School of Mathematical
and Physical Sciences

University of Reading

Department of Mathematics and Statistics

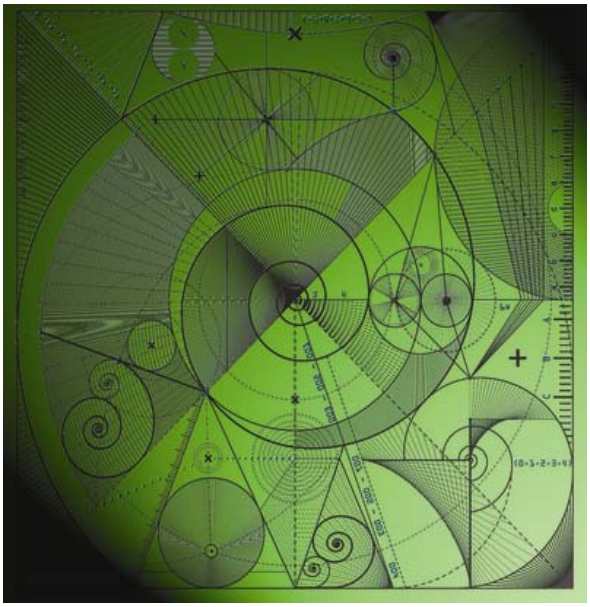
Preprint MPS-2011-05

18 April 2011

Capture-Recapture Estimation Based Upon the Zero-Truncated Exponentially Mixed Poisson

by

Sa-aat Niwitpong, Dankmar Böhning, Peter G.M. van der Heijden and Heinz Holling



Capture–Recapture Estimation Based Upon the Zero-Truncated Exponentially Mixed Poisson

Sa-aat Niwitpong*
Department of Applied Statistics, Faculty of Applied Science
King Mongut's University of Technology North–Bangkok, Thailand
email: snw@kmutnb.ac.th

Dankmar Böhning†
Department of Mathematics and Statistics
School of Mathematical and Physical Sciences
University of Reading, Reading, UK
email: d.a.w.bohning@reading.ac.uk

Peter G.M. van der Heijden
Department of Methodology and Statistics
Faculty of Social and Behavioral Sciences
Utrecht University, Utrecht, The Netherlands
email: p.g.m.vanderheijden@uu.nl

Heinz Holling
Statistics and Quantitative Methods
Faculty of Psychology and Sports Science
University of Münster, Münster, Germany
email: holling@psy.uni-muenster.de

April 18, 2011

*The paper was written while the first author was visiting the Department of Mathematics and Statistics at the University of Reading in the spring 2011 and would like to thank the department for any support that was received.

†The idea for this paper was developed while the second author was visiting the Department of Applied Statistics at the King Mongut's University North–Bangkok in the summers 2009 and 2010 and would like to thank the department for any support that was received.

Contents

1 Introduction and Background **4**

2 Maximum Likelihood Estimation **8**

3 Chao's Estimator Revisited **8**

4 An Estimator under Censoring **10**

5 Simulation Study **12**

5.1 Design 12

5.2 Results 12

6 Discussion **17**

Abstract

This note discusses the idea of using a censored likelihood to develop an improved capture-recapture estimator when heterogeneity can be validly described by an exponential mixture. Capture–Recapture methods aim to estimate the size of an elusive target population. Each member of the target population carries a count of identifications – the number of times it has been identified during the observational period. Only positive counts are observed and inference needs to be based on the observed count distribution. A widely used assumption for the count distribution is a Poisson mixture. If the mixing distribution can be described by an exponential density, the geometric distribution arises as the marginal. We use this result to show and exploit a number of beneficial properties. The zero-truncated geometric is a geometric distribution itself with support on the positive integers and the maximum likelihood estimator is available in closed–form. Since the maximum likelihood estimator is sensitive to model misspecification alternative estimators are considered including a version of Chao's estimator adapted and developed for the truncated geometric likelihood. Chao's estimator developed here gives a lower bound estimator which is valid under arbitrary mixing on the parameter of the geometric. However, Chao's estimator is also known for its relatively large variance (if compared to the maximum likelihood estimator), due to the fact that it only uses limited information stemming from counts of ones and twos only. Another estimator based on a censored geometric likelihood is suggested which uses the entire sample information but only for counts larger than 1 in a censored manner. The motivation behind this approach is the idea that violations of the geometric model assumption can be expected to be less influential than for the uncensored geometric likelihood. Simulation studies illustrate that the proposed censored estimator comprises a good compromise between the maximum likelihood estimator and Chao's estimator, e.g. between efficiency and bias.

