The Zelterman Estimate of Population Size under Heterogeneity

Dankmar Böhning

Quantitative Biology and Applied Statistics, School of Biological Sciences
University of Reading

November 19, 2007

Formulation of the Problem

- a population has N units of which n are identified by some mechanism (trap, register, police database, ...)
- probability of identifying an unit is $(1 - p_0)$
- so that $N = (1 - p_0)N + p_0 N = n + p_0 N$
- and the *Horvitz-Thompson* estimator follows:

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

Formulation of the Problem

- $\hat{N} = \frac{n}{1-p_0}$ is fine
-
- BUT: p_0 is assumed to be known
-
- usually an estimate of p_0 is required

Formulation of the Problem as Frequencies of Frequencies

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

- the identifying mechanism provides a count Y of repeated identifications (w.r.t. to a reference period), but zero counts are **not** observed
- leading to frequencies $f_1, f_2, ..., f_m$ where m is the largest observed count
- and f_j is the frequency of units with exactly j counts

Formulation of the Problem as Frequencies of Frequencies

we have:

- f_0 **is not observed**
- Recall that $N = f_0 + n = f_0 + f_1 + f_2 + \dots + f_m$, so that $\hat{f}_0$ leads to $\hat{N}$

McKendrick's Data on Cholera in India

McKendrick (1926) had the following frequency of households with j cases of Cholera in an Indian village:

f_0	f_1	f_2	f_3	f_4	n
-	32	16	6	1	55

How many households f_0 are affected by the epidemic, but have no cases?

Oremus' Capture-Recapture Data on Spinner Dolphins

- Oremus (2005) estimated the size of a small community of spinner dolphins around Moorea Island (Tahiti) in 2002
- the following repeated identifications were done in a 8-months period

f_0	f_1	f_2	f_3	n
-	42	7	2	51

What is the **size** of the community ?

Mathew's Data on Estimating the Dystrophin Density in the Human Muscle

- Cullen et al. (1990) attempted to locate dystrophin, a gene product of possible importance in muscular dystrophies, within the muscle fibres of biopsy specimens taken from normal patients
- Units (epitops) of Dystrophin cannot be detected by the electron microscope until they have been labelled by a suitable electron-dense substance; technique used gold-conjugated antibodies which adhere to the dystrophin

Mathew's Data on Estimating the Dystrophin Density in the Human Muscle

- not all units are labelled and it is important to account for all labelled and unlabelled units to achieve an unbiased estimate of the dystrophin density
- more than one anti-body molecule may attach to a dystrophin unit; observed then is a count variable Y counting the number of antibody molecules on each dystrophin unit
- $Y = 0$ means that unit is unlabelled and **not observed**

Mathews's Data on Estimating the Dystrophin Density in the Human Muscle

the frequency distribution of the **antibody count attached to dystrophin unit**:

f_0	f_1	f_2	f_3	f_4	f_5	n
-	122	50	18	4	4	198

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

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

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:

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

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 and arrive at:

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

Formulation of the Problem and the Idea for its 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

Formulation of the Problem and the Idea for its Solution

instead of providing an estimate $\hat{\lambda}(t)$ by means of **nonparametric mixture models** (Böhning and Schön 2005, *JRSSC*) interest is on **two alternatives**:

1. lower bound approach by Chao (1987, 1989, *Biometrics*)
   - monotonicity of the ratios of consecutive mixed Poisson probabilities
   - diagnostic device for presence of a mixed Poisson
2. robust approach of Zelterman (1988, *JSPI*)
   - likelihood framework for the Zelterman estimate
   - variance via Fisher information and covariate modelling via logistic regression (Böhning and Del Rio Vilas 2008, *JABES*)
   - bias investigation

Chao's Lower Bound Estimate

Poisson mixture for $j = 0, 1, 2, ...$

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

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

$$E(XY)^2 \leq E(X^2)E(Y^2)$$

where

$$X = \sqrt{\exp(-t)} \text{ and } Y = \sqrt{\exp(-t)}t,$$

and expected values are w.r.t. $\lambda(t)$:

$$\left(\int_0^\infty \exp(-t)t\lambda(t)dt\right)^2 \leq \int_0^\infty \exp(-t)\lambda(t)dt \int_0^\infty \exp(-t)t^2\lambda(t)dt$$

