Estimating the number of opiate users in Rotterdam using statistical models for incomplete count data *

December 1997

Filip Smit [1], Jaap Toet [2], Peter van der Heijden [3]

[1] Methods Section, Trimbos Institute, Netherlands Institute of Mental Health and Addiction, Utrecht, The Netherlands. [2] Dept. of Health Promotion, Municipal Health Service, Rotterdam, The Netherlands. [3] Dept. of Methods & Statistics, Faculty of Social Sciences, University of Utrecht, The Netherlands.

Summary
Often, the number of opiate users in a city is not known. This lack of knowledge hampers adequate policy making. However, the unknown size of the opiate using population can be estimated with help of statistics. We review seven statistical models and apply them to a real data-set. All models are based on a single data-set. This data-set is incomplete in the sense that only opiate users who apply for methadone maintenance are registered while all others are not. The only information which is available in this data-set is the number of visits of uniquely identifiable persons who are on methadone maintenance, plus a set of covariates like age, sex and so on for each of these persons. Still, from this single and incomplete data-set inferences about the total population size can be made. The estimates are compared and the pro's and con's of the estimators are evaluated.

Key words & phrases: population size estimators, prevalence, opiate use, capture-recapture, truncated Poisson, truncated Negative Binominal, regression, covariates, assumptions.

* **Please, cite as**: Smit F, Toet J, Van der Heijden P. (1997). Estimating the number of opiate users in Rotterdam using statistical models for incomplete count data. In: Hay G, McKeganey N, Birks E. Final report EMCDDA project methodological pilot study of local level prevalence estimates. Glasgow: Centre for Drug Misuse Research, University of Glasgow / EMCDDA.

1. Introduction

1.1 Background

The total number of persons in a city who are dependent on opiates is usually not known. Some opiate users might be known through health services, police contacts and the like, but others are never seen. Such a population is said to be 'partially observed' just like the proverbial ice berg. Not knowing the total size of a population hampers developing, implementing and evaluating health policies. Therefore, it is important to accurately estimate the unknown size of a population. Statistics can play a role here.

The utility of population size estimation has long been recognized. As a consequence there are many population size estimators; see Seber (1982, 1886, 1992), Pollock (1991) and Khorrazaty *et al*. (1977) for reviews. These estimators have by and large been developed in biometrics (fisheries and wild life studies) and their validity, when applied to human populations, is not well established. This is one reason to exercise some care and good judgement when selecting an estimator for studying human populations, specifically a population of opiate users (Simeone, Nottingham and Holland, 1993). An other reason presents itself when new estimators are developed and have not yet been tried very often. Either way, one likes to know how competing estimators perform and what they 'cost' in terms of data collection and computational effort. Predecessors in this form of methodology comparison are, for example, Simeone, Nottingham and Holland (1993), Brecht and Wickens (1993), Wickens (1993). This is not to say that all methodological issues have been resolved.

1.2 Purpose

In this study we want to compare different population size estimators when applied to a partially observed population of opiate users. To this end, we use a real life data-set which contains information on opiate users in Rotterdam in 1994. We will compare these estimators in terms of the 'quality' of their outcomes (congruence, criterion validity), assumptions (realism, validity) and 'costs' in terms of data requirements and computational effort. The rationale for this comparison is that each of these estimators is based on different assumptions and violation of these assumptions may render the estimates invalid. Also, some estimators require very little information whereas other estimators can only be calculated when more information is available. Possibly, there is a trade off between validity and the effort to obtain the required data. The purpose of this study, therefore, is to evaluate the balance between the 'quality' of the results and the 'costs' of the required data for these estimators.

1.3 Approach

In the method paragraph we will describe the population, the data source, the estimators and their underlying assumptions. The estimators are

1. two truncated Poisson estimators independently developed by Daniel Zelterman (1988) and Anne Chao (1987, 1989), and their stratified counterparts.

2. two estimators based on the truncated Poisson and the truncated Negative Binominal regressions with covariates which have recently been developed by Peter Van der Heijden *et al* (in progress).

