

Use of closed population models to estimate the number of injection drug users in Victoria, B.C.

Yuan Xu and Laura Cowen
Department of Mathematics and Statistics
University of Victoria, Victoria, B.C., Canada V8W 2Y2 *

June 7, 2011

Abstract

Closed mark-recapture experiments are widely used to estimate population size even though the assumption of “closed population” is often violated (immigrations and emigrations) compared to open population models. This work investigates various methods in closed population models (Petersen estimator, maximum likelihood estimation with mixtures, and Huggins model) to estimate the number of injection drug users (IDU) in Victoria, B.C. Model selection methods used to choose a final model as well as goodness-of-fit options are discussed. We use data from the 2003 and 2005’s I-Track survey program to compare these estimators.

Keywords: Capture-recapture; Closed populations; Lincoln-Petersen estimator; Huggins model; Mixture distribution; Injection drug user.

1 Introduction

The I-Track survey is a nationwide surveillance program which records the prevalence of HIV and hepatitis C over time but also looks at risk behaviour among injection drug users (IDU). The survey was conducted in 2003 and 2005 in Victoria, B.C. The aim of our work is to estimate the population size of IDU in Victoria. To simplify the study, we assume the population is closed to additions (birth and immigration) and losses (death and emigration), then the population size is considered constant over time.

The estimation of animal abundance is a state variable of key interest in studies of population dynamics. In the usual study cases, the captured animals are uniquely marked then returned (capture-recapture methods) in every sampling occasion. Identification codes are recorded for previously marked animals and then they are released back into the population. Let the subscript i of x_{ij} denote individual i and j the j th sampling occasion, $x_{ij} = 1$ denotes the animal was caught and $x_{ij} = 0$ denotes that it was not caught; then we can get a capture history of any animal in the study. Capture-recapture estimators are based on probabilistic models of events giving rise to the capture history.

We first use the well-known but simplest method, the Lincoln-Petersen (LP) estimator (Seber, 1982). Then maximum likelihood estimators (MLEs) are used to fit models and finite mixtures (Pledger, 2000) are introduced to allow for heterogeneity. Along with capture history, an individual’s sex is also known in the I-Track data. Therefore a conditional likelihood method (Huggins, 1989) is presented to allow for the modeling of capture probabilities in terms of sex. We used Akaike’s Information Criterion (AIC) (Burnham and Anderson, 2002) for model selection of each method. A likelihood ratio test was used to assess the goodness-of-fit.

*Mathematics and Statistics Technical Report #DMS-865-IR

2 Notation

As mentioned before, x_{ij} represents whether individual i is captured at occasion j or not. For example, the capture history $\{0\ 1\ 1\ 0\ 0\ 1\}$ means the individual was caught on sampling occasions 2, 3, and 6 in the study. The following is a description of some common statistics and parameters used in the text.

Statistics

K : number of sampling occasions.

n_j : the number of observed animals with capture history j .

m_j : the number of marked individuals caught at sampling occasion j .

M_j : the number of marked individuals at sampling occasion j .

M_{K+1} : the total number of distinct individuals caught during the study.

Parameters

N : population size.

p_{ij} : the probability of capture if the i th individual was caught on j th occasion.

p_c : capture probability for unmarked individuals.

p_r : capture probability for marked (recaptured) individuals.

The likelihood methods allow for time (probabilities vary from sample to sample), behaviour (a trap response of animals to the first capture), and heterogeneity (different animals have intrinsically different capture probabilities) effects on the capture probabilities. Otis et al. (1978) specified the following eight models which incorporated these three potential variations.

Models

M_0 : Capture probabilities are constant (p).

M_b : Behavioural response only (p_c, p_r).

M_t : Time response only ($p_j, j = 1, \dots, K$).

M_h : Individual capture heterogeneity only ($p_i, i = 1, \dots, N$).

M_{h_A} : Divide individuals into A groups; within each group, individuals have relatively homogenous capture probabilities.

M_{tb} : Behavioural and time responses ($p_{cj}, p_{rj}, j = 1, \dots, K$).

M_{bh} : Behavioural and capture heterogeneity responses ($p_{ci}, p_{ri}, i = 1, \dots, N$).

