DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 723–736, http://dx.doi.org/10.1515/JOS-2015-0042

B-Graph Sampling to Estimate the Size of a Hidden Population

Marinus Spreen[1] *and Stefan Bogaerts*[2]

Link-tracing designs are often used to estimate the size of hidden populations by utilizing the relational links between their members. A major problem in studies of hidden populations is the lack of a convenient sampling frame. The most frequently applied design in studies of hidden populations is respondent-driven sampling in which no sampling frame is used. However, in some studies multiple but incomplete sampling frames are available. In this article, we introduce the B-graph design that can be used in such situations. In this design, all available incomplete sampling frames are joined and turned into one sampling frame, from which a random sample is drawn and selected respondents are asked to mention their contacts. By considering the population as a bipartite graph of a two-mode network (those from the sampling frame and those who are not on the frame), the number of respondents who are directly linked to the sampling frame members can be estimated using Chao's and Zelterman's estimators for sparse data. The B-graph sampling design is illustrated using the data of a social network study from Utrecht, the Netherlands.

Key words: Network sampling; capture recapture; hidden populations.

1. Introduction

Estimating the sizes of hidden populations is important in the field of official statistics in order to provide local or national institutions with insights into the nature and extent of a social problem. Hidden populations are characterized by the lack of well-defined complete sampling frames due to the privacy-threatening nature of the variable that defines the study population (Spreen 1992; Heckathorn 1997). Privacy-threatening traits are often illegal activities and/or activities that are not socially accepted. Examples of illegal activities are drug trafficking, human trafficking, sexual abuse, child abuse, domestic violence, terrorist activities, criminal acts, and so on (e.g., Brugal et al. 1999; Holland et al. 2006; Surjadi et al. 2010; Kunst et al. 2010; Palusci et al. 2010). Depending on the culture and/or legal system of a nation, privacy-threatening traits can also be activities that are not socially accepted, like drug use, selling sex, buying sex, undeclared work, or tax evasion (Bogaerts and Daalder 2011). Requesting privacy-threatening information from members of hidden populations will lead to high rates of uncooperative individuals or unreliable answers (Heckathorn 1997). Two different data collection procedures can be distinguished for estimating the size of a hidden population. In capture-recapture procedures, official

[1] Stenden University of Applied Sciences-School of Social Work and Art Therapies, Rengerslaan 8 Leeuwarden 8917 DD, The Netherlands. Email: Marinus.Spreen@Stenden.com
[2] Tillburg University, Department of Developmental Psychology, Warandelaan 2, 5037 AB Tilburg, The Netherlands. Email: s.bogaerts@uvt.nl

registration sources are used as sampling frames to estimate hidden population sizes. In link-tracing or network sample procedures, social links between hidden population members are used as sampling frames for estimation purposes. The difference between these two procedures lies in the way data are collected. In capture-recapture procedures, hidden population members themselves are not sampled and interviewed, only registered. In link-tracing procedures, hidden population members are sampled and interviewed about their social links with other members of that hidden population.

In this article, a practical sampling design called the B-graph sampling design is introduced and illustrated. This design has been elaborated for research contexts in which one or multiple registration sources are available but each source on its own is considered too small to produce valid capture-recapture estimations. However, if pooling all available registration sources results in a substantial coverage of the unknown population according to local experts, this pooled source can be considered a plausible sampling frame to start a link-tracing data collection procedure. For example, all neighbourhood youth workers in a city agree that the number of names on the pooled list cover a substantial part of the total unknown population. For estimation purposes, the population of interest can be divided into two subpopulations, namely registered and unregistered persons. Drawing a probability sample from the (pooled) registered part of the hidden population and employing a link-tracing procedure by asking each sampled person to disclose his contacts with other hidden population members, the size of the unregistered subpopulation directly linked to registered persons can be estimated. Furthermore, if the assumption that each unregistered person of the study population has at least one direct link to a registered person is held to be plausible, each unregistered person has a positive probability of being included in the link-tracing sample. Thus the resulting estimate gives an indication of the total population size. The estimation problem of the number of persons directly linked to a known subset of persons is of interest in a variety of (forensic) social network studies. For instance, if the known set of persons is hooligans or gang members, the number of directly related unregistered hooligans or gang members can be estimated. If the known subset of persons is arrested problem youths in some city, their number of contacts with other youths may provide valuable information about the size of the problem.

