Dankmar Böhning
Institute for Social Medicine, Epidemiology, and Health Economy
Charité FU/HU Berlin, Germany

A Monotonicity Property for the NPMLE of the Population Size Based Upon Truncated Count Mixtures

13. July 2004

Special Session on
Capture-Recapture Procedures in Public Health
IBC, Cairns, Australia, 2004

OVERVIEW

- **I. Cholera Outbreak in India**
- **II. Truncated Count Distributions (Poisson)**
- **III. Truncated Poisson Mixtures**
- **IV. Application Study in Illicit Drug Use**
- **V. Main Result: A Monotonicity Property for the Population Size Estimator**

I. Cholera Outbreak in the Indian Village, 1915-1920

McKendrick (1926): data most likely collected during his military service as physician in India

How many houses are affected?

affected houses with $x > 0$ cases

affected houses with $x = 0$ cases

unaffected houses (also with $x = 0$ cases)

Formalizing

N unknown (affected houses) in population

n observed (affected houses with Cholera cases)

$1 - p_0$ probability for observing a house with cholera case

therefore:

$$N = \#\text{ observed} + \#\text{ unobserved}$$
$$= n + Np_0$$

$\Rightarrow$ **Horvitz-Thompson Estimator**:

$$\hat{N} = n/(1 - p_0)$$

however: $\hat{N}$ needs p_0!

II. Truncated Count Distribution

Estimation of p_0 requires more knowledge on *data structure* :

McKendrick (1926, following Meng 1997)

X= *Count of Cholera cases per household*

count x	1	2	3	4
frequency n_x	32	16	6	1

$(n = n_1 + ... + n_4 = 55)$

n_0 (count of affected households without cases): **missing**:

X has **(zero-) truncated count distribution**

Poisson

for a count variable Y :

$p_y = e^{-\lambda}\lambda^y/y!$ for $y = 0, 1, 2, 3, ...$

then

$$p_x = \left(e^{-\lambda}\lambda^x/x!\right) / \left(1 - p_0\right)$$
$$= \left(e^{-\lambda}\lambda^x/x!\right) / \left(1 - e^{-\lambda}\right) \text{ for } x = 1, 2, 3, ...$$

is associated **truncated** Poisson density

Note:

$$Y = 1, 2, 3, ... \Longleftrightarrow X = 1, 2, 3, ...$$
$$Y = 0 \Rightarrow X \; \textit{not} \text{ observed}$$

Apply to Cholera outbreak data: how many households are affected?

a) **MLE of λ**

$$L(\lambda) = \Pi_{i=1}^{n}\, p_{x_i} = \Pi_{i=1}^{n}\, \frac{e^{-\lambda}}{1-e^{-\lambda}}\lambda^{x_i}/x_i!$$

$$\hat{\lambda}:\ L(\hat{\lambda}) \geq L(\lambda)!$$

for **McKendrick-Daten**: $\hat{\lambda} = 0.972$

b) **how many affected households?**

$\hat{N} = n/(1-e^{-\hat{\lambda}}) = 55/(1-e^{-0.972}) = 88$

... **33 houses** are estimated to be additionally affected by outbreak

c) **Is truncated Poisson adequate?** diagnostics:

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - \hat{n}_i)^2}{n_i} = 0.197\ (2df)$$

Some Examples from Public Health

Estimating the amount of

- a health problem
  - drugs (Hser 1993)
  - HIV (Smit *et al.* 1997)
  - alcohol dependency among practicing physicians (Berliner Ärzteblatt 2004)
- illegal immigrants (v. d. Heijden *et al.* 2003)
- homeless population (Smit *et al.* 2002)

Wickens (*Journ. Drug Issues* 1993): review paper

Parameter Estimation with the EM-Algorithm

$s = 0 \times n_0 + 1 \times n_1 + 2 \times n_2 + 3 \times n_3 + ...$

known (independent of n_0)

i) conditional upon N :

$$\hat{\lambda}_N = \frac{1}{N} s$$

ii) conditional upon λ:

$$\hat{N}_\lambda = \frac{n}{1 - e^{-\lambda}}$$

iterate between i) and ii) !