We will not deal exhaustively with the related estimation procedures and refer the interested reader to the relevant literature. In the results paragraph we will present the outcomes of each of the estimators when applied to the data-set. Finally, in the discussion paragraph we will evaluate the pro's and con's of each of the estimators.

2. Methods

2.1 Data

Data on opiate users who apply for treatment are routinely collected and entered in the Roterdam Drug Information System (RODIS). These data are collected through three health service centres where opiate users are on methadone maintenance. The close corporation between the health centres effectively makes it one single institute which happens to have offices at three different locations. We therefore treat the RODIS-data as coming from a single 'data collecting agent'. The vistors of the centres can be classified as 'problematic opiate users'. In 1994, the case register contained information on 2029 persons. Per person a number of variables is known, like: sex, age, date of first contact, number of visits, marital status, educational level, source of income. A unique number is assigned to each person. This number serves as an identifier of that person. This identification helps to telly the number of visits of that particular person. In the following analyses we use the frequencies of these visits in the year 1994 as a key variable. We will now discuss the estimators and their underlying assumptions in more detail.

2.2 Zelterman's and Chao's truncated Poisson estimators

Zelterman's (1988) and Chao's (1987) estimators, but see also Chao (1989) and Wilson and Collins (1992), can be applied on data generated by counts of individuals who have been seen once, twice and so on. In our study, the health centre's administrator tellies the number of visits brought by persons who are on methadone maintenance. Persons who are never seen fall into the zero frequency class and are missing from the observed series of frequencies. Therefore, the frequencies of the visits are incomplete and are called 'truncated below one'. Naturally, the total population size equals the number of persons ever seen plus the number of persons never seen. The estimation problem, then, becomes to estimate the number of person never seen from the truncated series of persons ever seen. Both Zelterman's and Chao's estimators are based on this idea and both assume that the observed series of frequencies follows a Poisson distribution which is truncated below one. Since the calculations are so easy, we give the equations. Zelterman's (1988) estimator of the unknown population size, est(N), is given by

$$\text{est}(N) = S / [1-\exp(-2f_2 / f_1)]$$

and Chao's (1987) estimator is given by

$$\text{est}(N) = S + f_1^2 / (2f_2)$$

where,

f_1 = the number of persons falling in the first frequency class
f_2 = the number of persons falling in the second frequency class.
S = the sum of all frequencies.

We refer the interested reader to the cited literature for the calculations of the 95% confidence intervals.

Note, both estimators are primarily based on the lower frequency classes (f_1 and f_2). This emphasis on the lower frequencies classes makes sense. People rarely seen (only once or twice) are likely to bare a greater resemblance with persons never seen than people seen very often. In addition, the emphasis on the lower frequency classes makes the estimators robust in the presence of 'heterogeneity', e.g. persons seen very often may form a different subgroup as compared to persons rarely seen. The influence of the persons often seen is weighted down in both estimators and therefore heterogeneity, if present, is likely to exercises a relative small influence. Finally, the emphasis on the lower frequency classes results in an other bonus as well: both estimators are known to perform rather good even when we have few data.

For both estimators to be valid it must be assumed that
1. the population is 'closed'
2. the individual probabilities to be observed and re-observed are constant over time
3. the population of interest is homogeneous (no heterogeneity across individuals).

The first assumption, known as the 'closure assumption', asserts that the true population size, N, is unaffected by migration, birth and death during the period of interest. In this particular study, we have chosen a period of one year because we want to estimate the one year prevalence of opiate use in Rotterdam. We, therefore, must hope that the true population size is not too much affected by in-migration and out-migration. Keeping the study period short (say, one month) is one way of meeting the closure assumption. Evidently, it is hard to see how the population size of opiate users can change dramatically in a single month. Note, that a shorter period will result in fewer observations, but then again, both estimators are known to perform well when the data are sparse.

The second assumption -about constant capture probability- effectively denies the possibility that individuals show a behavioural response to the treatment they receive.

Whatever their experience with the methadone treatment, their probability to become a second time or a third time visitor is assumed to depend on a constant individual probability of being observed one time, two times, three times and so on. Clearly, with respect to the data generating process this assumption is a worrying one. At any rate, we do not find it particularly realistic. Only a cynic would say that methadone maintenance is so ineffective that this assumption is not at risk of being violated anyway. Again, one way of dealing with this assumption is to keep the time period of interest short, and bring it down from one year to, say, a single month. The influence of a single month of methadone maintenance is likely to be small and this may help to decrease the behavioural response problem.