Some key words: capture-recapture, Chao's estimator, censored estimator, censored likelihood, estimation under model misspecification, truncated likelihood, Mantel-Haenszel estimator

1 Introduction and Background

For integer N, we consider a sample of counts Y_1, Y_2, ..., $Y_N \in \{0, 1, 2, ..., \}$ arising with a mixture probability density function

$$g_y = \int_0^\infty p(y|\lambda)q(\lambda)d\lambda \tag{1}$$

where the mixing density $q(\lambda) = \frac{1}{\theta}\exp(-\frac{\lambda}{\theta})$ is exponential with parameter θ and the mixture kernel $p(y|\lambda)$ comes from the Poisson family $p(y|\lambda) = Po(y|\lambda) = \exp(-\lambda)\lambda^y/y!$. Whenever $Y_i = 0$ unit i remains unobserved, so that only a zero-truncated sample of size $n = \sum_{y=1}^m f_y$ is observed, where f_y is the frequency of counts with value $Y = y$ and m is the largest observed count. Hence, f_0 and consequently $N = \sum_{y=0}^m f_y$ are unknown. The purpose is to find an estimate of the size N. Since frequently the count variable Y represents repeated identifications of an individual in an observational period, the problem at hand is a special form of the capture-recapture problem (see Bunge and Fitzpatrick [2], Wilson and Collins [19] or Chao *et al.* [5] for a review on the topic).

The sample of counts Y_1, Y_2, ..., Y_N can occur in several ways. A target population which might be difficult to count consists out of N units. This population might be a wildlife population, a population of homeless people or drug addicts, software errors or animals with a specific disease. Furthermore, let an identification device (a trap, a register, a screening test) be available that identifies unit i at occasion t where $t = 1, .., T$ and T being potentially random itself. Let the binary result be y_{it} where $y_{it} = 1$ means that unit i has been identified at occasion t and $y_{it} = 0$ means that unit i has not been identified at occasion t. The indicators y_{it} might be observed or not, but it is assumed that $y_i = \sum_{t=1}^T y_{it}$ is observed if at least one $y_{it} > 0$ for $t = 1, ..., T$. Only if $y_{i1} = y_{i2} = ... = y_{iT} = 0$ and, consequently $y_i = 0$, the unit i remains *unobserved*. In this kind of situation the *clustering* occurs by repeated identifications of the same unit, the latter being the cluster.

Example. Before we go on, we illustrate the situation at hand with an example. In the social sciences capture–recapture methods are often employed to estimate the size of target populations which are difficult to enumerate because of their elusive character (van der Heijden *et al.* [9], Roberts and Brewer [16]). One example area is family violence which is largely a hidden activity (Palusci *et al.* [14], Oosterlee *et al.* [13]). Another area of interest is determining the size of a population with addiction problems. Hay and Smit [7] provide data on drug user contacts to a Scottish needle exchange programme in 1997. The system provided a record of the number of individuals accessing the service over the period from January to December 1997. The number of visited drug users over this 12 months was 647 and the frequency distribution of the number of times contacting a treatment centre is provided in Table 1.

