$ recovers the Lincoln-Petersen estimator $\widehat{N}_L$. Since finite-population and infill regimes for the two- and k-sample cases are similar in essence, we mainly discuss outfill asymptotics in this setting: for any finite k, we have $N, n_1, \ldots, n_k \to \infty$ with $n_i/N_i \to c_i \in (0,1)$ for $i = 1, \ldots, k$, and k_t may be finite or going to infinity. We assume the c_i's are bounded away from 0 and 1. Under outfill asymptotics with finite k, following from the delta method, the bias of the MLE is approximated by [34]

$$\mathbb{E}\left(\widehat{N}_{MLE}\right) - N \sim \frac{\left[\frac{k-1}{N} - \sum\left(\frac{1}{N-n_i}\right)\right]^2 + \left[\frac{k-1}{N^2} - \sum\left(\frac{1}{N-n_i}\right)^2\right]}{2\left[\frac{1}{N-\mathbb{E}[r]} + \frac{k-1}{N} - \sum\left(\frac{1}{N-n_i}\right)\right]^2},$$

which is $O(1)$, and the asymptotic variance is $O(N)$, approximated by [34]

$$\mathrm{Var}(\widehat{N}_{MLE}) \sim \left[\frac{1}{N - \mathbb{E}[r]} + \frac{k-1}{N} - \sum_{i=1}^{k}\left(\frac{1}{N-n_i}\right)\right]^{-1}.$$

Under outfill asymptotics with infinite sampling repetitions, we assume $\inf_{i\in[k]} p_i > 0$. Then the magnitude of bias is bounded above by $N - \mathbb{E}[r]$, and hence by $N\prod_{i=1}^{k}(1-p_i)$. The variance is $O(N\prod_{i=1}^{k}(1-p_i))$. Therefore, as long as k is increasing such that $N\prod_{i=1}^{k}(1-p_i) \to 0$, $\widehat{N}_{MLE}$ will be MSE consistent for N.

6. Discussion

Several features determine researchers' ability to learn about the size of a hidden set. First, the structure of the set – labeled units, ordering of the labels, or relational (network/graph) information – can permit researchers to learn about the number of remaining units when a subset is observed. Second, a feasible probabilistic query mechanism – random sampling, or observation conditional on a unit trait or attribute – must be available. Third, a statistical estimator that enjoys desirable statistical properties must be chosen. Some of these features may be under the control of researchers, while others may be intrinsic to the problem. Table 1 summarizes the models that have been discussed in this paper, as well as consistency results of estimators in each model.

How should empirical researchers evaluate the statistical properties of estimators, design a study or choose a sample size? Many of these tasks are based on asymptotic arguments, and statistical claims about the large-sample performance of hidden set size estimators depend on specification of an appropriate asymptotic (or even non-asymptotic) regime. It is crucial to identify how the sample size increases, especially in relation to the target population, when asymptotic approximation or comparison is involved in population size estimation tasks. When designing a study, this may include determining the minimum sample size that leads to a desired precision [95, 59], or selecting an "optimal" sampling strategy (e.g. one-time larger sample versus multi-time repeated smaller samples). In data analysis, this may include establishing valid approximation to biases and variances or comparing the efficiency of different statistical approaches [95, 2, 16, 86]. If the vast majority of the target population can be observed in one-step sampling, consistency under the trivial finite-population regime may be a goal when developing estimators. If the total population is fixed, and arbitrarily repeated i.i.d. samples can be obtained, then consistency under infill may justify the use of a statistical approach. If instead only one-time or finite-time sampling is permitted, in which the sample size is believed to reflect a proportion of the potentially large population, performance of estimators under outfill may be of more interest. We have shown that different asymptotic regimes can lead to dramatically different statistical properties. Some seemingly sensible estimators are inconsistent with different rates of MSE, and asymptotic claims for population size estimators under one regime may be of limited value for analyzing the general situation.