insert ii) into i):

$\lambda = (1 - e^{-\lambda}) s/n = (1 - e^{-\lambda}) \overline{x} = \Phi(\lambda)$

$\Rightarrow$ **fixed point equation** (follows also directly from likelihood of truncated Poisson!)

step i) and ii) are the steps in the

EM Algorithm

M-step (i) complete likelihood

$$\hat{\lambda} : \Pi_{i=0}^{m}(e^{-\hat{\lambda}}\hat{\lambda}^{i}/i!)^{n_i} \geq \Pi_{i=0}^{m}(e^{-\lambda}\lambda^{i}/i!)^{n_i},$$

conditional upon $n_0 = \hat{n}_0$, m largest, observed count

E-step (ii) expected value of **missing frequency of zero counts**

$$\hat{n}_0 = E(n_0|n_1, n_2, ..., n_m; \lambda) = n\frac{e^{-\lambda}}{1 - e^{-\lambda}},$$

conditional upon $\lambda = \hat{\lambda}$

III. Alternatives and Truncated Poisson Mixtures

Illegal immigrants in the Netherlands

van der Heijden *et al.* (*Stat. Mod.* 2003): police data set on **1880** illegale immigrants in the Netherlands

X Count of captures per person with consecutive expelling

count x	1	2	3	4	5
frequency n_x	1645	183	37	13	2
Poisson $\hat{n}_x$	1604,6	247,8	25,5	2,0	0,1

diagnostics:

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - \hat{n}_i)^2}{n_i} = 38,63 \ (3df)$$

Result: simple Poisson unsuitable

more general:

$f(y|\lambda)$ density for general count distribution for $y = 0, 1, 2, ..$

$\frac{1}{1-f(0|\lambda)} f(x|\lambda)$ associated truncated density for $x = 1, 2, ..$

$f(y|\lambda)$ *could* be:

- negative Binomial (Poisson-Gamma)
- double Poisson (Efron *JASA* 1986)
- weighted Poisson (Ridout *et al. Stat. Mod.* 2004)
- ...

EM Algorithm for Truncated Count Distributions

M-Step: complete data likelihood

$$\hat{\lambda} : \Pi_{i=0}^{m} f(i|\hat{\lambda})^{n_i} \geq \Pi_{i=0}^{m} f(i|\lambda)^{n_i},$$

conditional upon $n_0 = \hat{n}_0$

E-Step:

$$\hat{n}_0 = E(n_0|n_1, n_2, ..., n_m; \lambda) = n\frac{f(0|\lambda)}{1 - f(0|\lambda)},$$

conditional upon $\lambda = \hat{\lambda}$

Poisson Mixtures

Y non-parametric mixed Poisson density with k components:

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

for $y = 0, 1, 2, ...$ but now with parameter Q (mixing distribution)

$$Q = \begin{pmatrix} \lambda_1 & \lambda_2 & ... & \lambda_k \\ q_1 & q_2 & ... & q_k \end{pmatrix}$$

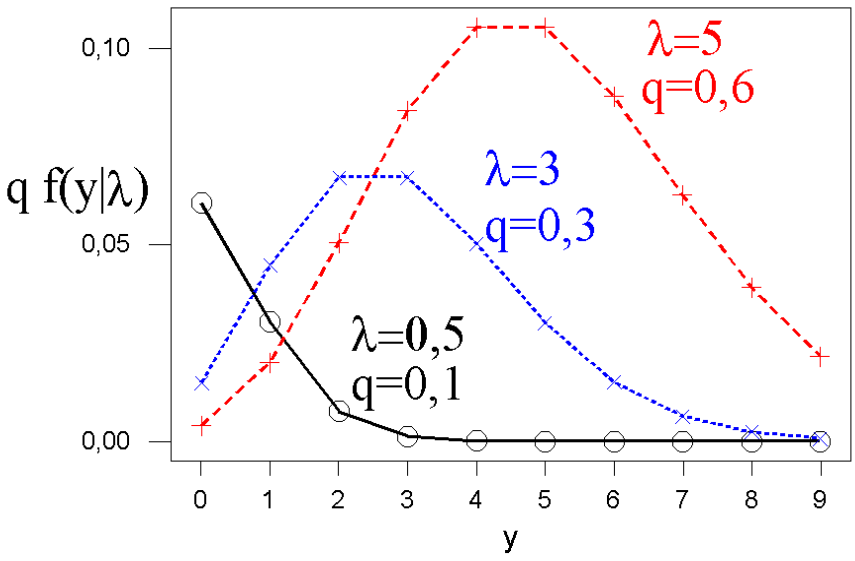
non-parametric mixture models more and more popular since the 80s