Table 1: Frequency of contacts per drug user of Scottish needle exchange in 1997 for $n = 647$ observed drug users

y	1	2	3	4	5	6	7	8	9	10	11+
f_y	175	85	50	47	37	38	32	16	17	17	133

The assumption of exponential mixing in (1) is attractive since it is a more general assumption than the conventional, homogeneous Poisson assumption. In addition, under exponential mixing the integral can easily solved so that for $y = 0, 1, ...$

$$g_y = \int_0^\infty p(y|\lambda)q(\lambda)d\lambda = (1-p)^y p \tag{2}$$

the *geometric* as the associated marginal arises, with parameter $p = 1/(1+\theta) \in (0, 1)$. The geometric distribution is a remarkably simple distribution and is popular in life time data analysis as a discrete survival distribution, although, despite its flexibility, has been widely ignored for modelling count distributions.

Example (continued). The geometric has the characteristic that $g_{y+1}/g_y =$

$(1-p)$, in other words the ratio of neighboring geometric probabilities is constant. An estimate of g_{y+1}/g_y is given by f_{y+1}/f_y which we see plotted in dependence of y for the data of the Scottish needle exchange program in Figure 2. There appears to be evidence of a fairly constant pattern.

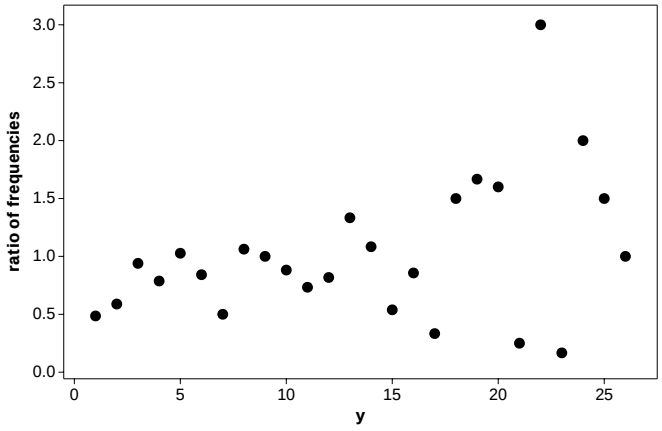


Figure 1: Ratio f_{y+1}/f_y of neighboring frequencies for the data of the Scottish needle exchange program

We also see in Figure 2 that the geometric distribution provides a much better fit than the Poisson distribution although the fit of the geometric is not perfect. It is exactly this situation for which the following estimators, in particular an estimator we call the *censored* estimator, are intended. The paper is organized

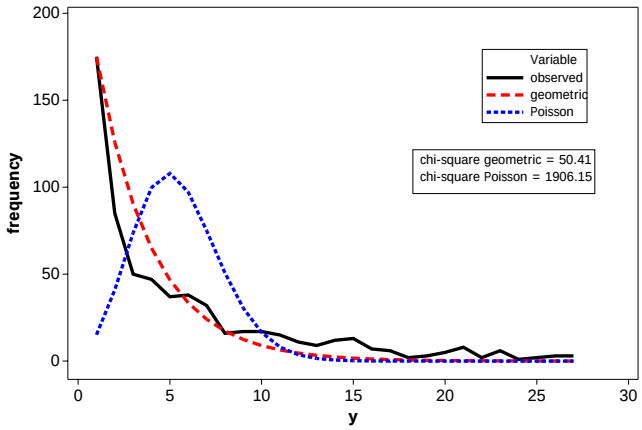


Figure 2: Observed frequencies with fitted frequencies under Poisson and geometric for the data of the Scottish needle exchange program

as follows. In section 2 we consider classical maximum likelihood estimation for the zero-truncated geometric including a form of Mantel-Haenszel estimation. In section 3, we develop Chao-estimation based upon a specific form of truncated likelihood. This estimator is appropriate for strong heterogeneity, but has the disadvantage of a large variance. In section 4 we develop an estimator that uses all available information but censors counts larger than 1. Finally, in section 5 we compare all estimators and demonstrate that the censored estimator is appropriate for mild or moderate forms of heterogeneity.