M_{th} : Time and capture heterogeneity responses ($p_{ij}, i = 1, \dots, N, j = 1, \dots, K$).

M_{tbb} : Time, behavioural, and capture heterogeneity ($p_{cij}, p_{rij}, i = 1, \dots, N, j = 1, \dots, K$).

M_{tbb_A} : Time, behavioural, and heterogeneity (dividing individuals into A groups) responses.

3 Methods

3.1 Lincoln-Petersen estimator (Seber, 1965, 1982)

The Lincoln-Petersen (LP) estimator for population size is well known and widely used. It is appropriate when there are just two sampling occasions. This is suitable for the I-Track data as $K = 2$. A total of n_1 individuals are marked and returned to the population at time 1, n_2 individuals are recaptured at time 2, and m_2 is the number of individuals which are caught twice. The estimation of the population size is

$$\hat{N} = \frac{n_1 n_2}{m_2}.$$

However, it can be shown that the LP estimator is biased and that the magnitude of the bias is inversely related to sample size (Chapman, 1951). The bias-adjusted estimator is given by

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

The variance for $\tilde{N}$ (Seber, 1970) can be estimated by

$$\widehat{var}(\tilde{N}) = \frac{(n_1 + 1)(n_2 + 1)(n_1 - m_2)(n_2 - m_2)}{(m_2 + 1)^2(m_2 + 2)}.$$

Confidence intervals for the LP estimator can be constructed in various ways. Seber (1982) suggests the following rules for choosing the method for approximating the confidence interval for N :

1. Binomial approximation: if the fraction of marks in the second sample is large ($m_2/n_2 > 0.1$);
2. Normal approximation: if the number of marks in the second sample is large ($m_2 > 50$);
3. Poisson approximation: otherwise.

Thus, the way we construct the confidence interval depends on the data.

3.2 Maximum Likelihood Estimators with mixtures (Pledger, 2000)

The likelihood for the closed-captures data is

$$L(N, p_{ij}|X) = \frac{N!}{(\prod_j n_j!)(N - M_{K+1})!} \prod_{i=1}^N \prod_{j=1}^K p_{ij}^{x_{ij}} (1 - p_{ij})^{1-x_{ij}}. \quad (1)$$

Maximum Likelihood Estimation (MLE) can be used to obtain estimates of the p_{ij} s. As N is a parameter, it can be estimated directly from the likelihood rather than indirectly using estimates of capture probabilities.

Finite mixtures are introduced to partition the individuals into two or more groups with relatively homogeneous capture probabilities. Assume there is a fixed number, A , of groups, and group membership is unknown. Each individual is assumed to come independently from group a with probability π_a ($a = 1, \dots, A, \sum \pi_a = 1$). The capture probability of individual i in group a at sample j , p_{ij} , is assumed to be θ_{jba} . Here the subscript b indicates the behaviour effects ($b = 1$ if individual i is captured before occasion j , otherwise $b = 0$). We model θ_{jba} with a linear logistic formulation as follows:

$$\ln\left(\frac{\theta_{jba}}{1 - \theta_{jba}}\right) = \mu + \tau_j + \beta_b + \eta_a + (\tau\beta)_{jb} + (\tau\eta)_{ja} + (\beta\eta)_{ba} + (\tau\beta\eta)_{jba}, \quad (2)$$

where μ is a constant, τ_j is the time response, β_b is the behaviour response, and η_a is the group response which also represents the heterogeneity effects. All the other variables in equation (2) are the interactions between the three responses.

Integrating equation (1) over each animal's independent group allocation gives the likelihood

$$L(N, \pi, \theta|X) = \frac{N!}{(N - M_{K+1})!} \times \prod_{i=1}^N \sum_{a=1}^A (\pi_a \prod_{j=1}^K \{(\theta_{jba})^{x_{ij}} (1 - \theta_{jba})^{1-x_{ij}}\}).$$

3.3 Conditional Likelihood estimation (Huggins, 1989, 1991)

The conditional likelihood model was first presented by Huggins (1989). This approach estimates the closed population size when the capture probabilities are heterogeneous by modeling the capture probabilities in terms of observable covariates such as age, sex, weight, etc. But because the covariates for uncaptured individuals are not known, Huggins (1989) constructed a likelihood conditional on the captured individuals so that characteristics of uncaptured individuals are not required, and the associated parameters such as N can be estimated. Huggins (1989) also modeled capture probabilities dependent on an individual's previous capture history, thus introducing behavioural effects into the model.

