A new design for estimating prevalence of diseases by Capture-Recapture

Jean-Benoit Hardouin, Anne Viallefont

A new design for estimating prevalence of diseases by Capture-Recapture

Jean-Benoit Hardouin* et Anne Viallefont*†
Laboratoire SABRES
Université de Bretagne Sud - Tohannic
rue Yves Mainguy
56000 Vannes - France

Abstract

The Capture-Recapture method was developed in animal ecology to estimate animal population parameters. Its applications in epidemiology, developed at the end of the 80's, set specific problems and generally lead to over-estimations of the sizes of the populations, due to the fact that the sampling lists are not independent and do not cover the same population. In this context, we propose an adaptation of the "Robust Design" to epidemiology. This method was proposed in zoology to permit a robust estimation of the population parameters. We obtain a new estimator that allows us to relax the assumption that the various lists sample the exact same population. We compare it to the classic estimator in different cases, in terms of bias and asymptotic variance, using simulations for the latter. The new estimator has various qualities (bias reduction, robustness with respect to lists sizes, efficiency) compared to the classic estimator, in all the considered cases.

keywords : Epidemiology, Robust Design, Log-Linear Models, Bias, Efficiency, Population size estimation.

*present adress : ORS du Centre - CHRO - 1 rue Porte Madeleine - BP 2439 - 45032 ORLEANS CEDEX 1 - France - phone : 33(0)2.38.74.42.36 - fax : 33(0)2.38.74.48.81 - jean-benoit.hardouin@univ-ubs.fr

†Corresponding author - present adress : Université Lumière Lyon 2 - Equipe de Recherche en Ingénierie des Connaissances - Bât L - 5 avenue Mendès France - 69676 BRON CEDEX - France - phone : 33(0)4.78.77.31.41 - fax : 33(0)4.78.77.23.75 - aviallef@univ-lyon2.fr

Introduction

The Capture-Recapture method, developed in animal population dynamics, was adaptated during the 80's to allow epidemiologists to estimate the prevalence of a disease [1][2]. This method is an alternative to prevalence studies, expensive and not always feasible. It is based upon log-linear models for a contingency table crossing the results of several lists, such as hospital records, health surveys, etc [3][4]. This method relies on two assumptions : first, the different lists are considered independent (for each patient, his/her enrolled on a list is a independant process of whetheher he/she is enrolled on another list); second, the different lists are supposed to sample exactly the same population. Undoubtedly, these two assumptions are rarely met, and difficult to test. This leads in general to an over-estimation of the size of the population and to a large variance of the estimator. This estimator is thus of little use in practice [5][6][7][8].

In this context, we propose a new estimator of the size of the population that allows the lists to sample different sub-populations within a population. The formal asymptotic expectation of this new estimator is given, and simulations are used to estimate its variance. The new estimator is less biased and more efficient than the classic one and it is more robust to uneven sizes of the sub-populations sampled by the lists.

1 Problem and proposition

1.1 Notations

k	Number of lists
N	Total size of the population covered by the union of the k lists (to estimate)
N_l	Total size of the population covered by the list l (unknown)
N_{11}	Total size of the population covered by the two lists (if $k = 2$) (unknown)
n	Total number of sampled individuals
n_l	Total number of individuals sampled on the list l
n_{11}	Total number of individuals sampled on the two lists (if $k = 2$)
q_l	Probability to sample an individual on list l
q_{11}	Probability to sample an individual on the two lists (if $k = 2$)
i_l	Status of individual on the list l ($i_l = 1$ for the individuals seen on the list l and $i_l = 0$ for the individuals not seen on this list)
$n_{i_1 \ldots i_l \ldots i_k}$	Total number of individuals with status i_l on list $l, l = 1 \ldots k$

1.2 Problem

In order to estimate the prevalence of a disease with Capture-Recapture in epidemiology, lists like registers, doctors lists, hospitals lists are often used as samples of an underlying population. Afterwards, a 2^k contingency table across all the samples is set up. Each cell of this table represents a profile of "capture". This table is, by definition, incomplete because the number of individuals who are never sampled ($n_{0...0}$) is unknown. Thus one works with $(2^k - 1)$ profiles [3].

A main assumption is made : all the lists are are supposed to sample exactly the same population.

