Transactions on Computational and Applied Mathematics (2024)
Clausius Scientific Press, Canada
DOI: 10.23977/tracam.2024.040117
ISSN 2616-1826 Vol. 4 Num. 1

Multiple Systems Estimation Based on Hidden Population Estimation

Yihua Zou

Faculty of Social Sciences, University of Southampton, Southampton, UK

Keywords: Multiple systems estimation (MSE); Log-linear model; Population size; Model selection; Bootstrap

Abstract: The multiple system estimation (MSE) method is a way of estimating the population size based on samples from two or more sources. Wildlife biologists first used it to estimate the number of wild animals, such as fish, insects and birds in a certain area, but it is now being used in the study of human disease and health, and its theory also along with the development of biostatistics and continually improved. In general, researchers will use the MSE to estimate some hidden populations that are not easily detected and consider log-linear models to explain the effects of list or covariates and their interactions on total population size and apply bootstrapping to calculate their confidence intervals.

1. Introduction

The earliest methods of estimating population numbers from marked animals can be traced back to Peterson's work in the 19th century [1]. Since then, the capture-recapture (CR), a method of estimating population size based on samples from two or more sources, has been widely used by animal ecologists. For example, Lincoln studied North American waterfowl in 1930, followed by Jackson, who studied tsetse flies in the Tanganyika region in 1933. Then later, the CR method was popularized and used by some statisticians to make population estimates. Abeni, Brancato and Perucci, for instance, used the primary method in 1994 with the human immunodeficiency virus type 1(HIV-1) surveillance system data in Lazio, Italy, to estimate the total number of people with HIV-1 [2].

The multiple systems estimation (MSE) method is an extension of the CR approach, which includes a range of statistical techniques specifically designed to estimate hidden populations from a multi-source incomplete contingency table. Nowadays, MSE has been widely applied in various fields, for example, to estimate wildlife populations [3], underreported transport survey points and the scale of modern slavery [4]. I will generally assume a Poisson distribution with the log-linear model for MSE.

After fitting the model with all the lists and covariates, we may find that not all the variables are significant; in other words, some variables may have no effect on the model. For this reason, it is necessary to perform model selection to remove some of the non-significant variables to ensure that our model is valid. Three ways of the log-linear model selection are best subset, forward stepwise and backward stepwise selection. However, a literature search and computer calculation found that the best subset regression method is very time consuming, especially for models with many lists and covariates. Therefore, this paper will only use the forward approaches.

2. Background

The capture-recapture (CR) method can be divided into two-sample CR method and multiple-sample CR (also called MSE) methods, depending on the number of data sources. The two-sample CR method is the simplest and was first applied in the ecological field, with Pertenson first linking it to fish tagging in 1894, though Dalh first actually used it in the fishing industry in 1917, and Lincoln in 1930 to estimate the size of duck populations [5]. In 1949, Sekar and Deming used the method to estimate birth rates, mortality rates and the extent of registration, marking the application of the CR method to human health [6]. However, the real use in epidemiological studies began in 1969 when Wittes and Sidel used it to estimate the rate of birth defects, and Lewis used to studying nosocomial infections in the same year [7]. At that time, this method was not well understood and applied. It was not until the 1980s that CMR method was widely used in various fields.

It is an attempt to estimate population size using samples from multiple sources. The MSE method was first proposed in 1938 when Schnabel estimated fish populations and made some assumptions about the process of sampling and tagging [8]. The MSE method was also applied to populations that were allowed to be born, die and migrate during the study period. As biostatistics has evolved, it has been realized that some basic assumptions may not hold. For example, there is the problem of heterogeneity - those individuals that are not tagged have a different probability of being captured in a given sample, tagged individuals behave differently from untagged individuals, and samples may not be independent of each other. However, since this is not the case in the list, we need to relax some of the assumptions [9].