In this review, we have focused on technical claims about the asymptotic properties of estimators, and have not discussed considerations for practical data collection. For example, the waiting time model does not accommodate censoring or truncation of observations, but could be easily extended to do so. Respondent recall bias in the network scale-up method may make the reported network degrees noisy estimates of the truth. The Horvitz-Thompson estimator relies on knowledge about marginal inclusion probabilities of each sampled individual, which may not be readily available when the size of the population is unknown. While improved data collection strategies may not be able to mitigate poor asymptotic properties – like inconsistency – under a particular regime,

Table 1. *Summary of models, estimators, and asymptotic results for estimating the size of a hidden set*

Problem	Trait	Sample	Estimator	Consistency
German tank	Consecutive integers	Uniform random draw from U	$\widehat{N}_{\mathrm{MLE}}, \widehat{N}_G, \widehat{N}_2, \widehat{N}_3$	Infill: consistent Outfill: MSE $O(1)$
Binomial N	Y_i ∼Bernoulli(p)	$\{i \in U : Y_i = 1\}$	$\widehat{N}_{\mathrm{MLE}}, \widehat{N}_{\mathrm{MME}}$, $Q_{(k)}, \widehat{N}'_{\mathrm{MLE}}$	Infill: consistent Outfill: $\widehat{N}_{\mathrm{MME}}, \widehat{N}'_{\mathrm{MLE}}$ have MSE $O(1)$[1]
Waiting times	T_i ∼Exponential(λ)	$\{i \in U : T_i < T\}$	Equivalent to Binomial N	
NSUM	Network degree d_i^U, d_i^V	Uniform random draw from $V \setminus U$	$\widehat{N}_{NS} = M\frac{\sum d_i^U}{\sum d_i^V}$	Infill: inconsistent Outfill: MSE $O(1)$
	Network degree d_i^s, d_i^U	Uniform random draw from U	$\widehat{N} = n\frac{\sum d_i^U}{\sum d_i^s}$	Infill: inconsistent[2] Outfill: MSE $O(1)$
HT	Cluster membership	Uniform random draw from each sampled cluster	$\widehat{N} = \sum 1/p_i$	Infill: consistent Outfill: MSE $O(N)$
k-sample CRC	Capture history	$\mathbb{P}(i \in s_j) = n_j/N, \forall i$	$\widehat{N}_L, \widehat{N}_C, \widehat{N}_{MLE}$	Infill: inconsistent Outfill $k = 2$: MSE $O(N)$ Outfill $k \to \infty$: $\widehat{N}_{MLE}$ consistent[3]
MBM	Similar to two-sample CRC			

[1] the error rate of $\widehat{N}_{MLE}$ and $Q_{(k)}$ are given in Section 4.1 in terms of relative error

[2] conditioning on $|E_s| > 0$

[3] if $N \prod_{i=1}^{k} (1 - p_i) \to 0$

better data may be able to reduce variance in finite samples.

While we have discussed many of the most popular settings and methods for estimating the size of a hidden set, there are several other settings we have not covered. Respondent-driven sampling (RDS), snowball sampling and link-tracing sampling generate samples from hidden networks, and modeling the stochastic process underlying such sampling mechanism can be used to estimate hidden population sizes [53, 28, 112, 113]. There is a large literature on CRC beyond what we have covered here. For example, there are approaches for CRC with an open population, with immigration, emigration, birth, and death [101, 122] or with heterogeneity in capture probabilities [90, 18]. CRC is also possible using data from network sampling designs [75]. We have also not discussed species number estimation [17], "count distinct" and streaming estimation problems [47, 20, 66], and genetic methods for population size estimation [29, 4]. In addition, we have not addressed the issue of entity resolution, or record de-duplication [98]. The results presented in this paper suggest that researchers employing methods for estimating the size of a hidden set should evaluate the performance of estimators under deliberately specified asymptotic assumptions.

Acknowledgements

This work was supported by NIH grants NICHD/BD2K DP2 HD091799, NIH/NCATS KL2TR000140, NIMH P30MH062294, the Center for Interdisciplinary Research on AIDS, and the Yale Center for Clinical Investigation. We are grateful to Peter M. Aronow, Heng Chen, Xi Fu, and Edward H. Kaplan for helpful comments.

Supplementary Material

Supplementary materials to "Estimating the size of a hidden finite set: large-sample behavior of estimators"
(doi: 10.1214/19-SS127SUPP; .pdf). Proof of theorems.

References

[1] Abdul-Quader, A. S., Baughman, A. L. and Hladik, W. (2014). Estimating the size of key populations: Current status and future possibilities. *Current Opinion in HIV and AIDS* **9** 107–114.

[2] Bailey, N. T. J. (1951). On estimating the size of mobile populations from recapture data. *Biometrika* **38** 293–306.