The following model is fitted to the data :

$$\log[E(n_{i_1...i_k})] = u + u_1(i_1) + ... + u_k(i_k) + u_{12}(i_1 i_2) + ... + u_{1...k}(i_1...i_k),$$

where the $u_j(1)$ $j = 1, ...k$ are the parameters associated to presence on the list j, and the $u_{1...l}(i_1...i_l)$ are the interaction parameters between the samples caracterised by non-nul ij's. Each term $u_{1...l}(i_1...i_l) = 0$ if at least one element $i_j = 0$ $\forall l = 1...k$ and $\forall j = 1...l$, and all the terms $u_{1...k}(i_1...i_k) = 0$ (because we assume there is no intercation between all the lists). This model is saturated since it includes $(2^k - 1)$ terms for $(2^k - 1)$ profiles.

Hereafter, we shall consider only two lists ($k = 2$). In this case, we make the assumption that the two lists are independent. The estimator $\hat{N}$ is $\frac{n_1 n_2}{n_{11}}$, and we obtain asymptotically [4]

$$E(\hat{N}) = \frac{E(n_1)E(n_2)}{E(n_{11})} = \frac{N_1 q_1 N_2 q_2}{N_{11} q_{11}}.$$

If the samples on the lists are independant ($q_{11} = q_1 q_2$) we have the Peterson Index

$$E(\hat{N}) = \frac{N_1 N_2}{N_{11}},$$

and if the two lists cover the exact same population, then $N_1 = N_2 = N_{11} = N$ and $\hat{N}$ is unbiased.

The asymptotic variance of $\hat{N}$ is

$$V(\hat{N}) = \frac{E(n_1)E(n_2)E(\hat{n}_{00})}{[E(n_{11})]^2} = \frac{E(\hat{N})E(\hat{n}_{00})}{N_{11} q_{11}},$$

with $E(\hat{n}_{00}) = (N_1 - N_{11})(1 - q_1) + (N_2 - N_{11})(1 - q_2) + N_{11}(1 - q_{11})$.

This variance very large whan the proportion of individuals seen on the 2 lists is small. It increases when $\frac{N_{11}}{N}$ decreases, i.e. when the sub-populations covered by the different lists are not exactly the same. When $N_1 = N_2 = N_{11} = N$ and $q_{11} = q_1 q_2$, we have $V(\hat{N}) = N\frac{1-q_1 q_2}{q_1 q_2}$, i.e. the variance only depend on the size of the population an on the "capture rates".

1.3 Proposition

In order to relax the assumption that all the lists cover the same underlying population, we propose a new experimental design. In this design, the size of the sub-populations covered by each list is first estimated, then the size of the total population, considered as the union of the various sub-populations covered by the lists, can be estimated.

To estimate the size of the sub-population covered by one list, we use the replicates appearing in the lists, as follows : we consider two periods (or more) on the list, and we consider the individuals sampled at each period as two independent samples for the underlying sub-population. We use the method described previously to estimate the size of this sub-population. The second step consists in estimating the size of the intersection of the underlying sub-population covered by several lists. Finally, we can estimate the size of the population of interest, considered as the union of the studied sub-populations. This design was inspired by Pollock's works on the "Robust Design"[9], proposed in animal ecology to solve the problem of estimation of the size of the studied population at several dates.

This new methodology relies on several new assumptions :

- The samples in a given list are independent, and the probability of appearing on a list at a given period is the same for all the individuals covered by this list.
- The sampling processes on the lists are independent (this assumption is useful to estimate the size of the intersection of the two sub-populations),
- The population of interest is the union of the different sub-populations,

In part 2.5, we study how violations of these new assumptions affect the results. Indeed, if an individual is present in a list at a given period, his probability to be present on the

same list for the same disease at another period may be lower than for another individual not already present on the list. We are in presence of a process similar to the trap-shyness in animal ecology [10].