Previous researchers have put forward various families of models to deal with the case of multiple lists, each of which makes different assumptions about the population. For instance, Gibbs sampling [10] and reversible jump Markov chain Monte Carlo methods [11] by Bayesian approach, log-linear models [12], discrete mixture models [13], and Rasch models [14]. Fortunately, the general log-linear model can be used to fit a multi-source contingency table [15]. Fienberg, in 1972, analyzed the relationships between captures or lists by using the log-linear model as it generated for the analysis of multidimensional contingency tables [16]. This solution is universal and applicable to population estimates of animal and human populations.

3. Methods

There are many population estimation methods, such as multiplier method, capture-recapture method, mapping methods, Delphi method and workbook method. Among them, the workbook method relies on existing official reports; the mapping methods and the Delphi method are based on population counting, while the multiplier method and the capture-recapture method are based on independent samples [17]. However, the accuracy of the multiplier method depends on the quality of the raw data [18], and different data sources will provide different estimates. Multiple systems estimation (MSE) is a generalization of the capture-recapture approach to obtain the volume of hidden populations. Due to the concealment of the hidden people, it is feasible and reasonable to apply the MSE method to estimate the size of the population.

3.1. Capture-recapture method and multiple systems estimation

3.1.1. Two sample capture-recapture methods

Begin with a brief description of the capture-recapture method using the example of estimating the number of fish in a pond. Suppose some fish can be caught from a pond, mark them in some way, and then release them back into the same pond. Tracking the number of fish that have been marked

and assume that it is n. After a period of time, make a second catch of the fish. Suppose M fish be caught in the second time, and find that m of them are tagged as being part of the first capture returned. According to Petersen–Lincoln method [19], we can obtain that the frequency of marked rate in the overall population is approximately equal to the proportion of the marked rate in the second sample, the formula form is

$$\frac{n}{N} \approx \frac{m}{M}$$

This will give us the estimate of the total number of fish in the pond (N),

$$\widehat{N} = \frac{n \times M}{m}$$

The method has been extended to include more than two sources [20]. For example, there are two lists ($S1$ and $S2$), which can be linked to obtain the number of individuals in $S1$ but not in $S2$ (n_{10}), that of individuals in $S2$ but not in $S1$(n_{01}), and that of individuals both in $S1$ and $S2$(n_{11}). These counts of the contingency table with the variables labelled $S1$ being distinguished into 'observed' and 'unobserved' categories, and likewise for list $S2$ (see Table 1). The estimate of total population size, denoted as N, is obtained by adding the observed number to the estimated number of individuals omitted from both lists (n_{00}).

Table 1: Simple two-list contingency table

		List 2 ($S2$)		
		Observed	Unobserved	Total
List 1 ($S1$)	Observed	n_{11}	n_{10}	$n_{1\cdot}$
	Unobserved	n_{01}	n_{00}	
	Total	$n_{\cdot 1}$		N
Where $n_{1\cdot}$ and $n_{\cdot 1}$ denotes the total number of individuals observed by list 1($S1$) and list 2 ($S2$).				

Then we can obtain the estimate of N by

$$\widehat{N} = \frac{n_{1\cdot} \times n_{\cdot 1}}{n_{11}}$$

In addition, the original data should fulfil the following four assumptions [21]:

1) *Closed system.* The study population should be closed, that is, there should be no significant changes in the total number of studies during the study period, such as a large number of migrations or mortality.

2) *Independent.* Each list should be independent of the other, meaning that the lists created do not affect each other, i.e., the inclusion of one list does not affect the possibility of inclusion in another.

3) *Homogeneity.* All the individuals should have the same probability of being included or captured.

4) *Identifiable.* All individuals need to be matched on all lists. In other words, we need to set at least one identifier. For example, if the last name, first name and gender are the same, we can mark him or her as the same person.

However, the independent assumption is an ideal situation. Log-linear model and multiple sources can help us with this problem.

3.1.2. Multiple systems estimation

MSE, also known as the multiple sample capture-recapture method, is a class of sampling

techniques for statistical inference that uses the overlap between several incomplete lists of data to estimate the total population.

