Statistical Methods & Applications (2005)
DOI: 10.1007/s10260-004-0103-5

SMA

Mixture models for capture-recapture count data

Dankmar Böhning[1], **Ekkehart Dietz**[1], **Ronny Kuhnert**[1], **Dieter Schön**[2]

[1] Biometry and Epidemiology, Division for International Health, Institute for Social Medicine, Epidemiology, and Health Economy, Charité, University Medicine Berlin, Fabeckstr. 60–62, Haus 562, 14195 Berlin, Germany (e-mail: dankmar.boehning@charite.de)
[2] Dachdokumentation Krebs, FG 21, Robert-Koch-Institut Berlin, Berlin, Germany

Abstract. The contribution investigates the problem of estimating the size of a population, also known as the missing cases problem. Suppose a registration system is targeting to identify all cases having a certain characteristic such as a specific disease (cancer, heart disease, ...), disease related condition (HIV, heroin use, ...) or a specific behavior (driving a car without license). Every case in such a registration system has a certain notification history in that it might have been identified several times (at least once) which can be understood as a particular capture-recapture situation. Typically, cases are left out which have never been listed at any occasion, and it is this frequency one wants to estimate. In this paper modelling is concentrating on the counting distribution, e.g. the distribution of the variable that counts how often a given case has been identified by the registration system. Besides very simple models like the binomial or Poisson distribution, finite (nonparametric) mixtures of these are considered providing rather flexible modelling tools. Estimation is done using maximum likelihood by means of the EM algorithm. A case study on heroin users in Bangkok in the year 2001 is completing the contribution.

Key words: Counting Distribution Model, capture-recapture, truncated count distribution, finite mixture models

1. Introduction

1.1. The missing cases problem

The following situation is very common in medicine and public health. A given disease registration system identifies n_{obs} cases of a particular disease of interest such as a specific cancer site. The question arises how many cases n are there in total, or, in other words, how many cases have been left out? Suppose that the system identifies a case with probability $1 - p_0$, so that $n = np_0 + n(1 - p_0)$. Note that

$n(1-p_0)$ is the expected number of cases identified by the registry which simply can be estimated by n_{obs}. number of observed cases. This leads to the estimating equation for n

$$n = np_0 + n_{obs}, \tag{1}$$

which in other words states that the population size is the sum of the unobserved and observed cases. The equation can easily be solved for n to provide the Horvitz-Thompson estimator

$$\hat{n}_{HTE} = n_{obs}/(1-p_0), \tag{2}$$

which is the number of observed cases adjusted for the probability of being included into the registration system.

1.2. Horvitz-Thompson approach

Yet, another more rigorous way to derive (2) is as follows: let Z_i be the indicator of identifying a diseased individual i in the population for $i = 1, 2, ..., n$ with $Z_i = 1$ meaning individual i is identified and $Z_i = 0$ otherwise. Then, $E(\sum_{i=1}^n Z_i) = \sum_{i=1}^n P(Z_i = 1) = n(1-p_0)$ and equating this expected value to the observed number of cases leads to (2) again, showing that (2) is a *moment* estimator. But (2) can be a maximum likelihood estimator as well. Consider the binomial likelihood $L(n) = \binom{n}{n_{obs}}(1-p_0)^{n_{obs}}p_0^{n-n_{obs}}$ of identifying exactly $\sum_{i=1}^n z_i = n_{obs}$ out of the n diseased individuals. Then, $L(n)$ is maximized for n being the integer part of (2). This result follows by considering the difference score function $U(n) = \frac{L(n)-L(n-1)}{L(n)} = \frac{n_{obs}-n(1-p_0)}{np_0}$ which is positive for integers n smaller than $n_{obs}/(1-p_0)$ and negative for integers n larger than $n_{obs}/(1-p_0)$. For details see Lindsay and Roeder (1987) or Bishop *et al.* (1975, Ch. 6).

If p_0 were known, then the problem is solved and the total of the population of cases could be simply estimated as (2), namely $n_{obs}/\hat{n}_{HTE}$. Unfortunately, p_0 is unknown for most applications and must be estimated. To accomplish this task the data collected by the identification mechanism are assumed to have a structure that can be used for modelling and predicting p_0. Frequently, two types of structures are used.

In the first type, a case is identified at prespecified occasions, times or sources of identification. This type of data is called capture-recapture data with different sources. The name stems from animal sampling designs in which a sample of animals is drawn (caught), tagged and released. This procedure is repeated several times, say m, and the m repetitions often assumed to be independent, so that at the end of the procedure a series of data for each animal is available representing the capture-recapture history of each individual animal. Usually, the capture-recapture history is provided as a m-vector consisting of 0s and 1s, where a 1 at the j-th position indicates presence in the jth recapture sample. The frequency of the vector $(0, 0, ..., 0)^T$, indicating a case which has been never identified, is not observed

and is the crucial part of the modelling and estimation. Frequently, log-linear modelling is used to model the arising m-dimensional frequency table. Recently, these approaches have been used for modelling completeness of a disease registry. In this case, different identification sources could be physicians, hospitals, or laboratories. See Chao (1998, 2001), Wittes and Sidel (1968), LaPorte *et al.* (1992), Hook and Regal (1995), IWGDMF (1995a,b), Comiskey and Barry (2001), Nannan and White (1997), Tilling (2001) or Schouten *et al.* (1994). for further details, and Sekar and Deming (1949) for an early public health application of the method.