[3] Bao, L., Raftery, A. E. and Reddy, A. (2015). Estimating the Sizes of Populations At Risk of HIV Infection From Multiple Data Sources Using a Bayesian Hierarchical Model. *Statistics and Its Interface* **8** 125–136. MR3322160

[4] Bellemain, E., Swenson, J. E., Tallmon, D., Brunberg, S. and Taberlet, P. (2005). Estimating population size of elusive animals with DNA from hunter-collected feces: Four methods for brown bears. *Conservation Biology* **19** 150–161.

[5] Bernard, H. R., Johnsen, E. C., Killworth, P. D. and Robinson, S. (1991). Estimating the size of an average personal network and of an event subpopulation: Some empirical results. *Social Science Research* **20** 109–121.

[6] Bernard, H. R., Killworth, P. D., Johnsen, E. C., Shelley, G. A. and McCarty, C. (2001). Estimating the ripple effect of a disaster. *Connections* **24** 18–22.

[7] Bernard, H. R., Hallett, T., Iovita, A., Johnsen, E. C., Lyerla, R., McCarty, C., Mahy, M., Salganik, M. J., Saliuk, T., Scutelniciuc, O., Shelley, G. A., Sirinirund, P., Weir, S. and Stroup, D. F. (2010). Counting hard-to-count populations: The network scale-up method for public health. *Sexually Transmitted Infections* **86** ii11–15.

[8] Bickel, P. J., Nair, V. N. and Wang, P. C. (1992). Nonparametric inference under biased sampling from a finite population. *The Annals of Statistics* **20** 853–878. MR1165596

[9] Blumenthal, S. and Dahiya, R. C. (1981). Estimating the Binomial Parameter n. *Journal of the American Statistical Association* **76** 903–909. MR0650902

[10] Böhning, D. and van der Heijden, P. G. (2009). A Covariate Adjustment for Zero-Truncated Approaches to Estimating the Size of Hidden and Elusive Populations. *The Annals of Applied Statistics* **3** 595–610. MR2750674

[11] Böhning, D., Suppawattanabodee, B., Kusolvisitkul, W. and Viwatwongkasem, C. (2004). Estimating the number of drug users in Bangkok 2001: A capture-recapture approach using repeated entries in one list. *European Journal of Epidemiology* **19** 1075.

[12] Borchers, D. L., Buckland, S. T., Goedhart, P. W., Clarke, E. D. and Hedley, S. L. (1998). Horvitz-Thompson estimators for double-platform line transect surveys. *Biometrics* **54** 1221–1237.

[13] Bouchard, M. (2007). A capture-recapture model to estimate the size of criminal populations and the risks of detection in a marijuana cultivation industry. *Journal of Quantitative Criminology* **23** 221–241.

[14] Brewer, K. R. W. (1979). A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association* **74** 911–915. MR0556487

[15] Brookmeyer, R. and Gail, M. H. (1988). A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association* **83** 301–308.

[16] Brownie, C. and Pollock, K. H. (1985). Analysis of multiple capture-recapture data using band-recovery methods. *Biometrics* **41** 411–420.

[17] Bunge, J., Willis, A. and Walsh, F. (2014). Estimating the Number of

Species in Microbial Diversity Studies. *Annual Review of Statistics and Its Application* **1** 427–445.

[18] Chao, A. (1987). Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. *Biometrics* **43** 783–791. MR0920467

[19] Chapman, D. G. (1951). Some Properties of the Hypergeometric Distribution With Applications to Zoological Sample Censuses. *University of California Publications in Statistics* **1** 131–160. MR0042093

[20] Chassaing, P. and Gerin, L. (2006). Efficient estimation of the cardinality of large data sets. In *Proceedings of the Fourth Colloquium on Mathematics and Computer Science* 419–422. Discrete Mathematics and Theoretical Computer Science.

[21] Chen, L., Karbasi, A. and Crawford, F. W. (2016). Estimating the Size of a Large Network and its Communities from a Random Sample. In *Advances in Neural Information Processing Systems 29* 3072–3080. Curran Associates, Inc.