In this article, we discuss three capture-recapture estimators by considering the hidden population as a bipartite graph of a two-mode network (registered and unregistered persons); for example, we focus on the social links between the two subpopulations. This approach is illustrated by data obtained from a social network study conducted among the population of opiate users in the city of Utrecht, the Netherlands (Ten Den et al. 1995). In the original study, three sampling procedures were applied: a random sample from the files of three drug-assistance organisations, a convenience fieldwork sample and a snowball sample to find unregistered opiate users. To illustrate the B-graph design, the three client lists are pooled into one sampling frame (excluding the respondents from the convenience and snowball sample), from which a random sample is drawn and a link-tracing procedure applied to sample unregistered opiate users. The statistical problem is to estimate the number of unregistered opiate users directly related to the clients of the aid agencies. The outline of the article is as follows. In Section 2, a brief review of capture-recapture techniques for hidden population size estimation using administrative sources and estimation techniques for research contexts in which sampling frames are lacking is given.

Section 3 introduces the proposed B-graph sampling design. Because newly mentioned users will be rather sparse in most contexts, we focus on size estimators based on multiple-capture techniques for sparse data in Section 4 (Chao 1987; 1988; 1989; Zelterman 1988; Böhning 2010). Finally, Section 5 is concerned with the illustration, and the article ends with some concluding remarks.

2. Review of Literature on Estimating Hidden Populations

According to Böhning and van der Heijden (2009), capture-recapture methods are conventionally used to estimate the size of a hidden population when only (multiple) registration sources are available. In particular, the so-called Petersen-Lincoln (PL) estimator has been widely applied in animal studies, but nowadays this estimation technique is also employed in social studies where two registration sources are available (McCullough and Hirth 1998; Chao et al. 2008). The PL estimator is based on the number of n_1 units captured in Source 1, the number of n_2 units captured in Source 2, and the number of m_2 units captured in both sources. By assuming that the two sources are independent of each other, the units not captured in one of the sources can be estimated because the odds ratio is close to unity (Brittain and Böhning 2009).

$$\hat{N}_{PL} = \frac{n_1 n_2}{m_2} \tag{1}$$

The standard procedure for estimating the size of an animal population in a two-sample capture-recapture study is to capture a first sample, mark the captured animals and release them. Subsequently, a second sample is captured, and the number of animals captured in the first, the second and both samples is used to estimate the size of the population with the PL estimator (1). The standard procedure for estimating the size of a human population where two registration sources are available mirrors the animal population procedure by considering persons on the lists to be "marked". Like the trapping samples in animal studies, the number of persons on the first, the second, and both lists are used to employ the PL estimator (1). Examples of registration sources are hospitals, treatment centres, pharmacies, police registers, birth registers, and so on. The assumptions for producing valid estimates by capture-recapture methods are more or less identical in animal and in human population studies. According to Chao (2001), the validity of a capture-recapture estimator for animal populations depends on:

1. Demographic closure assumption: there is no birth, death, or migration, so that the population size is stable over trapping times;
2. Equal catchability assumption: all animals have the same capture probability in each sample, although the probability can be allowed to vary among samples.

To fulfil the first assumption, in animal studies data are collected during a relatively short time period. The second assumption refers to the independence of the samples. Dependence between samples can occur through local list dependence and unequal catchabilities (Chao et al. 2008). Local list dependence occurs whenever captured animals are easier or more difficult to capture by next samples as a consequence of their trapping history. Unequal catchability refers to the process that samples are dependent because their capture probabilities are heterogeneous (Chao 2001).

To produce valid estimators in human populations, the following assumptions must be met (Brittain and Böhning 2009):

1. Independence between registration sources or lists,
2. The population must be closed,
3. Independence between individuals.

In most empirical situations, these assumptions are violated. For instance, in drug abuse studies the registration sources of addiction centres and police registers are often combined to estimate the size of the number of drug users who are not registered. However, both data sources may have administration flaws. If arrested drug users are structurally assigned to certain addiction centres, Assumption 1 is violated. If there is also a high death or removal rate, Assumption 2 is violated. If certain ethnic groups of drug users are treated by the same institution, Assumption 3 is violated. There is a growing amount of literature on how to deal with these types of dependencies (see the special issues of the Biometric Journal, 2008, volume 50, the AStA Advances in Statistical Analysis, 2009, volume 93 and Journal of Official Statistics, volume 31).