Chao's Lower Bound Estimate

$$\left(\int_0^\infty \exp(-t)t\lambda(t)dt\right)^2 \le \int_0^\infty \exp(-t)\lambda(t)dt \times 2\int_0^\infty \exp(-t)\frac{t^2}{2}\lambda(t)dt$$

$$p_1^2 \le p_0 2p_2$$

$$\Leftrightarrow \frac{p_1^2}{2p_2} \le p_0$$

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

$$\hat{f}_0 = \frac{f_1^2}{2f_2}$$

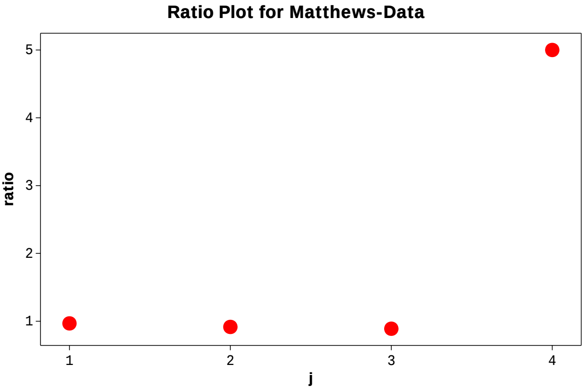
A Monotonicity Property

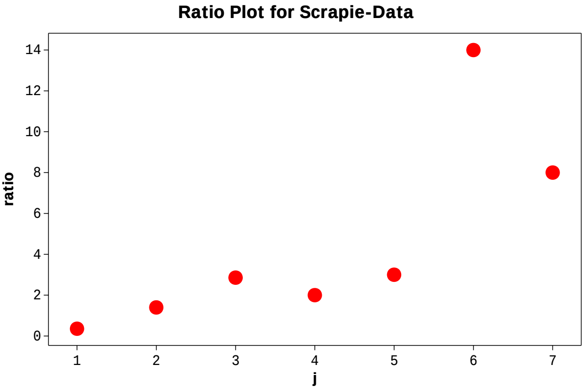
- Cauchy-Schwartz **more generally** applicable
- use $X = \sqrt{\exp(-t)t^{j-1}}$ and $Y = \sqrt{\exp(-t)t^{j+1}}$):
- $$E(XY)^2 = \left(\int_0^\infty \exp(-t)t^j \lambda(t)dt\right)^2$$
  $$\leq E(X^2)E(Y^2) = \int_0^\infty \exp(-t)t^{j-1}\lambda(t)dt \int_0^\infty \exp(-t)t^{j+1}\lambda(t)dt$$
- $$(j!p_j)^2 \leq (j-1)!p_{j-1}(j+1)!p_{j+1}$$
- $$\Leftrightarrow j\frac{p_j}{p_{j-1}} \leq (j+1)\frac{p_{j+1}}{p_j}$$
- says: ratios of consecutive mixed Poissons are **monotone**

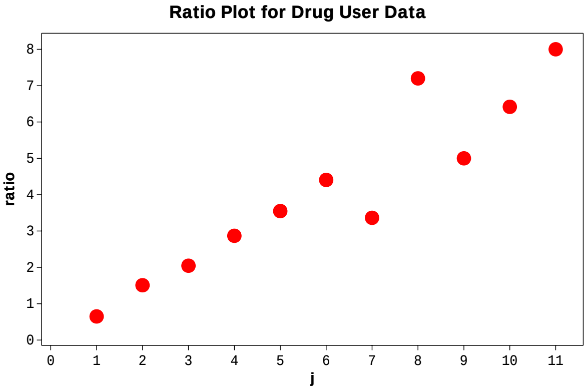
Application of the Monotonicity Property: Ratio Plot

- plot $j\frac{p_j}{p_{j-1}}$ against j for $j = 1, 2, ...$
- replace p_j by observed frequency f_j so that:
- plot $j\frac{f_j}{f_{j-1}}$ against j for $j = 2, 3, ..., m - 1$
- **monotonicity** indicative for a mixture (heterogeneity)

- conceptually related to the Poisson plot (Hoaglin 1980 *American Statistician*, Gart 1970, Rao 1971)
- difference: Poisson plot looks for a **horizontal line**, the ratio plot looks for a **monotone increasing** pattern


Ratio Plot for Matthews-Data
ratio
j
1
2
3
4
5


Ratio Plot for Scrapie-Data
ratio
0
2
4
6
8
10
12
14
1
2
3
4
5
6
7
j


Ratio Plot for Drug User Data
ratio
j
0
1
2
3
4
5
6
7
8
9
10
11

Conclusions from the Ratio Plot

- frequently, we find in count data sets **evidence** for heterogeneity in form of a mixture
- concept applicable for **both**: zero-truncated and untruncated count data (normalizing constant cancels out!)