We let z_{ij} be the indicator of the past capture history of individual i , i.e.,

$$z_{ij} = \begin{cases} 1 & \text{if individual } i \text{ has been captured before } j, \\ 0 & \text{otherwise.} \end{cases}$$

Then given individual i is captured at least once, the capture probability on occasion j is as follows:

$$\gamma_{ij} = \frac{p_{ij}}{1 - (1 - z_{ij}) \prod_{l=j}^K (1 - p_{il}^*)},$$

where p_{ij}^* is p_{ij} evaluated when individual i has not been captured before occasion j ($z_{ij} = 0$). We can label the captured individuals $i = 1, \dots, M_{K+1}$ and the uncaptured individuals $i = M_{K+1} + 1, \dots, N$. Then the conditional likelihood is as equation (3) which depends only on the captured individuals.

$$L = \prod_{i=1}^{M_{K+1}} \prod_{j=1}^K \gamma_{ij}^{x_{ij}} (1 - \gamma_{ij})^{1-x_{ij}} \quad (3)$$

Huggins (1991) supposed that $\ln[p_{ij}/(1 - p_{ij})]$ is a linear function of covariates corresponding to characteristics of the individuals, differences in behaviour response, and temporal variation. Every individual's sex is recorded in the I-Track data, thus we have the general model M_{tbb} as

$$\ln\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_0 + \beta_{sex}sex_i + \tau_j + \beta_b z_{ij}$$

Here the capture probabilities depend on the covariate sex (sex=1 indicates female, and 0 indicates male in the I-Track data), the temporal variation (τ_j), and the individual's past trapping history (z_{ij}).

The methods of Huggins (1989) can be applied directly to estimate the population size and obtain confidence intervals. Let

$$p_i(\underline{\beta}) = 1 - \prod_{j=1}^K (1 - p_{ij}^*)$$

where $\underline{\beta}$ is the vector of parameters associated with the model. Then $p_i(\underline{\beta})$ is the probability that individual i is captured at least once in the course of the trapping experiment. The unbiased estimator of the population size can be expressed as

$$\hat{N}(\underline{\beta}) = \sum_{i=1}^{M_{K+1}} (p_i(\underline{\beta}))^{-1},$$

and the estimator of the variance of $\hat{N}(\underline{\beta})$ is

$$s^2(\underline{\beta}) = \sum_{i=1}^{M_{K+1}} p_i(\underline{\beta})^{-2} [1 - p_i(\underline{\beta})].$$

3.4 Model Selection and Goodness-of-Fit

We investigated several models for the I-Track data for both likelihood methods. For model selection, we used Akaike's Information Criterion (AIC) (Burnham and Anderson, 2002). AIC is a model selection technique which is given by

$$AIC = -2\ln L(\hat{\beta}) + 2\kappa,$$

where $\ln L(\hat{\beta})$ is the log-likelihood for the model evaluated at the maximum conditional likelihood estimator $\hat{\beta}$ and κ is the number of parameters of $\hat{\beta}$. Given a data set, several models can be ranked according to their AIC, with the one having the lowest AIC being the best.

However, when the sample size of the data is very small, we used the second-order information criterion AICc. AICc can be written as

$$AICc = -2\ln L(\hat{\beta}) + 2\kappa\left(\frac{n}{n - \kappa - 1}\right)$$

where n is the sample size. In our data set, we applied AICc as the model selection criterion since the sample size is small.

After selecting the model with the lowest AICc value for each method, we tested the goodness-of-fit of these models. Usually, the full model M_{tbb} would fit the data best. Therefore, the model we selected was contrasted with the full model to test its goodness-of-fit. We utilized a likelihood ratio test (LRT) (Burnham and Anderson, 2002) to compare the fit of two models. The test statistic of the LRT is given as follows:

$$LRT = -2\ln\left(\frac{L(\beta_0)}{L(\beta_1)}\right).$$

Here, $L(\beta_0)$ is the likelihood for the model of interest chosen by AICc and $L(\beta_1)$ is the likelihood for the full model. It can be shown that this test statistics is asymptotically chi-square distributed with the differences of the number of parameters between the two models being the degree of freedom (denoted by k). Then we can get a p -value for the test which equals $P(LRT > \chi_k^2)$.