[22] Chen, H.-S., Simpson, D. G. and Ying, Z. (2000). Infill asymptotics for a stochastic process model with measurement error. *Statistica Sinica* **10** 141–156. MR1742105

[23] Cheng, S., Eck, D. J. and Crawford, F. W. (2019). Supplementary materials to "Estimating the size of a hidden finite set: large-sample behavior of estimators". DOI: 10.1214/19-SS127SUPP.

[24] Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley New York. MR0474575

[25] Corn, P. S. and Fogleman, J. C. (1984). Extinction of Montane Populations of the Northern Leopard Frog (Rana pipiens) in Colorado. *Journal of Herpetology* **18** 147–152.

[26] Craig, C. C. (1953). On the utilization of marked specimens in estimating populations of flying insects. *Biometrika* **40** 170–176.

[27] Crawford, F. W. (2016). The graphical structure of respondent-driven sampling. *Sociological Methodology* **46** 187–211.

[28] Crawford, F. W., Wu, J. and Heimer, R. (2018). Hidden population size estimation from respondent-driven sampling: A network approach. *Journal of the American Statistical Association* **113** 755–766. MR3832224

[29] Creel, S., Spong, G., Sands, J. L., Rotella, J., Zeigle, J., Joe, L., Murphy, K. M. and Smith, D. (2003). Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Molecular Ecology* **12** 2003–2009.

[30] Cressie, N. (2015). *Statistics for Spatial Data*. John Wiley & Sons. MR3559472

[31] Cressie, N. and Lahiri, S. N. (1993). The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis* **45** 217–233. MR1221918

[32] Cruyff, M. J. and van der Heijden, P. G. (2008). Point and Interval Estimation of the Population Size Using a Zero-Truncated Negative Binomial Regression Model. *Biometrical Journal* **50** 1035–1050. MR2649393

[33] Daniel, W. W. (1999). *Biostatistics: A Foundation for Analysis in the Health Sciences*, 7th ed. Wiley New York.

[34] Darroch, J. N. (1958). The multiple-recapture census: I. Estimation of a closed population. *Biometrika* **45** 343–359. MR0119360

[35] Dávid, B. and Snijders, T. A. (2002). Estimating the size of the homeless population in Budapest, Hungary. *Quality & Quantity* **36** 291–303.

[36] Davies, G. and Dawson, S. (2014). A Framework for Estimating the Number of Extremists in Canada Technical Report, Canadian Network for Research on Terrorism, Security, and Society Working Paper Series No. 14-08.

[37] Erdős, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae* **6** 290–297. MR0120167

[38] Eubank, R. and LaRiccia, V. (1992). Asymptotic comparison of Cramer-von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *The Annals of Statistics* **20** 2071–2086. MR1193326

[39] Ezoe, S., Morooka, T., Noda, T., Sabin, M. L. and Koike, S. (2012). Population size estimation of men who have sex with men through the network scale-up method in Japan. *PLoS One* **7** e31184.

[40] Feehan, D. M. and Salganik, M. J. (2016). Estimating the size of hidden populations using the generalized network scale-up estimator. *Sociological Methodology* **46** 153–186.

[41] Feldman, D. and Fox, M. (1968). Estimation of the Parameter n in the Binomial Distribution. *Journal of American Statistical Association* **63** 150–158. MR0224196

[42] Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59** 591–603. MR0383619

[43] Frey, J. C. and Kaplan, E. H. (2010). Queue inference from periodic reporting data. *Operations Research Letters* **38** 420–426. MR2719234

[44] Friedman, T. and Towsley, D. (1999). Multicast session membership size estimation. In *Proceedings of the 18th Annual Joint Conference of the IEEE Computer and Communications Societies. INFOCOM'99* **2** 965–972. IEEE.

[45] Frischer, M., Hickman, M., Kraus, L., Mariani, F. and Wiessing, L. (2001). A comparison of different methods for estimating the prevalence of problematic drug misuse in Great Britain. *Addiction* **96** 1465–1476.

[46] Funk, W. C., Almeida-Reinoso, D., Nogales-Sornosa, F. and Bustamante, M. R. (2003). Monitoring population trends of Eleutherodactylus frogs. *Journal of Herpetology* **37** 245–256.

[47] Fusy, E. and Giroire, F. (2007). Estimating the Number of Active Flows in a Data Stream over a Sliding Window. In *Proceedings of the Meeting on Analytic Algorithmics and Combinatorics. ANALCO '07* 223–231. Society for Industrial and Applied Mathematics, Philadelphia, USA. MR2498130

[48] Godfrey, C., Eaton, G., McDougall, C. and Culyer, A. (2002). *The Economic and Social Costs of Class A Drug Use in England and Wales, 2000.* Home Office London.

