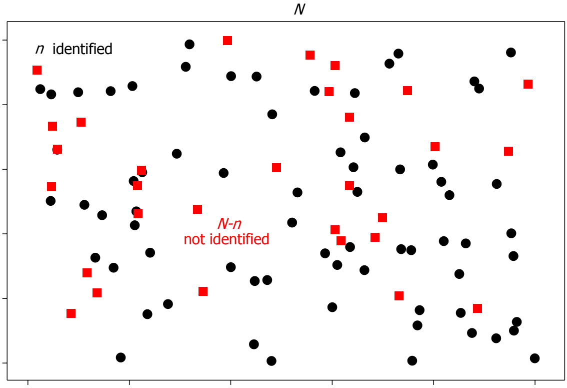
Capture-Recapture Estimation of Population Size by Means of Truncated Likelihood and Empirical Bayesian Smoothing

Dankmar Böhning

Applied Statistics, School of Biological Sciences
University of Reading

University of Kent, Canterbury, 12 November 2009

an **elusive** population has N units of which n are identified by some mechanism (trap, register, police database, ...)



Formulation of the Problem

- probability of identifying an unit is $(1 - p_0)$
- so that $N = \underbrace{(1 - p_0)N}_{\textit{observed}} + \underbrace{p_0 N}_{\textit{hidden}} \approx n + p_0 N$
- and the **Horvitz-Thompson** estimator follows:

$$\hat{N} = \frac{n}{1 - p_0}$$

- usually an **estimate** of p_0 is required

Formulation of the Problem as Frequencies of Frequencies

Frequencies of Frequencies

a common setting for estimating p_0 is the **Frequencies of Frequencies** setting:

Identifying Mechanism

the identifying mechanism provides a count Y of **repeated identifications** (w.r.t. to a reference period)

Illustration of the CR-Concept

Table: *Illustration with Case Data from Software Inspection (Wohlin et al. 1995)*

	Reviewers				
Error i	R1	R2	...	R22	Marginal Y_i
1	1	0	...	1	2
2	1	1	...	0	4
3	0	0	...	1	2
4	0	0	...	0	**0**
5	0	1	...	0	1
...	...	...	...	...	...
38	1	1	...	0	7

Formulation of the Problem as Frequencies of Frequencies

Marginal distribution

marginal distribution of Y is leading to frequencies $f_1, f_2, ..., f_m$ of the counts $1, 2, .., m$ (m is the largest observed count)

estimating f_0 on the basis of $f_1, f_2, ..., f_m$

zero counts are **not** observed: hence f_0 **is unknown**
Recall that $N = f_0 + n = f_0 + f_1 + f_2 + ... + f_m$, so that $\hat{f}_0$ leads to $\hat{N}$

Illustration of the frequencies of frequencies situation at hand of the software inspection data

Table: Zero-truncated count distribution of software errors

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
-	5	1	5	1	3	2	0	5	4	2

f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}	f_{19}	f_{20}	n
3	1	0	2	0	1	0	0	0	1	36

Application Areas

- Epidemiology and Medicine
- Biology and Agriculture
- Social Science and Criminology
- Research on Terrorism
- Systems Engineering
- Text Analysis and Language studies

Hser's Data on Estimating Hidden Intravenous Drug Users in Los Angeles 1989

- intravenous drug users in L.A. county were entered into the California Drug Abuse Data System (CAL-DADS)
- the data below refer to the frequency distribution of the episode count per drug user in 1989

the frequency distribution of the **episode count per drug user** for the year 1989:

f_0	f_1	f_2	f_3	f_4	f_5	f_6
-	11,982	3,893	1,959	1,002	575	340

f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	n
214	90	72	36	21	14	20,198

Screening for colorectal adenomatous polyps

- In 1990, the Arizona Cancer Center initiated a multicenter trial to determine whether wheat bran fiber (WBF) can prevent the recurrence of colorectal adenomatous polyps (Alberts *et al.* (2000) and Hsu (2007)).
- Subjects with previous history of colorectal adenomatous polyps were recruited and randomly assigned to one of two treatment groups, low fiber and high fiber.

Screening for colorectal adenomatous polyps

- The researchers noted that adenomatous polyp data are often subject to unobservable measurement error due to misclassification at colonoscopy. It can be assumed that patients with a positive polyp count were diagnosed correctly, whereas it is unclear how many persons with zero-count of polyps were false-negatively diagnosed.
- Thus we approach the data as if zero-counts were not observed, and we try to estimate the undercount from the non-zero frequencies.
- the maximum polyp count in a patient is 77.

Screening for colorectal adenomatous polyps

Table: *Arizona polyps data: count distribution of recurrent adenomatous polyps per patient, separated for low and fiber group*

Count of polyps	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8+
low fiber group									
No. of subjects	*(285)*	145	66	39	17	8	8	7	9
high fiber group									
No. of subjects	*(381)*	144	61	55	37	17	5	4	15

Del Rio Vilas's Data on Estimating Hidden Scrapie in Great Britain 2005

- sheep is kept in holdings in Great Britain (and elsewhere)
- the occurrence of scrapie is monitored in the Compulsory Scrapie Flocks Scheme (CSFS) summarizing abbatoir survey, stock survey and the statutory reporting of clinical cases
- CSFS established since 2004

the frequency distribution of the **scrapie count within each holding** for the year 2005:

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	n
-	84	15	7	5	2	1	2	2	118

Microbial diversity in the Gotland Deep.