Start with a brief introduction to the contingency table. Suppose there are S_k available sources for $k = 1, \dots, K$, and x_j binary covariates for $j = 1, \dots, p$. There will be 2^K combination of different sources. Then we can construct an incomplete contingency table with $2^K \times 2^p$ cells. Each entry, representing the number of individuals, is called cell count. There will be $n_U = 2^p$ missing values (unobserved cell counts, $\boldsymbol{n}$, a $1 \times p$ vector) of the contingency table we need to estimate. Moreover, the sum of all cell counts is the total population size (T).

For example, let $\boldsymbol{x}$ denotes the combination of the lists and covariates, suppose there are two sources and two characteristics (see Table 2), then $\boldsymbol{x} = \{0,1,0,1\}$ represents the individual that only occurs in the second (S_2) list with the second binary covariate equals to 1. $\boldsymbol{N} = \{N_1, N_2, N_3, N_4\}$ is a 1×4 vector we want to predict, and y_{ijkl} denote the number of each individual observed by the combination of sources and covariates for $\mathrm{i, j, k, l} \in \mathrm{R} \cup \{0000\}$.

Table 2: Multi-source contingency table with covariates

List 1 (S_1)	List 2 (S_2)	Cov.1 = 0		Cov.1 = 1		Total
		Cov.2 = 0	Cov.2 = 1	Cov.2 = 0	Cov.2 = 1	
unobserved	unobserved	N_1	N_2	N_3	N_4	$\sum_{i=1}^{4} N_i$
observed	unobserved	y_{1000}	y_{1001}	y_{1010}	y_{1011}	$y_{2\cdot}$
unobserved	observed	y_{0100}	y_{0101}	y_{0110}	y_{0111}	$y_{3\cdot}$
observed	observed	y_{1100}	y_{1101}	y_{1110}	y_{1111}	$y_{4\cdot}$
Total		$y_{\cdot 1}$	$y_{\cdot 2}$	$y_{\cdot 3}$	$y_{\cdot 4}$	T
where $\sum_{i=1}^{4} N_i$ denotes the number of individuals that do not appear in both lists, $y_{2\cdot}$ and $y_{3\cdot}$ denotes the total number of individuals observed by list 1 (S_1) and list 2 (S_2), respectively, and $y_{4\cdot}$ denotes the overlap between S_1 and S_2.						

To predict the missing cell counts $\boldsymbol{N} = \{N_1, N_2, N_3, N_4\}$, Fienberg introduced the log-linear model in 1972, specifying that the expected number of cells in log form.

3.2. Model fitting

We assume a Poisson distribution and model the log of the cell count μ_i, that is, for cell $i = 1, 2, \dots, n$, let y_i denotes the cell count and assume they are independent,

$$y_i \mid \boldsymbol{\beta} \sim Poisson(\mu_i), \tag{1}$$

the probability distribution function of Y_i is

$$f_{Y_i}(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = exp\{y_i \log(\mu_i) - \mu_i - \log(y_i!)\} \tag{2}$$

and the link function

$$g(y_i) = log(\mu_i) = \boldsymbol{X}_i^T \boldsymbol{\beta} \tag{3}$$

where

$$\boldsymbol{X}_i^T = \begin{bmatrix} 1 \\ X_{i1} \\ \vdots \\ \mathrm{X_{iK}} \\ X_{i(K+1)} \\ \vdots \\ X_{i(K+p)} \end{bmatrix} \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{\mathrm{K}} \\ \beta_{\mathrm{K+1}} \\ \vdots \\ \beta_{K+p} \end{bmatrix}$$

Then the log linear model (also called Poisson model) that contains main effects but no two-way or higher order interactions can be formulated for the i-th unit of a sample of n units:

$$\log E(Y_i) = log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \cdots + \beta_K X_{iK} + \cdots + \beta_{K+1} X_{i(K+1)} + \cdots + \beta_{K+j} X_{i(K+j)} + \cdots + \beta_{K+p} X_{i(K+p)} \text{ for } i = 1, \dots, n \quad (4)$$

where

Y_i is the response variable, which follows Poisson (μ_i), $i = 1, \dots, n$;

β_l are unknown parameters, $l = 1, \dots, K, K+1, \dots, K+p$ and β_0 is the intercept;

X_{il} are the predictors, $i = 1, \dots, n,\ l = 1, \dots, K, K+1, \dots, K+p$.