2 Maximum Likelihood Estimation

We first consider conventional maximum likelihood estimation. For $y = 1, 2, ..,$ let $g_y^+ = g_y/(1-p) = (1-p)^{y-1}p$ be the associated zero-truncated geometric. Then the log–likelihood is given as

$$\log L(p) = \sum_{y=1}^{m}(y-1)f_y \log(1-p) + n\log(p) = S\log(1-p) + n(\log p - \log(1-p)), \tag{3}$$

where $S = \sum_{y=1} yf_y$. It is easy to verify that (3) leads to the score–equation

$$\frac{n}{p} = \frac{S-n}{1-p},$$

which is uniquely solved for $\hat{p}_{ML} = n/S$. Since $e_0 = E(f_0|p) = Np = (e_0+n)p$ we have that $e_0 = np/(1-p)$, so that $\hat{e}_0 = n\hat{p}_{ML}/(1-\hat{p}_{ML})$ and $\hat{N}_{ML} = n + e_0 = n/(1-\hat{p}_{ML})$. Note that $\hat{N}_{ML}$ can be simply written as

$$\hat{N}_{ML} = \frac{n}{1-n/S} = \frac{nS}{S-n}.$$

Since $g_{y+1}/g_y = 1-p$ it is intuitively reasonable to consider a weighted estimator of the form $\sum_{y=1}^{m-1} w_y f_{y+1}/f_y$. With $w_y = f_y$ we get the Mantel-Haenszel estimator

$$1 - \hat{p}_{MH} = \frac{\sum_{y=1}^{m-1} f_{y+1}}{\sum_{y=1}^{m-1} f_y} = \frac{n-f_1}{n-f_m}, \tag{4}$$

which, with $\hat{N}_{MH} = n/(1-\hat{p}_{MH}) = n(n-f_m)/(n-f_1)$, will not only be less affected by zero frequencies, but also is expected to behave more robust towards misspecification of the geometric than the maximum likelihood estimator.

3 Chao's Estimator Revisited

Clearly, the geometric model might not hold for the entire target population. Hence it seems more appropriate to consider additional heterogeneity

$$\int_0^1 g_y(p)q(p)dp = \int_0^1 (1-p)^y p\ q(p)dp \tag{5}$$

The importance of the mixture (5) can be seen in the fact that it is a natural model for modeling population heterogeneity. There appears to be consensus (see for example Pledger [15] for the discrete mixture model approach and Dorazio and Royle [6] for the continuous mixture model approach) that a simple model $g_y(p)$ is not flexible enough to capture the variation in the re-capture probability for the different members of most real life populations. Every item might be different, as might be every animal or human being. However, recently there has been also a debate on the identifiability of the binomial mixture model (see Link [11], [12] and Holzmann *et al.* [10]). Furthermore, using the nonparametric maximum likelihood estimate (NPMLE) of the mixing density in constructing an estimate of the population size leads to the *boundary problem* implying often unrealistically high values for the estimate of the population site (Wang and Lindsay [17], Wang and Lindsay [18]). Hence, a renewed interest has re-occurred in the lower bound approach for population size estimation suggested by Chao [3]. In the lower bound approach there is neither need to specify a mixing distribution, nor is there need to estimate it. In this sense it is completely non-parametric. To give some details on the lower bound approach recall that for two random variables U and V we have the Cauchy-Schwarz inequality $E(UV)^2 = E(U^2)E(V^2)$. Now, choose $U = (1-p)\sqrt{p}$ and $V = \sqrt{p}$, then

$$E(UV)^2 = \left(\int_0^1 (1-p)pq(p)dp\right)^2 \leq \int_0^1 (1-p)^2 pq(p)dp \int_0^1 pq(p)dp = E(U^2)E(V^2).$$

Now, the LHS can be estimated by f_1^2/N^2, whereas the RHS can be estimated by $(f_0/N)(f_2/N)$ from where Chao's lower bound estimator $f_0 = f_1^2/f_2$ follows. In total, we have that

$$\hat{N}_C = n + f_1^2/f_2.$$

We note that this lower bound estimator is specific for the geometric mixture kernel in (5) and differs from the original lower bound estimator $n + f_1^2/(2f_2)$

which was developed for the Poisson mixture kernel and is clearly too small for the situation considered here.