[49] Goodman, L. A. (1952). Serial Number Analysis. *Journal of the American Statistical Association* **47** 622–634. MR0053462

[50] Gum, B., Lipton, R. J., LaPaugh, A. and Fich, F. (2005). Estimating the maximum. *Journal of Algorithms* **54** 105–114. MR2108424

[51] Hadfield, M. G., Miller, S. E. and Carwile, A. H. (1993). The Decimation of Endemic Hawai'ian Tree Snails by Alien Predators. *American Zoologist* **33** 610–622.

[52] Hall, W. D., Ross, J. E., Lynskey, M. T., Law, M. G. and Degenhardt, L. J. (2000). How many dependent heroin users are there in Australia? *The Medical Journal of Australia* **173** 528–531.

[53] Handcock, M. S., Gile, K. J. and Mar, C. M. (2014). Estimating hidden population size using respondent-driven sampling data. *Electronic Journal of Statistics* **8** 1491. MR3263129

[54] Heimer, R. and White, E. (2010). Estimation of the number of injection drug users in St. Petersburg, Russia. *Drug and Alcohol Dependence* **109** 79–83.

[55] Hickman, M., Hope, V., Platt, L., Higgins, V., Bellis, M., Rhodes, T., Taylor, C. and Tilling, K. (2006). Estimating prevalence of injecting drug use: A comparison of multiplier and capture–recapture methods in cities in England and Russia. *Drug and Alcohol Review* **25** 131–140.

[56] Hook, E. B. and Regal, R. R. (1995). Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiologic Reviews* **17** 243–264.

[57] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* **47** 663–685. MR0053460

[58] Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association* **77** 89–96. MR0648029

[59] Jensen, A. (1981). Sample sizes for single mark and single recapture experiments. *Transactions of the American Fisheries Society* **110** 455–458.

[60] Joglar, R. L. and Burrowes, P. A. (1996). Declining amphibian populations in Puerto Rico. In *Contributions to West Indian Herpetology: A tribute to Albert Schwartz* (R. Powell and R. W. Henderson, eds.) 371–380. The Society for the Study of Amphibians and Reptiles, Ithaca, NY.

[61] Johnston, L. G., McLaughlin, K. R., El Rhilani, H., Latifi, A., Toufik, A., Bennani, A., Alami, K., Elomari, B. and Handcock, M. S. (2015). Estimating the size of hidden populations using respondent-driven sampling data: Case examples from Morocco. *Epidemiology* **26** 846.

[62] Johnston, L. G., McLaughlin, K. R., Rouhani, S. A. and Bartels, S. A. (2017). Measuring a hidden population: A novel technique to estimate the population size of women with sexual violence-related pregnancies in South Kivu Province, Democratic Republic of Congo. *Journal of Epidemiology and Global Health* **7** 45–53.

[63] Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika* **52** 225–247. MR0210227

[64] Kadushin, C., Killworth, P. D., Bernard, H. R. and Beveridge, A. A. (2006). Scale-up methods as applied to estimates of heroin use. *Journal of Drug Issues* **36** 417–440.

[65] Kahn, W. D. (1987). A Cautionary Note for Bayesian Estimation of the Binomial Parameter *n*. *The American Statistician* **41** 38–40. MR0882767

[66] Kane, D. M., Nelson, J. and Woodruff, D. P. (2010). An Optimal Algorithm for the Distinct Elements Problem. In *Proceedings of the Twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. PODS '10* 41–52. Association for Computing Machinery, New York, USA.

[67] Kaplan, E. H. (2010). Terror queues. *Operations Research* **58** 773–784. MR2683476

[68] Kaplan, E. H. (2012). Estimating the duration of Jihadi terror plots in the United States. *Studies in Conflict & Terrorism* **35** 880–894.

[69] Kaplan, E. H. and Soloshatz, D. (1993). How many drug injectors are there in New Haven? Answers from AIDS data. *Mathematical and Computer Modelling* **17** 109–115.