3.5 Software

For the LP estimator, we used the Chapman-type estimator in the “*mrclosed*” function in “FSA” package of R program (*R Development Core Team*, 2010) to obtain LP estimators of N . Running this package requires R version 2.11-1 or greater.

Program MARK (White and Burnham, 1999) was used to fit the I-Track data for the likelihood based methods. MARK is a Windows-based program for analysis of data from marked individuals. It computes the estimates of model parameters via numerical maximum likelihood techniques. To download MARK online, go to <http://www.phidot.org/forum/index.php>.

4 Data Analysis

4.1 Estimation

First we used the LP estimator for the I-Track data. In our dataset, $n_1 = 254$, $n_2 = 250$, and $m_2 = 19$; then $\tilde{N} = 3199$. By using the methods of Seber (1982), the standard error of the LP estimator is 643. Since $m_2/n_2 = 0.076 < 0.1$ and $m_2 = 19 < 50$, we used the Poisson approximation to construct a confidence interval. The approximate 95% confidence interval is (2086, 5144) individuals. Notice that one of the key assumptions of the Lincoln-Petersen estimator is that all individuals are equally catchable in each sample.

But this assumption is easily violated when time, behaviour, or heterogeneity affect the probability of capture.

Table 1 represents the results of the likelihood with mixtures (Pledger, 2000) method from MARK (White and Burnham, 1999) output. For the heterogeneity effects, we assumed the mixtures A was 2 at first; within each group, individuals have relatively homogeneous capture probabilities. After trying $A \geq 3$, we found there was no need to introduce another group as the residual deviance for M_{h_3} was the same (7.899) as for M_{h_2} . Actually, model M_{tbh_2} and M_{tb} have the lowest AICc values of -4226 and -4222. However, the estimates of N are equal to $M_{K+1} = 485$. These nonsensical results occur because program MARK requires a constraint on the p_{iK} value and these two full models do not have such constraints; therefore they are meaningless. Disregarding these models we find model M_0 has the lowest AICc value in Table 1. In addition, the AICc weight suggests M_0 fits the data approximately four times better than M_b and M_t ($0.662/0.164 = 4.037$)

Table 1: Model selection results using Pledger's (2000) maximum likelihood estimation with mixtures method for the I-Track data, p denotes the capture parameters and κ is the number of parameters in each model.

Model	p	AICc	AICc weight	$\hat{N}$	κ
M_0	p	-4218	0.524	3329.55	2
M_b	p_c, p_r	-4216	0.193	2600.37	3
M_t	p_j	-4216	0.193	3329.34	3
M_{h_2}	p_i	-4214	0.071	3329.81	3
M_{bh_2}	p_{ci}, p_{ri}	-4210	0.010	2600.30	4
M_{th_2}	p_{ij}	-4210	0.010	3329.66	4

The results in Table 2 show the model selection results based on Huggins' (1989) method. There are 2 groups which reflect the heterogeneity effects. $A \geq 3$ was also tested to see if any higher heterogeneity should be presented. But the residual deviance of M_{h_3} stayed the same (806.37) which implied no further exploration needed. Again there are two erratic estimations of N : 1008.84 and 705.45, which is given by M_{tb} and M_{tbh_2} and are not presented in Table 2. In M_{tb} and M_{tbh_2} , as mentioned before, the respectively last capture probability p_{iK} is unconstrained. The model with the lowest AICc is M_0 and its AICc weight is much higher than any other model weight.

Table 2: Model selection results using Huggins' (1989) method for the I-Track data, p denotes the capture parameters and κ is the number of parameters in each model.