2 Construction and properties of the new estimator

2.1 Complementary notations

X_{lj}	Dichotomic variable representing the fact that one individual is seen at the date j on the list l
$n_{l(i_1 i_2)}$	Total number of individuals, with status i_1 in period 1 and i_2 in period 2, on list l
p_{lj}	Probability to sample an individual during the period j on list l
Υ_l	Degree of "trap-response" between the two periods for list l
δ	Degree of dependency between the two lists (for the N_{12} individuals potentially sampled by the two lists)

2.2 Construction of the new estimator

Several samples are made in each sub-population at different dates. Hereafter, we will consider only two sub-populations and two samples per sub-population (*i.e.* two periods of time). Classical Capture-Recapture methods provide estimates of the size of each sub-population sampled by each list :

$$\tilde{N}_l = \frac{n_{l(1.)} n_{l(.1)}}{n_{l(11)}},$$

where $n_{l(.j_2)} = n_{l(0j_2)} + n_{l(1j_2)}$ and $n_{l(j_1.)} = n_{l(j_10)} + n_{l(j_11)}$.

We estimate the probability of presence of an individual in list l as

$$\tilde{q}_l = \frac{n_l}{\tilde{N}_l};$$

then, we can estimate the probability of presence of an individual in two lists, assuming independence of the lists as

$$\tilde{q}_{11} = \tilde{q}_1 \tilde{q}_2.$$

We can estimate the size of the sub-population sampled by the two lists as

$$\tilde{N}_{11} = \frac{n_{11}}{\tilde{q}_{11}}.$$

Finally, we can estimate the size of the population as

$$\tilde{N} = \tilde{N}_1 + \tilde{N}_2 - \tilde{N}_{11}.$$

2.3 Asymptotic properties of the new estimator in the "classic" case

In the classic use of capture-recapture method in epidemiology, the lists are supposed independent. With the new estimator this assumption has still to hold. In a first approach, we consider that this assumption is fulfilled and that there is no "trap-response" between the different samples on each list. In parts 2.4 and 2.5, we will consider the general case when when there is a process of "trap-shyness" between the different samples on each list and when the assumption of independence is not met.

For each sub-population covered by a list, we realise two independant samples on two periods. The subpopulation covered by the two samples is exactly the same and thus :

$$E(\tilde{N}_l) = \frac{E(n_{l(1.)})E(n_{l(.1)})}{E(n_{l(11)})} = \frac{(N_l p_{l1} N_l p_{l2})}{N_l p_{l1} p_{l2}} = N_l,$$

thus the two estimators $\tilde{N}_l$ are unbiased.

In the case of independence, it is easily shown that the new estimator is unbiased. In this case, we also have :

$$E(n_{11}) = N_{11} q_{11} = N_{11} q_1 q_2,$$

and $\forall l = \{1, 2\}$

$$E(\tilde{q}_l) = E\left(\frac{n_l}{\tilde{N}_l}\right) = E\left(\frac{n_l n_{l(11)}}{n_{l(1.)} n_{l(.1)}}\right) = E\left(\frac{N_l q_l N_l p_{l1} p_{l2}}{N_l p_{l1} N_l p_{l2}}\right) = E(q_l) = q_l,$$

and

$$E(\tilde{q}_{11}) = E(\tilde{q}_1 \tilde{q}_2) = E(\tilde{q}_1)E(\tilde{q}_2) = q_1 q_2 = q_{11}.$$

Asymptotically, we obtain :

$$\begin{aligned} E(\tilde{N}_{11}) &= E\left(\frac{n_{11}}{\tilde{q}_{11}}\right) \\ &= E\left(\frac{N_{11} q_{11}}{\tilde{q}_1 \tilde{q}_2}\right) \end{aligned}$$

$$
\begin{aligned}
&= E\left(\frac{N_{11}q_{11}\tilde{N}_1\tilde{N}_2}{n_1 n_2}\right)\\
&= E\left(\frac{N_{11}q_1q_2n_{1(1.)}n_{1(.1)}n_{2(1.)}n_{2(.1)}}{N_1q_1N_2q_2n_{1(11)}n_{2(11)}}\right)\\
&= E\left(\frac{N_{11}N_1p_{11}N_1p_{12}N_2p_{21}N_2p_{22}}{N_1N_2N_1q_1N_2q_2}\right)\\
&= E\left(\frac{N_{11}p_{11}p_{12}p_{21}p_{22}}{p_{11}p_{12}p_{21}p_{22}}\right)\\
&= E(N_{11})\\
&= N_{11},
\end{aligned}
$$

i.e. the estimator of the size of the intersection is also asymptotically unbiased when the three assumptions made earlier are fulfilled. Finally, the asymptotic expectation of our new estimator is :

$$E(\tilde{N}) = E(\tilde{N}_1) + E(\tilde{N}_2) - E(\tilde{N}_{11}) = N_1 + N_2 - N_{11} = N.$$

2.4 Properties of the new estimator in the case of "trap-response" between the samples of a given list

In practice, the probability that an individual is seen at a date on a list may depend on whether this individual was seen at other dates on the same list ("trap-response").