$k = 1, \dots, K$ is the number of lists, and $j = 1, \dots, p$ is the number of covariates.

(4) can be rewritten as a matrix form:

$$\log \mu_{\boldsymbol{i}} = \boldsymbol{X_i^T \beta} \quad (5)$$

where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}, \boldsymbol{X^T} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1l} \\ 1 & X_{21} & X_{22} & \dots & X_{2l} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nl} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_l \end{bmatrix},$$

Here $\boldsymbol{Y}$ is the $n \times 1$ response matrix, with i entry Y_i, $\boldsymbol{y}$ is a vector of expected cell counts, $\boldsymbol{X}$ is the $(K+p+1) \times n$ input matrix, $\boldsymbol{\beta}$ is the $(K+p+1) \times 1$ matrix of parameters and $\boldsymbol{\mu}$ is the $n \times 1$ matrix of cell counts.

Then we can obtain the likelihood function for the parameters $\boldsymbol{\beta}$

$$L(\boldsymbol{\beta}) = \prod_{\boldsymbol{i}} f_{Y_i(y_i)}$$

$$= \prod_{\boldsymbol{i}} exp\{y_i\, log(\mu_i) - \mu_i - log(y_i!)\}$$

$$= exp\{\sum_i y_i\, log(\mu_i) - \sum_i \mu_i - \sum_i log(y_i!)\}$$

Hence, the log-likelihood for the log-linear model is

$$l(\boldsymbol{\beta}) = \sum_i y_i \log(\mu_i) - \sum_i \mu_i - \sum_i \log(y_i!)$$

$$= \textstyle\sum_i y_i\, \boldsymbol{X_i^T \beta} - \sum_i \exp\{\boldsymbol{X_i^T \beta}\} - \sum_i \{\log(y_i\}!). \quad (6)$$

We can ignore the last term $\sum_i \log(y_i!)$ as it does not involve any parameter.

To maximize the log-likelihood, we need to take the partial derivative with respect to $\boldsymbol{\beta}$ and set to

zero

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_i (\boldsymbol{y}_i - \exp\{\boldsymbol{X}_i^T \boldsymbol{\beta}\}) \boldsymbol{X}_i = 0 \tag{7}$$

Let $\widehat{\boldsymbol{\beta}}$ denote the maximum likelihood estimator of the regression parameters coefficients. Iteratively reweighted least squares can be used to find the $\widehat{\boldsymbol{\beta}}$. Once the coefficient estimates $\widehat{\boldsymbol{\beta}}$ are obtained, we can get the fitted regression model

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{exp}\{\boldsymbol{X}^T \widehat{\boldsymbol{\beta}}\}. \tag{8}$$

Therefore, we can obtain the predicted unobserved population $\widehat{\boldsymbol{N}} = \{N_{\boldsymbol{1}}, \dots, N_{\boldsymbol{2^p}}\}$

$$\widehat{N} = exp\{A^T \hat{\beta}\} \tag{9}$$

where $\boldsymbol{A}$ is a $(l+1) \times 2^p$ vector representing all the combination of the lists and covariates for the unobserved value, and $\widehat{\boldsymbol{N}}$ is a $l \times 1$ vector, for $l = 1, \dots, K, K+1, \dots, K+p$.

Therefore, the total population size T is the sum of the observed population and the estimated population,

$$T = \sum_{i=1}^{n} y_i + \sum_{j=1}^{\boldsymbol{2^p}} N_j$$

where $\sum_{i=1}^{n} y_i$ denotes the sum of the number of the individuals which observed in each list.

