

The Petersen–Lincoln Estimator and its Extension to Estimate the Size of a Shared Population

Anne Chao^{*1}, H.-Y. Pan², and Shu-Chuan Chiang³

¹ Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan 30043

² Department of Applied Mathematics, National Chia-Yi University, Chia-Yi, Taiwan 60004

³ Institute of Public Health & Division of Preventive Medicine, National Yang-Ming University, Taipei, Taiwan, 11221; Ching Yun University, Jung-Li, Taiwan 32097; Sunrise Clinic, Taoyuan, Taiwan 32041

Received 4 April 2008, revised 4 June 2008, accepted 30 July 2008

Summary

The Petersen–Lincoln estimator has been used to estimate the size of a population in a single mark release experiment. However, the estimator is not valid when the capture sample and recapture sample are not independent. We provide an intuitive interpretation for “independence” between samples based on 2×2 categorical data formed by capture/non-capture in each of the two samples. From the interpretation, we review a general measure of “dependence” and quantify the correlation bias of the Petersen–Lincoln estimator when two types of dependences (local list dependence and heterogeneity of capture probability) exist. An important implication in the census undercount problem is that instead of using a post enumeration sample to assess the undercount of a census, one should conduct a prior enumeration sample to avoid correlation bias. We extend the Petersen–Lincoln method to the case of two populations. This new estimator of the size of the shared population is proposed and its variance is derived. We discuss a special case where the correlation bias of the proposed estimator due to dependence between samples vanishes. The proposed method is applied to a study of the relapse rate of illicit drug use in Taiwan.

Key words: Capture-recapture; Correlation bias; Dependence; Dual-system Estimator; Heterogeneity, Mark-recapture.

1 Introduction

The Petersen–Lincoln estimator in animal populations (or the dual-system estimator in human populations) has been widely used to estimate population size in two-sample closed capture-recapture experiments. Seber (1982) indicated that the idea of the two-sample capture-recapture technique can be traced back to a 1786 paper by Pierre Simmon Laplace. Stigler (1986) noted that Laplace applied the method in 1802 to estimate the human population of France. Laplace’s method (Stigler, 1986; Cochran, 1978) was based on a new-birth register in France and a complete census (including new births) of a few selected communities. Thus, the population size of France could be estimated. Laplace also provided a variance of his estimate. Stephan (1948) made an enlightening comment about Laplace’s method. “Even though the method of estimate was crude and the measure of precision not wholly valid, Laplace’s effort was much more successful than the complete census of France that was attempted at the same time.” This is an important motivation for capture-recapture studies; i.e., careful sampling with proper marking can provide more accurate estimate about the population size than an incomplete census.

* Corresponding author: e-mail: chao@stat.nthu.edu.tw, Phone: +886 3571 5131 ext 33161, Fax: +886 3572 8318

The idea of capture-recapture was later applied to biology and ecology. White et al. (1982) described the significant contribution of the two pioneers: Carl George Johannes Petersen and Frederick Lincoln. Petersen was a fishery biologist. He invented a brass tag that could be attached to mark fish in order to study their migrations. In his 1896 work on plaice, Petersen recognized that the proportion of his marked fish caught by fisherman constituted a basis for estimating population size. This work represented a landmark in the historical development of capture-recapture models. Lincoln conducted a life-long study of birds and was in charge of bird banding work for the Biological Survey (now U.S. Fish and Wildlife Service) in the 1920's. His 1930 paper on estimating waterfowl abundance based on returned bands collected from shooters founded the banded recovery models.

The Petersen–Lincoln estimator for animal populations can be described as follows. Assume the size of a population is N . A first sample of n_1 animals is captured, marked and released back to the population. Thus the marked rate in the population is n_1/N . A second sample of n_2 animals is subsequently drawn and there are m_2 previously marked. Equating the proportion of the marked rate in the population to the marked rate in the second sample suggests that $m_2/n_2 \approx n_1/N$, which yields the following Petersen–Lincoln estimator for the population size:

$$\hat{N} = n_1 n_2 / m_2. \quad (1.1)$$

Based on a hypergeometric model (in which n_1 and n_2 are regarded as fixed), Chapman (1951) derived the following estimator to adjust the bias that arises mainly due to small value of m_2 :

$$\tilde{N} = (n_1 + 1)(n_2 + 1)/(m_2 + 1) - 1. \quad (1.2)$$

Under the same model, both estimators have approximately the same variance given by

$$(n_1 + 1)(n_2 + 1)(n_1 - m_2)(n_2 - m_2)/[(m_2 + 1)^2(m_2 + 2)]. \quad (1.3)$$

A critical assumption for the validity of the Petersen–Lincoln and Chapman estimators is that the two samples are independent. In animal studies, a more restrictive assumption is the “equal-catchability assumption”; i.e., in each fixed sample all animals have the same probability of being caught. Seber (1982, Chapter 3) thoroughly discussed the validity of the Petersen–Lincoln estimator using various models. In Section 2.1, we provide an intuitive approach to the concept of independence between samples based on 2×2 categorical data formed by capture/non-capture in each of the two samples.

Local list dependence and unequal catchabilities are two sources of dependences that lead to bias for the Petersen–Lincoln and Chapman's estimators. This type of bias is referred to as “correlation bias” in the literature. In Section 2.2, a general measure of “dependence” is defined based on our intuitive interpretation of independence. In Section 2.3, we use the measure of dependence to quantify the correlation bias in some dependent models. We also review some relevant results on correlation bias.

This paper considers an extension of the Petersen–Lincoln method. Assume that there are two populations with some individuals in common. Suppose a two-sample capture-recapture experiment is conducted in each of the two populations. Our purpose is to estimate the number of individuals in the intersection of the two populations (i.e., the size of the shared population). In Section 3, we formulate a Petersen–Lincoln type estimator for the estimating target and derive its variance and correlation bias. Under some specific conditions, we can show that the correlation bias vanishes. In Section 4, we apply the proposed estimator to estimate the number of drug users in a recidivist population. The estimate is helpful in understanding the rate of relapse or recidivism in drug use. Concluding remarks and discussion are provided in Section 5.

2 Correlation Bias

2.1 An intuitive concept of “independence”

Since the concept of “independence” in capture-recapture is important, we provide an intuitive interpretation in this subsection. First, we summarize the capture-recapture data as a 2×2 table formed by

Table 1 A 2×2 categorical table for capture-recapture data.