We can define Υ_l, the "Degree of trap-response" within the list, as

$$P(X_{l2} = 1/X_{l1} = 1) = \Upsilon_l P(X_{l2} = 1) = \Upsilon_l p_{l2}.$$

We see easily that this "trap-response" is a reciprocal process because

$$P(X_{l1} = 1/X_{l2} = 1) = \frac{P(X_{l1} = 1, X_{l2} = 1)}{P(X_{l2} = 1)} = P(X_{l2} = 1/X_{l1} = 1)\frac{P(X_{l1} = 1)}{P(X_{l2} = 1)} = \Upsilon_l p_{l2}\frac{p_{l1}}{p_{l2}} = \Upsilon_l p_{l1}.$$

In fact, we think that data coming from lists should often present an apparent "trap-shyness"; *i.e.* if an individual is seen on a list at a date, his probability to be seen at an other date decreases. In this case, Υ_l is a real number between 0 and 1, depending on the strength of the "trap-shyness" ($\Upsilon_l = 1$ is the no trap-reponse case for the list l).

We still have

$$E(\tilde{N}_l) = \frac{E(n_{l(1.)})E(n_{l(.1)})}{E(n_{l(11)})},$$

now we have

$$E(n_{l(1.)}) = N_l p_{l1},$$

$$E(n_{l(.1)}) = N_l p_{l2},$$

and

$$E(n_{l(11)}) = N_l P(X_{l1} = 1) P(X_{l2} = 1/X_{l1} = 1) = N_l p_{l1} \Upsilon_l p_{l2},$$

thus

$$E(\tilde{N}_l) = E\left(\frac{n_{l(1.)} n_{l(.1)}}{n_{l(11)}}\right) = E\left(\frac{N_l p_{l1} N_l p_{l2}}{N_l p_{l1} \Upsilon_l p_{l2}}\right) = E\left(\frac{N_l}{\Upsilon_l}\right) = \frac{N_l}{\Upsilon_l};$$

Asymptotically

$$\begin{aligned} E(\tilde{q}_l) &= E\left(\frac{n_l}{\tilde{N}_l}\right) \\ &= E\left(\frac{N_l q_l n_{l(11)}}{n_{l(1.)} n_{l(.1)}}\right) \\ &= E\left(\frac{N_l q_l N_l p_{l1} \Upsilon_l p_{l2}}{N_l p_{l1} N_l p_{l2}}\right) \\ &= E(q_l \Upsilon_l) \\ &= q_l \Upsilon_l, \end{aligned}$$

and

$$E(\tilde{q}_{11}) = E(\tilde{q}_1) E(\tilde{q}_2) = q_1 \Upsilon_1 q_2 \Upsilon_2 = q_{11} \Upsilon_1 \Upsilon_2.$$

Thus asymptotically

$$\begin{aligned} E(\tilde{N}_{11}) &= E\left(\frac{n_{11}}{\tilde{q}_{11}}\right) \\ &= E\left(\frac{N_{11} q_{11}}{\tilde{q}_1 \tilde{q}_2}\right) \\ &= E\left(\frac{N_{11} q_1 q_2 \tilde{N}_1 \tilde{N}_2}{n_1 n_2}\right) \\ &= E\left(\frac{N_{11} q_1 q_2 n_{1(1.)} n_{1(.1)} n_{2(1.)} n_{(.1)}}{N_1 q_1 N_2 q_2 n_{1(11)} n_{2(11)}}\right) \\ &= E\left(\frac{N_{11} N_1 p_{11} N_1 p_{12} N_2 p_{21} N_2 p_{22}}{N_1 N_2 N_1 p_{11} \Upsilon_1 p_{12} N_2 p_{21} \Upsilon_2 p_{22}}\right) \\ &= E\left(\frac{N_{11}}{\Upsilon_1 \Upsilon_2}\right) \\ &= \frac{N_{11}}{\Upsilon_1 \Upsilon_2}, \end{aligned}$$

so

$$E(\tilde{N}) = E(\tilde{N}_1) + E(\tilde{N}_2) - E(\tilde{N}_{11}) = \frac{N_1}{\Upsilon_1} + \frac{N_2}{\Upsilon_2} - \frac{N_{11}}{\Upsilon_1\Upsilon_2} = \frac{\Upsilon_2 N_1 + \Upsilon_1 N_2 - N_{11}}{\Upsilon_1\Upsilon_2}.$$