Finally, the homogeneity assumption dictates that the probabilities of being observed and re-observed should not differ too much across groups of individuals. In theory this assumption should not cause too much worries, in the sense that both models will *under*estimate the true population size in the presence of heterogeneity. So, if heterogeneity is suspected, then one may reason that the estimates are lower bounds of the true population size. Alternatively, one may prefer to stratify the data-set and then carry out subgroup analysis on groups that are more homogeneous and finally pool these estimates into a single estimate of N.

In sum, both estimators of Zelterman and Chao appear to be fairly realistic with respect to the underlying assumptions, but we are not sure about the constant recapture assumption. The logistics of the data collection are easy to manage as only counts of visits are required. This is also an advantage with regard to privacy regulations. Finally, the estimators are computationally easy and these computations will therefore not result in appreciable costs.

2.3 Van der Heijden's et al (in progress) truncated Poisson and truncated Negative Binominal models with covariates

A truncated Poisson regression analysis forms the core of Van der Heijden's *et al* (in progress) estimation model and covariates can be included in the model as a matter of course. In this way between-subject heterogeneity, if any, can be taken into account. The model allows formal testing whether or not sources of heterogeneity -and interactions between these sources- should be included in the model. Once an appropriate model has been fitted to the data, one can compute the probability that a person with a set of covariates has been observed never. Finally, summing the numbers of persons that have been seen plus the estimated numbers of persons never seen produces an estimate of N. The corresponding estimation of the Poisson coefficients and computations for est(N) are somewhat involved and can not be detailed here. The interested reader is referred to Van der Heijden *et al* (in progress).

The model allows that for each group of individuals who share the same characteristics -in terms of the covariates- the unobserved number of persons can be computed. This is in a sense equivalent to the subgroup analysis on a stratified data-set that can be carried out using Zelterman's and Chao's estimators, but it is more efficient since we do not have to assume that all possible interaction terms are present.

The truncated Negative Binominal estimator with covariates follows the same basic strategy, only this time a truncated Negative Binominal regression model is used. Occasionally, there is an advantage in this approach. Under the Poisson model it is assumed that the variance equals the Poisson mean and this assumption may be violated. Under the Negative Binominal model a separate parameter, δ, is estimated which captures variance in excess (or, in lack) of what is expected under the Poisson model, i.e. over- and under-dispersion. In other words, the Negative Binominal regression model helps to account for heterogeneity that could not be modelled explicitly using the covariates. We will return to this issue in the discussion. The presence of dispersion can be formally tested, and if present, can be modelled. In our study on opiate users we found evidence for a significant degree of over dispersion in the data and so we prefer the truncated Negative Binominal model.

Both estimators assume that
1. the population is 'closed', and
2. the individual probabilities to be observed and re-observed are constant over time.

Note, the homogeneity assumption has been relaxed now. Subject to data availability, heterogeneity can be explicitly modelled and accounted for. This is an improvement over Zelterman's and Chao's estimators. However, one has to bear in mind that under the truncated Poisson model with covariates some heterogeneity may be present which has not been adequately modelled. This unmodelled heterogeneity may attenuate est(N). Fact is, additional covariates in the model will produce higher estimates of N. At first glance, this may look like an undesirable feature of the model. However, we like to remind the reader that Zelterman's and Chao's estimators will produce *under*estimates of N in the presence of heterogeneity. So we really should expect from regression-type estimators that est(N) becomes larger when more covariates capture between subject heterogeneity.

All this is slightly different with the truncated Negative Binominal regression estimator. This model can deal with observed heterogeneity (i.e. the covariates) and, in addition to that, with unobserved heterogeneity (i.e. the dispersion parameter).

The closure assumption remains in full force since in- and out-migration of the population can not be accounted for by the model. The effect of migration on the true population size is smaller, of course, during a brief period of time and so it might be advisable to keep the time period under consideration short with respect to the population dynamics. Deciding what 'short' means, falls outside the reign of statistics and remains a matter of intelligent judgement.