[70] Karami, M., Khazaei, S., Poorolajal, J., Soltanian, A. and Sajadipoor, M. (2017). Estimating the Population Size of Female Sex Worker Population in Tehran, Iran: Application of Direct Capture–Recapture Method. *AIDS and Behavior* **27** 1–7.

[71] Karanth, K. U. and Nichols, J. D. (1998). Estimation of tiger densities in India using photographic captures and recaptures. *Ecology* **79** 2852–2862.

[72] Karon, J. M., Song, R., Brookmeyer, R., Kaplan, E. H. and Hall, H. I. (2008). Estimating HIV incidence in the United States from HIV/AIDS surveillance data and biomarker HIV test results. *Statistics in Medicine* **27** 4617–4633. MR2528570

[73] Katzir, L., Liberty, E. and Somekh, O. (2011). Estimating sizes of social networks via biased sampling. In *Proceedings of the 20th International Conference on World Wide Web* 597–606. ACM.

[74] Khalid, F. J., Hamad, F. M., Othman, A. A., Khatib, A. M., Mohamed, S., Ali, A. K. and Dahoma, M. J. (2014). Estimating the number of people who inject drugs, female sex workers, and men who have sex with men, Unguja Island, Zanzibar: Results and synthesis of multiple methods. *AIDS and Behavior* **18** 25–31.

[75] Khan, B., Lee, H.-W. and Dombrowski, K. (2018). One-step Estimation of Networked Population Size with Anonymity Using Respondent-Driven Capture-Recapture and Hashing. *PLoS One* **13** e0195959.

[76] Killworth, P. D., McCarty, C., Bernard, H. R., Shelley, G. A. and Johnsen, E. C. (1998a). Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach. *Evaluation Review* **22** 289–308.

[77] Killworth, P. D., Johnsen, E. C., McCarty, C., Shelley, G. A. and Bernard, H. R. (1998b). A social network approach to estimating seroprevalence in the United States. *Social Networks* **20** 23–50.

[78] Kimber, J., Hickman, M., Degenhardt, L., Coulson, T. and Van Beek, I. (2008). Estimating the size and dynamics of an injecting drug user population and implications for health service coverage: Comparison of indirect prevalence estimation methods. *Addiction* **103** 1604–1613.

[79] Lahiri, S. N. (1996). On inconsistency of estimators based on spatial data under infill asymptotics. *Sankhyā: The Indian Journal of Statistics, Series A* **58** 403–417. MR1659130

[80] Lincoln, F. C. (1930). *Calculating Waterfowl Abundance on the Basis of Banding Returns*. U.S. Department of Agriculture, Washington, D.C.

[81] Lwanga, S. K. and Lemeshow, S. (1991). *Sample Size Determination in Health Studies: A Practical Manual*. Geneva: World Health Organization.

[82] Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71** 135–146. MR0738334

[83] Massoulié, L., Le Merrer, E., Kermarrec, A.-M. and Ganesh, A. (2006). Peer Counting and Sampling in Overlay Networks: Random Walk Methods. In *Proceedings of the 25th Annual ACM Symposium on Principles of Distributed Computing. PODC '06* 123–132. ACM.

[84] McCormick, T. H., Salganik, M. J. and Zheng, T. (2010). How many people do you know?: Efficiently estimating personal network size. *Journal of the American Statistical Association* **105** 59–70. MR2757192

[85] McFarland, W., Wilson, E. and Raymond, H. F. (2018). How Many Transgender Men Are There in San Francisco? *Journal of Urban Health* **95** 129–133.

[86] Mills, L. S., Citta, J. J., Lair, K. P., Schwartz, M. K. and Tallmon, D. A. (2000). Estimating animal abundance using noninvasive DNA sampling: promise and pitfalls. *Ecological Applications* **10** 283–294.

[87] Nikfarjam, A., Shokoohi, M., Shahesmaeili, A., Haghdoost, A. A., Baneshi, M. R., Haji-Maghsoudi, S., Rastegari, A., Nasehi, A. A., Memaryan, N. and Tarjoman, T. (2016). National population size estimation of illicit drug users through the network scale-up method in 2013 in Iran. *International Journal of Drug Policy* **31** 147–152.

[88] Paz-Bailey, G., Jacobson, J., Guardado, M., Hernandez, F., Nieto, A., Estrada, M. and Creswell, J. (2011). How many men who have sex with men and female sex workers live in El Salvador? Using respondent-driven sampling and capture-recapture to estimate population sizes. *Sexually Transmitted Infections* **87** 279–282.