We note that if the degree of "trap-response" is the same for the two lists ($\Upsilon_1 = \Upsilon_2 = \Upsilon$) then

$$E(\tilde{N}) = \frac{\Upsilon(N + N_{11}) - N_{11}}{\Upsilon^2}.$$

In this case, the bias of the estimator does not depend any longer on the sizes of the subpopulations N_1 and N_2 but only of the size of the population and of the size of the intersection between the two sub-populations.

We can represent the asymptotic relative bias of the new estimator as a function of the parameters. We will consider that the degree of "trap-response" (Υ) is the same for the two lists. When the lists are independent ($\delta = 1$), we know that the expectation of the new estimator only depends on the size of the population (N), on the size of the intersection between the two lists (N_{11}) and on the degree of trap-response (Υ). Figure 1 represents the bias of the new estimator in this case in fonction of the degree of trap-response. Different values are used for the size of the intersection (see table 1), the size of the population being fixed at $N = 2000$. We note that the new estimator is unbiased when $\Upsilon = 1$, it is positively biased when $\frac{N_{11}}{N} < \Upsilon < 1$ (minor trap-shyness) and negatively biased otherwise (trap-happiness or important trap-shyness). With a fixed degree of trap-response $\Upsilon > \frac{N_{11}}{N}$, the absolute value of the bias decreases as N_{12} increases.

2.5 Properties of the new estimator in the general case

In practice, the probability that an individual belonging to the intersection of the sampled populations is seen at a date on a given list certainly depends on whether this individual was seen at the same date on the other list (dependence between the two lists) : if this individual is already seen at a date on a given list, his probability to be seen at the same date on another list decreases.

We can define δ the degree of dependence between the lists as

$$P(X_{lj} = 1/X_{lj'} = x_{lj'}, X_{l'j} = 1) = \delta P(X_{lj} = 1/X_{lj'} = x_{lj'})$$

with l' the complementary of l in $\{1, 2\}$ and j' the complementary of j in $\{1, 2\}$, $\forall\, l, j \in \{1, 2\} \times \{1, 2\}$.

We note that this dependance between the two lists only concerns the individuals of the population covered by the two lists (N_{11} individuals).

The expectation of the estimators are in this case very hard to compute because each sample depends on two other samples and this phenomenon is reciprocal. We have prefered to run simulations to estimate the expectation of the estimator in this case. But the construction of simulate samples needs full of times because it's an iterative process.

So we have realised simulations that mimic the phenomenon of dependence, as defined in part 3.

2.6 Estimating the variance of the new estimator

Formal estimates of the asymptotic variance of $\tilde{N}$ can not be calculated without further assumptions on the distribution of the estimators. Thus, we chose to run simulations to estimate the variance of the new estimator in different cases.

3 Simulations

We have realised several series of 1000 simulations, in different cases, using the Monte Carlo method. The size of the simulated underlying population was 2000 individuals (value to estimate). Eight cases were taken into account according to whether the sub-populations covered by the various lists were equal in size, whether they had a large intersection and whether the lists were independent (table 1).

For each case, we have defined 9 scripts according to the probability of capture on each sample of the two sub-populations p_1 and p_2 (table 2).

We then have simulated 9 new scripts in each case, with "trap-response" between the different samples of a list. These scripts are numbered from 10 to 18 and the probabilities p_i are the same than for the scripts 1 to 9.

To simulate trap-response between the different samples of a list, we have chosen $\Upsilon_i = \Upsilon = 0.8 \ \forall i = 1, 2$. For each list we have simulated N_1 and N_2 independent Bernoulli variables with probability p_1 and p_2 (variables X_{11} and X_{21}). The variables X_{12} and X_{22} have been simulated in function of the result for each individual for the variables X_{11} and X_{21} respectively : if $X_{l1} = 1$ then $P(X_{l2} = 1/X_{l1} = 1) = \Upsilon_l p_l = 0.8 p_l$ and if $X_{l1} = 0$ then

$P(X_{l2} = 1/X_{l1} = 0) = \frac{p_l(1-\Upsilon p_l)}{1-p_l}$ [to verify that $P(X_{l2} = 1) = p_l$].