The Idea of Zelterman (1988)

- 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 $j = 1$

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

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

Bias for Zelterman in a 2-component mixture model

assume that

$$p_j = (1 - p)Po(j; \lambda) + pPo(j; \mu)$$

for $j = 0, 1, 2, ...$

bias of

$$\hat{N} = \frac{n}{1 - \hat{p}_0}$$

is determined by bias in $\hat{p}_0$

frequently: evidence for a 2-component mixture model

Example	Non-parametric mixture model
McKendrick	homogeneity
Dolphins	homogeneity
Matthews	**2**-component
Scrapie	**2**-component
Drug Use L.A.	**3**-component

Bias for Zelterman in a 2-component mixture model

- for Zelterman:

$$\hat{p}_0 = \exp(-\hat{\lambda}) = \exp(-2\frac{f_2}{f_1})$$

- replacing frequencies by expected values

$$E(\hat{p}_0) \approx \exp(-2\frac{p_2}{p_1})$$

- **bias** of $\hat{p}_0$

$$E(\hat{p}_0) - p_0 \approx \exp(-2\frac{p_2}{p_1}) - [(1-p)e^{-\lambda} + pe^{-\mu}]$$

Bias for Zelterman in a 2-component mixture model

- **bias** of $\hat{p}_0$

$$\exp\left(-2\frac{p_2}{p_1}\right) - [(1-p)e^{-\lambda} + pe^{-\mu}]$$

$$= \exp\left(-\frac{(1-p)\lambda^2 e^{-\lambda} + p\mu^2 e^{-\mu}}{(1-p)\lambda e^{-\lambda} + p\mu e^{-\mu}}\right) - [(1-p)e^{-\lambda} + pe^{-\mu}]$$

- $$\rightarrow_{\mu\rightarrow\infty} e^{-\lambda} - (1-p)e^{-\lambda} = pe^{-\lambda}$$

- **small amount of contamination = small bias**

Bias for Zelterman in a 2-component mixture model

- **bias** of $\hat{p}_0$ for large μ

$$pe^{-\lambda} > 0$$

- Zelterman **overestimates** (upper bound)
- **small** amount of contamination = **small** bias

For Comparison: Bias of simple MLE in a 2-component mixture model

- **bias** of $\hat{p}_0 = \exp(-\bar{Y})$

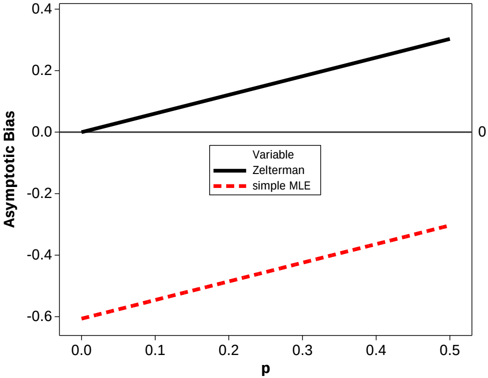
$$e^{-[(1-p)\lambda+p\mu]} - [(1-p)e^{-\lambda} + pe^{-\mu}]$$

- ≤ 0 by Jensen's inequality
- so that simple homogeneity model **underestimates** for all mixture models
- and for large μ

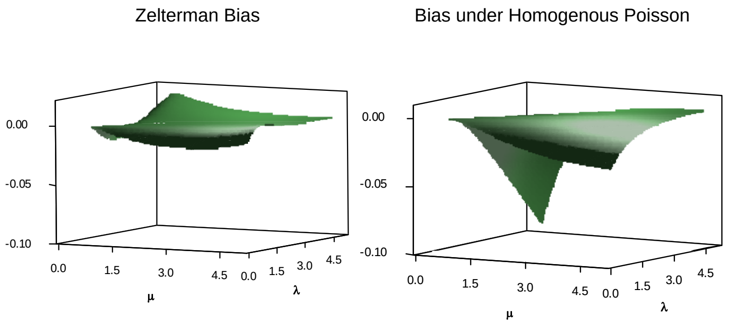
$$\textbf{bias of } \hat{p}_0 = -(1-p)e^{-\lambda}$$