The 'constant capture probability over time' assumption is still the most worrying one when we study opiate users who are observed in the context of methadone maintenance. The very treatment they receive may exercise an influence on the probability of being seen again.

In brief, regression-type estimators are in improvement over Chao's and Zelterman's estimators in the sense that between subject heterogeneity can be handled explicitly. Also, the regression-type estimators produce some insight in the factors which are associated with the frequency of contacts, and this may be interesting in its own right. In the same vein, the presence or absence of interactions between the covariates can be formally tested, which is an improvement over subgroup analyses. However, we must pay a price for these benefits. More data, in the form of covariates on all observed individuals, are required. Finally, the estimation procedures are 'expensive' in the sense that they are complicated and non-standard statistical software must be used.

2.4 Hypotheses

We have no way of knowing the true number of opiate users in Rotterdam. There is, however, another estimate for the number of opiate users in Rotterdam in 1994 (Wiessing *et al.*,1995). This estimate is based on the multiplier method and indicates a population size in the range of 3500 - 4000 persons. We must bear in mind that this is an estimate and that the true number is, in fact, unknown. In the absence of a gold standard it is impossible to say how accurate a particular estimator is. All we can do is compare one estimator with the other and use these comparisons as the base for some inference about the 'quality' of the estimators. To that end we will formulate several hypotheses. The hypotheses, or rather expectations, are:

1. Both Zelterman's and Chao's estimators will produce about the same estimates of N. This is a well known result and it is only likely that we will reproduce the same result here.

2. Both Zelterman's and Chao's estimators will produce higher estimates than the estimate of N based on the homogeneous Poisson estimator. This hypothesis is based on the assumption that heterogeneity in the population will not severely affect Zelterman's and Chao's robust estimators, but will result in an underestimation of the true N by the homogenous Poisson estimator. We will not discuss the homogeneous Poisson estimator; it will only serve as a bench mark.

3. The pooled estimators of Zelterman and Chao will be about the same and will be higher than the respective unpooled estimates. This expectation is motivated as follows: the pooled estimators will capture heterogeneity in the population better than the unpooled ones, and this, in turn, will result in less underestimation.

4. The truncated Poisson regression estimator will be higher than the pooled estimators of Chao and Zelterman respectively, because it is better in accounting for heterogeneity than the pooled estimators of Chao and Zelterman.

5. Likewise, the truncated Negative Binominal regression estimator will be higher than the pooled estimators of Zelterman and Chao respectively, because it is better in accounting for heterogeneity than the pooled estimators of Zelterman and Chao.

Note, we have no hypothesis about how Van der Heijden's estimators compare one to the other. The estimators are new and we have no theory that helps to formulate a hypothesis.

3. Results

3.1 Results from Zelterman's and Chao's estimators

For the 1994 RODIS-data we obtain an estimate of 3727 opiate users in Rotterdam when using Zelterman's (1988) Truncated Poisson estimator. The 95% confidence interval ranges from 3497 to 3990. Table 1 gives estimates for the population when it is stratified by sex and age groups (10 year bands).

Table 1 Observed numbers, estimated numbers within 95% CI's using Zelterman's (1988) Truncated Poisson estimator of N by sex and age groups

	males				females			
	obs(n)	low	est(n)	high	obs(n)	low	est(n)	high
age groups								
15-24 yrs	94	134	177	252	65	77	101	148
25-34 yrs	708	1138	1258	1408	306	448	518	614
35-44 yrs	585	961	1080	1233	153	279	368	540
45-54 yrs	90	142	195	311	20	29	61	∞
55-64 yrs	8	6	11	35	0	0	0	0
all ages	1485	2524	2716	2940	544	894	1011	1162

With regard to Table 1 we like to make the following remarks. When all est(N) are summed over all strata, then the sum total is 3769, which is only marginally higher than the direct estimate of 3727 and well within the confidence interval of 3497 - 3990. So the pooled estimate does not differ significantly from the direct estimate. This, in part, reflects the robustness of the estimator in the presence of heterogeneity. Note that in one instance (the males in the 55-64 age group) the lower bound of the 95% CI is lower than the observed number. Conceptually, this does not make sense, but is the result of a obs(n) which is too small for the asymptomatic character of these 95% CIs. Note also that in an other occasion the upper bound explodes into infinity (the females in the 45-54 year band). These freak results are known to happen when using Zelterman's CI's.