To simulate dependence between the lists, we chose $\delta = 0.8$. For the sake of computational simplicity, we first simulated the variables X_{11} and X_{12} as described above. The variable X_{21} was simulated in function of the result of the variable X_{11} and the variable X_{22} was simulated in function of the result of the variables X_{21} and X_{12}. The probabilities we used are given in table 3. The probabilities $P(X_{12} = 1/X_{11} = 0)$, $P(X_{21} = 1/X_{11} = 0)$ and $P(X_{22} = 1/X_{12} = 0, X_{21} = 0)$ are computed to verify that $\forall\ l \in \{1,2\}, \forall\ j \in \{1,2\}, P(X_{lj} = 1) = p_l$.

4 Results

In tables 4 and 5, we give the results of the simulations, when the lists are independent (Cases I to IV). The first table considers the scripts when there is no "trap-response" (scripts 1 to 9) and the second one, the scripts when there is "trap-shyness" (scripts 10 to 18). In tables 6 and 7, we give the results for the cases when the lists are dependent (cases V to VIII), distributed in the same way between the two tables. For each case, we only give the minimum and the maximum of the asymptotic formal results for the expectations and for the standard error of the classic estimator and the range of the expectation and of the standard error of the two estimators obtained by the simulations over all the scripts. The asymptotic results for the expectation of the new estimator was given only when the two lists are independent.

According to these results, we can consider two main situations : two sub-populations with a large intersection (cases I, III, V and VII) ; two sub-populations with a small intersection (cases II, IV, VI, VIII). In the first situation, the bias of the two estimators are generally positive (except in case I when the two sub-populations largely covered the population) and relatively small (with a maximum of 25.5% for the classic estimator and of 20.8% for the new estimator on all the scripts). We note that, in these cases, the new estimator is unbiased when the lists are independent (cases I and III) and when there is no "trap-response" between the dates of sampling in a sub-population (scripts 1 to 9). We also note that the new estimator is unbiased in the case I when there is "trap-shyness". This result is due to chance, as it comes from the fact that, when $\Upsilon_i = \Upsilon\ \forall i$, the new

estimator is unbiased if $\Upsilon = 1$ (no trap-response) or if $\Upsilon = \frac{N_{11}}{N}$. In the case I, we have $\frac{N_{11}}{N} = \frac{1600}{2000} = 0.8$ and we have chosen $\Upsilon = 0.8$, so that, in this case, the new estimator happens to be unbiased.

In these cases, the variances fall in the same range. The variance of the new estimator is, in general, slightly higher than the one of the classic estimator when the probabilities of capture at each date on the sub-populations are small and inversely, when these probabilities are large.

When the sub-populations sampled have a small intersection, the classic estimator is very biased, due to the unrealistic assumption that all the lists cover the same population. The bias of the new estimator seem to be always much lower than that of the classic estimator. As for the variance, that of the new estimator is much lower than that of the classic estimator. In this case, the new estimator is thus much better than the classic one.

5 Discussion

The new estimator had various advantages compared with the classic one. Indeed, its properties are similar to those of the classic one when the assumptions of independance of the lists and of sampling of the same population are respected or not strongly violated. When this assumptions are strongly violated, the new estimator is better than the classic one as regards of bias and consistency, because the strong assumption of unicity of the population sampled by the differents lists is relaxed.

The new estimator could easily be generalised to more than two lists, and more than two dates per list, although, in practice, its application will certainly be limited to two or three lists and two or three dates per list.

The main difficulty that remains to routinely use this estimator seems to be the choice of cut for splitting the different lists at different dates. However, epidemiologists often make several samples on the lists and delete the replicates to use the classic estimator. These deletions of information prevented us from testing our estimator on published data.

Thus, the pratical use of this design should not pose any problem because it does not impose additional work. Furthermore, the computation of the new estimator permits to estimate the size of each sub-population, a piece of information that was never obtained

with the former method.