In the second type, a case is identified at several occasions, but only this repeated counting information is known. For example, the city police records how often a drug dealer has been caught dealing drugs, or a surveillance systems counts how many times a heroin user went into a treatment institution. We call this type of capture-recapture data *repeated counting data*. For each case, acount can be provided, on how often this case was identified by the identification mechanism. The repeated counting data provide a frequency n_1 for observing exactly 1 count, n_2 for observing exactly 2 repeated counts, ..., n_m for observing exactly m repeated counts where m is the largest observed repeated count. The frequency of 0 counts, those cases which have never been identified, is missing information and need to be estimated. See Mao and Lindsay (2003), van der Heijden, Bustami *et al.* (2003), van der Heijden, Cruy *et al.* (2003), Scollnik (1997), Meng (1997) or Wilson and Collins (1992) for further details.

Clearly, from data structures of capture-recapture with different sources type the data structures of repeated counting type can be constructed, but not vice versa. Both data structures leave n_0 unknown, though different methods are used to estimate it.

1.3. The occurrence of the counting distribution

The counting distribution simply occurs when counting the number of notifications for a given case. Again, there will be *no* frequency n_0 of zeros observed, where, in general, n_i is the frequency of exactly i notifications from m possible occurrences. In statistical terms, we are dealing with a *zero-truncated* count distribution.

To illustrate the situation we will first discuss a historic example, the cholera epidemic in India. The example stems from Mao and Lindsay (2003) and has been discussed previously in Blumenthal *et al.* (1978), Scollnik (1997), and others. A cholera epidemic affected a village with 223 households in India. Let n_i be the number of households with exactly i cases. The data are: $n_1 = 32$, $n_2 = 16$, $n_3 = 6$, $n_4 = 1$, so that $n_{obs} = 55$. Originally, the data were presented by McKendrick (1926) in his paper presentation to the Edinburgh Mathematical Society. It can be assumed that McKendrick was confronted with the data of the cholera epidemic during the period of his service in India. It is interesting to note that there is also a number $n_0 = 168$ reported, the frequency of houses with *no* cholera cases. However, McKendrick knew that some unknown percentage of these houses were affected by the cholera epidemic, though no cases were observed in these houses. It should be recalled that cholera is a water-borne disease. In this case, the epidemic was caused by a specific, contaminated well, and houses supplied

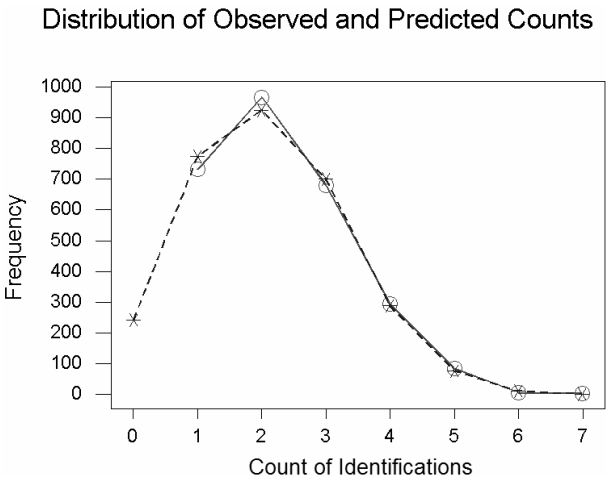

Fig. 1. Observed (circle) and fitted (asterisk) counting distribution

with drinking water from that contaminated well were exposed to developing the disease. McKendrick was interested in modelling the count Y of cholera cases in a cholera-affected household. For any case count $Y_i > 0$ the associated i-th household is clearly cholera affected. If $Y_i = 0$ the household might be affected or not, so McKendrick ignored the 168 households with zero cases since they are not helpful in determining the number of affected houses with no cases, and developed a (moment) estimator for the number n_0 of affected households with no cases from the distribution of non-zero case-counts.

We will concentrate here on modelling the distribution of case notification counts. Suppose $f(y, \theta)$ is a suitable distributional model for the counting distribution of the number of times Y that a particular case has been identified by the surveillance system. In the case of the cholera epidemic, Y is the number of cholera cases in a particular household that has been registered by the village health worker. To estimate p_0, one could simply replace p_0 by $f(0, \theta)$ (see also Fig. 1) and, from (2), one is lead to

$$\hat{n} = \frac{n_{obs}}{1 - f(0, \theta)}. \tag{3}$$

2. Simple models: Binomial and Poisson

Suppose the identification mechanism can identify a case with a maximum of m times, then one could think of a suitable distribution for Y as the *Binomial* given as

$$f(y, \theta) = \binom{m}{y} \theta^y (1 - \theta)^{m-y}, y = 0, 1, ..., m \tag{4}$$

with specifically, $p_0 = f(0,\theta) = (1-\theta)^m$, or, when m is large or unknown, the *Poisson*

$$f(y,\lambda) = \lambda^y e^{-\lambda}/y!, y = 0, 1, ... \tag{5}$$

with specifically, $p_0 = f(0,\lambda) = e^{-\lambda}$. We have denoted in the Poisson case the parameter by λ, while for the binomial case the parameter is denoted by θ. Note that both parameters are connected via $\lambda = m\theta = E(Y)$.