Model	p	AICc	AICc weight	$\hat{N}$	κ
M_0	p	808.372	0.514	3342.31	1
M_t	p_j	810.346	0.199	3342.10	2
M_b	p_c, p_r	810.346	0.199	2804.96	2
M_{h_2}	p_i	812.390	0.070	3696.01	2
M_{bh_2}	p_{ri}, p_{ci}	816.400	0.009	3407.83	3
M_{th_2}	p_{ij}	816.400	0.009	3342.12	3

4.2 Goodness-of-Fit

We used a likelihood ratio test (LRT) to access the goodness-of-fit of each model in the likelihood with mixtures (Pledger, 2000) and the conditional likelihood (Huggins, 1989) methods. The model selected by AICc with mixtures was compared with the full model M_{tbh_2} to see whether they fit the data as well

as M_{tbb_2} . For the MLE method, the p – value of M_0 is too small (p – value < 0.001) which shows strong evidence against the hypothesis that M_0 fits the data as well as M_{tbb_2} . The goodness-of-fit results for Huggins’ (1989) method compares model M_0 with model M_{tbb_2} ; again there is evidence to reject the hypothesis that model M_0 fits the data as well as model M_{tbb_2} (p – value < 0.001).

4.3 Final results

We found that model M_0 was selected by AICc as the model of interest in both the likelihood with mixtures and conditional likelihood methods. The lack of fit of these models may be due to overdispersion in the data and we should explore this issue further in the future. Combined with the results of the LP estimator, Table 3 gives the final estimates for each method. The results are very similar to each other.

Table 3: Population size estimate ($\hat{N}$), estimated standard error, and approximate 95% confidence intervals obtained from the Lincoln-Petersen estimator (LP), model M_0 using Pledger’s (2000) MLE method, and model M_0 using Huggins’ (1989) method.

Method	$\hat{N}$	Std.err	95%CI
LP	3199	643	(2086, 5144)
MLE	3330	706	(2246, 5078)
Huggins	3342	709	(2254, 5098)

5 Discussion

In this paper, we use three different methods to estimate the population size of IDU in Victoria, BC. The Lincoln-Petersen estimator is the simplest method. One of the key assumptions of the LP estimator is the probability that individual are equally catchable in each sample. However, time, behaviour, or heterogeneity responses often cause violation to this assumption as they can affect the probability of capture. Therefore, we introduced the likelihood with mixtures and conditional likelihood methods. These two methods relate the three responses with the capture probability in order to have a better estimate of N .

There was concern that the number of recaptures ($m_2 = 19$) in the I-Track study was lower than expected, resulting in population estimates that were higher than the accepted estimate of 2000 individuals (Stajduhar *et al*, 2004). In the I-Track program, survey respondents were asked to provide a unique identifier – an initial and a significant date. This unique identifier was then encrypted through a computer program which generated an ID. However, we cannot know that each resurvey respondent in 2005 provided the same identifier as in 2003. This is a form of tag loss and would be a violation of one of the model assumptions. If someone changed their identifier (e.g., through marriage), then the number of individuals who were captured on both occasions should be more than $m_2 = 19$. The Figure 1 shows how $\hat{N}$ changes based on the LP method when the value of recaptured individuals changes. We can see that when m_2 gets larger, the population size estimate will be smaller. When m_2 is less than 30, N increases dramatically. Therefore, we can further investigate the accuracy of a respondent’s ID number. If there are more individuals who were resurveyed, then the estimated size of the IDU population may reduce significantly.

The lack of fit for the likelihood methods needs exploration. Overdispersion of the data and a single variance inflation factor (c) can be estimated from the goodness-of-fit chi-square statistic (χ^2) of the full model and its degrees of freedom, $\hat{c} = \chi^2/df$. Then QAICc (Burnham and Anderson, 2002) can be the criterion for model selection. Model assumptions and their effect on the fit of the models is another avenue to investigate in future work.

In both the MLE and conditional likelihood estimation, the behaviour effect has been modeled at two levels:

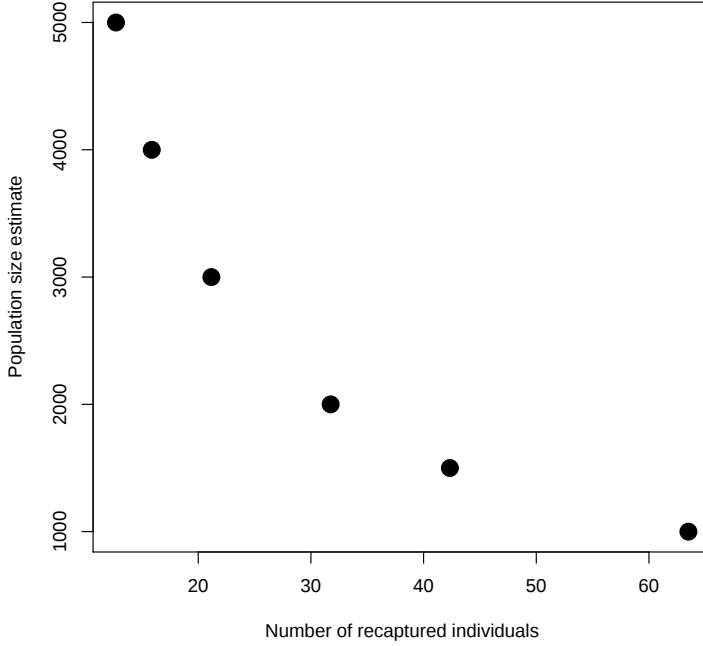


Figure 1: Population size estimates of the number of injection drug users in Victoria, B.C. obtained using the Lincoln-Petersen estimator by varying the number of recaptured individuals (m_2).

one for a first capture and the other for recapture. However, other choices could be specified. Pollock (1975) gives a general model in which capture probability only depends on previous capture history.

Although we chose $A = 2$, it does not imply that there really are two groups of individuals. The grouping allows us to use heterogeneity and correct for the bias in $\hat{N}$.

The assumption of “closed-population” can be easily violated since there are birth, death, immigrations, and emigrations in the real world. When a third I-Track survey is completed, further study can focus on fitting the data in “open-population capture-recapture” models. The Cormack-Jolly-Seber (CJS) model (Cormack, 1964; Jolly, 1965; Seber 1965) and the Jolly-Seber (JS) model (Jolly, 1965; Seber, 1965) are the two primary methods for open-population. Schwarz and Arnason (1996) gave a general methodology for analysis with CJS and JS models. But some assumptions of CJS and JS models can also be violated. Cowen and Schwarz (2006) discussed tag loss problems which is one of the main violations in mark-recapture experiments.

6 Acknowledgements

We thank the Vancouver Island Health Authority for sharing the I-Track data.

References

- [1] Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed., Springer.
- [2] Chapman, D. G. (1951). Some properties of the hypergeometric distribution with application to zoological censuses. *Univ. Cal. Publ. Stat.* **1**, 131-160.
- [3] Cormack, R. M. (1964). Estimates of survival from the sighting of marked animals. *Biometrika* **51**, 429-438.
- [4] Cowen, L., and Schwarz, C. J. (2006). The Jolly-Seber model with tag loss. *Biometrics* **62**, 699-705.
- [5] Huggins, R. M. (1989). On the Statistical analysis of capture experiments. *Biometrika* **76**, 133-140.
- [6] Huggins, R. M. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics* **47**, 725-732.
- [7] Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-Stochastic model. *Biometrika* **52**, 225-247.
- [8] Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs* **62**, 1-135.
- [9] Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* **56**, 434-442.
- [10] Pollock, K. H. (1975). A K-sample tag-recapture model allowing for unequal survival and catchability. *Biometrika* **37**, 521-529.
- [11] R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- [12] Schwarz, Carl J., and Arnason, A. N. (1996). A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics* **52**, 860-873.
- [13] Seber, G. A. F. (1965). A note on the multiple recapture census. *Biometrika* **52**, 249-259.
- [14] Seber, G. A. F. (1970). The effects of trap response on tag-recapture estimates. *Biometrika* **26**, 13-22.
- [15] Seber, G. A. G. (1982). *The Estimation of Animal Abundance and Related Parameters*, 2nd ed., London Blackburn Press.
- [16] Stajduhar, K. Poffenroth, L., Wong, E., Archibald, C., Sutherland, D., and Rekart, M. (2004). Missed opportunities: injection drug use and HIV/AIDS in Victoria, Canada. *International Journal of Drug Policy* **15**, 171-181.
- [17] White, G. C., and Burnham, K. P. (1999). Program MARK: survival estimation from populations of marked animals. *Bird Study* **46**, 120-139.