References

[1] Stephen Craig. Capture-recapture methods in epidemiological studies. *Infection Control and Hospital Epidemiology*, 17: 262–266, 1996.

[2] Daniel J. McCarty, Eugene S. Tull, Claudia S. Moy, C. Kent Kwoh, and Ronald E. LaPorte. Ascertainment corrected rates : Applications of capture-recapture methods. *International Journal of Epidemiology*, 22(3): 559–565, 1993.

[3] Richard M. Cormack. Log-linear models for capture-recapture. *Biometrics*, 45: 395–413, 1989.

[4] Stephen E. Fienberg. The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59: 591–603, 1972.

[5] Richard M. Cormack. Can capture-recapture models used in epidemiology give sensible estimates of prevalence rates? In *Guest communication, fiftieth birthday of the british branch of the International Society of Biometry*, 7-12 April 1998, Edinburgh.

[6] Jean-Claude Desenclos and Bruno Hubert. Limitations to the universal use of capture-recapture methods. *International Journal of Epidemiology*, 23(6): 1322–1323, 1994.

[7] Jørgen Hilden. Ascertainment corrected rates : Applications of capture-recapture methods. *International Journal of Epidemiology*, 23(4): 865–866, 1994.

[8] Laure Papoz, Beverley Balkau, and Joseph Lelouch. Case counting in epidemiology : Limitations of methods based on multiple data sources. *International Journal of Epidemiology*, 25(3): 474–478, 1996.

[9] Kenneth H. Pollock. A capture-recapture design robust to unequal probability of capture. *Journal of Wildlife Management*, 46(3): 757–760, 1982.

[10] Kenneth H. Pollock. A k-sample tag-recapture model allowing for unequal survival and catchability. *Biometrika*, 62: 577–583, 1975.

Table 1: Simulated cases

Case	I	II	III	IV	V	VI	VII	VIII
same sub-population sizes	X	X			X	X		
large intersection	X		X		X		X	
independence	X	X	X	X				
N_1	1800	1100	1800	1800	1800	1100	1800	1800
N_2	1800	1100	1100	400	1800	1100	1100	400
N_{11}	1600	200	900	200	1600	200	900	200

Table 2: Probabilities used in simulations

scripts without "trap-response"	scripts with "trap-shyness"	p_1	p_2
1	10	0.5	0.5
2	11	0.5	0.3
3	12	0.5	0.1
4	13	0.3	0.3
5	14	0.3	0.1
6	15	0.1	0.1
7	16	0.3	0.5
8	17	0.1	0.5
9	18	0.1	0.3

Table 3: Probabilities used in simulations with dependance between the two lists

Probability	population covered only by list 1	population covered only by list 2	population covered by the two lists
Size	$N_1 - N_{11}$	$N_2 - N_{11}$	N_{11}
$P(X_{11} = 1)$	p_1	0	p_1
$P(X_{12} = 1/X_{11} = 1)$	$\Upsilon_1 p_1$	$-$	$\Upsilon_1 p_1$
$P(X_{21} = 1/X_{11} = 1)$	$-$	$-$	δp_2
$P(X_{22} = 1/X_{11} = 1, X_{21} = 1)$	$-$	$-$	$\delta \Upsilon_2 p_2$
$P(X_{22} = 1/X_{11} = 0, X_{21} = 1)$	$-$	$\Upsilon_2 p_2$	$\Upsilon_2 p_2$
$P(X_{22} = 1/X_{11} = 1, X_{21} = 0)$	$-$	$-$	δp_2

Table 4: Results of the simulations : N=2000, independent lists, no "trap-response"

cases		asymptotic expectations	range of asymptotic standard errors	ranges of simulations expectations	range of simulations standard errors
I	classic	2025	21-219	2023-2054	21-235
	new	2000		1999-2013	18-237
II	classic	6050	157-1153	6068-7046	360-3662
	new	2000		2001-2170	40-474
III	classic	2200	37-316	2201-2244	43-260
	new	2000		2000-2049	27-279
IV	classic	3600	121-890	3613-4266	196-2443
	new	2000		2000-2135	40-440