2.1. Maximum likelihood estimation

Suppose that $f(y,\theta)$ is a model for which the maximum likelihood estimator is readily available if n_0 the frequencies of zeros is given. On the other hand, if θ is given, then also $\hat{n}$ is provided by (2), or $\hat{n}_0 = \hat{n} - n_{obs}$. Thus, one could construct an algorithm as follows:

Algorithm 1

Step 0. Choose some initial value for $\hat{\theta}^{(0)}$.

Step 1. Compute $\hat{n}^{(j+1)} = \frac{n_{obs}}{1-f(0,\hat{\theta}^{(j)})}$, and $\hat{n}_0^{(j+1)} = \hat{n}^{(j+1)} - n_{obs}$.

Step 2. Use the complete frequency table $\hat{n}_0^{(j+1)}, n_1, ..., n_m$ to compute a new maximum likelihood estimator $\hat{\theta}^{(j+1)}$, and go back to Step 1.

Algorithm 1 is a version of the EM algorithm (Dempster *et al.* 1977, see also Meng (1997) for historical review and Dietz and Böhning (2000) for a general approach to zero-modified models). It should be noted that Algorithm 1 provides an estimate which seeks to maximize the following likelihood of truncated densities:

$$\prod_{i=1}^{m} \left(\frac{f(i,\theta)}{1-f(0,\theta)} \right)^{n_i}.$$

In the EM-terminology, this is called the *observed, incomplete* data likelihood. The corresponding *unobserved, complete* data likelihood is

$$f(0,\theta)^{n_0} \prod_{i=1}^{m} f(i,\theta)^{n_i}$$

which is maximized in Step 2 of Algorithm 1 with $n_0 = \hat{n}_0^{(j+1)}$, the expected value of n_0 conditional upon the previous value of θ.

2.2. Special cases

These steps of Algorithm 1 take specific versions for the binomial and the Poisson. We consider now two simple, special cases of this algorithm.

Algorithm 1 for the binomial

Step 1. Compute $\hat{n}^{(j+1)} = \frac{n_{obs}}{1-(1-\hat{\theta}^{(j)})^m}$.

Step 2. Use the complete frequency table $\hat{n}_0^{(j+1)}, n_1, ..., n_m$ to compute $\hat{\theta}^{(j+1)} = [0\hat{n}_0^{(j+1)} + 1n_1 + 2n_2 + ... + mn_m]/(\hat{n}^{(j+1)}m)$, and go back to Step 1.

The steps are very similar in the Poisson case.

Algorithm 1 for the Poisson

Step 1. Compute $\hat{n}^{(j+1)} = \frac{n_{obs}}{1-exp(-\hat{\lambda}^{(j)})}$.

Step 2. Use the complete frequency table $\hat{n}_0^{(j+1)}, n_1, ..., n_m$ to compute $\hat{\lambda}^{(j+1)} = [0\hat{n}_0^{(j+1)} + 1n_1 + 2n_2 + ... + mn_m]/\hat{n}^{(j+1)}$, and go back to Step 1.

Note that for both, the binomial and the Poisson, the Step 2 is very simple, the arithmetic mean of the observed proportions and the arithmetic mean of the observed counts, respectively. Note also that in the mean computation only the denominator changes during iteration, the numerator is always $1n_1 + 2n_2 + ... + mn_m$. For both cases, moment estimators are also available in closed form which might be used instead, or alternatively, as initial values for Step 0. For details, see also Meng (1997). We point out here that any sequence generated by Algorithm 1 for the Binomial or Poisson case converges to the unique MLE.

3. Allowing heterogeneity: Mixtures of Poissons and Binomials

We generalize now the concepts of the previous section to a more flexible framework. Let $f(y, \theta)$ denote some simple, parametric density such as the binomial or Poisson. Then, the (finite) mixture distribution

$$f(y, Q) = \sum_{j=1}^{k} f(y, \theta_j) q_j \tag{6}$$

arises as the marginal distribution with respect to some latent variable Z having distribution Q, where the discrete mass distribution Q, the mixing distribution, gives non-negative weights q_j ($\sum_{j=1}^{k} q_j = 1$) to θ_j. The finite mixture model is sometimes named *latent class* model in the capture-recapture setting and has the characteristic that the heterogeneity distribution is discrete. This is in contrast to continuous models for the heterogeneity distribution like the beta-binomial or normal-logits distribution. For a recent review on parametric mixture models for the beta-binomial capture-recapture situation see Dorazio and Royle (2003). In the discussion section we will briefly compare the nonparametic Poisson mixture approach to the Poisson-Gamma model.