	Sample 2 = 1	Sample 2 = 0	Total
Sample 1 = 1	m_2	$n_1 - m_2$	n_1
Sample 1 = 0	$n_2 - m_2$	–	$N - n_1$
Total	n_2	$N - n_2$	N

Table 2 The cell probabilities for the i -th individual.

	Sample 2 = 1	Sample 2 = 0	Row total
Sample 1 = 1	$E(X_{i1}X_{i2})$	$E[X_{i1}(1 - X_{i2})]$	$E(X_{i1})$
Sample 1 = 0	$E[(1 - X_{i1})X_{i2}]$	$E[(1 - X_{i1})(1 - X_{i2})]$	$E(1 - X_{i1})$
Column total	$E(X_{i2})$	$E(1 - X_{i2})$	1

two samples (Sample 1 and Sample 2); there are two categories in each sample (capture or non-capture). For notational simplicity, denote capture and non-capture by 1 and 0, respectively. Then the two-sample data can be expressed as in Table 1 (where N , n_1 , n_2 and m_2 are defined in the Introduction) with one missing cell.

We first formulate our model. Define $X_{ij} = I$ [the i -th individual is captured in Sample j], $i = 1, 2, \dots, N$ and $j = 1, 2$, where $I[\cdot]$ denotes an indicator function. The capture history of each individual is classified into one of the four categories in Table 1. Thus the capture probability of the i -th individual in Sample j is $E(X_{ij})$, and the probability of being caught in both samples is $E(X_{i1}X_{i2})$. The corresponding cell probabilities of the four categories for the i -th animal can be expressed in terms of the expectations as given in Table 2.

For each $i = 1, 2, \dots, N$, we construct a table as Table 2 and there are a total of N tables. Averaging the corresponding cell probabilities over the N tables, we obtain Table 3, where $\mu_{11} = N^{-1} \sum E(X_{i1}X_{i2})$ denotes the *average* capture probability that any animal is caught in both samples. Similar interpretations hold for μ_{10} , μ_{01} and μ_{00} . In the same table, we have μ_1 and μ_2 as the *average* capture probability of being captured in Sample 1 and Sample 2 respectively. In our model, n_1 and n_2 are random variables with $n_j = \sum_{i=1}^N I(X_{ij} = 1)$ and $E(n_j) = N\mu_j$, $j = 1, 2$.

Table 3 The average cell probabilities and average marginal probabilities.

	Sample 2 = 1	Sample 2 = 0	Row total
Sample 1 = 1	$\mu_{11} = N^{-1} \sum_{i=1}^N E(X_{i1}X_{i2})$ $= N^{-1}E(m_2)$	$\mu_{10} = N^{-1} \sum_{i=1}^N E[X_{i1}(1 - X_{i2})]$ $= N^{-1}E(n_1 - m_2)$	$\mu_1 = N^{-1} \sum_{i=1}^N E(X_{i1})$
Sample 1 = 0	$\mu_{01} = N^{-1} \sum_{i=1}^N E[(1 - X_{i1})X_{i2}]$ $= N^{-1}E(n_2 - m_2)$	$\mu_{00} = N^{-1} \sum_{i=1}^N E[(1 - X_{i1})(1 - X_{i2})]$	$1 - \mu_1$ $= 1 - N^{-1} \sum_{i=1}^N E(X_{i1})$
Column total	$\mu_2 = N^{-1} \sum_{i=1}^N E(X_{i2})$	$1 - \mu_2 = 1 - N^{-1} \sum_{i=1}^N E(X_{i2})$	1

Based on the average cell and marginal probabilities in Table 3, the concept of independence can be intuitively interpreted in the following two equivalent ways:

- (a) The marked rate in the second sample (i.e., μ_{11}/μ_2) is equal to the population marked rate (i.e., μ_1). In terms of the data in Table 1, one should have that the sample fraction of tagged individuals is representative of the population fraction of tagged individuals; i.e., $m_2/n_2 \approx n_1/N$.
- (b) In the second sample, the capture rate for marked animals (i.e., μ_{11}/μ_1) is equal to that for animals randomly selected from the population (i.e., μ_2). Namely, animals captured in Sample 1 have the same probability of being captured in Sample 2 as are animals randomly selected from the population. In other words, marked and unmarked have the same rate of being captured in Sample 2. Under this assumption, we have $m_2/n_1 \approx n_2/N$.

Both interpretations of “independence” imply exactly the same condition that $\mu_{11} = \mu_1\mu_2$, that is, the average probability of being captured in both samples is equal to the product of two average marginal probabilities. Note that the four conditions: $\mu_{11} = \mu_1\mu_2$, $\mu_{10} = \mu_1(1 - \mu_2)$, $\mu_{01} = (1 - \mu_1)\mu_2$ and $\mu_{00} = (1 - \mu_1)(1 - \mu_2)$ are equivalent in the sense that any one of the four equalities implies the other three. Thus one condition (we use $\mu_{11} = \mu_1\mu_2$) is sufficient to define independence, although there are four categories.

The classic assumption to ensure independence in animal applications is called the “equal-catchability” assumption. This means that all animals have the same capture probabilities in any sample. Under this assumption, we see from (b) that the two samples are independent. However, this assumption is too restrictive. As will be rigorously discussed in Section 2.3, “equal-catchability” assumption for the second sample suffices and there are situations where samples are independent but the capture probabilities are heterogeneous.

2.2 A measure of dependence

From the two interpretations of independence, it is clear that a measure of dependence should be a function in terms of $\mu_{11} - \mu_1\mu_2$ or $\mu_{11}/(\mu_1\mu_2)$. We define a dependence measure, the coefficient of covariation (CCV), between Sample 1 and Sample 2 as

$$\begin{aligned} \gamma_{12} &= \frac{\mu_{11}}{\mu_1\mu_2} - 1 = \frac{N \sum_{i=1}^N E(X_{i1}X_{i2})}{[\sum_{i=1}^N E(X_{i1})][\sum_{i=1}^N E(X_{i2})]} - 1 \\ &= \frac{N \sum_{i=1}^N E[(X_{i1} - \mu_1)(X_{i2} - \mu_2)]}{[\sum_{i=1}^N E(X_{i1})][\sum_{i=1}^N E(X_{i2})]}. \end{aligned} \quad (2.1)$$

A generalized CCV for multiple samples is similarly constructed in Chao and Tsay (1998) and Chao et al. (2001). From (2.1), the magnitude of γ_{12} measures the degree of dependence between the two samples. The two samples are independent if and only if $\gamma_{12} = 0$; the two samples are positively (negatively) dependent if $\gamma_{12} > 0$ ($\gamma_{12} < 0$). Compared to the previous independence interpretations (a) and (b), a positively dependent sample is interpreted in the following two ways:

- (a*) the marked rate in the sample is larger than the marked rate in the population; i.e., $\mu_{11}/\mu_2 > \mu_1$.
- (b*) In Sample 2, the marked animals have higher probability of being captured than animals that are randomly selected from the population; i.e., $\mu_{11}/\mu_1 > \mu_2$.