In some empirical research contexts, registration sources are simply lacking or of such poor quality (for example, incomplete registration systems) that valid capture-recapture estimation is debatable. In such situations, link-tracing sampling procedures can be applied (Spreen 1992). Link-tracing designs use existing relational structures within the study population for sampling purposes. Up-to-date respondent-driven sampling (RDS) is the link-tracing procedure applied most frequently to estimate hidden populations sizes when (proper) sampling frames are lacking (Heckathorn 1997; Salganik and Heckathorn 2004; Volz and Heckathorn 2008). In RDS, the hidden trait to be estimated is viewed as a network phenomenon because it is assumed "that those best able to access members of hidden populations are their own peers" (Heckathorn 1997, 178). The sampling procedure starts with the recruitment of individuals (called "seeds") from the target population. This recruitment is nonrandom. The recruited individuals are offered dual incentives: they are financially rewarded for completing the interview and for recruiting other individuals (typically 3-5 persons) into the study. Subsequently, the newly recruited persons are asked to become recruiters themselves and are also rewarded financially. To estimate the size $\hat{y}$ of a hidden population, Volz and Heckathorn (2008) defined the RDS estimator (Formula 7, p. 85) as:

$$\hat{y} = \frac{1}{\sum_{i \in S} \frac{1}{d_i}} \sum_{i \in S} \frac{y_i}{d_i}, \tag{2}$$

where S is the set of all sampled persons and d_i the number of persons mentioned by i (degree).

The RDS estimator takes account of the network structure within the hidden population by weighing each interviewed respondent with the number of persons he or she is linked to in the network. These individual degree weights are assumed to be arbitrary positive inclusion probabilities which can be expanded to reach the level of the whole population (Särndal et al. 1992). According to Volz and Heckathorn (2008), it is usually prudent to exclude the initial recruits of the sample because they are not randomly found, although the estimator will be asymptotically unbiased.

Other link-tracing design-based estimators for hidden population sizes are the Frank and Snijders estimators (1994). Like RDS, their sampling design is based on the assumption that the population of interest can be viewed as a social network. In their theoretical (one-wave snowball) design, a random sample of n persons (vertices) is drawn from an unknown network and the selected persons are asked to mention other persons (their degree) they know in the network. Frank and Snijders propose the following estimator:

$$\hat{v}_{F-S} = \frac{(n-1)T_{01}}{T_{00}} + n, \tag{3}$$

where n is the size of the initial sample, T_{00} the number of times initial respondents mentioned each other, and T_{01} the number of times newly mentioned fellow hidden population members are mentioned by initial respondents. Estimator (3) can be understood in terms of capture-recapture, where capture is interpreted as drawn in the initial sample and recapture as mentioned by initial respondents. Frank and Snijders (1994) considered the initial sample to be a Bernoulli sample, which is not feasible in practical research. To relax this assumption, they recommend using some variant of targeted sampling (Watters and Biernacki 1989). To approximate a Bernoulli initial sample to a reasonable extent, Frank and Snijders (1994) recommend using several unrelated sources of well-defined social meeting places during the sampling phase. There are other examples of link-tracing designs in literature, such as multiple-wave snowball designs (Goodman 1961; Frank 1979), random-walk designs (Klovdahl 1989), and adaptive sampling designs (Thompson and Frank 2000), which we will not discuss.

RDS and the Frank-Snijders estimators are both elaborated for situations in which sampling frames are lacking. For situations in which various scattered sampling frames are available, B-graph sampling can be used.

3. B-Graph Sampling

Consider a hidden population in some well-defined geographic area for which it is assumed that its members know each other because of the hidden activity. For instance, a group of hooligans know each other because they operate as group against other groups of hooligans, drug users know each other for economic reasons (e.g., procuring drugs, knowing the market), terrorists know each other for political reasons, homeless people know each other from the street, and so forth.

Hidden populations are often registered by multiple administrative sources. For instance, a population of drug users may be registered as clients of a local drug-assistance institution but also as detainees by the police. In this situation, the Petersen-Lincoln estimator (1) for two sample closed experiments can be employed using both registration resources to estimate the number of unregistered drug users. Obviously, the quality of the estimate is dependent on different issues. For instance, administration flaws may render the accuracy of the registration systems too questionable to be valid for capture-recapture estimation. In such situations, one may consider a B-graph sampling procedure. A B-graph sampling design consists of the following steps. In Step 1, it is decided whether the hidden activity to be estimated leads to relations and/or administrative records by different institutions. Step 2 consists of collecting all available administrative records of all relevant institutions; all collected individual records are turned into one sampling frame and a local team of fieldworkers evaluate whether the persons on the list cover a substantial part of the population. Most of the time, local field workers have a good overview of their caseloads and neighbourhood (Heckathorn 1997). If the constructed sampling frame is considered to cover a substantial part of the population, the unknown total population can be considered as a bipartite graph (Figure 1).

For argument's sake, the four uncoloured vertices represent registered hidden population members pooled into one sampling frame from different sources, that is, sampling frame $\alpha = \{1, 2, 3, 4\}$. The unknown hidden populations members are coloured vertices, that is, subset $\beta = \{5, 6, 7\}$. Note that all coloured vertices have at least one link to an uncoloured vertex, that is, all unregistered hidden population members have a positive probability of being included in a sample when the registered hidden population members are asked to give their relations with unregistered hidden network members. In this article, the problem of estimating the number of unknown hidden population members (coloured vertices) is considered.