- The data on microbial diversity shown in the table below stem from a recent work by Stock *et al.* (2009).

Table: *Protistan diversity in the Gotland Deep: Frequency counts of observed species.*

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
-	48	9	6	2	0	2	0	2	1	1

Microbial diversity in the Gotland Deep.

- Microbial ecologists are interested in estimating the number of species N in particular environments.
- Unlike butterflies, microbial species membership is not clear from visual inspection, so individuals are defined to be members of the same species (or more general taxonomic group) if their DNA sequences (derived from a certain gene) are identical up to some given percentage, 95% in this case.
- Here the study concerned protistan diversity in the Gotland Deep, a basin in the central Baltic Sea. The sample was collected in May 2005, resulting in the data displayed in the above table. The maximum observed frequency was 53.

How many words did Shakespeare know?

- Efron and Thisted (1987, *Biometrika*): How many words did Shakespeare know, but not use?
- important question in text analysis and estimation of language knowledge

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	...	n
-	14,376	4,343	2,292	1,463	1,043	837	638	..	31,534

Formulation of the Problem and the Idea for its Solution

Suppose we can find some model for the count probabilities

$$p_j = p_j(\lambda)$$

then estimate λ by some method (truncated likelihood) and then use the model for p_0:

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

Formulation of the Problem and the Idea for its Solution

Only to illustrate: Poisson model for the count probabilities

$$p_j = p_j(\lambda) = \exp(-\lambda)\lambda^j/j!$$

then estimate λ maximizing the zero-truncated Poisson likelihood

$$\prod_{j=1}^{m}\left(\frac{p_j}{1-p_0}\right)^{f_j} = \prod_{j=1}^{m}\left(\frac{1}{1-\exp(-\lambda)}\exp(-\lambda)\lambda^j/j!\right)^{f_j}$$

and yield estimate for N

$$\hat{N} = \frac{n}{1-\hat{p}_0} = \frac{n}{1-\exp(-\hat{\lambda})}$$

What speaks against this simple solution?

However: using a simple Poisson model for the count probabilities

$$p_j = p_j(\lambda) = \exp(-\lambda)\lambda^j/j!$$

is **not** appropriate, since

- every unit is different
- there is population heterogeneity

so that more **realistic**

$$p_j = p_j(\lambda) = \int_0^\infty \exp(-t)t^j/j!\lambda(t)dt$$

where $\lambda(t)$ stands for the heterogeneity distribution of the Poisson parameter t

Effects of Heterogeneity?

Table: *Simulation using* $Y \sim 0.5Po(1) + 0.5Po(t)$ *and* $N = 100$

t	estimator	mean	SD	RMSE
1	MLE-hom	**101.91**	12.98	13.12
2	MLE-hom	**94.07**	7.02	9.19
3	MLE-hom	**88.19**	4.96	12.81
4	MLE-hom	**85.34**	4.30	15.30
5	MLE-hom	**83.47**	3.71	16.94

Effect of Heterogeneity on an estimator under homogeneity:

underestimation because of Jensen's inequality applied to $\exp(x)$:

$$\frac{n}{1-p_0} = \frac{n}{1-\int_0^\infty \exp(-t)\lambda(t)dt}$$

$$\geq \frac{n}{1-\exp\left(-\int_0^\infty t\lambda(t)dt\right)}$$

$$= \frac{n}{1-\exp(-\mu)},$$

where $\mu = \int_0^\infty t\lambda(t)dt$

Simple nonparametric estimates under heterogeneity

under heterogeneity

instead of providing an estimate $\hat{\lambda}(t)$ in

$$p_j(\lambda) = \int_0^\infty \exp(-t)t^j/j!\lambda(t)dt$$

by means of

- **parametric** Poisson-Gamma (Chao and Bunge 2002 *Biometrics*)
- or **nonparametric mixture models** (Böhning and Schön 2005, *JRSSC*, B"ohning and Kuhnert 2006, *Biometrics*)

interest is on the lower bound approach by **Chao** (1987, 1989, *Biometrics*)

mixed Poisson

consider

$$p_j = \int_0^\infty \exp(-t) t^j / j! \lambda(t) dt$$

with unknown $\lambda(t)$ for $t > 0$. Then, by the **Cauchy-Schwarz** inequality

$$\frac{p_1}{p_0} \leq \frac{2p_2}{p_1} \leq \frac{3p_3}{p_2} ... \leq \frac{(j+1)p_{j+1}}{p_j} \leq ....$$

in particular, for $j = 0$

$$\frac{p_1^2}{2p_2} \leq p_0$$

leads to **Chao's lower bound estimate** (truely nonparametric)

$$\hat{f}_0 = \frac{f_1^2}{2f_2} \text{ or } \hat{N} = n + \hat{f}_0$$

Comparing the Estimators

Table: *Simulation using* $Y \sim 0.5Po(1) + 0.5Po(\lambda)$ *and* $N = 100$

λ	estimator	mean	SD	RMSE
1	MLE-hom	101.91	12.98	13.12
	Chao	**103.82**	18.73	19.12
2	MLE-hom	94.07	7.02	9.19
	Chao	**99.10**	12.22	12.25
3	MLE-hom	88.19	4.96	12.81
	Chao	**96.61**	9.77	10.34
4	MLE-hom	85.34	4.30	15.30
	Chao	**97.03**	10.00	10.43
5	MLE-hom	83.47	3.71	16.94
	Chao	**97.98**	10.24	10.43