The mixing distribution can be interpreted as the heterogeneity distribution of the listing parameter in the population. Whereas the simple Binomial- or Poisson-model requires specific assumptions such as independence of observations and homogeneity of the listing parameter, mixture models are more flexible models

in capturing these phenomena. Suppose that the listing parameter varies in the population according to some unobserved variable. Then, the associated marginal distribution will be a mixture of the component densities where the component membership is described by the latent variable. But also *positive* dependencies between recaptures (several identifications in the listing system) will be adjusted for, since autocorrelation of repeated identifications will lead to overdispersion which can be explained by means of a mixture model. This shows that mixture models provide not only a richer class of possible distributions, but also can be well motivated. The nonparametric binomial and Poisson mixture has been discussed perviously including Norris and Pollock (1996, 1998) though their approach is different in that for *every* value of n the associated nonparametric maximum likehood estimate is found (see Laird 1978, Lindsay 1983, or Böhning 2000 for details) and the resulting profile likelihood then maximized in n.

3.1. Maximum likelihood estimation for mixtures

As before, if Q is given, then n can be estimated as $n_{obs}/(1-f(0,Q))$, and if n_0 is provided Q is estimated by maximum likelihood. The latter step is particular easy, when $f(y,\theta)$ is the binomial or Poisson. For the binomial the mixture (4) takes the form

$$f(y,Q) = \sum_{j=1}^{k} q_j \binom{m}{y} \theta_j^y (1-\theta_j)^{m-y}$$

and for the Poisson

$$f(y,Q) = \sum_{j=1}^{k} q_j \lambda_j^y e^{-\lambda_j}/y!$$

and the predicted population sizes are $\sum_{j=1}^{k} q_j(1-\theta_j)^m$ and $\sum_{j=1}^{k} q_j e^{-\lambda_j}$, respectively.

Algorithm 2

Step 0. Choose some initial value for the mixing distribution, $Q^{(0)}$.

Step 1. Compute $\hat{n}^{(j+1)} = \frac{n_{obs}}{1-f(0,Q^{(j)})}$, and $\hat{n}_0^{(j+1)} = \hat{n}^{(j+1)} - n_{obs}$.

Step 2. Use the complete frequency table $\hat{n}_0^{(j+1)}, n_1, ..., n_m$ to compute a new maximum likelihood estimator $Q^{(j+1)}$, set $j = j+1$ and go back to Step 1.

In Step 2 of Algorithm 2 the NPMLE of the mixing distribution Q needs to be calculated algorithmically itself. This can be done by one of the algorithms discussed in Böhning (2000) or using the popular EM algorithmic framework for mixtures of distributions (McLachlan and Peel 2000), in which case we have a *nested* EM algorithm. The EM-steps for mixtures are well-known. Define $e_{il}^{(j)} = f(i,\theta_l^{(j)})q_l^{(j)}/f(i,Q^{(j)})$ and let $n_0 = \hat{n}_0^{(j+1)}$. Then:

Step 2.1 $q_l^{(j+1)} = \frac{1}{\hat{n}^{(j+1)}} \sum_{i=0}^{m} n_i e_{il}^{(j)}$

Step 2.2 Find solution $\theta_l^{(j+1)}$ for l-th component scoring equation in θ_l

$$\sum_{i=0}^{m} n_i e_{il}^{(j)} \frac{\partial}{\partial \theta_l} f(i, \theta_l)$$

for $l = 1, 2, ..., k$. For simple component densities, the solution of the scoring equation in Step 2.2 is available in closed form, for the binomial density it is

$$\theta_l^{(j+1)} = \sum_{i=0}^{m} i \; n_i e_{il}^{(j)} / \sum_{i=0}^{m} n_i m \; e_{il}^{(j)}$$

and for the Poisson we have

$$\lambda_l^{(j+1)} = \sum_{i=0}^{m} i \; n_i e_{il}^{(j)} / \sum_{i=0}^{m} n_i e_{il}^{(j)}.$$

To execute the EM algorithm for mixtures in Step 2, two ways can be followed: either one iterates as long as the maximum likelihood estimator is approximated closely enough for the $\hat{n}_0^{(j+1)}$ at hand, or we just do *exactly one* E- and M-step and go back to Step 1. This is a form of GEM algorithm as discussed in McLachlan and Krishnan (1997). Both versions of the EM algorithm have the monotonicity property though the second version appears computationally more efficient.

3.2. Confidence interval estimation

Confidence interval estimation is not an easy task for capture-recapture studies as pointed out by several authors including Chao (1989) and Cormack (1992). For the modelling approach using Poisson mixtures, Bootstrap resampling techniques were used as described in van der Heijden, Bustami *et al.* (2003). If p_0 in (2) were known, the only source of variation in the estimator $\hat{n}_{HTE}$ would arise from sampling the n_{obs} out of n. If p_0 is estimated using the truncated Poisson mixture model, there is a second source of random variation arising. To mimic both sources of variation the Bootstrap is realized in the following fashion. Firstly, $n_{obs}^{(b)}$ is sampled from a Binomial distribution with success parameter $p = n_{obs}/\hat{n}$ and sample size parameter $\hat{n}$, where $\hat{n}$ as well as the parameters in the truncated mixture are estimated from the original data set. Secondly, frequencies $n_1^{(b)}, n_2^{(b)}, ...$ are sampled from the truncated Poisson mixture model with parameters as estimated in the original data set, namely $Y_l^{(b)} \sim (\sum_j \hat{q}_j Po(\hat{\lambda}_j))/(1 - \sum_j \hat{q}_j \exp(-\hat{\lambda}_j))$ for $l = 1, ..., \sum_i n_{obs}^{(b)}$. As before, $n_i^{(b)} = \#\{Y_l^{(b)} = i | l = 1, ..., n_{obs}^{(b)}\}$. For each of these B resamples, $n_0^{(b)}$ is estimated using the EM algorithm for the truncated mixture model, and these resample data are used to compute standard errors and confidence intervals. It was found that the statistics of interest stabilized beyond $B = 1,000$, so that