In Step 3 of a B-graph sample, a simple random sample S of s vertices from sampling frame α is drawn. Each sampled $i \in S$ is asked to mention his or her relations with other hidden population members according to a predefined inclusion criteria. As a result, a sample of subset β is observed. Throughout this article, we assume that this observation is without measurement error (each respondent completely discloses his contacts in the hidden network). The total number of observed distinct unregistered hidden population members (coloured vertices) in the final sample is denoted $m(S)$. The number of unregistered $u \in \beta$ mentioned exactly t times by the s selected registered hidden population members is denoted f_t, that is, $\sum_{t=1}^{s} f_t = m(S)$. As an illustration, consider Figure 2, in which a sample S of $s = 2$ uncoloured vertices from α is drawn from the bipartite graph of Figure 1. The selected uncoloured vertices are vertices 2 and 4.

In Figure 2, the total number of distinct vertices $u \in \beta$ observed is $m(S) = 2$, that is, vertices 6 and 7. Vertex 7 is involved two times with a vertex $i \in S$, while vertex 6 is involved one time, that is $f_2 = 1$ and $f_1 = 1$, respectively. Using this sample information, multiple capture-recapture estimators for the size of vertex set β can be employed.

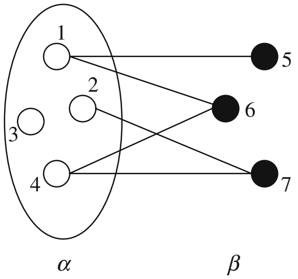

Fig. 1. Bipartite graph example

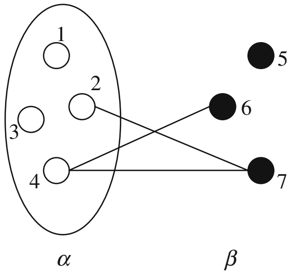

Fig. 2. Sample of bipartite graph

4. Multiple Capture-Recapture Estimators

Data collected from a B-graph sample can be understood as a multiple-capture sample in which each unregistered hidden population member (coloured vertex) captured via $i \in S$ is regarded as an independent trapping sample. Using this assumption, multiple-capture census estimators as discussed in Fienberg (1972), Bishop et al. (1988), and Cormack (1989; 1992) can be applied. However, the larger a population, the more sparse the total times unregistered hidden population members of subset β will be captured via registered members $i \in S$. Dependent on the sampling design and the assumptions about the population, various refinements of multiple-capture models have been introduced, especially for sparsely distributed animal populations. For a general review, we refer the reader to Seber (1986). The review article of Wilson and Collins (1992) merits special attention; it discusses the performance of 14 capture-recapture estimators. In this article, we discuss three capture-recapture estimators whose model assumptions are closely related to the assumptions of the proposed B-graph design: the moment estimator of Chao and a modified version of this estimator (Chao 1987; 1988; 1989) and the truncated Poisson estimator of Zelterman (1988).

Chao (1989) considered estimators for animal population size studies in which capture frequencies of the animals are low. In this study, we focus on the heterogeneity model-based estimator proposed by Chao (1989). This estimator has the following assumptions:

1. The animal population is closed, so there are no changes due to birth, death, emigration or immigration during the sampling period,
2. The probability of capturing an animal is independent of that animal's previous history,
3. Different animals are allowed to have different probabilities of capture.

The proposed B-graph design for human populations meets the assumptions of Chao's estimator. The 'closure' assumption refers to the definition of the inclusion criteria of the hidden population: who belongs to the population? To produce valid estimations, the definition of the hidden population must at least be bounded by strict relational, time and geographic criteria, that is, can you give me your friendly relations with people who have the same hidden variable in common as you, whom you have met during the last three months and who live in your town? The second assumption refers to the sampling procedure of the proposed B-graph design. The probability of an unregistered hidden population member $u \in \beta$ being mentioned by a registered hidden population member is independent of the previous capture history of $u \in \beta$. In the B-graph design, a simple random sample is drawn from the sampling frame. This implies that a multiple capture of an unregistered hidden population member is independent of the registered persons by whom he or she is mentioned. In Chao's terminology, the capture probability of unregistered $u \in \beta$ is independent of the sequence of the samples. The third assumption also applies to the proposed B-graph design. Each unregistered hidden population member $u \in \beta$ is assumed to have at least one contact with a registered hidden population member; this leads to positive inclusion probabilities for all $u \in \beta$ when drawing a sample from α. Accordingly, by random sampling from sampling frame α each $u \in \beta$ has a chance of being mentioned by an $i \in S$. However, different vertices have different probabilities of being mentioned, that is, the higher the degree of $u \in \beta$ in the total population, the higher the probability of being mentioned in the final sample.