We can similarly interpret negative dependence.

If the two samples are positively correlated, then from (a*) we have $\mu_{11}/\mu_2 > \mu_1$ in the population. Data would show that $m_2/n_2 > n_1/N$, yielding $N > n_1n_2/m_2 = \hat{N}$. Similarly, from (b*) data would show that $m_2/n_1 > n_2/N$, giving $N > n_1n_2/m_2$. Both interpretations conclude that the Petersen–Lincoln estimator underestimates. Conversely, it overestimates for negatively dependent samples. See Section 2.3 for the magnitude of the bias.

Correlation bias may arise from:

- (1) The capture probability of an individual in one sample depending on its previous capture history, known as local dependence (also called local list dependence or simply list dependence).

For example, marking or trapping affects future capture probability and thus animals may become either trap happy or trap shy because of a behavioral response to capture usually due to identical trapping methods. Local independence implies that for any given individual i , $E(X_{i1}X_{i2}) = E(X_{i1})E(X_{i2})$.

- (2) Heterogeneity among individuals. Even if the two samples are *locally* independent for any given individual, they may become *globally* dependent if capture probabilities are heterogeneous. A simple example is given in Section 2.3. Individual capture probabilities may vary with age, gender, environmental factors, weight (for animal populations), socio-economic status (for human populations), or other unobserved individual characteristics and thus may be heterogeneous.

These two types of dependence are confounded and cannot be separated in a data analysis unless more information is available.

2.3 Quantifying correlation bias for some models

2.3.1 A general model with both list dependence and heterogeneity

What is the correlation bias for the Petersen–Lincoln estimator if samples are dependent? It follows from the definition of the dependence measure that

$$N = \frac{E(n_1)E(n_2)}{E(m_2)}(1 + \gamma_{12}) \approx E(\hat{N})(1 + \gamma_{12}). \quad (2.2)$$

Here we use a first-order approximation that $E(n_1n_2/m_2) \approx E(n_1)E(n_2)/E(m_2)$. It is asymptotically valid when N tends to be large and capture probabilities are fixed. From (2.2), an approximate correlation bias of the Petersen–Lincoln estimator is

$$\text{Correlation bias} = E(\hat{N}) - N \approx -\gamma_{12}E(\hat{N}). \quad (2.3)$$

Chao and Tsay (1998) derived the above conclusion in a general model with both local dependence and heterogeneity. When there is only heterogeneity, Eq. (2.3) was discovered by Seker and Deming (1949) and Cormack (1966).

The magnitude of the correlation bias is proportional to the extent of dependence. Negative (positive) bias arises for positively (negatively) dependent samples, which theoretically justifies the previous intuitive arguments.

From (2.1) and (2.3), *no correlation bias arises if the equal-catchability assumption holds for the second sample*. This condition is usually rephrased as “the second sample is a representative sample” or “the second sample is a random sample”. If all animals have the same catchability in the second sample, which implies that marked and unmarked animals have equal capture probabilities, then capture in the second sample is independent of marking status. That is, $E(X_{i2} | X_{i1}) = E(X_{i2})$, implying no local dependence. Then it follows from (2.1) that the CCV is zero if $E(X_{i2})$ for all i are identical. In this case, no correlation bias arises even if the first sample is highly selective or extremely heterogeneous. This can be understood intuitively. For example, we may only tag large fish in the first sample, but fish of any size are captured in an equally likely manner in the second sample. Then marked and unmarked fish have identical chances of being caught in the recapture sample, which justifies the use of the Petersen–Lincoln estimator.

On the other hand, when the equal-catchability assumption holds for the first sample, the CCV does not necessarily vanish because local dependence may exist. There is a sequential (temporal) ordering for the two samples, thus it is not meaningful to consider $E(X_{i1} | X_{i2})$ and the condition $E(X_{i1} | X_{i2}) = E(X_{i1})$ is meaningless. *When there is no local dependence, heterogeneity for only one sample would not cause correlation bias*. An equal-catchability in the first sample means that the tag rates are nearly the same in any groups of different sizes. As long as tagging does not affect recapture

sampling, even if we only recapture large fish, the fraction of tagged fish in the recapture sample is still approximately equal to the fraction tagged in the whole population, which validates the use of the Petersen–Lincoln estimator. Related implication and discussion for the application to the census undercount estimation are given in Section 5.

2.3.2 Behavioral response model

We consider a special model with local dependence, a two-sample behavioral response model (model \mathbf{M}_b). This model assumes that in the first sample, the capture probability is p for all animals. In the second sample, the probability of capturing a previously uncaptured individual is p but capturing a previously captured individual becomes $c = \phi p$. Then $\mu_{11} = pc$, $\mu_{10} = p(1 - c)$, $\mu_{01} = (1 - p)p$. Consequently, $\mu_1 = p$, $\mu_2 = (1 - p + c)p$. From (2.1), the CCV can be computed as $\gamma_{12} = (\phi - 1)(1 - p)/[(1 - p) + \phi p]$. Thus the two samples are positively dependent if $\phi > 1$ (trap-happy), and negatively dependent if $\phi < 1$ (trap-shy). From (2.3), the Petersen–Lincoln estimator is biased downwards in trap-happy cases whereas it is biased upwards in trap-shy cases. This is a well-known general result in the multiple capture-recapture experiments (e.g., see O’Brien et al., 1985; Pollock et al., 1990). Our formula in (2.3) provides the magnitude of the bias with $\gamma_{12} = (\phi - 1)(1 - p)/[(1 - p) + \phi p]$.