several books: Titterington, Smith & Makov (1985), McLachlan & Basford (1988), Lindsay (1995), Böhning (2000, *C.A.MAN*), McLachlan & Peel (2000)

important reason: mixture models unobserved population heterogeneity

population has k subpopulations:
j-th component: $e^{-\lambda_j}\lambda_j^y/y!$ with weight q_j

for example, population with three components:



$\boldsymbol{Z}$ unobserved k-vector **with exactly one 1** in j-th position

then:

$$f(y,\boldsymbol{z}) = f(y|\boldsymbol{z})f(\boldsymbol{z}) = f(y|\lambda_j)q_j$$

marginal density is mixture:

$$\Sigma_z f(y,\boldsymbol{z}) = \Sigma_{j=1}^{k} f(y|\lambda_j)q_j$$

$$= \sum_{j=1}^{k} e^{-\lambda_j}\lambda_j^y/y!\; q_j$$

ML Estimation with the EM Algorithm for (untruncated) Poisson mixtures

sample $n_0, n_1, n_2, ..., n_m$ available with Likelihood:

$$\textbf{observed: } \prod_{i=0}^{m} \left(\sum_{j=1}^{k} f(i|\lambda_j)\ q_j \right)^{n_i}$$

$$\textbf{complete: } \prod_{i=0}^{m} \left(\prod_{j=1}^{k} \{f(i|\lambda_j) q_j\}^{z_{ij}} \right)^{n_i}$$

$f(i|\lambda_j) = e^{-\lambda_j} \lambda_j^i / i!$

McLachlan, Peel (2001), McLachlan, Krishnan (1997)

E-Step:

$$e_{ij} = E(Z_{ij}|n_0, n_1, ..., n_m; Q) \overset{Bayes}{=} \frac{f(i|\lambda_j) q_j}{\Sigma_l f(i|\lambda_l) q_l}$$

M-Step:

$$\hat{\lambda}_j = [\sum_{i=0}^{m} e_{ij}\ i\ n_i] / [\sum_{i=0}^{m} e_{ij}\ n_i]$$

$$\hat{q}_j = [\sum_{i=0}^{m} n_i\ e_{ij}] / N$$

for Poisson component $j = 1, ..., k$.

Truncated Poisson Mixtures: Nested EM Algorithm

Böhning and Schön (*JRSS C* 2004)

Data: $n_1, n_2, ..., n_m$
Model: $f(i|Q) = \Sigma_{j=1}^{k} e^{-\lambda_j} \lambda_j^i / i!\ q_j$

E-Step: $Q = \hat{Q}$ fixed,

exterior cycle

$\hat{n}_0 = E(n_0|n_1, n_2, ..., n_m; Q) = n \frac{f(0|Q)}{1-f(0|Q)}$

M-Step:

$$\hat{Q} : \Pi_{i=0}^{m} f(i|\hat{Q})^{n_i} \geq \Pi_{i=0}^{m} f(i|Q)^{n_i}$$

cond. upon $n_0 = \hat{n}_0, N = \hat{N} = \hat{n}_0 + n$

solved by **EM Algorithm for Mixtures**:

interior cycle

E-Step:

$$e_{ij} = \frac{f(i|\lambda_j) q_j}{\Sigma_l\, f(i|\lambda_l) q_l}$$

M-Step:

$$\hat{\lambda}_j = [\sum_{i=0}^{m} e_{ij}\ i\ n_i] / [\sum_{i=0}^{m} e_{ij}\ n_i]$$

$$\hat{q}_j = [\sum_{i=0}^{m} n_i\ e_{ij}] / N$$

IV. Application Study in Illicit Drug Use

Böhning, Busaba *et al.* (*EJE* 2004)