For situations where s is not too small (≥ 5) and most unregistered hidden populations members are observed only one or two times, the following estimator of Chao (1988) can be employed:

$$\hat{m}_C = m(S) + \left[\frac{f_1^2}{2f_2}\right]. \tag{4}$$

Chao (1987, 1988) also proposed a biased-corrected version to correct for overestimation bias:

$$\tilde{m}_C = m(S) + \left[\frac{f_1(f_1 - 1)}{2(f_2 + 1)}\right] \tag{5}$$

The computation of the 95-percent confidence intervals of (4) and (5) are found in Chao (1989).

The idea behind Estimator (4) is that unregistered hidden population members of subset β with small capture probabilities (they have few relations in the network with members that are registered) are likely to be not mentioned (frequency class f_0) or only mentioned very few times by $i \in S$. This emphasis on the lower frequency classes makes Estimator (4) robust in the presence of heterogeneity. The influence of unregistered hidden population members mentioned very often is weighted down so that the presence of heterogeneity exercises a small influence on the estimate (Smit et al. 1997).

Based on the intuitive notion that 'individuals never seen are more similar to those rarely seen than those captured many times', Zelterman (1988, 227) formulated, independently of Chao, an estimator for the relative frequency of the unobservable zero class in a truncated Poisson distribution, that is,

$$\hat{m}_Z = \frac{m(S)}{1 - Q_1} \tag{6}$$

where $Q_1 = \exp\left[-2f_2/f_1\right]$.

The 95-percent confidence interval is given in Zelterman (1988).

Estimators (4) and (6) will produce about the same estimates, because both assume that the observed series of frequencies follows a Poisson distribution which is truncated below one (Smit et al. 1997). In a simulation study by Böhning (2010), in which the performance of Chao's estimator was compared with Zelterman's estimator, the author showed that the estimators are close if the ratio f_2/f_1 is small. He also showed that the biased-corrected estimator (5) of Chao performs best for small samples and small amounts of heterogeneity.

5. Illustration

To illustrate the B-graph sampling design, data from a social network study of the opiate-using population in the city of Utrecht, the Netherlands (Ten Den et al. 1995; Jansson and Spreen 1998) are used for secondary analyses. Utrecht is one of the largest cities (about 320,000 inhabitants) in the Netherlands and is geographically located in the middle of the country. At the time of the study, the opiate-using population in Utrecht caused a lot of nuisance for the general public, but there was also concern about specific health issues such as the relation between injecting drugs and contagious hepatitis, HIV, and sexually transmitted diseases. The goal of the study was to gain an insight into the nature of opiate use, such as types of users injecting drugs, lifestyles of opiate users, and so on. Another goal of this study was to gain insight in the total number of opiate users in Utrecht. Therefore several estimation techniques were used.

In Utrecht, local authorities managed several drug-assistance institutions that kept registration files of their clients, but worked more or less independently of each other. In the original study, the resulting sample of 101 opiate users was gathered by a random sample of 51 users from the registers of three drug-assistance organisations, by a convenience field work sample in which 37 users were found, and by a snowball sample in which 13 users were found via other users. Each interviewed opiate user was asked to mention other opiate users in Utrecht. Due to privacy reasons and to prevent a high rate of nonresponse, each opiate user was asked to give the first two letters of his or her first and family name, nickname, age, neighbourhood, and whether he or she was known as a client of the drug assistance by his or her fellow drug users. The identification of the respondents was done by a team of experienced field workers. Based on this sample, several estimation techniques were applied to estimate the prevalence of opiate users in Utrecht. It was possible to compute a Peterson-Lincoln estimate by using the registration files of the police and the largest drug-assistance organisation in Utrecht. The Petersen-Lincoln estimate was about 1,100 users. Furthermore, two extrapolation estimators (Smit et al. 1996) based on the registers of the largest drug-assistance organisation and the police were computed. Based on the first source, the estimate for the total population was about 1,000 users; for the second source (police data), the estimate was about 900 users. Finally, 69 users (51 of the random sample and 18 of the users found during field work) were evaluated as collected independently of each other, and served as the initial "random" sample for the Frank-Snijders estimators. Two network estimators of Frank and Snijders (1994) were reported (without standard errors) and resulted in estimates of 759 and 936 users. Finally, the researchers combined all different estimators and decided that the most likely estimate for the population size of the Utrecht opiate users population was about 950 users (Ten Den et al. 1995). The final report of Ten Den et al. (1995) did not provide the confidence intervals of the estimates.