2.3.3 Fixed-effect heterogeneous model

Hook and Regal (1993) provided an interesting example in epidemiological applications to show why heterogeneity may cause dependence. Here we give a simple example in terms of the average capture probabilities in Table 3. Assume that an animal population is half male and half female. The cell probabilities for each male or female are shown in Table 4. Conditional on each individual (male or female), it is readily seen that the two samples are locally independent. However, in the averaged table, the samples are clearly dependent. This phenomenon is similar to Simpson’s paradox in categorical data analysis. Namely, aggregating two independent 2×2 tables may result in a dependent table. In this example, a stratified analysis by gender is suggested. Data for males and females should be treated separately as discussed in Section 4 for drug use data.

Consider a case where there is no local dependence and only heterogeneity is present. The cell probabilities μ_{11} , μ_{10} and μ_{01} in Table 3 are reduced to $\mu_{11} = N^{-1} \sum_{i=1}^N E(X_{i1}) E(X_{i2}) = N^{-1} \sum_{i=1}^N P_{i1} P_{i2}$ where $P_{ij} = E(X_{ij})$ is the probability of capturing animal i in Sample j . The two marginal average probabilities become $\mu_1 = N^{-1} \sum_{i=1}^N P_{i1}$ and $\mu_2 = N^{-1} \sum_{i=1}^N P_{i2}$. The dependence measure CCV from (2.1) is simplified to

$$\gamma_{12} = \frac{1}{N} \frac{\sum_{i=1}^N P_{i1} P_{i2}}{\mu_1 \mu_2} - 1 = \frac{1}{N} \sum_{i=1}^N \frac{(P_{i1} - \mu_1)(P_{i2} - \mu_2)}{\mu_1 \mu_2} . \tag{2.4}$$

Thus, the two heterogeneous samples are independent if and only if the covariance between the two sets of probabilities, $\{P_{i1}; i = 1, 2, \dots, N\}$ and $\{P_{i2}; i = 1, 2, \dots, N\}$, is zero. Therefore, heterogeneity

Table 4 An example of cell probabilities to induce dependence by heterogeneity.

	Male		Female		Average	
	Sample 2 = 1	Sample 2 = 0	Sample 2 = 1	Sample 2 = 0	Sample 2 = 1	Sample 2 = 0
Sample 1 = 1	1/4	1/4	1/12	1/4	1/6	1/4
Sample 1 = 0	1/4	1/4	1/6	1/2	5/24	9/24

for only one sample would not cause correlation bias when there is no local dependence. The two samples are positively (negatively) dependent if the covariance is positive (negative). In the special case that $P_{i1} = P_{i2} \equiv P_i$, which is a special case of what is referred to as Model M_h in the capture-recapture literature, the CCV reduces to $[\sum_{i=1}^N (P_i - \bar{P})^2 / N] / \bar{P}^2$, the square of the coefficient of variation (CV) of $\{P_1, P_2, \dots, P_N\}$, where $\bar{P} = N^{-1} \sum_{i=1}^N P_i$. As a result, any two samples must be positively correlated under Model M_h . The parameter CV has been used to measure the extent of heterogeneity in ecological models (Lee and Chao, 1994) and now we have shown this is also closely related to the CCV of any two samples.

2.3.4 Random-effect heterogeneous model

Some authors have found a random-effect model more appealing for heterogeneous populations. As in Section 2.3.3, we define P_{ij} as the probability of capturing animal i in Sample j . A random-effect model assumes that $\{(P_{11}, P_{12}), (P_{21}, P_{22}), \dots, (P_{N1}, P_{N2})\}$ are a random sample from a two-dimensional distribution $F_{P_1, P_2}(p_1, p_2)$. The CCV in this case becomes

$$\gamma_{12} = \frac{E(P_1 P_2)}{\mu_1 \mu_2} - 1 = \frac{E[(P_1 - \mu_1)(P_2 - \mu_2)]}{\mu_1 \mu_2} = \frac{\text{cov}(P_1, P_2)}{\mu_1 \mu_2}, \quad (2.5)$$

where $\mu_j = E(P_j)$ denotes the average capture probability for the j -th sample.

Under a random-effect model, samples are independent if the covariance between the variables, P_1 and P_2 , is 0. Researchers in fishery sciences have suggested that correlation bias due to heterogeneity could be reduced if two different sampling schemes were used (e.g., trapping and then resighting, or netting and then angling) mainly because there is almost no covariance between the distributions for two distinct sampling methods. This was justified by Seber (1982). It also could be seen from Eq. (2.5).

In the two-sample capture-recapture data, only three cells are observable (Table 1). However, there are four parameters: N , two mean capture probabilities and a dependence measure. The data are insufficient for estimating dependence unless additional covariates are available. All existing methods unavoidably encounter this problem and adopt the independence assumption. This independence assumption has become the main weak point for the two-sample capture-recapture method. We have shown from the above discussions for models in sections 2.3.1–2.3.4 that the classical assumption of equal-catchability in each sample can be relaxed for the validity of the Petersen–Lincoln estimator. In some special cases of heterogeneous capture probabilities, the use of the Petersen–Lincoln estimator is justified because the correlation bias vanishes.

3 Extension to Estimate the Size of a Shared Population

Assume that there are two populations (Population A with size N_1 and Population B with size N_2) and that there are N_{12} individuals common to the two populations. The population of interest is the intersection of Populations A and B (i.e., the shared population) as shown in Figure 1. Without loss of generality, we assume the individuals in the intersection are indexed by $1, 2, \dots, N_{12}$ in both populations. A two-sample capture-recapture experiment is conducted in Population A. For notational convenience, we label the two samples from Population A as Sample A1 and Sample A2. See Figure 1 for an example. Another two-sample (Samples B1 and B2) capture-recapture data set is collected from Population B. We assume that marks are consistent for all sampling and each individual can be uniquely identified across all four samples. Can the information of the two sets of capture-recapture data be used to estimate the number of individuals common to both populations?