The Idea for a robust approach of Zelterman (1988, *JSPI*)

- he noted that

$$\lambda = \frac{\lambda^{j+1}}{\lambda^j} = (j+1)\frac{\lambda^{j+1}/(j+1)!}{\lambda^j/j!}$$

$$\lambda = (j+1)\frac{Po(j+1;\lambda)}{Po(j;\lambda)}$$

- leading to the proposal

$$\hat{\lambda}_j = (j+1)\frac{f_{j+1}}{f_j}$$

- and in particular for $j = 1$

$$\hat{\lambda} = \hat{\lambda}_1 = 2\frac{f_2}{f_1}$$

The idea for a robust approach of Zelterman

$\hat{\lambda} = 2\frac{f_2}{f_1}$ is **robust** in the sense that

- it is **not affected** by any changes in counts larger than 2
- count distribution need only to behave **like** a Poisson for counts of 1 or 2

Zelterman larger than Chao?

Table: *Simulation using* $Y \sim 0.5Po(1) + 0.5Po(\lambda)$ *and* $N = 100$

λ	estimator	mean	SD	RMSE
1	MLE-hom	101.91	12.98	13.12
	Chao	103.82	18.73	19.12
	Zelterman	104.51	21.48	21.95
2	MLE-hom	94.07	7.02	9.19
	Chao	99.10	12.22	12.25
	Zelterman	101.49	16.22	16.29

Table: *Simulation using* $Y \sim 0.5Po(1) + 0.5Po(\lambda)$ *and* $N = 100$

λ	estimator	mean	SD	RMSE
3	MLE-hom	88.19	4.96	12.81
	Chao	96.61	9.77	10.34
	Zelterman	102.23	15.31	15.47
4	MLE-hom	85.34	4.30	15.30
	Chao	97.03	10.00	10.43
	Zelterman	107.85	19.84	21.34
5	MLE-hom	83.47	3.71	16.94
	Chao	97.98	10.24	10.43
	Zelterman	115.19	23.12	27.66

Zelterman larger than Chao?

- $$\hat{N}_Z = \frac{n}{1-\exp(-\hat{\lambda})} = n + \frac{n}{\exp(\hat{\lambda})-1} \approx n + \frac{n}{1+\hat{\lambda}+\frac{1}{2}\hat{\lambda}^2-1}$$
- $$= n + \frac{n}{\hat{\lambda}+\frac{1}{2}\hat{\lambda}^2} = n + \frac{n}{\frac{2f_2}{f_1}+\frac{1}{2}\left(\frac{2f_2}{f_1}\right)^2} = n + \left(\frac{f_1^2}{2f_2}\right)\frac{n}{f_1+f_2}$$
- $$\geq n + \left(\frac{f_1^2}{2f_2}\right) = \hat{N}_C$$
- **yes,** if $\hat{\lambda}$ is **small** (Böhning *SJOS* 2009)

Zelterman Estimation as a Result of a Truncated Poisson Likelihood

Zelterman estimate truncates all counts different from 1 or 2: write

$$1 - p = p_1 = \frac{\exp(-\lambda)\lambda}{\exp(-\lambda)\lambda + \exp(-\lambda)\lambda^2/2} = \frac{1}{1+\lambda/2}$$

$$p = p_2 = \frac{\exp(-\lambda)\lambda^2/2}{\exp(-\lambda)\lambda + \exp(-\lambda)\lambda^2/2} = \frac{\lambda/2}{1+\lambda/2}$$

and consider associated **binomial** log-likelihood

$$f_1 \log(p_1) + f_2 \log(p_2) = f_1 \log(1-p) + f_2 \log(p)$$

which is maximized for $\hat{p} = \hat{p}_2 = \frac{f_2}{f_1+f_2}$, or

$$\hat{\lambda} = \frac{2\hat{p}_2}{1-\hat{p}_2} = \frac{2f_2}{f_1}$$

Making Zelterman right

where Zelterman is **right**:

the Zelterman estimate of λ comes out as the MLE from a 2-truncated Poisson likelihood

$$\hat{\lambda} = 2f_2/f_1$$

where Zelterman is **wrong**:

it should use

$$E(f_0|\lambda, f_1, f_2) = \frac{Po(0|\lambda)}{Po(1|\lambda) + Po(2|\lambda)}(f_1 + f_2) = \frac{(f_1 + f_2)}{\lambda + \lambda^2/2}$$

$$E(f_0|\hat{\lambda}, f_1, f_2) = \frac{(f_1 + f_2)}{\hat{\lambda} + \hat{\lambda}^2/2} = \frac{f_1^2}{2f_2}$$

- Zelterman should use

$$\hat{N} = n + E(f_0|\lambda = 2f_2/f_1, f_1, f_2) = n + \frac{f_1^2}{2f_2},$$

  entirely **identical to Chao's estimator**
- but instead uses

$$\hat{N} = \frac{n}{1 - \exp(-2f_2/f_1)}$$

  resulting in a potentially **strong overestimation** if heterogeneity is strong

Truncated Poisson Likelihoods offer Flexibility

a likelihood framework offers generalizations:

- extending Chao's estimator: finding best lower bounds
- capture-recapture modelling between robustness and efficiency
- include higher counts to improve efficiency

Robustness vs. Efficiency

original observed counts $f_1, f_2, ..., f_m$ with f_0 **unobserved**
the following sequential truncation is considered:

1. f_1, f_2 (**most robust, least efficient**)
2. f_1, f_2, f_3
3. ....
4. $f_1, f_2, ..., f_{m-1}$
5. $f_1, f_2, ..., f_{m-1}, f_m$ (**most efficient, least robust**)

note that 1) is the Chao approach, whereas 5) corresponds to the conventional maximum likelihood approach

Maximum Likelihood Estimators

original observed counts $f_1, f_2, ..., f_m$ with f_0 **unobserved**
the following sequential truncation is considered:

1. MLE-2 (Chao): f_1, f_2 (**most robust, least efficient**)
2. MLE-3: f_1, f_2, f_3
3. MLE-4: f_1, f_2, f_3, f_4
4. ....
5. MLE-(m-1): $f_1, f_2, ..., f_{m-1}$
6. MLE-m (homogeneity): $f_1, f_2, ..., f_{m-1}, f_m$ (**most efficient, least robust**)

Associated Likelihoods

original observed counts $f_1, f_2, ..., f_m$ with f_0 **unobserved**
the following sequential truncation is considered with log-Likelihoods:

1. f_1, f_2: $f_1 \log p_1 + f_2 \log p_2$
2. f_1, f_2, f_3: $f_1 \log p_1 + f_2 \log p_2 + f_3 \log p_3$
3. ....
4. $f_1, f_2, ..., f_{m-1}$: $f_1 \log p_1 + f_2 \log p_2 + ... + f_{m-1} \log p_{m-1}$
5. $f_1, f_2, ..., f_{m-1}, f_m$: $f_1 \log p_1 + f_2 \log p_2 + ... + f_m \log p_m$

with

$$p_i = \exp(-\lambda)\lambda^i/i!/\sum_{x=1}^{j} \exp(-\lambda)\lambda^x/x!$$

Generalized Chao Estimator MLE-3 has a Closed Form

truncate all counts different from 1, 2, and 3:

$$p_1 = \frac{\exp(-\lambda)\lambda}{\exp(-\lambda)\lambda + \exp(-\lambda)\lambda^2/2 + \exp(-\lambda)\lambda^3/6} = \frac{1}{1 + \lambda/2 + \lambda^2/6}$$

$$p_2 = \frac{\exp(-\lambda)\lambda^2/2}{\exp(-\lambda)\lambda + \exp(-\lambda)\lambda^2/2 + \exp(-\lambda)\lambda^3/6} = \frac{\lambda/2}{1 + \lambda/2 + \lambda^2/6}$$

$$p_3 = \frac{\exp(-\lambda)\lambda^3/6}{\exp(-\lambda)\lambda + \exp(-\lambda)\lambda^2/2 + \exp(-\lambda)\lambda^3/6} = \frac{\lambda^2/6}{1 + \lambda/2 + \lambda^2/6}$$

Generalized Chao Estimator as a Result of a Truncated Poisson Likelihood

and consider associated **trinomial** log-likelihood

$$\log L(\lambda) = f_1 \log(p_1) + f_2 \log(p_2) + f_3 \log(p_3)$$

which is maximized for

$$\hat{\lambda} = -\frac{3}{2}\frac{f_1 - f_3}{f_2 + 2f_1} + \sqrt{\frac{6(f_2 + 2f_3)}{f_2 + 2f_1} + \left(\frac{3}{2}\frac{(f_1 - f_3)}{f_2 + 2f_1}\right)^2} \geq 0$$

and, finally

$$\hat{N} = n + E(f_0|\hat{\lambda}, f_1, f_2, f_3)$$

EM Algorithm

consider **arbitrary truncation count** J, $2 \leq J \leq m$:

observed, incomplete likelihood

$$\prod_{j=1}^{J} p_j^{f_j}$$

with

$$p_j = \exp(-\lambda)\lambda^j / j! / \sum_{x=1}^{J} \exp(-\lambda)\lambda^x / x!$$

EM Algorithm

unobserved, complete likelihood

$$\prod_{j=0}^{m} p_j^{f_j}$$

with

$$p_j = \exp(-\lambda)\lambda^j / j!$$

Robustness vs. Efficiency: MLE-3

M-Step

suppose **all** counts $f_0, f_1, f_2, ..., f_m$ were observed
then the parameter of the Poisson is easily available by maximizing the Poisson likelihood

$$\hat{\lambda} = \sum_{x=0}^{m} x \times f_x / \sum_{x=0}^{m} f_x$$

E-Step

1. $e_0, f_1, f_2, e_3, ..., e_m$
2. $e_0, f_1, f_2, f_3, e_4, ..., e_m$
3. ...
4. $e_0, f_1, f_2, ...f_{m-2}, e_{m-1}, e_m$

E-Step details

consider an arbitrary truncation count J:

$$e_0, f_1, f_2, ..., f_J, e_{J+1}, ..., e_m$$

clearly, for $x = 0$ or $x > J$

$$e_x = E(f_x | f_1, f_2, ..., f_J, \lambda) = Po(x|\lambda)N$$

$$= Po(x|\lambda)[e_0 + f_1 + f_2 + ... + f_J + e_{J+1} + ... + e_m]$$

E-Step

$$e_0 + \sum_{x=J+1}^{m} e_x$$

$$= [1 - \sum_{x=1}^{J} Po(x|\lambda)][f_1 + f_2 + ... + f_J] + [1 - \sum_{x=1}^{J} Po(x|\lambda)][e_0 + \sum_{x=J+1}^{m} e_x]$$

hence

$$e_0 + \sum_{x=J+1}^{m} e_x = \frac{1 - \sum_{x=1}^{J} Po(x|\lambda)}{\sum_{x=1}^{J} Po(x|\lambda)} [f_1 + f_2 + ... + f_J]$$