Characteristically, Chao's (1987) estimator compares well with those of Zelterman, but her CI's behave better. Using her estimate we find 3565 opiate users in Rotterdam in the year 1994 as a best guess. This is well within the 95% CI of Zelterman's (1988) estimator. The 95% CI of Chao's estimator is 3348 - 3818, which has, of course, a substantial overlap with Zelterman's 95% CI. In table 2 we present the observed and estimated numbers under Chao's model.

Table 2 Observed numbers, estimated numbers within 95% CI's using Chao's (1989) Truncated Poisson estimator of N by sex and age groups

	males				females			
	obs(n)	low	est(n)	high	obs(n)	low	est(n)	high
age groups								
15-24 yrs	94	130	167	241	65	83	104	151
25-34 yrs	708	1080	1192	1338	306	441	506	601
35-44 yrs	585	929	1041	1189	153	274	359	504
45-54 yrs	90	138	187	285	20	33	73	226
55-64 yrs	8	8	11	33	0	0	0	0
all ages	1485	2402	2583	2799	544	875	985	1131

The pooled estimate under Chao's model is 3640 persons which is slightly higher than her direct estimate of 3565, but stays well with in the latter's CI-bounds. This, again underscores that both Zelterman's and Chao's estimators are fairly robust in the presence of heterogeneity. Table 2 shows that Chao's CI's do not produce lower bounds lower than the observed number, while her upper limits do not explode into infinity as Zelterman's CI's sometimes do. Finally, note that Chao's 95% CI's are symmetric around est(N) while Zelterman's are not; and her CI's are usually more narrow than Zelterman's.

3.2 Van der Heijden's et al (in progress) estimators

As outlined in the methods paragraph we first fitted a truncated Poisson model. Initially, variables like SEX (1=male, 0=female), MAR (1=married, 0=not married)), DUT (1=Dutch nationality, 0=otherwise), AGE (in years), INC (1=income from work, 0=otherwise), TOG (1=living together with a partner, 0=otherwise) and SUR (1=of Surinam origine, 0=otherwise) were included in the model. Since not all terms turned out to be significant, a more parsimonious model was obtained by only including, SEX, DUT, AGE, TOG and SUR. In the process of model specification it was also checked whether or not interaction terms should be added to the model and we checked if a quadratic term for AGE had to be included in the model. As it turned out, the simple model without interaction- and quadratic terms fitted well. Under this model N was estimated to be 2991 persons.

Table 3 Coefficients of the truncated Poisson and the truncated Negative Binominal regression equations (all coefficients significant at p<.05; obs(N)=2029)

	Poisson	Neg.Bin.
Cons	0.37	-0.14
SEX	0.19	0.22
DUT	0.22	0.26
TOG	0.11	0.14
SUR	0.27	0.32
AGE	-0.02	-0.02
δ	-	0.98

In comparison with the homogeneous Truncated Poisson model without covariates (see Table 4, below) this estimate with covariates is only slightly higher. So, the additional covariates appear to capture some heterogeneity and increase est(N) somewhat, but relative to Zelterman's and Chao's estimators Van der Heijden's Truncated Poisson estimator with covariates appears to underestimate N for these data. As a possible explanation we speculated that overdispersion biased est(N) downward.

This speculation turned out to have a sound basis when we fitted a Truncated Negative Binominal model to the data. The same parsimonious model was fitted and the parameter that captured the dispersion, δ, turned out to be significantly different from zero (δ=.98; T=3.98; p=.000). Under this truncated Negative Binominal model with covariates N is estimated to be 5006, which is the highest estimate found so far.

4. Summary of results and discussion

Table 4 summaries the point estimates, est(N), and the related estimated prevalence, est(p), as a percentage of the Rotterdam population (close to 600,000). In addition, it gives the estimate of N under the homogeneous truncated Poisson model.