To illustrate the B-graph sampling design, we were able to use the random sample of size 44 from the largest drug-assistance organisation. We call this the Regular Drug

Assistance (RDA). Note that our purpose is to estimate the number of opiate users who are not clients of the RDA but directly related to a user who is a client. In other words: how many opiate users in Utrecht are not known to the RDA but could be contacted via the RDA's clients? This is important information for the effectiveness of all kinds of health measures.

In Utrecht at the time of the study, 427 drug abusers were recorded as clients of the RDA, that is, $\alpha = \{1, 2 . . . , 427\}$. A simple random sample without replacement S of size $s = 44$ was drawn and each $i \in S$ was asked to mention his/her contacts with other opiate users. This way, a respondent could mention not only other opiate users already registered by the RDA but also opiate users who were not registered on the RDA list. For each mentioned opiate user, the respondent gave individual and identifying characteristics. The criteria for opiate users to be included in the sample were:

1. the mentioned opiate user is a resident of the city of Utrecht or resides in Utrecht at least (at a minimum of) four days a week;
2. the mentioned opiate user has used opiates a minimum of 25 times in the past six months;
3. the respondent and the mentioned opiate user must know each other by first and family name.

Of the 44 selected clients, six refused to provide information about their opiate-using contacts. The remaining 38 clients mentioned 98 other opiate users who were not on the RDA list, that is, $m(S) = 98$. The 38 respondents reported 107 relations with the 98 mentioned opiate users. As a result, the observed frequency distribution of the sampled B-graph was rather sparse (see Table 1).

In Table 2 the three multiple-capture estimates and their 95-percent confidence intervals are given.

As expected, the estimates of the Chao and Zelterman estimator are close to each other, 538 and 535 respectively, because the ratio f_2/f_1 is rather small. Taking into account the 95-percent confidence intervals of the model-based estimators, we observe some differences. The underlying assumptions of the Chao and Zelterman estimators applied to this specific study can be regarded as plausible. The population can be considered closed, because respondents report only other opiate users whom they know by name and live in Utrecht and the practical sample was done in a time frame of three months. The number to be estimated can be understood as the number of opiate users directly connected to the clients of the RDA. The probability of capturing unregistered opiate user k via registered opiate user i is independent of registered user h because i and h are randomly selected from the register. The probability of capturing an unregistered opiate user is dependent on his or her amount of contacts with registered opiate users. The 95-percent confidence regions are rather large, but this is characteristic for sparse frequency distributions. The confidence

Table 1. Capture frequency distribution of mentioned opiate users

F_t	1	2
Counts	89	9

Table 2. Results of different estimators of opiate using population directly linked to clients of the RDA-lists

Estimator	Lower bound	Point estimate	Upper bound
Chao	306	538	1,031
Chao modified	293	490	886
Zelterman	340	535	1,307

region of the Zelterman estimator in particular is known to produce anomalous values caused by small standard errors close to zero (Wilson and Collins 1992).

Following various simulation studies, Chao (1989) concluded that her proposed moment estimator performed best for sparse populations. Furthermore, in a simulation study by Wilson and Collins (1992), Chao's estimator performed best in heterogeneous populations. Böhning (2010) showed in a simulation study that Chao's modified estimator performs best for small samples and small amounts of heterogeneity. In Table 2, the modified estimator has a smaller variance than the other two. However, these simulation results are based on slightly different sampling schemes. Based on the three estimators, we may conclude that a reasonable estimate of the number of opiate users directly linked to RDA opiate users in Utrecht is in the range of $500-550$. Compared to the estimations of the population size from the original study, the B-graph sampling design gives comparable point estimates ($500+427=927$; $550+427=977$; $490+427=917$), implying that the proportion of opiate users in Utrecht who are at a social distance of Step 2 from clients of the RDA (they know clients only via unregistered opiate users) is probably very small.

6. Discussion

In studies of hidden populations, sampling frames are often lacking, but sometimes the nature of the hidden trait will lead to the emergence of networks. In such research situations, Frank and Snijders (1994) proposed estimators that can be applied when one may assume an initial sample of individuals found independently of one another that resembles a random sample of the total network. Heckathorn (1997) elaborated RDS in which the recruitment of respondents is done by respondents, showing that "RDS produces samples that are independent of the initial subjects from which sampling begins" (p. 176). However, often partial sampling frames are available in studies of hidden populations. In this article, an alternative sampling design is introduced that makes use of the partial sampling frames by pooling them into one sampling frame. If this sampling frame is considered to cover a substantial part of the unknown hidden population by the local experts, one may draw a random sample of this sampling frame, asked the respondents who they know in the hidden population and estimate the number of persons who are not on the sampling frame. This proposed B-graph sampling design has some challenging features for hidden population research. First, in a lot of studies it is often interesting to know how many people with hidden activities are directly related to the registered group of known people. By random sampling from the registered population and application of the B-graph design, each member of the unknown directly related population has a chance to be in the sample. For instance, if a health organisation wants to know how many other possible "future" clients they can reach via their own clients for health education purposes, the B-graph design can be used. This way, "recruit" markets of criminal organisations, radicals, youth or street gangs, or networks of paedophiles can also be estimated. Furthermore, if capture-recapture estimates based on administrative sources are possible, a comparison can be made, revealing the size of the proportion of that part of the populations that is very difficult for institutions to reach. Another advantage of the B-graph design is that more qualitative information about the population of interest is collected, such as the quality of relations, lifestyles, and so on.