$B = 5,000$ was considered to be sufficient in all Bootstrap calculations. Following van der Heijden, Bustami *et al.* (2003) confidence intervals were calculated as asymptotic normal intervals.

Van der Heijden, Bustami *et al.* (2003) compared coverage properties based upon the Bootstrap approach with the desired confidence levels for the homogeneous Poisson case and found that the Bootstrap confidence intervals achieved good coverage probabilities.

4. An application for estimating the number of heroin users in Bangkok 2001

4.1. Data sources and characteristics

In a surveillance study on drug use in Bangkok (Thailand) data were analyzed for the year 2001 (Böhning, Suppawattanabodee, Kusolvisitkul, and Viwatwongkasem 2004). The study used all data on drug use from 61 health treatment centers in the Bangkok Metropolitan region collected by the Office of the Narcotics Control Board (ONCB), Ministry of Prime Minister, which occurred from October 1 to December 31, 2001. All private and public health treatment centers in the Bangkok Metropolitan region licensed by the Ministry of Public Health to treat drug dependence were included in the study. Each patient entering the surveillance system receives a unique identification number that is used to enter information about the patient every time the patient initialized a new treatment episode. From the available data source it was possible to construct the information on the frequency of episodes for each patient in the sampling period which will serve as the key element in the modelling process.

4.2. Estimating the number of heroin users

Here, the modelling of the distribution of the counts of the treatment episodes is considered. Figure 2 shows three curves for the group of heroin users: the distribution of observed counts, the distribution of predicted counts under the homogeneous Poisson model and under the mixed Poisson model. The homogeneous Poisson has a bad goodness-of-fit value $\chi^2 = 3245.20$ with 13 *df* (p-value = 0.0000). The Poisson mixture gives an acceptable goodness-of-fit value with $\chi^2 = 5.65$ and 2 *df* (p-value $= 0.0593$).

The estimate of the unobserved number of heroin users is provided by a four-component Poisson mixture model as $10,219$ with 95% CI: (7,046 $-$ 13,392), (see Table 1). The count distribution of treatment episodes changes with age, so that also an age-adjusted estimate of the unobserved number of heroin users is computed leading to $11,296$ with 95% CI: (8,964 $-$ 13,628), (see Table 1 again). Together with the observed number of $7,048$ the total number of heroin users in B k Metropolis is estimated as $18,344$ with 95% CI: (16,006 $-$ 20,710). It should be pointed out that incorporating covariates into some form of generalized linear modelling – as suggested by Pledger (2000) – could further improve the model building, in particular, when sample sizes are small. Here, we believe that the stratified approach is acceptable, given the size of the observed data.

Table 1. Number of unobserved heroin users with 95% CI estimated using the Poisson mixture model

		Discrete mixture		
Age group	$\hat{n}_0$(95%CI)	$\hat{\lambda}_j$	$\hat{p}_j$	k
unstratified	10,219	0.214	0.705	4
	(7,046 − 13,392)	2.130	0.187	
		5.850	0.105	
		12.200	0.003	
I	754	0.384	0.736	4
	(0 − 1,530)	2.967	0.173	
		7.008	0.089	
		14.563	0.003	
II	3,685	0.157	0.704	4
	(2,493 − 4,877)	1.975	0.184	
		5.755	0.106	
		11.631	0.005	
III	4,607	0.122	0.766	4
	(3,360 − 5,854)	2.084	0.156	
		5.719	0.074	
		11.524	0.004	
IV	2,250	0.362	0.701	3
	(886 − 3,614)	2.459	0.182	
		5.936	0.117	
Age-stratified	11,296♠			
	(8,964 − 13,628)			

♠ sum of estimates of n_0 for the four age groups.