Finally, we are interested in whether our explanatory variables have an effect on the response variables. So we need to test the hypothesis of $H_0: \beta = 0$ versus $H_1: \beta \neq 0$, the statistics is

$$z = \frac{\hat{\beta}}{s.e.(\hat{\beta})}, \tag{10}$$

where $s.e.(\hat{\beta})$ is the standard error of $\hat{\beta}$.

If the null hypothesis is true, then the statistics z approximates a standard normal distribution, that is

$$z = \frac{\hat{\beta}}{s.e.(\hat{\beta})} \sim N(0,1) \tag{11}$$

Where $N(0,1)$ is standard normal distribution with mean equals to zero and variance equals to one.

3.3. Model selection

3.3.1. Statistics

When we model a set of data, we have many variables at our disposal and choosing different combinations of variables can lead to different models and results. We need to choose the best model by estimating the performance of counting different models. Therefore, we need to consider some methods of selecting subsets of variables. There are various statistics that can be used to judge the quality of a model, for example, *Mallow's* C_p, *Akaike information criterion* (AIC), *Bayesian information criterion* (BIC), *residual sum of squares* (RSS) and *adjusted R squared*.

AIC is a measure of the goodness of fit of a statistical model and was created and developed by Japanese statistician Hirotugu Akaike [22]. The AIC is based on the concept of entropy. In general, AIC can be expressed as

$$AIC = -2\, l(\widehat{\boldsymbol{\beta}}) + 2p,$$

where $l(\boldsymbol{\beta})$ is the log-likelihood function of the parameters $\widehat{\boldsymbol{\beta}}$ based on the maximum likelihood estimators (MLE), p is the number of unknown parameters. And the second term is called the penalty.

When using AIC for model selection, the smaller the AIC value, the better the model. Therefore, we will select the model with the smallest AIC value in the set of models under consideration.

3.3.2. Best subset selection

To select a subset of the variables of the full model, a straightforward way is called best subset selection. Best subset regression finds for each $k \in \{1, 2, \ldots, p\}$ the subset of size k that gives the smallest AIC. Here is the algorithm,

Algorithm 1. Best Subset selection
1) Let M_0 denote the null model, which contains an intercept but no predictors. 2) Fit all $\binom{p}{k}$ models that include exactly k predictors. Pick the model with the smallest residual sum of squares (RSS) or largest R^2 among these $\binom{p}{k}$ models, and denote it as M_k. 3) Choose the model using AIC from among $M_0, \ldots, M_p$.

Although the best subset selection is straightforward, it may be subject to some computational constraints. The number of possible models we have to consider ($\binom{p}{0} + \binom{p}{1} + \cdots + \binom{p}{p} = 2^p$) will be huge as p increases, which can be very time-consuming.

3.3.3. Forward and backward stepwise selection

Considering the computational efficiency, the best subset selection is not very suitable for very large p, then we need to seek another way. Another way to select a subset of predictors is called *stepwise selection*. This method contains forward stepwise and backward stepwise procedures.

To perform the *forward method*, we will begin with a null model. Additional predictor that most improves the fit is then added to the model one by one until all the p predictors are included in the model. The details of this method are as below.

Algorithm 2. Forward Stepwise Selection
1) Generate a null model, denoted by M_0. 2) For $k \in \{0, \ldots, p-1\}$, consider all $p-k$ models, augmenting the predictors in M_k with one extra predictor. Choose the best model, having the smallest RSS or highest R^2, among these $p-k$ models and denote it as M_{k+1}. 3) Select a single best model using AIC.

3.4. Confidence interval

After fitting the model, it is necessary to calculate the confidence intervals (CI). The main function of CI is to demonstrate the accuracy of the sample study estimate as a population value [23]. Jerzy Neyman introduced confidence intervals in 1937, since then, statisticians and scientists gradually began to use this idea [23]. Despite this, confidence intervals were rarely used. It was not until the late 1980s that medical journals began requiring confidence intervals to be reported [24]. It is customary to write down the *confidence level* as $1 - \alpha$, which means that there is a $1 - \alpha$ probability that the true value will fall within the range we have calculated, and α is a very small positive number called *significant level*.