Description of Study

Objective: Estimate the Number of Drug Users in Bangkok (Thailand)

Data Source: all 61 treatment institutions licensed for treating drug addicts

Period: 1. Okt. 2001 - 31. Dec. 2001

Target Variable: X Count of Treatment Episodes per Patient

observed number of drug users:
$n = 11,222$ with 26,638 episodes

Study Results

Demography:

Drug	n	Age (SD)	% male	EA (SD)
Heroin	7048	30,8 (8,2)	96,0	2,9 (2,52)
Metamph.	3334	22,3 (5,9)	91,8	1,1 (0,66)
Other	812	34,3 (11,5)	94,7	2,6 (2,68)

Distribution of Count of Episodes for Heroin Users:

x	1	2	3	4	5	6	7	8	9
n_x	2955	1186	803	611	416	338	278	180	125
x	10	11	12	13	14	15	16	17	18+
n_x	74	38	20	14	11	4	1	3	5

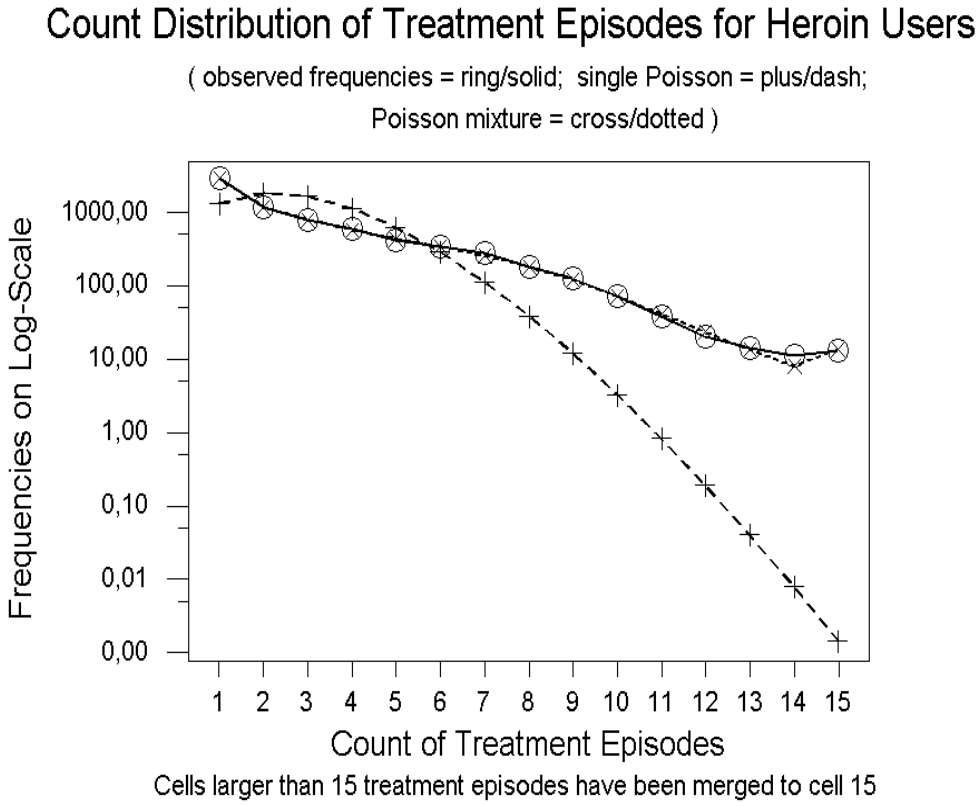
n_x: Frequency of **Heroin**-using Patients with exactly x treatment episodes