5. Discussion

5.1. Model evaluation

Given a parametric class of models, which model should be selected? Several selection criteria have been suggested. McLachlan and Peel (2000, p. 202–219) provide an overview in the context of mixture models. The criteria are constructed in the way that the log-likelihood is penalized with a function of the model complexity, and criteria differ in the way they measure model complexity. We consider Akaike's information criterion

$$AIC\alpha = 2L(\hat{Q}_k) - \alpha(2k-1),$$

with $\alpha = 2$. Miloslavsky and van der Laan (2003) suggest to consider values of α other than 2 to steer the penalizing effect. Here, we just look at $\alpha = 2$. A further criterion is the Bayesian Information Criterion defined as

$$BIC = 2L(\hat{Q}_k) - (2k-1)\log(n_{obs}),$$

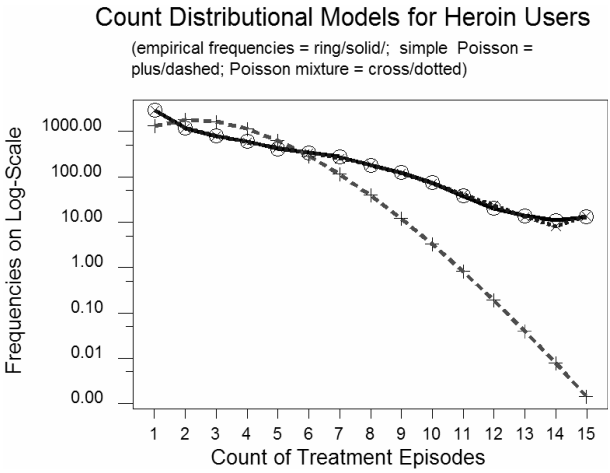

Fig. 2. Count distribution of treatment episodes for heroin users

Table 2. Number of unobserved heroin users with 95% CI estimated using the poisson mixture model

k	$\hat{\lambda}_j$	$\hat{q}_j$	log-likelihood	AIC	BIC
1	2.75	1.00	−15,462	−30,927	−30,934
2	0.88	0.75	−13,214	−26,434	−26,455
	5.40	0.25			
3	0.41	0.69	−13,134	−26,279	−26,313
	2.97	0.22			
	6.80	0.09			
4	0.21	0.70	−13,120	**−26,255**	**−26,303**
	2.13	0.19			
	5.84	0.10			
	12.20	0.01			

which is known to penalize complex models more strongly than the Akaike-type criteria. For the heroin data, Table 2 provides clear and consistent evidence that four components are required. No further components are possible since the four-component estimate is already the nonparametric maximum likelihood estimate, so that no further increase in the likelihood is possible.

5.2. Continuous mixing distribution

As alternative to finite, discrete mixture models to model unobserved heterogeneity it is sometimes suggested to use a continuous, parametric mixing distribution. In the situation that the largest number of possible identifications of a particular case is known, the beta-binomial model is a potential candidate (see Dorazio and Royle 2003 for details). In the light of the particular application discussed here we will consider the Poisson-Gamma mixture in more detail. For this model the

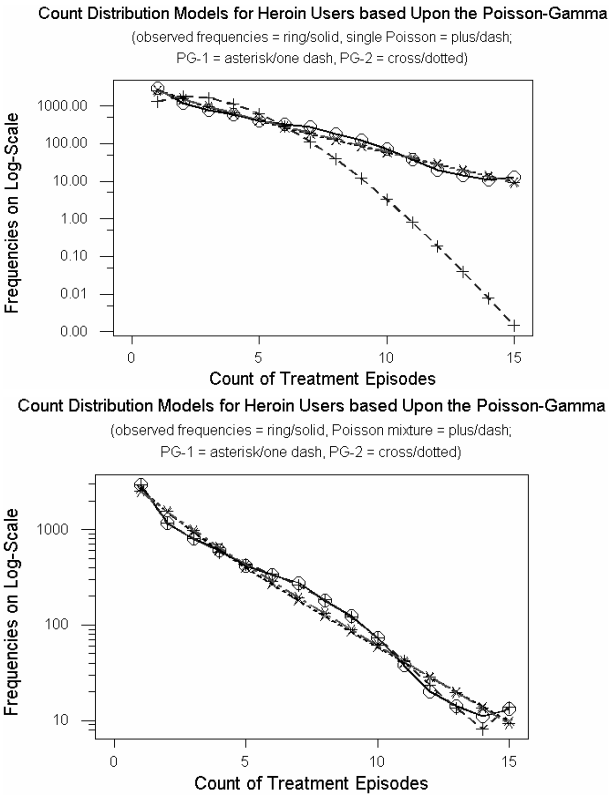

Fig. 3. Count distribution models of treatment episodes for heroin users based upon the Poisson-Gamma. Top: with simple Poisson. Bottom without simple Poisson

heterogeneity is modelled via a Gamma-distribution leading to the marginal

$$f(y|\theta,\kappa) = \int_0^\infty e^{-\lambda}\lambda^y/y!q(\lambda|\theta,\kappa)d\lambda, \tag{7}$$

where $q(\lambda|\theta,\kappa) = \theta^{-\kappa}\lambda^{\kappa-1}e^{-\lambda/\theta}/\Gamma(\kappa)$ is the *Gamma* density. The mean and the variance of the Gamma are $\mu = \kappa\theta$ and $\tau^2 = \kappa\theta^2 = \mu\theta = \mu^2/\kappa$, respectively, so that different, equivalent reparameterizations are possible. (7) can be simplified to

$$f(y|\mu,\kappa) = \frac{\Gamma(y+\kappa)}{\Gamma(y+1)\Gamma(\kappa)}\alpha^\kappa(1-\alpha)^y, \tag{8}$$

with $\alpha = \frac{\kappa}{\kappa+\mu}$, showing in particular that (7) is a *negative binomial* density. Because of the Gamma-function involved in (8), maximum likelihood estimation for κ is not trivial, even in the untruncated situation. The MLE for μ is the sample mean $\overline{y}$, and the moment estimator for τ^2 is $S^2 - \overline{y}$, where S^2 is the sample variance. This suggests to approximate the *M-step* in the EM algorithm using the maximum likelihood/moment – estimators $\mu = \bar{y}$ and $\hat{\tau}^2 = S^2 - \overline{y}$ where it is assumed that n_0 is given. In the *E-step*, $\hat{n}_0$ is simply worked out as $\hat{\alpha}^{\hat{\kappa}}$ with