Table 5: Results of the simulations : N=2000, independent lists, "trap-shyness"

cases		asymptotic expectations	range of asymptotic standard errors	ranges of simulations expectations	range of simulations standard errors
I	classic	2025	16-216	2023-2049	16-236
	new	2000		1979-1999	23-361
II	classic	6050	131-1140	6068-7231	296-3749
	new	2437		2442-2651	59-813
III	classic	2200	31-312	2198-2256	35-352
	new	2218		2214-2258	27-384
IV	classic	3600	101-879	3608-4311	167-2548
	new	2437		2441-2506	60-605

Table 6: Results of the simulations : N=2000, dependent lists, no "trap-response"

cases		range of asymptotic expectations	range of asymptotic standard errors	ranges of simulations expectations	range of simulations standard errors
V	classic	2145-2228	18-240	2145-2274	22-270
	new			2115-2192	17-248
VI	classic	6504-6882	170-1279	6615-7692	420-4043
	new			2026-2167	41-484
VII	classic	2345-2422	37-348	2343-2511	45-452
	new			2072-2159	27-305
VIII	classic	3862-4028	129-983	3911-4808	234-2643
	new			2023-2156	41-521

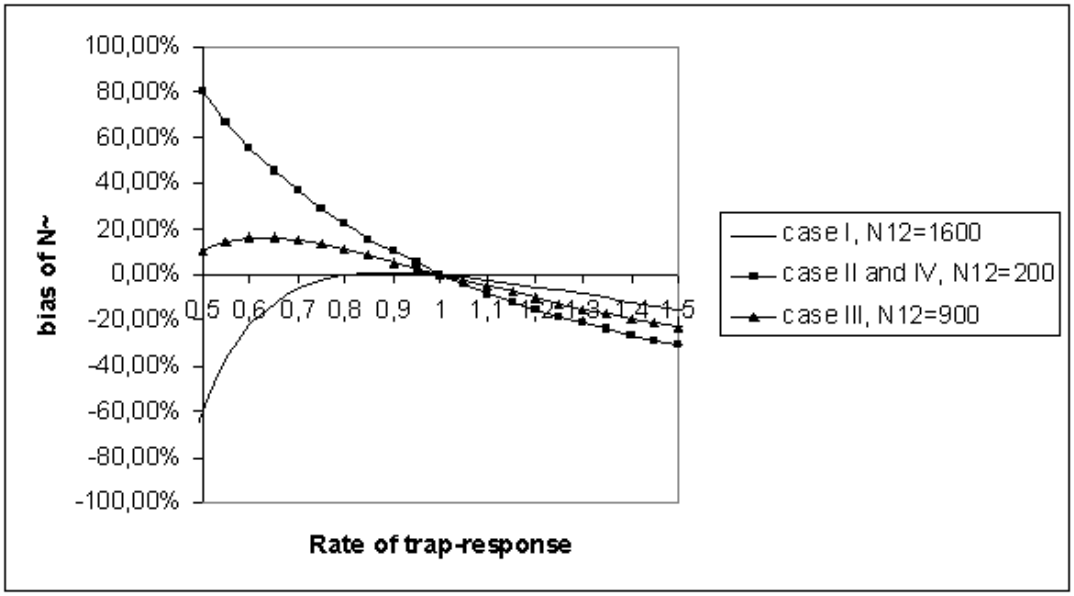
Table 7: Results of the simulations : N=2000, dependent lists, "trap-shyness"

cases		range of asymptotic expectations	range of asymptotic standard errors	ranges of simulations expectations	range of simulations standard errors
V	classic	2137-2234	13-238	2137-2269	17-269
	new			2118-2230	25-381
VI	classic	6651-6780	144-1268	6777-7930	387-4345
	new			2492-2737	64-779
VII	classic	2343-2430	30-345	2343-2498	38-438
	new			2309-2415	31-401
VIII	classic	3944-3997	131-987	3907-4777	238-2688
	new			2477-2629	42-440

Captions for figures

Figure 1 Bias of $\tilde{N}$ as a function of N_{11} and of the degree of "trap-response" Υ when the lists are independent ($\delta = 1$)

Figure 1:


100,00%
80,00%
60,00%
40,00%
20,00%
0,00%
-20,00%
-40,00%
-60,00%
-80,00%
-100,00%
bias of N~
0,5 0,6 0,7 0,8 0,9 1 1,1
Rate of trap-response
case I, N12=1600
case II and IV, N12=200
case III, N12=900