The study was motivated by a project to assess the prevalence and relapse rate of drug use after treatment in an area of Taiwan from 1999 to 2002. Each year in the study period, researchers involved in the project compiled two incomplete lists of drug-users. One purpose was to estimate the number

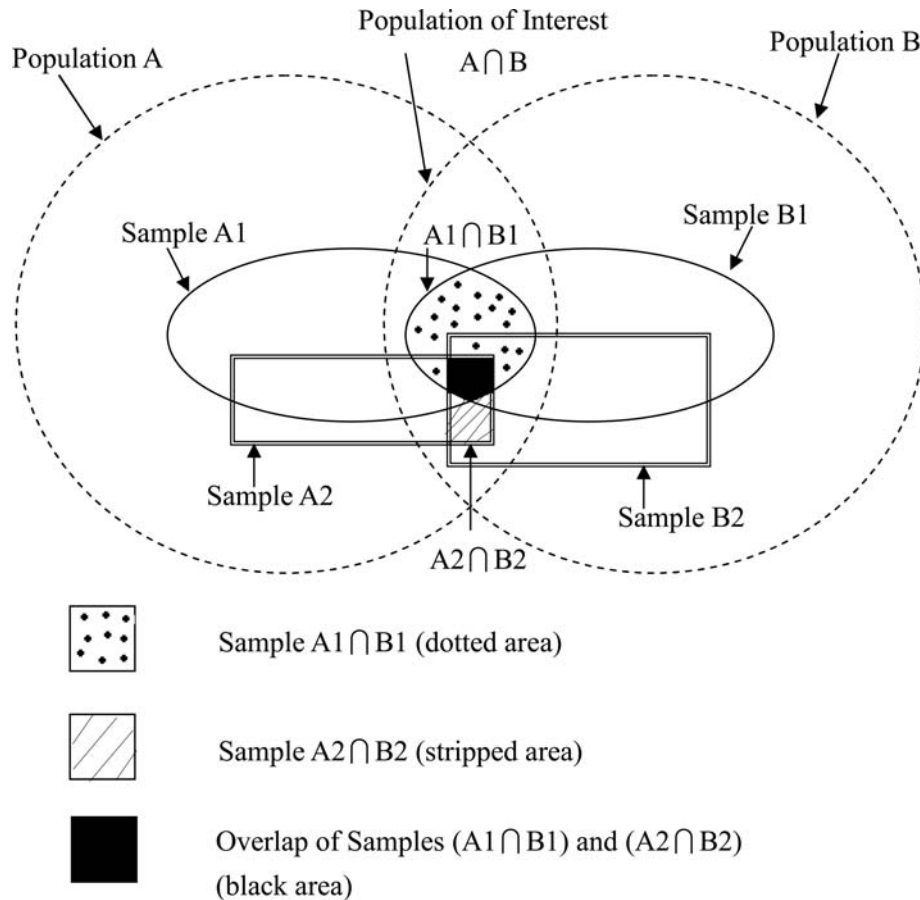


Figure 1 A diagram shows the intersection of two populations and two samples from each population. (See Sections 3 and 4 for detailed explanation.)

of individuals who were drug users across any two years. For illustration, we regard the drug-use population in 1999 as Population A and drug-use population in 2000 as Population B. The number of drug users in the shared population is our estimating target. See Section 4 for further data details and analysis.

To estimate the size of the shared population, we have to “construct” two samples from the shared population to apply the Petersen-Lincoln method. The only way to identify individuals in a sample as coming from the *shared* population is by its appearance in a sample from Population A *and* a sample from Population B. Consequently, we construct the following two samples from the shared population:

- (1) Sample $A1 \cap B1$: this first sample includes those individuals captured in Sample A1 from Population A and Sample B1 from Population B. (This sample is shown as the dotted area in Figure 1.) We denote the sample size of this sample by n_1^* in order to distinguish it from n_1 in the case of one population.
- (2) Sample $A2 \cap B2$: this second sample includes those individuals captured in Sample A2 from Population A and Sample B2 from Population B. (This sample is shown as the striped area in Figure 1.) We denote the sample size of this sample by n_2^* .

The overlapped individuals in the two samples (the black area in Figure 1) are those who are captured in all of the four samples. We denote the overlap by m_2^* . Based on Samples $A1 \cap B1$ and $A2 \cap B2$

from the shared population, we have a similar framework to the Petersen–Lincoln framework with n_1^* , n_2^* and m_2^* corresponding to n_1 , n_2 and m_2 respectively.

For Population A, define $X_{ij} = I$ [the i -th individual is captured in sample j], $i = 1, 2, \dots, N_1$ and $j = 1$ (for Sample A1), $j = 2$ (for Sample A2). For Population B, we similarly define $Y_{ij} = I$ [the i -th individual is captured in sample j], $i = 1, 2, \dots, N_2$ and $j = 1$ (for Sample B1), $j = 2$ (for Sample B2). Then

$$n_1^* = \sum_{i=1}^{N_{12}} I[X_{i1} = Y_{i1} = 1]; \quad n_2^* = \sum_{i=1}^{N_{12}} I[X_{i2} = Y_{i2} = 1], \tag{3.1}$$

and

$$m_2^* = \sum_{i=1}^{N_{12}} I[X_{i1} = Y_{i1} = X_{i2} = Y_{i2} = 1]. \tag{3.2}$$

Replacing n_1 , n_2 and m_2 by n_1^* , n_2^* and m_2^* respectively in the Petersen–Lincoln estimator, we obtain an estimator of the size of the shared population as

$$\hat{N}_{12} = \frac{n_1^* n_2^*}{m_2^*}, \tag{3.3}$$

with a corresponding Chapman estimator

$$\tilde{N}_{12} = \frac{(n_1^* + 1)(n_2^* + 1)}{(m_2^* + 1)} - 1. \tag{3.4}$$

Although the two estimators have exactly the same forms as those in one population, modifications are needed to derive their variance formulas and dependence measures. A variance estimator based on (1.3) in which n_1 and n_2 are fixed would underestimate because it does not take account of the variation of the two random variables n_1^* and n_2^* in our two-population model. Instead, we adopt an asymptotic method proposed in Jolly (1965) to derive an asymptotic variance formula as follows.