Table 4 Estimates of N and the prevalence of opiate use by model (Rotterdam 1994)

Model	est(N)	est(p)
Homogeneous Truncated Poisson estimator	2937	.49
Zelterman's (1988) Truncated Poisson estimator	3727	.62
Zelterman's (1988) pooled estimator	3769	.63
Chao's (1989) Truncated Poisson estimator	3565	.59
Chao's (1989) pooled estimator	3640	.61
Van der Heijden's *et al* (in progress) Poisson estimator + covars	2991	.50
Van der Heijden's *et al* (in progress) Neg.Bin. estimator + covars	5006	.83

At this point we remind the reader that in the study of Wiessing *et al* (1995) the number of opiate users in Rotterdam in 1994 was estimated to be 3500 - 4000, which is somewhere in the middle range of our own estimates. Still, we observe some variance across the estimates. We will now return to each of the hypotheses and see if we can explain this variance.

1. We find support for the hypothesis that Zelterman's and Chao's estimators produce about the same results. We obtained 3727 and 3565 respectively. Note also that the 95% CIs show substantial overlap: 3497-3990 and 3348-3818. From this we conclude that Zelterman's and Chao's estimators do indeed produce about the same estimates of N.

2. We also find support for the hypotheses that both Zelterman's and Chao's estimators produce higher estimates than what will be obtained under the homogenous truncated Poisson model. We obtained $3727 \approx 3565 > 2937$. Since the latter is well below the lower 95% CI limits of Zelterman's and Chao's estimators, we accept this hypothesis. This supports the idea that when we must assume homogeneity, as one is forced to when using the homogenous truncated Poisson estimator, then this leads to an estimate which is too low relative to the more robust estimators of Zelterman and Chao when, in fact, there is heterogeneity.

3. We find no support neither to accept nor to reject the hypothesis that the pooled estimators of Zelterman and Chao result in higher estimates than the unpooled ones. We found 3769 and 3640 for the pooled estimators, and we found 3727 and 3565 for the unpooled ones. As expected, the pooled estimators of Zelterman and Chao are about the same and both are higher than the respective unpooled estimates. This supports the idea that the pooled estimators somewhat better capture heterogeneity in the population and this, in turn, results in less underestimation. However, and in all fairness, we must also point to the fact that the pooled estimates do not fall outside the 95% CIs of the unpooled estimates. So, the difference, if present, is in the expected direction, but for these data we have no statistically significant finding. It is also clear that, for these data, Chao's estimator is lower than Zelterman's, even to the extent that her pooled estimator is lower than his unpooled estimator. So, the support for our expectation is only tentative.

4. In contrast to our expectation we find that the truncated Poisson regression estimator is lower, not higher, than the pooled estimators of Zelterman and Chao: $2991 < 3769 \approx 3640$. This is probably indicative that his estimator does *not* deal with heterogeneity any better than the pooled estimators of Zelterman and Chao with respect to these data. Here we like to recall what has been said in the methods paragraph: the truncated Poisson regression model can not deal with that part of heterogeneity which has not been modelled through the covariates. Perhaps the presence of dispersion has biased results here.

5. We expected, and found support for the idea, that the truncated Negative Binominal regression estimator is likely to be higher than the pooled estimators of Chao and Zelterman respectively, because it better deals with heterogeneity than the pooled estimators of Chao and Zelterman. With regard to hypothesis 4 we note again that the dispersion parameter, δ, turned out to be significantly different from zero (δ=.98; T=3.98; p=.000).

In the absence of a gold standard it is impossible to say what estimator is the best in terms of its outcome. Having said that, the conclusions based on the confirmed and rejected hypotheses can be summarized as follows.

- The homogenous Poisson estimator would not be our favourite. It produces an outcome too low relative to the other estimates. This apparent underestimation comes not as a surprise and is the likely result of its inability to adequately deal with heterogeneity in these data. This, in turn, is indicative of the importance of estimators that do better cope with heterogeneity, when, of course, we suspect the presence of heterogeneity..