$\hat{\alpha} = \overline{y}/S^2$ and $\hat{\kappa} = \hat{\mu}^2/\hat{\tau}^2$. Again, the algorithm toggles between M- and E-step until convergence. We call this procedure PG-1. Alternatively, we might construct maximum likelihood estimates of μ and τ^2 from the maximum likelihood estimates $\hat{\lambda}_1, ..., \hat{\lambda}_k$ and $\hat{q}_1, ..., \hat{q}_k$ of the finite mixture model leading to

$$\hat{\mu} = \sum_{j=1}^{k} \hat{q}_j \hat{\lambda}_j \text{ and } \hat{\tau}^2 = \sum_{j=1}^{k} \hat{q}_j (\hat{\lambda}_j - \hat{\mu})^2.$$

We call this procedure PG-2. Note that PG-2 is non-iterative, but based upon the results of the nonparametric maximum likelihood procedure. It is mainly included here for comparison. For the Bangkok heroin data, PG-1 delivers an estimate of $\hat{n}_0 = 4,938$ which is only half the size of the estimate of 10,219 from the nonparametric procedure. Here, the goodness-of-fit value is $\chi^2 = 283.06$ (p-value=0.0000). It is not surprising that the estimate of $\hat{n}_0 = 7,364$ delivered by PG-2 is closer to the one of the nonparametric procedure since the parameter estimates are constructed from the mixing distribution. In this case, the goodness-of-fit value is $\chi^2 = 234.74$ (p-value=0.0000). A graphical representation of both models is presented in Figure 3. Although the fit is much improved upon the fit for the simple Poisson (see Figure 3, top), both models, PG-1 and PG-2, experience lack-of-fit at various data points. This becomes clear when the simple Poisson fit is removed (see Figure 3, bottom). In addition, the figure shows the superior fit of the nonparamteric mixture model.

5.3. Unconditional and conditional likelihood

For capture-recapture modelling two likelihood methods are possible. One is based upon the full, unconditional likelihood

$$L(n, \theta) = \frac{n!}{(n - n_{obs})! n_1! \times ... \times n_m!} f(0, \theta)^{n - n_{obs}} \prod_{i=1}^{m} f(i, \theta)^{n_i}, \qquad (9)$$

where $f(i, \theta)$ represents again the distributional model for count i. The likelihood (9) can be factored into two other likelihoods such that

$$L(n, \theta) = L_b(n, \theta) \times L_c(\theta)$$

where

$$L_b(n, \theta) = \binom{n}{n_{obs}} (1 - f(\theta))^{n_{obs}} f(0, \theta)^{n - n_{obs}} \qquad (10)$$

and

$$L_c(\theta) = \frac{n_{obs}!}{n_1! \times ... \times n_m!} \prod_{i=1}^{m} \left(\frac{f(i, \theta)}{1 - f(0, \theta)} \right)^{n_i}. \qquad (11)$$

The conditional likelihood does not involve the unknown population size parameter and is conceptually easier to treat. This approach was used here. The justification of the conditional procedure uses the fact that if θ is given, the binomial

likelihood (10) is maximized as discussed in Sect. 1.2. For various cases it could be shown that conditional and unconditional MLE coincide (see Bishop *et al.* 1975; Sanathanan 1972, 1977; Chao and Bunge 2002), though they need not to be identical. Norris and Pollock (1996, 1998) use a profile likelihood method aiming to maximize the full likelihood. The algorithm developed in this paper – to motonically maximimize the conditional likelihood – will also monotonically increase the full likelihood. For the homogenuous Poisson case, a straight-forward argument shows that conditional and unconditional MLE for the population size agree.

Acknowledgement. This paper developed from an invited presentation given at the *Workshop "Mixture Models between Theory and Applications" Rome, September 13, 2002* organized by Marco Alfò, Roberto Rocci, Luca Tardella, Maurizio Vichi, and Cecilia Vitiello of *Universitàà degli Studi di Roma "La Sapienza", Dipartimento di Statistica, Probabilitàà e Statistiche Applicate*. The first author would like to express his sincerest thanks for this invitation. The authors would like to express their sincere thanks to the Editor, Associate Editor and an unknown Reviewer for their helpful and clarifying comments.