Estimating the Number of Heroin Users:

k	$\hat{\lambda}_j$	$\hat{q}_j$	log-likelih.	AIC	BIC	$\hat{N}$
1	2,75	1,00	-15462	-30927	-30934	7543
2	0,88	0,75	-13214	-26434	-26455	10226
	5,40	0,25				
3	0,41	0,69	-13134	-26279	-26313	13350
	2,97	0,22				
	6,80	0,09				
4	0,21	0,70	-13120	**-26255**	**-26303**	17278
	2,13	0,19				
	5,84	0,10				
	12,20	0,01				

$$AIC = 2 \times \text{log-likelihood} - (2k-1)2$$

$$BIC = 2 \times \text{log-likelihood} - (2k-1)\log(n)$$



$$\text{log-likelihood} = \sum_{i=1}^{m} n_i \log \left\{ \frac{f(i|\hat{Q})}{1 - f(0, \hat{Q})} \right\}$$

Diagnostics:

simple Poisson ($k = 1$):

$$\chi^2 = 3245.20 \text{ mit } 13df$$

Poisson mixture ($k = 4$):

$$\chi^2 = 5.65 \text{ mit } 7df.$$

V. A Monotonicity Property for the Population Size Estimator

Result: Böhning and Schön (*JRSS C* 2004)

$\hat{N}_k$ MLE of population size w.r.t. a truncated Poisson mixture with k components, $k = 1, 2, ...$ Then:

$$\hat{N}_k \geq \hat{N}_1$$

likely, the **more general statement** is also true:

$$\hat{N}_{k+1} \geq \hat{N}_k$$

Sketch of Proof:

i) let $(\hat{N}_k^{(j)})_{j\geq 0}$ be a sequence of the nested EM Algorithm with assoc. sequence $(\hat{Q}^{(j)})_{j\geq 0}$. For $j = 0$ the statemenent is valid. Remains to show:

$$\hat{N}_k^{(j)} \geq \hat{N}_1^{(j)} \Rightarrow \hat{N}_k^{(j+1)} \geq \hat{N}_1^{(j+1)}.$$

ii)

$$\overline{y}_k^{(j+1)} = (n_1\ 1 + n_2\ 2 + ... + n_m\ m)/\hat{N}_k^{(j)}$$

Now, because $\hat{N}_k^{(j)} \geq \hat{N}_1^{(j)}$:

$$\overline{y}_k^{(j+1)} \leq \overline{y}_1^{(j+1)} \overset{e^{-x} decreasing}{\Leftrightarrow} e^{-\overline{y}_k^{(j+1)}} \geq e^{-\overline{y}_1^{(j+1)}}$$

iii) Property of the EM Algorithm for Poisson mixtures: $\Sigma_{\iota=1}^k \hat{\lambda}_\iota^{(j+1)} q_\iota^{(j+1)} = \overline{y}_k^{(j+1)}$.

$$\Rightarrow e^{\Sigma_{\iota=1}^k -\hat{\lambda}_\iota^{(j+1)} q_\iota^{(j+1)}} = e^{-\overline{y}_k^{(j+1)}} \geq e^{-\overline{y}_1^{(j+1)}}$$

iv) Apply **Jensen's inequality** for convex fcts:

$$\Rightarrow \sum_{\iota=1}^k e^{-\hat{\lambda}_\iota^{(j+1)}} q_\iota^{(j+1)} \geq e^{-\Sigma_{\iota=1}^k -\hat{\lambda}_\iota^{(j+1)} q_\iota^{(j+1)}} \geq e^{-\overline{y}_1^{(j+1)}}$$

v) $\frac{1}{1-x}$ strictly increasing in $(0, 1)$:

$$\hat{N}_k^{(j+1)} = \frac{n}{1 - \Sigma_{\iota=1}^k e^{-\hat{\lambda}_\iota^{(j+1)}} q_\iota^{(j+1)}} \geq \frac{n}{1 - e^{-\overline{y}_1^{(j+1)}}} = \hat{N}_1^{(j+1)}$$

Concluding Remarks

Open Problems and Research Questions

- Standard errors and confidence intervals
- Suitable modification of resampling techniques
- Validation studies
- Comparison to other approaches (Pollock-Norris or Zelterman)
- ...