E-Step

finally:

$$
\begin{aligned}
e_x = \quad & Po(x|\lambda)[e_0 + f_1 + f_2 + \ldots + f_J + e_{J+1} + \ldots + e_m] \\
= \quad & Po(x|\lambda)[f_1 + f_2 + \ldots + f_J] \\
& + Po(x|\lambda)\frac{1-\sum_{x'=1}^{J} Po(x'|\lambda)}{\sum_{x'=1}^{J} Po(x'|\lambda)}[f_1 + f_2 + \ldots + f_J] \\
= \quad & \frac{Po(x|\lambda)}{\sum_{x'=1}^{J} Po(x'|\lambda)}[f_1 + f_2 + \ldots + f_J] \\
= \quad & \frac{\lambda^x/x!}{\sum_{j=1}^{J} \lambda^j/j!}[f_1 + f_2 + \ldots + f_J]
\end{aligned}
$$

e_0 for MLE-2 and MLE-3

$J = 2$ (Chao)

- $$e_0 = \frac{1}{\sum_{j=1}^{J} \lambda^j / j!}[f_1 + f_2 + ... + f_J] = \frac{f_1 + f_2}{\lambda + \lambda^2/2}$$

$J = 3$ (Generalized Chao)

- $$e_0 = \frac{1}{\sum_{j=1}^{J} \lambda^j / j!}[f_1 + f_2 + ... + f_J] = \frac{f_1 + f_2 + f_3}{\lambda + \lambda^2/2 + \lambda^3/6}$$

Estimating N

N is now estimated as

$$\hat{N} = e_0 + \sum_{i=1}^{m} f_i = E(f_0|\hat{\lambda}) + \sum_{i=1}^{m} f_i$$

Comparing the Estimators by Simulation

design

- sample $Y_i \sim 0.5Po(1) + 0.5Po(\lambda)$ for $i = 1, ..., N$ and $N = 100$ for $\lambda = 1, 2, 3, 4, 5$
- determine $f_0, f_1, ..., f_m$ from sample $y_1, ..., y_N$
- drop f_0
- determine MLE-2 (Chao), MLE-3, MLE-4, and MLE-m (homogenous) with associated sample size estimates
- repeat $B = 1,000$ times
- determine BIAS, SD, RMSE for MLE-2 (Chao), MLE-3, MLE-4, and MLE-m

Comparing the Estimators

Table: *Simulation using* $Y \sim 0.5Po(1) + 0.5Po(\lambda)$ *and* $N = 100$

λ	estimator	mean	SD	RMSE
1	MLE-2(Chao)	103.82	18.73	19.12
	MLE-3	102.49	14.35	14.56
	MLE-4	103.58	13.07	13.55
	MLE-hom	101.91	12.98	**13.12**
2	MLE-2(Chao)	99.10	12.22	12.25
	MLE-3	96.59	8.73	9.38
	MLE-4	96.74	7.71	**8.37**
	MLE-hom	94.07	7.02	9.19

Table: *Simulation using* $Y \sim 0.5Po(1) + 0.5Po(\lambda)$ *and* $N = 100$

λ	estimator	mean	SD	RMSE
3	MLE-2(Chao)	96.61	9.77	10.34
	MLE-3	93.23	6.52	**9.40**
	MLE-4	91.73	5.62	10.00
	MLE-hom	88.19	4.96	12.81
4	MLE-2(Chao)	97.03	10.00	10.43
	MLE-3	92.68	6.41	**9.73**
	MLE-4	89.86	5.15	11.37
	MLE-hom	85.34	4.30	15.30
5	MLE-2(Chao)	97.98	10.24	10.43
	MLE-3	93.10	6.35	**9.37**
	MLE-4	89.28	5.18	11.91
	MLE-hom	83.47	3.71	16.94

Application to Study Data

	data				size estimators $\hat{N}$		
Data set	f_1	f_2	f_3	n	Zelt.	Chao	MLE-3
Drugs L.A.	11982	3893	1959	20198	42268	38637	33434
Polyps-l.	145	66	39	299	500	458	416
Polyps-h.	144	61	55	338	592	508	433
Scrapie	84	15	7	118	393	353	270
Terr. A.	286	114	101	785	1429	1144	983
Mic. Div.	48	9	6	84	269	212	154

Problems with the NPMLE

under heterogeneity:

$$p_j(\lambda) = \int_0^\infty \exp(-t)t^j/j! \; \lambda(t)dt$$

- **nonidentifiability** of the population size under arbitrary mixing
- **boundary problem**

Example by Link (2003) on lack of identifiability

under binomial mixture:

$$p_j(\lambda) = \int_0^1 \binom{4}{j} t^j (1-t)^{4-j} \, \lambda(t) dt$$

$j = 0, 1, 2, 3, 4.$

two mixing distributions:

- uniform $\lambda(t) \sim U(a, b)$ with $a = 0.026$ and $b = 0.80$
- discrete two-component mixture $0.576421 \times \delta_{0.286245} + 0.423579 \times \delta_{0.676474}$

the following table from Link (2003)