References

Bishop YMM, Fienberg SE, Holland PW (1975) Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge

Blumenthal S, Dahiya R, Gross A (1978) Estimating complete sample-size from an incomplete Poisson sample. Journal of the American Statistical Association 73, 182–187

Böhning Suppawattanabodee B, Kusolvisitkul W, Viwatwongkasem C (2004) Estimating the number of drug users in Bangkok 2001: A capture-recapture approach using repeated entries in one list. European Journal of Epidemiology (to appear)

Böhning D (2000) Computer-assisted analysis of mixtures and applications. Meta-analysis, disease mapping and others. Chapman & Hall/CRC, Boca Raton

Chao A (2001) An overview of closed capture-recapture models. Journal of Agricultural, Biological, and Environmental Statistics 6, 158–175

Chao A (1998) Capture-recapture. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, vol.1. Wiley, ♣, pp 482–486

Chao A (1989) Estimating population size for sparse data in capture-recapture experiments. Biometrics 45, 427–438

Chao A, Bunge J (2002) Estimating the number of species in a stochastic abundance model. Biometrics 58, 531–539

Comiskey CM, Barry JM (2001) A capture-recapture study of the prevalence and implications of opiate use in Dublin. European Journal of Public Health 11, 198–200

Cormack RM (1992) Interval estimation for mark-recapture studies of closed populations. Biometrics 48, 567–576

Dorazio RM, Royle JA (2003) Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics 59, 351–364

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B 39, 1–38

Dietz E, Böhning D (2000) On estimation of the Poisson parameter in zero-modified Poisson models. Computational Statistics & Data Analysis 34, 441–459

Hook EB, Regal R (1995) Capture-recapture methods in epidemiology: methods and limitations. Epidemiologic Reviews 17, 243–264

International Working Group for Disease Monitoring and Forecasting (1995a) Capture-recapture and multiple record systems estimation I: history and theoretical development. American Journal of Epidemiology 142, 1047–1058

International Working Group for Disease Monitoring and Forecasting (1995b) Capture-recapture and multiple record systems estimation II: Applications in human diseases. American Journal of Epidemiology 142, 1059–1068

LaPorte RE, McCarty DJ, Tull ES, Tajima N (1992) Counting birds, bees, and NCDs. Lancet 339, 494–495

Laird NM (1978) Nonparametric maximum likelihood estimation of a mixing distribution. Journal of the American Statistical Association 73, 805–811

Lindsay BG, Roeder K (1987) A unified treatment of integer parameter models. Journal of the American Statistical Association 82, 758–764

Mao CX, Lindsay BG (2003) Tests and diagnostics for heterogeneity in the species problem. Computational Statistics and Data Analysis 41, 389–398

McKendrick AG (1926) Application of mathematics to medical problems. Proceedings of the Edinburgh Mathematical Society 44, 98–130

McLachlan G, Krishnan T (1997) The EM algorithm and extensions. Wiley, New York

McLachlan G, Peel D (2000) Finite mixture models. Wiley, New York

Meng X-L (1997) The EM algorithm and medical studies: a historical link. Statistical Methods in Medical Research 6, 3–23

Miloslavsky M, van der Laan MJ (2003) Fitting of mixtures with unspecified number of components using cross validation distance estimate. Computational Statistics and Data Analysis 41, 413–428

Nannan DJ, White F (1997) Capture-recapture: Reconnaissance of a demographic technique in epidemiology. Health Canada 18(4)

Norris JL III, Pollock KH (1998) Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. Environmental and Ecological Statistics 5, 391–402

Norris JL III, Pollock KH (1996) Nonparametric MLE under two closed capture-recapture models with heterogeneity. Biometrics 52, 639–649

Pledger S (2000) Unified maximum likelihood estimates for closed capture-recapture models using mixtures. Biometrics 56, 434–442

Sanathanan L (1972) Estimating the size of a multinomial population. Annals of Mathematical Statistics 42, 58–69

Sanathanan L (1977) Estimating the size of a truncated sample. Journal of the American Statistical Association 72, 669–672

Schouten LJ, Straatmann H, Kiemeney LA, Gimbrere CH, Verbeek AL (1994) The capture-recapture method for estimation of cancer registry completeness: a useful tool? International Journal of Epidemiology 23, 1111–1116

Scollnik D (1997) Inference concerning the size of the zero class from an incomplete Poisson sample. Communication in Statistics – Theory and Methods 26, 221–236a

Sekar C, Deming WE (1949) On a method of estimating birth and death rates and the extent of registration. JASA 44, 101–115

Tilling K (2001) Capture-recapture methods – useful or misleading? International Journal of Epidemiology 30, 12–14

van der Heijden PGM, Bustami R, Cruyff M, Engbersen G, van Houwelingen HC (2003) Point and interval estimation of the population size using the truncated Poisson regression model. Statistical Modelling – An International Journal 3, 305–322

van der Heijden PGM, Cruyff M, van Houwelingen H C (2003) Estimating the size of a criminal population from police records using the truncated Poisson regression model. Statistica Neerlandica 57, 1–16

Wilson RM, Collins MF (1992) Capture-recapture estimation with samples of size one using frequency data. Biometrika 79, 543–553

Wittes JT, Sidel VW (1968) A generalization of the simple capture-recapture model with applications to epidemiological research. Journal of Chronic Diseases 21, 287–301