- **small** amount of contamination = **large** bias

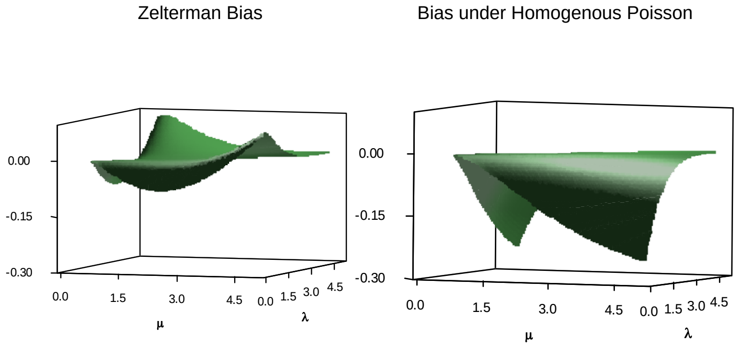
- next graph shows:
- $$\textbf{bias} \text{ of Zelterman } \hat{p}_0 = pe^{-\lambda}$$
- $$\textbf{bias} \text{ of simple MLE } \hat{p}_0 = -(1-p)e^{-\lambda}$$


0.4
0.2
0.0
-0.2
-0.4
-0.6
0
Asymptotic Bias
Variable
Zelterman
simple MLE
0.0
0.1
0.2
0.3
0.4
0.5
p

- next graphs show:
- exact **bias** of Zelterman $\hat{p}_0$
$$\exp\left(-\frac{(1-p)\lambda^2 e^{-\lambda} + p\mu^2 e^{-\mu}}{(1-p)\lambda e^{-\lambda} + p\mu e^{-\mu}}\right) - [(1-p)e^{-\lambda} + pe^{-\mu}]$$
- exact **bias** of simple MLE $\hat{p}_0$
$$e^{-[(1-p)\lambda + p\mu]} - [(1-p)e^{-\lambda} + pe^{-\mu}]$$


Zelterman Bias
Bias under Homogenous Poisson
0.00
-0.05
-0.10
0.0
1.5
3.0
4.5
μ
λ

p = 0.05


Zelterman Bias
0.00
-0.15
-0.30
0.0
1.5
3.0
4.5
μ
0.0
1.5
3.0
4.5
λ
Bias under Homogenous Poisson
0.00
-0.15
-0.30
0.0
1.5
3.0
4.5
μ
0.0
1.5
3.0
4.5
λ

p = 0.5

Population Size Estimates for the Examples

Example	n	simple MLE	Chao	Zelterman
McKendrick	55	87	87	87
Dolphins	51	157	177	180
Matthews	198	315	347	354
Scrapie	118	188	353	393
Drug Use L.A.	20,198	26,425	38,637	42,268

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 and Brittain 2007)

Zelterman larger than Chao?

Example	n	Chao	Zelterman	$\frac{f_2}{f_1}$	$\frac{n}{f_1+f_2}$
McKendrick	55	87	87	0.51	1.15
Dolphins	51	177	180	0.17	1.04
Matthews	198	347	354	0.41	1.15
Scrapie	118	353	393	0.18	1.19
Drug Use L.A.	20,198	38,637	42,268	0.33	1.27

Zelterman and non-parametric mixture

Example	n	Chao	Zelterman	NPMLE of mixture
McKendrick	55	87	87	88 (1)
Dolphins	51	177	180	149 (1)
Matthews	198	347	354	361 (2)
Scrapie	118	353	393	375 (2)
Drug Use L.A.	20,198	38,637	42,268	39,173 (2)
				56,836 (3)

Zelterman Estimation offers Flexibility

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

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

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

and consider associated **binomial** log-likelihood

$$f_1 \log(p_1) + f_2 \log(p_2)$$

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

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

Zelterman Estimation offers Flexibility

a likelihood framework offers generalizations:

- (correct) variance estimate of the Zelterman estimator (Fisher information) (Böhning 2008,*Statistical Methodology*)
- extension of the estimator for **case data**
- incorporation of **covariates** (binomial logistic regression with log-link function to the Poisson parameter) (Böhning and van der Heijden 2008)
- efficiency