Table: *untruncated and truncated count distributions*

model	probability	count j				
		0	1	2	3	4
uniform	p_j	0.227	0.255	0.243	0.190	0.085
	$p_j/(1-p_0)$	-	0.329	0.315	0.246	0.110
2 pt. mixture	p_j	0.154	0.279	0.266	0.208	0.093
	$p_j/(1-p_0)$	-	0.329	0.315	0.246	0.110

Consequences of lack of identifiability

- suppose $n = 100$ observed
- using uniform: $\hat{N} = n/0.227 = 440$
- using 2 point mixture: $\hat{N} = n/0.154 = 650$
- very **different values**, but both distributions are indistinguishable as truncated, observable distributions

Problems with the NPMLE

under heterogeneity:

$$p_j(\lambda) = \int_0^\infty \exp(-t)t^j/j!\lambda(t)dt$$

estimation under heterogeneity: the NPMLE

maximize zero-truncated Poisson mixture likelihood in Q

$$L(Q) = \prod_{j=1}^{m} \left(\frac{p_j}{1-p_0}\right)^{f_j} = \prod_{j=1}^{m} \left(\sum_{\ell=1}^{k} \frac{Po(j|t_\ell)\lambda_\ell}{1-\sum_i \exp(-t_i)\lambda_i}\right)^{f_j}$$

where

$$Q = \begin{pmatrix} t_1 & t_2 & \dots & t_k \\ \lambda_1 & \lambda_2 & \dots & \lambda_k \end{pmatrix}$$

Problems with the NPMLE

boundary problem:

$$f(0|\hat{Q}) \geq f_0/N$$

where

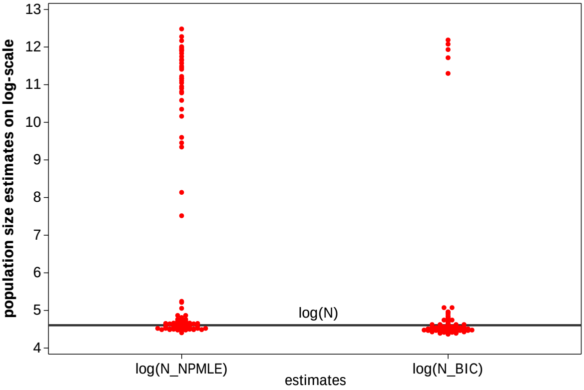
$$f(0|\hat{Q}) = \sum_{\ell} \exp(-t_\ell)\lambda_\ell$$

(Wang and Lindsay 2005, 2008; Harris 1991)

Illustration of Severity of Boundary Problem

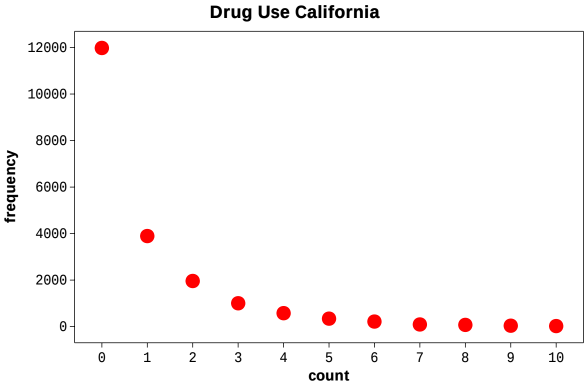
Table: *Simulation using* $Y \sim 0.5Po(1) + 0.5Po(t)$ *and* $N = 100$

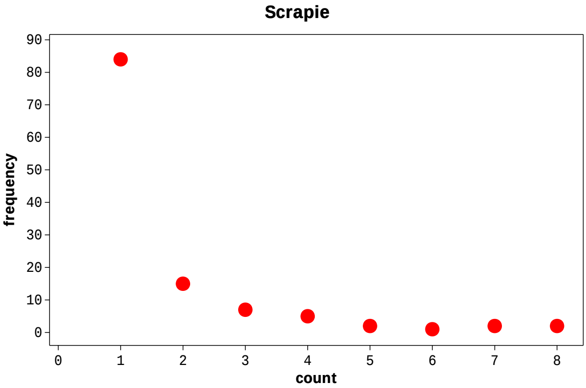
t	estimator	mean	SD
1	Chao	102	17
	NPMLE	484	12098
2	Chao	99	12
	NPMLE	4599	35028
3	Chao	97	10
	NPMLE	12517	52425
4	Chao	97	9
	NPMLE	11715	54501
5	Chao	98	10
	NPMLE	4657	33069

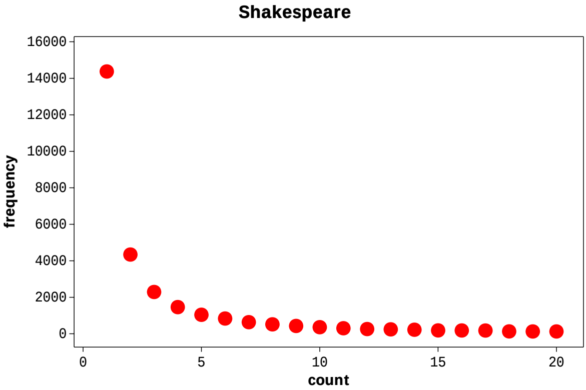

population size estimates on log-scale
13
12
11
10
9
8
7
6
5
4
log(N)
log(N_NPMLE)
log(N_BIC)
estimates

Where do we go from here?

looking at frequency distribution does **not** help!