$$\widehat{\text{Var}}(\hat{N}_{12}) \approx \hat{N}_{12} + \hat{N}_{12}(\hat{N}_{12} - n_1^* - n_2^*)/m_2^*, \tag{3.5a}$$

$$\begin{aligned} \widehat{\text{Var}}(\tilde{N}_{12}) \approx & U^2 n_2^* + V^2 n_1^* + U^2 V^2 m_2^* + 2UVm_2^*(1 - U - V) \\ & - \{V^2(n_1^*)^2 + U^2(n_2^*)^2 + U^2V^2(m_2^*)^2 + 2UVn_1^*n_2^*\}/\tilde{N}_{12} \\ & + \{2UVm_2^*(Vn_1^* + Un_2^*)\}/\tilde{N}_{12}, \end{aligned} \tag{3.5b}$$

where $U = (n_1^* + 1)/(m_2^* + 1)$ and $V = (n_2^* + 1)/(m_2^* + 1)$. Based on the estimated variance, a 95% confidence interval can be constructed by applying a log-transformation (Chao, 1987). Extending (2.1) to a shared population and using (3.1) and (3.2), we define the dependence measure between Sample A1 \cap B1 and Sample A2 \cap B2 as

$$\Gamma_{12} = \frac{N_{12} \sum_{i=1}^{N_{12}} E(X_{i1}Y_{i1}X_{i2}Y_{i2})}{[\sum_{i=1}^{N_{12}} E(X_{i1}Y_{i1})][\sum_{i=1}^{N_{12}} E(X_{i2}Y_{i2})]} - 1. \tag{3.6}$$

Thus we have the following identity

$$N_{12} = \frac{E(n_1^*) E(n_2^*)}{E(m_2^*)} (1 + \Gamma_{12}). \tag{3.7}$$

An analogous correlation bias formula to Eq. (2.3) is

$$\text{Correlation bias} = E(\hat{N}_{12}) - N_{12} \approx -\Gamma_{12}E(\hat{N}_{12}).$$

We consider a special case with no local list dependence between Sample A1 \cap B1 and Sample A2 \cap B2 in the sense that $E(X_{i1}Y_{i1}X_{i2}Y_{i2}) = E(X_{i1}Y_{i1}) E(X_{i2}Y_{i2})$ for each individual i . If we can further

assume that Sample A1 is a random sample from Population A such that all individuals have the same probabilities β_1 of being caught and also Sample B1 is a random sample from Population B with a constant capture probability of β_2 , then we have $\sum E(X_{i1}Y_{i1}) = N_{12}\beta_1\beta_2$ and $\sum E(X_{i1}Y_{i1})E(X_{i2}Y_{i2}) = \beta_1\beta_2\sum E(X_{i2}Y_{i2})$ in Eq. (3.6), yielding $\Gamma_{12} = 0$. Consequently, no correlation bias exists in this special case for the proposed estimator. In this scenario, we have conducted an extensive simulation study to investigate the performance of our method. Details are omitted here. A general conclusion is that when data are sufficient to provide enough recapture information, our proposed estimator and its associated variance estimator work satisfactorily. However, a general guideline about how large the sample sizes should be in order to have sufficient information is still unclear to us.

We could also use samples A1 and B2 to form the first sample from the shared population, and use A2 and B1 to form the second sample. Similar derivations lead to an alternative estimator \hat{N}_{12}^* . The choice between the two estimators depends on which generates the smaller estimate of Γ_{12} . For example, in the previously discussed scenario when the correlation bias vanishes for \hat{N}_{12} , whereas this bias exist for the alternative estimator; and in terms of bias, \hat{N}_{12} is thus preferable to \hat{N}_{12}^* .

4 Application to Drug Data

The capture-recapture methodology has been extensively applied in epidemiology and health sciences to estimate the size of a target population based on several incomplete lists of individuals (Hook and Regal, 1995; IWGDMF, 1995a, b; Chao et al., 2001; Böhning et al., 2004). The ascertainment data discussed in this paper relate to the prevalence and recidivism rates of (mostly) heroin and methamphetamine use in Taoyuan County, Taiwan. In Taiwan, illegal drug users who are arrested and return a positive urinalysis for illegal drugs are sent for detoxification or maintenance treatment in prison under the Statute for Narcotics Hazard Control (valid since 1998).

The data were collected by Dr. Shu-Chuan Chiang and colleagues and details were provided in Chiang et al. (2007). The study period covered from 1999 to 2002. There are two incomplete lists of drug-users in each year: (1) Prison or judicial list (P-list), all records from the Detoxification Unit of Taoyuan Prison; and (2) Hospital list (H-list), all records from local hospitals.

For the P-list, we assume within males (or females) that the arrested illegal drug users, including some severe cases that might have met the criteria of substance dependence or associated medical complications, were a random sample drawn from the population of illegal drug users in the community. As indicated by Chiang et al. (2007), this assumption may hold true given several factors related to illegal drug use in Taiwan. First, heroin and methamphetamine use is illegal and users regardless of their characteristics are arrested. Second, only those users identified by positive urinalysis are sentenced to detoxification in the prison system, not non-drug-users arrested for either smuggling or possessing drugs. This policy means that the subjects recruited from detoxification units in the judicial system are a more homogenous group as drug users. Third, the chance of an illegal drug user being arrested appears independent of the subject's geographical and demographic background (e.g. local residence, age, education, occupation, drug type and onset age) except for gender. The suspected male drug users tend to have higher chance than females to be arrested. Thus, a stratified analysis by gender was conducted. For comparison, we also analyzed the un-stratified data. It turned out that the un-stratified and stratified analyses yield consistent estimates; see below for details.

With respect to the H-list, patients with an illegal drug-using history were recruited on the basis of clinical diagnostic codes. All patients who met either methamphetamine- or heroin-associated diagnosis (such as abuse, dependence, or associated psychosis) were included. All the participating hospitals adopted the diagnostic system of the International Classification of Diseases-10 (ICD-10, WHO, 1992), and the author (Chiang), a board-certified psychiatrist specializing in drug abuse treatment, performed a chart review to confirm the diagnosis for any unspecified substance-associated disorder found in the computerized data system. The Ministry of Justice, the Taoyuan Prison and the Institutional Review boards of all participating hospitals granted access to the personal data and relevant information of all study subjects.

Table 5a The number of individuals in each list by gender and year.

Population		P-list	H-list	Overlap
1999	Male	3131	329	45
	Female	338	69	9
	Both	3469	398	54
2000	Male	2550	413	56
	Female	439	77	6
	Both	2989	490	62

Table 5b The number of individuals in each list from the shared population by gender.