Several ways of calculating CI have been proposed. A robust way to calculate confidence intervals is to use the *parametric bootstrap* method. Bootstrap methods are used in a variety of contexts, the most common mode is to provide a measure of the accuracy of the parameter estimates. Bootstrap

estimation is a method of estimating parameters using repeated sampling, which began to develop after the popularization of computers, as it is extremely cumbersome to do by hand without the assistance of computers for repeated sampling. The main benefit of bootstrapping is that it can be performed relatively straightforward.

The algorithm for calculating the confidence interval by parametric bootstrapping is as follows.

Algorithm 3. Bootstrapping for Confidence Interval
1) Generate multinomially distributed random cell counts based on the original data size and the per centage of the observed population size, denoted as G_i. 2) For each of the G_i sample, fit a log-linear model and use stepwise selection to choose the *best* model (here best means the model with the lowest AIC or values). 3) Use the selected model to predict the unobserved (N) and obtain the estimated total population size (T). 4) Repeat steps 1-3 B times (the larger B the better, usually be in the hundreds or the thousands, as the number of bootstrap repetitions defines the variance of the estimates), obtaining B estimated total population size. 5) The $(1-\alpha)$ confidence interval is the middle $(1-\alpha)$ of the B predicted values.

In short, bootstrapping for CI is resampling the original data several times and selecting the best model to obtain the point estimates of the unobserved variables. Then pick the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ per centiles of the distribution formed by bootstrap estimates as the lower and upper bound of the CI, respectively.

4. Discussion

As some specific populations are hidden and difficult to reach by conventional means, their population size often needs to be estimated using mathematical modelling. Multiple systems estimation with the log-linear model is one of the common methods. In this paper, we use the AIC criteria to perform forward stepwise selection to choose the best model based on the MSE method to estimate the total population size, and use the bootstrap method to calculate its confidence interval.

Although there is some bias in using the MSE method to estimate the total population size, this study uses existing data to make estimates that are short time, low cost and easy to operate.

However, despite the advantages of simplicity and ease of use, the MSE method has more demanding conditions for application. The following four assumptions should be met when it is used: the study population remains constant over the study period, each list is independent of the other, each individual within the population has the same probability of being sampled and each list is within the same study population. And the difference in observed objects will significantly influence the estimation results. Besides, the confidence interval calculated by bootstrapping may be ridiculous as it takes into account the uncertainty of the model.

This study is able to estimate the total population by using the MSE method, however, I need to emphasize that the results should be considered exploratory until further research is undertaken. Furthermore, I need to collect more data to get a larger sample size to improve the reliability of the model. In addition, for the confidence interval, due to the lack of stability of the classic bootstrap method, we can further consider adopting the Bayesian bootstrap method and the parameter empirical Bayesian (PEB) bootstrap method to narrow the confidence interval and make it more reasonable.

References

[1] Nichols, J. D. (1992). Capture-Recapture Models. BioScience, 42(2), pp.94-102. doi: 10.2307/1311650
[2] Abeni, D., Brancato, G. and Perucci, C. (1994). Capture-Recapture to Estimate the Size of the Population with Human