Drug Use California
frequency
12000
10000
8000
6000
4000
2000
0
0
1
2
3
4
5
6
7
8
9
10
count


Scrapie
frequency
count
0
10
20
30
40
50
60
70
80
90
0
1
2
3
4
5
6
7
8


Shakespeare
frequency
16000
14000
12000
10000
8000
6000
4000
2000
0
0
5
10
15
20
count

Where do we go from here?

looking at ratios of neighboring frequencies **does** help:

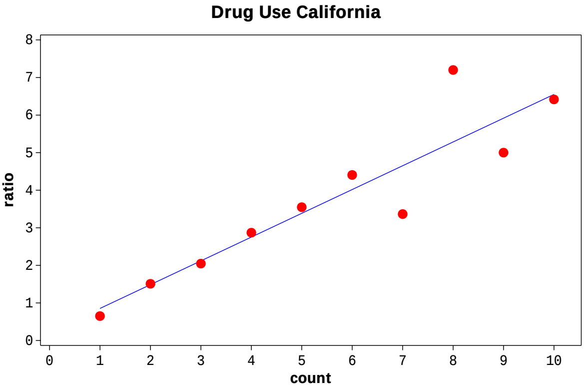
ratio plot

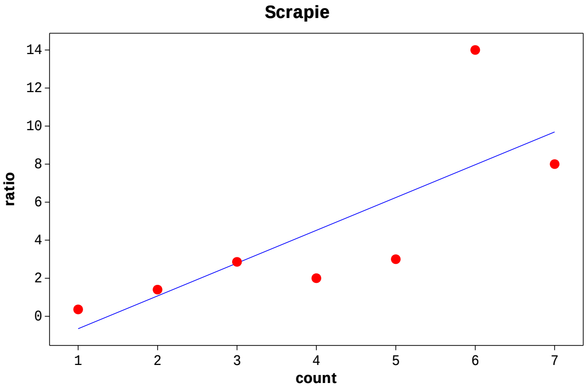
$$y \longrightarrow r_y = (y+1)\frac{f_{y+1}}{f_y}$$

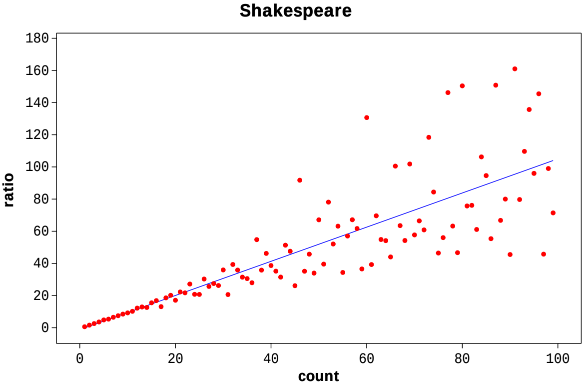
because

$$y \longrightarrow (y+1)\frac{p_{y+1}}{p_y}$$

is monotone nondecreasing


Drug Use California
ratio
count
0
1
2
3
4
5
6
7
8
9
10


Scrapie
ratio
count
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7


Shakespeare
ratio
180
160
140
120
100
80
60
40
20
0
0
20
40
60
80
100
count

Benefits

looking at ratios of neighboring frequencies is beneficial because

- no identifiability problem since $\frac{p_{j+1}}{p_j} = \frac{p_{j+1}/(1-p_0)}{p_j/(1-p_0)}$
- no boundary problem involved

Justification by means of Empirical Bayes

conventional Horvitz-Thompson

$$\hat{N} = \frac{n}{1 - \exp(-\lambda)}$$

better (each unit gets its own parameter):

$$\hat{N} = \frac{f_1}{1 - \exp(-\lambda_1)} + \frac{f_2}{1 - \exp(-\lambda_2)} + \frac{f_3}{1 - \exp(-\lambda_3)} + \ldots$$

$$= \sum_{x=1}^{n} \frac{1}{1 - \exp(-\lambda_x)}$$

but: how to choose or estimate λ_x for $x = 1, 2, 3, \ldots$?

Justification by means of Empirical Bayes

We think of the mixing distribution $\lambda(t)$ as a prior distribution on t so that

$$\lambda_x = E(t|x) = \int_0^\infty t \frac{Po(x|t)\lambda(t)}{\int_0^\infty Po(x|\theta)\lambda(\theta)d\theta} dt \qquad (1)$$

is the *posterior mean* w.r.t the prior $\lambda(t)$ and Poisson likelihood for observation x.

Note that (1) can be further simplified to

$$\lambda_x = E(t|x) = \frac{\int_0^\infty tPo(x|t)\lambda(t)dt}{\int_0^\infty Po(x|t)\lambda(t)dt}$$

$$= \frac{\int_0^\infty te^{-t}t^x/x!\lambda(t)dt}{\int_0^\infty e^{-t}t^x/x!\lambda(t)dt}$$

$$(x+1)\frac{\int_0^\infty Po(x+1|t)\lambda(t)dt}{\int_0^\infty Po(x|t)\lambda(t)dt}$$

$$= (x+1)\frac{p_{x+1}}{p_x},$$

where $p_x = \int_0^\infty Po(x|t)\lambda(t)d(t)$ is the **marginal density** of X