[89] Petersen, C. G. J. (1894). *On the Biology of Our Flatfishes and on the Decrease of Our Flat-Fish Fisheries: With Some Observations Showing How to Remedy the Latter and Promote the Flat-Fish Fisheries in Our Seas East of the Skaw*. Centraltrykkeriet.

[90] Pollock, K. H. (1982). A capture-recapture design robust to unequal probability of capture. *The Journal of Wildlife Management* **46** 752–757.

[91] Pollock, K. H., Nichols, J. D., Brownie, C. and Hines, J. E. (1990). Statistical inference for capture-recapture experiments. *Wildlife Monographs* **107** 3–97.

[92] Quaye, S., Raymond, H. F., Atuahene, K., Amenyah, R., Aberle-Grasse, J., McFarland, W., El-Adas, A. and Ghana Men Study Group (2015). Critique and lessons learned from using multiple methods to estimate population size of men who have sex with men in Ghana. *AIDS and Behavior* **19** 16–23.

[93] Rich, A. J., Lachowsky, N. J., Sereda, P., Cui, Z., Wong, J., Wong, S., Jollimore, J., Raymond, H. F., Hottes, T. S., Roth, E. A., Hogg, R. S. and Moore, D. M. (2018). Estimating the Size of the MSM Population in Metro Vancouver, Canada, Using Multiple Methods and Diverse Data Sources. *Journal of Urban Health* **95** 188–195.

[94] Robles, S. C., Marrett, L. D., Clarke, E. A. and Risch, H. A. (1988). An application of capture-recapture methods to the estimation of completeness of cancer registration. *Journal of Clinical Epidemiology* **41** 495–501.

[95] Robson, D. and Regier, H. (1964). Sample size in Petersen mark–recapture experiments. *Transactions of the American Fisheries Society* **93** 215–226.

[96] Ruggles, R. and Brodie, H. (1947). An empirical approach to economic intelligence in World War II. *Journal of the American Statistical Association* **42** 72–91.

[97] Sabin, K., Zhao, J., Calleja, J. M. G., Sheng, Y., Garcia, S. A., Reinisch, A. and Komatsu, R. (2016). Availability and quality of size estimations of female sex workers, men who have sex with men, people who inject drugs and transgender women in low-and middle-income countries. *PLoS One* **11** e0155150.

[98] Sadosky, P., Shrivastava, A., Price, M. and Steorts, R. C. (2015). Blocking Methods Applied to Casualty Records from the Syrian Conflict. *arXiv preprint arXiv:1510.07714*.

[99] Safarnejad, A., Nga, N. T. and Son, V. H. (2017). Population Size Estimation of Men Who Have Sex with Men in Ho Chi Minh City and Nghe An Using Social App Multiplier Method. *Journal of Urban Health* **94** 339–349.

[100] Salganik, M. J., Fazito, D., Bertoni, N., Abdo, A. H., Mello, M. B. and Bastos, F. I. (2011). Assessing Network Scale-up Estimates for Groups Most at Risk of HIV/AIDS: Evidence From a Multiple-Method Study of Heavy Drug Users in Curitiba, Brazil. *American Journal of Epidemiology* **174** 1190.

[101] Schwarz, C. J. and Arnason, A. N. (1996). A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics* **52** 860–873. MR1411736

[102] Schwarz, C. J. and Seber, G. A. F. (1999). Estimating Animal Abundance: Review III. *Statistical Science* **14** 427–456.

[103] Scollnik, D. P. (1997). Inference concerning the size of the zero class from an incomplete Poisson sample. *Communications in Statistics – Theory and Methods* **26** 221–236.

[104] Seber, G. A. F. (1973). *The Estimation of Animal Abundance and Related Parameters*, 2nd ed. Oxford University Press.

[105] Shelton, J. F. (2015). Proposed utilization of the network scale-up method to estimate the prevalence of trafficked persons. In *Forum on Crime and Society*, **8** 85–94. United Nations Publications.

[106] Stein, M. L. (2012). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media.

[107] Talluri, K. (2009). A Finite-Population Revenue Management Model and a Risk-Ratio Procedure for the Joint Estimation of Population Size and Parameters Technical Report, Universitat Pompeu Fabra, Barcelona, Spain.