- If we were to make a choice, we would prefer Zelterman's, or for that matter, Chao's estimator. Both estimators (and their pooled counterparts) produce about the same results and are also in line with the study of Wiessing *et al*. Both estimators have additional advantages. They are easy to calculate -although the 95% CIs of Zelterman require a computer to solve an equation iteratively. More importantly, they are based on readily available data. In principle, only counts of observations of opiate users by a single agency are needed. If covariates are present, then they can be used to stratify the data-set. Further, both estimators are not based on assumptions that are totally unrealistic, but they do assume a constant (re)capture probability of each individual over time.

- We expected, and hoped, that the truncated Poisson estimator with covariates would be an improvement over Zelterman's and Chao's estimators, but the results of this study have cast a shadow of doubt over the appropriateness of this estimator with respect to the analysis of this particular data-set. A certain degree of unobserved heterogeneity might be present in the data, and then the truncated Poisson regression model is not the first choice.

- The truncated Negative Binominal estimator with covariates produced results which were in line with our expectations. This gives strength to the idea that this estimator is a suitable candidate for these data. As compared to the estimators of Zelterman and Chao it came up with an higher estimate, which was expected. It's draw back is its computational complexity and the fact that it needs more data than Zelterman's and Chao's estimators.

In sum, Zelterman's and Chao's estimators have some appeal, and so has Van der Heijden's *et al* truncated Negative Binominal estimator with covariates. Choosing between these estimators is mainly a matter data availability and should be guided by a justifiable concern about heterogeneity. Further, all these estimators share a feature: they are all based on a single data-set. This can be seen as a major improvement over more 'classical' capture-recapture estimators which are based on two, three, or even more samples.

Literature

Brecht, M-L., Wickens, Th. D. (1993) Application of multiple capture methods for estimating drug use prevalence. *Journal of Drug Issues* 2, 229-50.

Chao, A. (1988) Estimating animal ambudance with capture frequency data. *Journal of Wildlife Management* 52, 295-300.

Chao, A. (1989) Estimating population size for sparse data in capture-recapture experiments. *Biometrics* 45, 427-38.

Cormack, R.M., Jupp P.E. (1991) Inference for Poisson and multinominal models for capture-recapture experiments. *Biometrika* 78, 911-6.

Domingo-Salvany, A, Hartnoll R.L., Maguire, A., Suelves J.M., Antó J.M. (1995) Use of capture-recapture to estimate the prevalence of opiate addiction in Barcelona, Spain, 1989. *American Journal of Epidemiology*. 141, 567-74.

El Khorazaty, M.N., Imrey P.B., Koch G.G., Wells, H.B. (1977) Estimating the total number of events with data from multiple record systems: a review of methodological strategies. *International Statistical Review*. 45, 129-57.

Pollock, K.H. (1991) Modelling capture-recapture and removal statistics for estimation of demographic parameters of fish and wildlife populations: past, present and future. *Journal of the American Statistical Association*. 86, 225-38.

Seber, G.A.F. (1982) *The estimation of animal abundance and related parameters*. London: Charles Griffin.

Seber, G.A.F. (1986) A review of estimating animal abundance. *Biometrics*. 42, 267-92.

Seber, G.A.F. (1992) A review of estimating animal abundance II. *International Statistical Review*. 60, 129-66.

Van der Heijden, P.G.M., Zelterman D., Engbertsen G.B.M., Van der Leun, J. (in prep.) Estimating the number of illegals in the Netherlands with the truncated Poisson regression model.

Simeone, R.S., Nottingham, W.T., Holland, L. (1993) Estimating the size of a heroine using population: An examination of the use treatment admissions data. *The International Journal of the Addictions*. 28, 107-28.

Wickens Th. D. (1993) Quantitative methods for estimating the size of a drug using population. *Journal of Drug Issues*. 2, 185-216.

Wiessing, L.G., Toet, J., Houweling, H. *et al*. (1995) *Prevalentie en risicofactoren van HIV-infectie onder druggebruikers in Rotterdam*. Bilthoven, RIVM / Rotterdam, GGD Rotterdam.

Wilson R.M., Collins, M.F. (1992) Capture-recapture estimation with samples of size one using frequency data. *Biometrika* 79, 543-53.

Zelterman, D. (1988) Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *Journal of Statistical Planning and Inference*. 18, 225-37.