An empirical Bayes version of the Horvitz-Thompson estimator

choice of λ_x:

$$\lambda_x = E(t|x) = (x+1)\frac{p_{x+1}}{p_x}$$

to achieve

$$\hat{N} = \sum_{x=1}^{m} \frac{f_x}{1 - \exp[-\lambda_x]} = \sum_{x=1}^{m} \frac{f_x}{1 - \exp[-(x+1)p_{x+1}/p_x]}$$

with $p_x = \int_0^\infty Po(x|t)\lambda(t)dt$

empirical Bayes:

p_x can be estimated by the relative, empirical frequency f_x/N so that

$$\widehat{E(t|x)} = \hat{\lambda}_x = (x+1)\frac{f_{x+1}}{f_x}$$

provides an estimate of the posterior mean $E(t|x) = \lambda_x$

important:

- the unknown denominators N cancel out
- idea is a special case of the nonparametric, empirical Bayes estimator (Robbins 1955, Carlin and Louis 1996).

Robbins approach:

hence, using

$$\widehat{E(\lambda|x)} = \hat{\lambda}_x = (x+1)\frac{f_{x+1}}{f_x}$$

the empirical Bayes approach (Robbins) leads to

$$\hat{N} = \sum_{x=1}^{m} \frac{f_x}{1 - \exp[-(x+1)\frac{f_{x+1}}{f_x}]}$$

Empirical Bayesian Smoothing

$$\hat{N} = \sum_{x=1}^{m} \frac{f_x}{1 - \exp[-(x+1)\frac{p_{x+1}}{p_x}]}$$

with

$$p_x = \int_0^\infty Po(x|t)\lambda(t)dt$$

offers **options:**

1. Robbins
2. nonparametric smoothing with discrete mixture model
3. parametric smoothing with Gamma-mixing distribution
4. nonparametric smoothing with empirical distribution function

$$\hat{p}_x = \sum_{y=1}^{m} Po(x|y)\frac{f_y}{n}$$

Empirical Bayesian Smoothing

1. Robbins (no need for estimating $\lambda(t)$!!!)
2. nonparametric smoothing with discrete mixture model (computational expensive!)
3. parametric smoothing with Gamma-mixing distribution (computational instable)
4. nonparametric smoothing with empirical distribution function (not a good estimate of the mixing distribution)

Software Inspection

Table: Zero-truncated count distribution of software errors

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
-	5	1	5	1	3	2	0	5	4	2

Table: *Estimate* $\hat{N}$

conventional				empirical Bayes			
Chao	k	FM	BIC	FM	Robbins	$\Gamma(t)$	EDF
49	1	36	244.1	36	50	37	37
	2	38	**211.4**	37			
	3	124,279	215.2	40			
	4	84,946	219.7	40			

FM = finite mixture, k = number of components in FM, $\Gamma(t)$ = Gamma density

Drug Use in California 1989

f_0	f_1	f_2	f_3	f_4	f_5	f_6
-	11,982	3,893	1,959	1,002	575	340

f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	n
214	90	72	36	21	14	20,198

Drug Use in California 1989

Table: *Estimate* $\hat{N}$

conventional				empirical Bayes			
Chao	k	FM	BIC	FM	Robb.	$\Gamma(t)$	EDF
38,637	1	26,426	57,944	26,426	34,776	35,572	26,434
	2	39,183	52,262	33,757			
	3	58,224	**52,083**	34,756			
	4	424,168	52,085	34,766			

FM = finite mixture, k = number of components in FM, $\Gamma(t)$ = Gamma density

Hidden Scrapie in Great Britain

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	n
-	84	15	7	5	2	1	2	2	118

Table: *Estimate* $\hat{N}$

conventional				empirical Bayes			
Chao	k	FM	BIC	FM	Robb.	$\Gamma(t)$	EDF
353	1	170	313.9	170	320	313	164
	2	274	**260.0**	310			
	3	1,111	263.2	320			

FM = finite mixture, k = number of components in FM, $\Gamma(t)$ = Gamma density

Table: *Simulation using* $Y \sim 0.5Po(1) + 0.5Po(t)$ *and* $N = 100$

t	estimator	mean	SD	RMSE
1	Chao	102	17.3	17.4
	EB-FM	101	12.5	12.5
	EB-Robbins	107	23.1	24.2
2	Chao	99	11.7	11.7
	EB-FM	95	8.1	9.5
	EB-Robbins	100	12.2	12.2
3	Chao	97	10.6	11.0
	EB-FM	92	7.3	10.8
	EB-Robbins	97	9.1	9.6
4	Chao	97	9.9	10.3
	EB-FM	91	7.0	11.4
	EB-Robbins	95	8.3	9.7
5	Chao	98	10.4	10.6
	EB-FM	93	7.8	10.5
	EB-Robbins	96	8.7	9.6

Conclusions

- application of conventional mixture models for CR is problematic
- inference based upon ratios offers benefits
- Horvitz-Thompson estimator can be corrected and generalized for nonparametric mixture models count specific parameters can be estimated via posterior means
- using as priors estimated mixture models
- finally, **a simple solution is a beautiful solution**: the nonparametric empirical Bayes estimator

$$\hat{N} = \sum_{x=1}^{m} \frac{f_x}{1 - \exp[-(x+1)\frac{f_{x+1}}{f_x}]}$$

where to find things:

- paper available soon on this: Böhning, Kuhnert, and Del Rio Vilas (2009)
- Software by Kuhnert (2009): CR_Smooth
- references, publications, talks:
- www.reading.ac.uk/∼sns05dab