	Two P-lists Sample A1 \cap B1 (n_1^*)	Two H-lists Sample A2 \cap B2 (n_2^*)	Overlap (m_2^*)
Male	323	101	3
Female	21	19	1
Both	344	120	4

As briefly described in Section 3, one purpose of the project was to study the relapse rate of drug use in the Taoyuan area. Therefore, the number of drug-users common to the two populations; i.e., the shared population across any two years would be most informative for the study. Data in 1999 and 2000 are used here for illustration. We regard the drug-use population in 1999 and in 2000 as Population A and Population B respectively. The P-list and H-list conducted in 1999 correspond respectively to Sample A1 and Sample A2 as defined in Section 3; and the P-list and H-list conducted in 2000 correspond to Sample B1 and Sample B2. The goal is to estimate the number of individuals in the shared population. All the data considered in this paper are shown in Tables 5a and 5b and the estimation results are given in Table 6.

We first examine within males and females the Petersen–Lincoln and Chapman’s estimates separately for each year. For single-population estimation, the differences between the two estimates are limited, thus conclusions are generally similar. We thus only discuss the Petersen–Lincoln estimates below. In 1999, there were 3131 males recorded in the P-list, 329 males in the H-list and 45 in both lists. The Petersen–Lincoln estimate is 22891 with an estimated standard error (s.e.) of 3018 using formula (1.3). Based on a log-transformation (Chao, 1987), we can construct a 95% confidence interval of (17820, 29760). For the females, there were 338 drug users in the P-list and 69 in the H-list with 9 in both lists. The Petersen–Lincoln estimate is 2591 (s.e. 653) with a 95% confidence interval of (1640, 4280). We have the sum of the two estimates as 25482 with an approximate s.e. of $(3018^2 + 653^2)^{0.5} = 3088$; the sum is very close to the un-stratified estimate of 25568 (see Table 6). In 2000, the estimated size of male drug users declined whereas the estimated size of females was doubled. The sum of the two stratified estimates is $18806 + 5634 = 24440$ with an approximate s.e. of $(2234^2 + 1641^2)^{0.5} = 2772$. The un-stratified estimate is 23623. All estimates indicate for each year that a substantial portion of the individuals were missed in the data.

For the estimation of the shared population, the number of individuals who appeared in at least one list for both years is 526. This number is suspected to be severely underestimated. Our proposed method can be applied to infer the undercount. Using the two P-lists within males, we have Sample A1 \cap B1 from the shared population and there are $n_1^* = 323$ individuals in this list. We then use the two H-lists within males to form Sample A2 \cap B2 which includes $n_2^* = 101$ individuals. Since there

Table 6 Point and interval estimation of the recidivist population for data in Tables 5a and 5b.

Population		Petersen-Lincoln estimate	s.e.	95% confidence interval	Chapman estimate	s.e.	95% confidence interval
1999	Male	22 891	3018	(17 820, 29 760)	22 468	3018	(17 410, 29 350)
	Female	2 591	653	(1640, 4280)	2 372	653	(1450, 4110)
	Both	25 568	3099	(20 290, 32 530)	25 172	3099	(19 910, 32 160)
2000	Male	18 806	2234	(15 000, 23 820)	18 527	2234	(14 730, 23 550)
	Female	5 634	1641	(3290, 9960)	4 902	1641	(2670, 9430)
	Both	23 623	2691	(19 000, 29 620)	23 302	2691	(18 690, 29 310)
Shared	Male	10 874	6156	(4010, 30 880)	8 261	3514	(3810, 18 560)
	Female	399	379	(105, 1995)	219	108	(100, 573)
	Both	10 320	5044	(4290, 25 830)	8 348	3268	(4080, 17 670)

are only $m_2^* = 3$ individuals in all four lists, the discrepancy between the Petersen–Lincoln and Chapman’s estimates is clearly seen in Table 6. The use of Chapman’s estimator is suggested. The estimate for the size of the shared male population is $\tilde{N}_{12} = 8261$ based on Eq. (3.4).

As discussed earlier, each P-list within males (or females) can be regarded as a random sample. The individuals in the H-list are more severe addicts, thus each of the two H-lists is generally selective and heterogeneous and the two H-lists may be locally dependent. Despite this heterogeneity, the correlation bias of our estimator vanishes, as derived in Section 3. In our analysis, if un-diagnosed heterogeneity does present in any of the two P-lists, statistical inference becomes difficult; see Link (2003).

The asymptotic variance formula in (3.5b) for the Chapman estimator produces an estimated s.e. of 3514, leading to a 95% confidence interval of (3810, 18560). For females, there is only one individual in all of the four lists, the Chapman estimate is 219 which is inevitably subject to a relatively large s.e. of 108. The sum of the two stratified estimates is $8261 + 219 = 8480$ with an approximate s.e. of $(3514^2 + 108^2)^{0.5} = 3516$, and a 95% confidence interval of (3990, 18700). The un-stratified Chapman estimator yields an estimate of 8348 (s.e. 3268) with a 95% interval of (4080, 17670). Again, the stratified and un-stratified analyses result in very close point and interval estimates. These estimates imply a substantially high relapse rate of drug use in the study area. Chiang et al. (2006) indicated that the most frustrating aspect of substance abuse treatment is the high rate of relapse after treatment. More efforts have been expended to study the efficacy of detoxification programs at detention centers in Taiwan.

5 Concluding Remarks and Discussion

Extending the concept of the Petersen–Lincoln estimator to the case of two populations, we have proposed a simple method to estimate the size of a shared population. The correlation bias of the proposed shared population size estimator is also derived. The proposed methodology is applied to estimate the number of individuals common to two drug-use populations. Our method can be similarly extended to estimate the number of individuals common to multiple populations. We remark that relevant individual covariates collected in our data (e.g., geographical and demographical variables mentioned in Section 4) could also be incorporated in our analysis by using a conditional logistic model (Huggins, 1989). Model fitting and theoretic developments along this direction are under investigation.

In our applications, one advantage is that for each population we believe we had a random sample. Hence there is no correlation bias for our proposed estimator of the shared population. If a random

sample is not feasible, a lower bound for the size of the shared population can be obtained under some sampling schemes; see Pan et al. (2007) for a proposed lower bound.