The B-graph sampling design can only be applied to populations with a network structure; the hidden activity must lead to network formation. As with capture-recapture or RDS studies, the practical problem of accurately identifying population members also remains for the B-graph design. Selected members have to disclose their relations. This is not a straightforward activity. Network members can be identified by a number of characteristics, such as the first two or three letters of first and family name, sex, age, neighbourhood, and so on. Reasons to work with identification variables are often to protect the privacy of users but also to reduce nonresponse. However, the remark of Chao et al. (2008, 957) for animal size studies also applies to human population size studies:

> "Careful sampling with proper marking (identifying) can provide more accurate estimates about the population size than an incomplete census."

7. References

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1988. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA, and London: The MIT Press.

Bogaerts, S. and A. Daalder. 2011. "Measuring Childhood Abuse and Neglect in a Group of Female Indoor Sex Workers in the Netherlands. A Confirmatory Factor Analysis of the Dutch Version of the Childhood Trauma Questionnaire-Short Form." *Psychological Reports* 108: 856–860. Doi: http://dx.doi.org/10.2466/02.10.13.16.PR0.108.3.856-860.

Böhning, D. 2010. "Some General Comparative Points on Chao's and Zelterman's Estimators of the Population Size." *Scandinavian Journal of Statistics* 37: 221–236. Doi: http://dx.doi.org/10.1111/j.1467-9469.2009.00676.x.

Böhning, D. and P. van der Heijden. 2009. "Recent Developments in Life and Social Science Applications of Capture-Recapture Methods." *AStA Advances in Statistical Analysis* 93: 1–3. Doi: http://dx.doi.org/10.1007/s10182-008-0097-7.

Brittain, S. and D. Böhning. 2009. "Estimators in Capture-Recapture Studies With Two Sources." *AStA Advances in Statistical Analysis* 93: 23–47. Doi: http://dx.doi.org/10.1007/s10182-008-0085-y.

Brugal, M.T., A. Domingo-Salvany, A. Maguire, J.A. Cayla, J.R. Villalbi, and R. Hartnoll. 1999. "A Small Area Analysis Estimating the Prevalence of Addiction to Opioids in Barcelona." *Journal of Epidemiology and Community Health* 53: 488–494. Doi: http://dx.doi.org/10.1136/jech.53.8.488.

Chao, A. 1987. "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability." *Biometrics* 43: 783–791. Doi: http://dx.doi.org/10.2307/2531532.

Chao, A. 1988. "Estimating Animal Abundance with Capture Frequency Data." *Journal of Wildlife Management* 52: 295–300. Doi: http://dx.doi.org/10.2307/3801237.

Chao, A. 1989. "Estimating Population Size for Sparse Data in Capture-Recapture Experiments." *Biometrics* 45: 427–438. Doi: http://dx.doi.org/10.2307/2531487.

Chao, A. 2001. "An Overview of Closed Capture-Recapture Models." *Journal of Agricultural, Biological, and Environmental Statistics* 6: 158–175. Doi: http://dx.doi.org/10.1198/108571101750524670.

Chao, A., H.Y. Pan, and S.C. Chiang. 2008. "The Petersen-Lincoln Estimator and its Extension to Estimate the Size of Shared Population." *Biometrical Journal* 50: 957–970. Doi: http://dx.doi.org/10.1002/bimj.200810482.

Cormack, R.M. 1989. "Log-Linear Models for Capture-Recapture." *Biometrics* 45: 395–413. Doi: http://dx.doi.org/10.2307/2531485.

Cormack, R.M. 1992. "Interval Estimation for Mark-Recapture Studies of Closed Populations." *Biometrics* 48: 567–576. Doi: http://dx.doi.org/10.2307/2532310.

Ten Den, C., B. Bieleman, E. De Bie, and J. Snippe. 1995. *Pijn in het hart*. Groningen and Rotterdam: Intraval.

Fienberg, S.E. 1972. "The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables." *Biometrika* 59: 591–603. Doi: http://dx.doi.org/10.1093/biomet/59.3.591.

Frank, O. 1979. "Estimation of Population Totals by Use of Snowball Samples." In *Perspectives on Social Network Research*, edited by P.W. Holland and S. Leinhardt, 319–348. New York: Academic Press.