[108] UNAIDS and World Health Organization (2010). Guidelines on Estimating the Size of Populations most at Risk to HIV Technical Report, Geneva, Switzerland.

[109] van der Heijden, P. G., Cruyff, M. and van Houwelingen, H. C. (2003). Estimating the size of a criminal population from police records using the truncated Poisson regression model. *Statistica Neerlandica* **57** 289–304.

[110] van der Heijden, P. G., Bustami, R., Cruyff, M. J., Engbersen, G. and van Houwelingen, H. C. (2003). Point and interval estimation of the population size using the truncated Poisson regression model. *Statistical Modelling* **3** 305–322.

[111] van der Heijden, P. G., de Vries, I., Böhning, D. and Cruyff, M. (2015). Estimating the size of hard-to-reach populations using capture-recapture methodology, with a discussion of the International Labour Organization's global estimate of forced labour. In *Forum on Crime and Society*, **8** 109–136. United Nations Publications.

[112] Vincent, K. and Thompson, S. (2014). Estimating the size and distribution of networked populations with snowball sampling. *arXiv preprint arXiv:1402.4372*.

[113] Vincent, K. and Thompson, S. (2017). Estimating population size with link-tracing sampling. *Journal of the American Statistical Association* **112** 1286–1295.

[114] Vuylsteke, B., Sika, L., Semdé, G., Anoma, C., Kacou, E. and Laga, M. (2017). Estimating the number of female sex workers in Côte d'Ivoire: Results and lessons learned. *Tropical Medicine and International Health* **22** 1112–1118.

[115] Wang, J., Yang, Y., Zhao, W., Su, H., Zhao, Y., Chen, Y., Zhang, T. and Zhang, T. (2015). Application of network scale up method in the estimation of population size for men who have sex with men in Shanghai, China. *PLoS One* **10** e0143118.

[116] Wesson, P., Handcock, M. S., McFarland, W. and Raymond, H. F. (2015). If You Are Not Counted, You Don't Count: Esti-

mating the Number of African-American Men Who Have Sex with Men in San Francisco Using a Novel Bayesian Approach. *Journal of Urban Health* **92** 1052–1064.

[117] Wilson, R. M. and Collins, M. F. (1992). Capture-recapture estimation with samples of size one using frequency data. *Biometrika* **79** 543–553.

[118] Wilson, J. M., Sullivan, B. A. and Hollis, M. E. (2016). Measuring the "Unmeasurable" Approaches to Assessing the Nature and Extent of Product Counterfeiting. *International Criminal Justice Review* **26** 259–276.

[119] Witte, J. S., Gauderman, W. J. and Thomas, D. C. (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *American Journal of Epidemiology* **149** 693–705.

[120] Wittes, J. and Sidel, V. W. (1968). A generalization of the simple capture-recapture model with applications to epidemiological research. *Journal of Chronic Diseases* **21** 287–301.

[121] Yip, P. S. F., Bruno, G., Tajima, N., Seber, G. A. F., Buckland, S. T., Cormack, R. M., Unwin, N., Chang, Y.-F., Fienberg, S. E., Junker, B. W., LaPorte, R. E., Libman, I. M. and McCarty, D. J. (1995). Capture-recapture and multiple-record systems estimation II: Applications in human diseases. *American Journal of Epidemiology* **142** 1059–1068.

[122] Young, L. J. and Young, J. H. (1998). Capture-Recapture: Open Populations. In *Statistical Ecology* 357–389. Springer.

[123] Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99** 250–261.

[124] Zhang, D., Wang, L., Lv, F., Su, W., Liu, Y., Shen, R. and Bi, P. (2007a). Advantages and challenges of using census and multiplier methods to estimate the number of female sex workers in a Chinese city. *AIDS Care* **19** 17–19.

[125] Zhang, D., Lv, F., Wang, L., Sun, L., Zhou, J., Su, W. and Bi, P. (2007b). Estimating the population of female sex workers in two Chinese cities on the basis of the HIV/AIDS behavioural surveillance approach combined with a multiplier method. *Sexually Transmitted Infections* **83** 228–231.

[126] Zheng, T., Salganik, M. J. and Gelman, A. (2006). How many people do you know in prison? Using overdispersion in count data to estimate social structure in networks. *Journal of the American Statistical Association* **101** 409–423.