We have reviewed the correlation bias associated with the classic Petersen–Lincoln estimator. One conclusion is that no correlation bias arises if the equal-catchability assumption holds for the second sample (i.e., the first sample could be highly selective or heterogeneous). This has an implication in the application of the census undercount estimation. For example, a post enumeration sample (PES) is often conducted after the census (Hogan, 1993; Chao and Tsay, 1998) in order to estimate the undercount of the census by the Petersen–Lincoln method. Generally, all individuals in a census can be assumed to have similar probability of being recorded and thus can be regarded as a random sample. Nevertheless, the correlation bias may arise if there is local dependence between the census and the PES (see Section 2.3). In contrast, if the census is conducted after a prior enumeration sample, then there is no local dependence and correlation bias vanishes even if heterogeneity exists in the prior sample. From this point of view, we suggest a reverse ordering of the census and sample to eliminate correlation bias. In other words, the extent of undercount can be accurately estimated if the census is conducted after an enumeration sample.

Investigations into sensitive or stigmatized behaviors, such as child prostitution, drug abuse or mental illness are very difficult to conduct. Denial and under-reporting, resulting from the fear of exposure of anti-social behaviors, are unavoidable and usually hard to adjust for. Understanding prevalence is essential to the development of intervention and further prevention strategies. Methods to estimate prevalence (i.e. the capture-recapture method) for such populations and relapsed sub-populations; i.e., estimation of a shared population as proposed in this paper and Pan et al. (2007) could help to assess the effectiveness of interventions as well as modify future policies. The application of the capture-recapture method to drug use populations has been used in many studies (IWGDMF, 1995a, b; Böhning et al., 2004) and the estimation of relapsed or shared populations merits more research.

Acknowledgements *The authors thank reviewers for carefully reading the manuscript and providing very helpful and thoughtful comments and suggestions, which significantly improved the presentation and exposition. The suggestion of providing a figure (as Figure 1 in Section 3 of the paper) from one of the reviewers is specially appreciated. Part of this paper was presented in the opening keynote speech in the Conference of Recent Developments in Capture-Recapture Methods and their Applications (held on July 12–13, 2007, School of Biological Sciences, University of Reading, UK). The author (Chao) thanks Dr. Dankmar Böhning and the University of Reading for the invitation to the conference. The National Science Council of Taiwan, under contracts 96-2118-M007-001 (for Chao), 96-2118-M415-003 (for Pan), and the Department of Health, Executive Yuan (for Chiang) are acknowledged for the support for this research. The authors are also grateful to the contribution from Taoyuan Prison, Taoyuan Women Prison, Ju-Shan Psychiatric Hospital, Taoyuan Mental Hospital, and Taoyuan Veterans Hospital.*

Conflict of Interests Statement

The authors have declared no conflict of interest.

References

- Böhning, D., Suppawattanakodee, 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**, 1075–1083.
- Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**, 783–791.
- Chao, A. and Tsay, P. K. (1998). A sample coverage approach to multiple-system estimation with application to census undercount. *Journal of the American Statistical Association* **93**, 283–293.
- Chao, A., Tsay, P. K., Lin, S. H., Shau, W. Y., and Chao, D. Y. (2001). The applications of capture-recapture models to epidemiological data. *Statistics in Medicine* **20**, 3123–3157.
- Chapman, D. G. (1951). Some properties of the hypergeometric distribution with applications to zoological censuses. *University of California Publications in Statistics* **1**, 131–160.

- Chiang, S. C., Chan, H. Y., Chen, C. H., Sun, H. J., Chang, H. J., Chen, W. J., Lin, S. K., and Chen, C. K. (2006). Recidivism among incarcerated male heroin users and methamphetamine users in Taiwan. *Psychiatry and Clinical Neurosciences* **60**, 444–451.
- Chiang, S. C., Chen, C. Y., Chang, Y. Y., Sun, H. J., and Chen, W. J. (2007). Prevalence of heroin and methamphetamine male users in the northern Taiwan, 1999–2002: capture-recapture estimates. *BMC Public Health* **7**, 292–299.
- Cochran, W. G. (1978). Laplace's ratio estimators. In H. A. David (ed), *Contributions to Survey Sampling and Applied Statistics*, New York: Academic Press, pp. 3–10.
- Cormack, R. M. (1966). A test for equal catchability. *Biometrics* **22**, 330–342.
- Hogan, H. (1993). The 1990 post enumeration survey: operational and results. *Journal of the American Statistical Association* **88**, 1047–1060.
- Hook, E. B. and Regal, R. R. (1993). Effects of variation in probability of ascertainment by sources ("variable catchability") upon "capture-recapture" estimates of prevalence. *American Journal of Epidemiology* **137**, 1148–1166.
- Hook, E. B. and Regal, R. R. (1995). Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic Reviews* **17**, 243–264.
- Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika* **76**, 133–140.
- International Working Group for Disease Monitoring and Forecasting (IWGDMF). (1995a). Capture-recapture and multiple record systems estimation I. History and theoretical development. *American Journal of Epidemiology* **142**, 1047–1058.
- International Working Group for Disease Monitoring and Forecasting (IWGDMF). (1995b). Capture-recapture and multiple record systems estimation II. Application in human diseases. *American Journal of Epidemiology* **142**, 1059–1068.
- Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration stochastic model. *Biometrika* **52**, 225–247.
- Lee, S. M. and Chao, A. (1994). Estimating population size via sample coverage for closed capture-recapture models. *Biometrics* **50**, 88–97.
- Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* **59**, 1123–1130.
- O'Brien, T. G., Pollock, K. H., Davidson, W. R., and Kellogg, F. E. (1985). A comparison of capture-recapture with capture-removal for quail populations. *Journal of Wildlife Management* **49**, 1062–1065.
- Pan, H. Y., Chao, A., and Foissner, W. (2007). A non-parametric lower bound for the number of species shared by multiple communities. Under revision, *Journal of Agricultural, Biological and Environmental Statistics*.
- Pollock, K. H., Nichols, J. D., Brownie, C. and Hines, J. E. (1990). Statistical inference for capture-recapture experiments. *Wildlife Monographs* **107**, 1–97.
- Seber, G. A. F. (1982). *The Estimation of Animal Abundance*. 2nd edition. Griffin, London.
- Seker, C. and Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association* **44**, 101–115.
- Stephan, F. F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association* **43**, 12–39.
- Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. The Belknap Press of Harvard University Press, Cambridge.
- White, G. C., Anderson, D. R., Burnham, K. P., and Otis, D. L. (1982). *Capture-Recapture and Removal Methods for Sampling Closed Populations*. Los Alamos National Lab, LA-8787-NERP, Los Alamos.
- World Health Organization (WHO) (1992). *International Statistical Classification of Diseases and Related Health Problems*. 10th edition, Geneva, WHO.