Immunodeficiency Virus Type 1 Infection. Epidemiology, 5(4), pp.410-414.
[3] King, R. and Brooks, S. P. (2008). On the Bayesian Estimation of a Closed Population Size in the Presence of Heterogeneity and Model Uncertainty. Biometrics, 64(3), pp.816-824.
[4] Klingwort, J., Buelens, B. and Schnell, R. (2021). Capture-Recapture Techniques for Transport Survey Estimate Adjustment Using Permanently Installed Highway-Sensors. Social Science Computer Review, 39(4), pp.527-542. doi: 10.1177/0894439319874684.
[5] Le Cren, E. D. (1965). A Note on the History of Mark-Recapture Population Estimates. Journal of Animal Ecology, 34(2), pp.453-454.
[6] Sekar, C. 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(245), pp.101-115.
[7] Wittes, J. T., Colton, T. and Sidel, V. W. (1974). Capture-Recapture Methods for Assessing the Completeness of Case Ascertainment when Using Multiple Information Sources. Journal of chronic diseases, 27(1), pp. 25-36.
[8] Schnabel, Z. E. (1938). The Estimation of Total Fish Population of a Lake. The American Mathematical Monthly, 45(6), p.348-352
[9] Silverman, B. (2014). Modern slavery: An application of multiple systems estimation. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/386841/Modern_Slavery_an_application_of_MSE_revised.pdf (Accessed: 10 August 2022)
[10] George, E. I. and Robert, C. P. (1992). Capture-Recapture Estimation via Gibbs Sampling. Biometrika, 79(4), pp.677-683.
[11] Arnold, R., Hayakawa, Y. and Yip, P. (2010). Capture-Recapture Estimation Using Finite Mixtures of Arbitrary Dimension. Biometrics, 66(2), pp.644-655. doi: 10.1111/j.1541-0420.2009.01289.x
[12] Fienberg, S. E. (1972). The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables. Biometrika, 59(3), pp.591. doi: 10.2307/2334810
[13] Basu, S. and Ebrahimi, N. (2001). Bayesian Capture-Recapture Methods for Error Detection and Estimation of Population Size: Heterogeneity and Dependence. Biometrika, 88, pp.269-279.
[14] Fienberg, S. E., Johnson, M. S. and Junker, B. W. (1999). Classical Multilevel and Bayesian Approaches to Population Size Estimation Using Multiple Lists. Journal of the Royal Statistical Society: Series A (Statistics in Society), 162(3), pp.383-405. doi: 10.1111/1467-985X.00143.
[15] Cormack R. (1989). Log-Linear Models for Capture-Recapture. Biometrics, 45(2), pp.395.
[16] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (2007). Discrete multivariate analysis theory and practice: theory and practice. Dordrecht: Springer.
[17] Xu, C., Jing, F., Lu, Y., Ni, Y., Tucker, J., Wu, D., Zhou, Y., Ong, J., Zhang, Q. and Tang, W. (2022). Summarizing methods for estimating population size for key populations: a global scoping review for human immunodeficiency virus research. AIDS Research and Therapy, 19(1). doi: 10.1186/s12981-022-00434-7.
[18] Rich, A. J., Lachowsky, N. J., Sereda, P., Cui, Z., Wong, J., Wong, S., Jollimore, J., Raymond, H. F., Hottes, T. S., Roth, E. A., Hogg, R. S. and Moore, D.M. (2018). Estimating the Size of the MSM Population in Metro Vancouver, Canada, Using Multiple Methods and Diverse Data Sources. Journal of Urban Health: Bulletin of the New York Academy of Medicine, 95(2), pp.188-195. doi: 10.1007/s11524-017-0176-8.
[19] Krebs, C. J. (1998). Ecological methodology. 2nd ed. Menlo Park: Benjamin/Cummings.
[20] Hook E. and Regal R. (1995). Capture-Recapture Methods in Epidemiology: Methods and Limitations. Epidemiologic Reviews, 17(2), pp.243-264.
[21] International Working Group for Disease Monitoring and Forecasting. (1995). Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development. American Journal of Epidemiology. 142(10), pp.1047-1058. doi: 10.1093/oxfordjournals.aje.a117558.
[22] Akaike H. (1974). A new look at the statistical model identification. IEEE TranS1Actions on Automatic Control, 19(6), pp.716-723. doi: 10.1109/TAC.1974.1100705.
[23] Altman, D., Machin, D., Bryant, T. and Gardner, M. (2013). Statistics with Confidence: Confidence Intervals and Statistical Guidelines. Hoboken: Wiley.
[24] Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236(767), pp.333-380.