Frank, O. and T.A.B. Snijders. 1994. "Estimating the Size of Hidden Populations Using Snowball Sampling." *Journal of Official Statistics* 10: 53–67.

Goodman, L.A. 1961. "Snowball Sampling." *Annals of Mathematical Statistics* 32: 148–170.

Heckathorn, D.D. 1997. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Social Problems* 44: 174–199. Doi: http://dx.doi.org/10.2307/3096941.

Holland, R., R. Vivancos, V. Maskrey, J. Sadler, D. Rumball, I. Harvey, and L. Swift. 2006. "The Prevalence of Problem Drug Misuse in a Rural County of England." *Journal of Public Health* 28: 88–95. Doi: http://dx.doi.org/10.1093/pubmed/fdl009.

Jansson, I. and M. Spreen. 1998. "The Use of Local Networks in a Study of Heroin Users: Assessing Average Local Networks." *Bulletin de Méthodologie Sociologique* 59: 49–61. Doi: http://dx.doi.org/10.1177/075910639805900105.

Klovdahl, A.S. 1989. "Urban Social Networks: Some Methodological Problems and Possibilities." In *The Small World*, edited by M. Kochen, 176–210. Norwood, NJ: Ablex.

Kunst, M.J.J., F.W. Winkel, and S. Bogaerts. 2010. "Prevalence and Predictors of Posttraumatic Stress Disorder Among Victims of Violence Applying for State Compensation." *Journal of Interpersonal Violence* 2010: 1631–1654. Doi: http://dx.doi.org/10.1177/0886260509354591.

McCullough, D.R. and D.H. Hirth. 1998. "Evaluation of the Petersen-Lincoln Estimator for a White-tailed Deer Population." *Journal of Wildlife Management* 52: 534–544. Doi: http://dx.doi.org/10.2307/3801606.

Palusci, V.J., S.J. Wirtz, and T.M. Covington. 2010. "Using Capture-Recapture Methods to Better Ascertain the Incidence of Fatal Child Maltreatment." *Child Abuse & Neglect* 34: 396–402. Doi: http://dx.doi.org/10.1016/j.chiabu.2009.11.002.

Salganik, M.J. and D.D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34: 193–239. Doi: http://dx.doi.org/10.1111/j.0081-1750.2004.00152.x.

Särndal, B.E., B. Swensson, and J.H. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Seber, G.A.F. 1986. "A Review of Estimating Animal Abundance." *Biometrics* 42: 267–292. Doi: http://dx.doi.org/10.2307/2531049.

Smit, F., W. Brunenberg, and P. van der Heijden. 1996. "Het Schatten van Populatiegroottes." *Tijdschrift voor Sociale Gezondheidszorg* 74: 171–176.

Smit, F., J. Toet, and P. van der Heijden. 1997. "Estimating the Number of Opiate Users in Rotterdam Using Statistical Models for Incomplete Count Data." *In European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) Methodological Pilot Study of Local Prevalence Estimates*, Lisbon: EMCDDA.

Spreen, M. 1992. "Populations, Hidden Populations, and Link-Tracing Designs: What and Why?" *Bulletin de Méthodologie Sociologique* 36: 34–58. Doi: http://dx.doi.org/10.1177/075910639203600103.

Surjadi, B., J. van Horn, S. Bogaerts, and R. Bullens. 2010. "Internet Offending: Sexual and Non-Sexual Functions within a Dutch Sample." *Journal of Sexual Aggression* 16: 47–58. Doi: http://dx.doi.org/10.1080/13552600903470054.

Thompson, S.K. and O. Frank. 2000. "Model-Based Estimation with Link-Tracing Sampling Designs." *Survey Methodology* 26: 87–98.

Volz, E. and D.D. Heckathorn. 2008. "Probability Based Estimation Theory for Respondent Driven Sampling." *Journal of Official Statistics* 24: 79–97.

Watters, J.K. and P. Biernacki. 1989. "Targeted Sampling: Options for the Study of Hidden Populations." *Social Problems* 36: 416–430. Doi: http://dx.doi.org/10.2307/800824.

Wilson, R.M. and M.F. Collins. 1992. "Capture-Recapture Estimation with Samples of Size One Using Frequency Data." *Biometrika* 79: 543–553. Doi: http://dx.doi.org/10.1093/biomet/79.3.543.

Zelterman, D. 1988. "Robust Estimation in Truncated Discrete Distributions with Application to Capture-Recapture Experiments." *Journal of Statistical Planning and Inference* 18: 225–237. Doi: http://dx.doi.org/10.1016/0378-3758(88)90007-9.

Received September 2013
Revised November 2014
Accepted January 2015