It is interesting to see that a truncated likelihood approach yields Chao's estimator. Since the Chao estimator uses only frequencies with counts of 1 and 2, a truncated sample *consisting only out of counts of ones and twos* might be considered. We call this the *binomial truncated* sample. The associated truncated Poisson probabilities are

$$q_1 = \frac{(1-p)p}{(1-p)p + (1-p)p^2} = 1/(2-p) \text{ and } q_2 = (1-p)/(2-p).$$

This truncated sample leads to a binomial log-likelihood $f_1 \log(q_1) + f_2 \log(q_2)$ which is uniquely maximized for $\hat{q}_2 = 1 - \hat{q}_1 = f_2/(f_1 + f_2)$. Since $q_2 = (1-p)/(2-p)$ the estimate $\hat{p} = (f_1 - f_2)/f_1$ for the geometric density parameter p arises. We show in the appendix that under binomial truncated sampling $e_0 = E(f_0|p; f_1, f_2) = \frac{f_1+f_2}{(1-p)(2-p)}$ which leads to the estimated value

$$\hat{e}_0 = \frac{f_1 + f_2}{(1-\hat{p})(2-\hat{p})} = \frac{f_1 + f_2}{(1 - \frac{f_1-f_2}{f_1})(2 - \frac{f_1-f_2}{f_1})} = \frac{f_1 + f_2}{\frac{f_2}{f_1}\frac{2f_1-f_1+f_2}{f_1}} = \frac{f_1^2}{f_2}.$$

From here Chao's estimator $N_C = n + f_1^2/f_2$ follows. Note that the likelihood framework into which we have embedded the Chao estimator offers potential. For example, we can derive easily asymptotic variance formula and also extend the estimator with respect to covariates.

4 An Estimator under Censoring

One of the critical points in Chao's estimator is that it disregards the information contributed from counts larger than two. A compromise between retaining robustness as well as efficiency appears to be an approach based upon *censoring* which we try to develop here. Occasionally, we find the hint in the literature that members of the target population which have been identified only once behave quite differently from members of the target population which have been

identified more frequently. Hence also from this, more substantial aspect the approach appears justified. Consider the conventional zero-truncated geometric

$$g_y^+ = \frac{p(1-p)^y}{1-p} = p(1-p)^{y-1},$$

for $y = 1, 2, ...$ Then, if we consider all observations larger than 1 to be censored, $P(Y = 1) = g_1^+ = p$ and $P(Y > 1) = \sum_{y=2}^{\infty} g_y^+ = 1 - p$, using the log-likelihood $f_1 \log p + (n - f_1) \log(1 - p)$. The maximum likelihood estimate for p is simply $\hat{p}_{Cen} = f_1/n$. Here, it is easy to work out $e_0 = E(f_0|p) = Ng_0 = (e_0 + n)p$, from where $e_0 = np/(1 - p)$ follows. Hence we have $\hat{e}_0 = n\frac{f_1/n}{1-f_1/n}$ and

$$\hat{N}_{Cen} = n + \frac{f_1}{1 - f_1/n} = \frac{n}{1 - f_1/n} = \frac{n^2}{n - f_1}$$

follows. Note the close similarity to the Mantel-Haenszel estimator $\hat{N}_{MH} = n(n - f_m)/(n - f_1)$ with identity for $f_m = 0$. Hence we expect that $\hat{N}_{Cen}$ and $\hat{N}_{MH}$ to be close since typically f_m will be small (often only equal to 1).

Example (continued). Before we continue comparing and evaluating these estimators more systematically on empirical grounds we illustrate their numerical behavior for the data of the Scottish needle exchange program. We had seen before that the geometric provides a reasonable, but not perfect fit to the data. Hence we expect that there is residual heterogeneity so that the maximum likelihood estimator can be expected to underestimate. Indeed, $\hat{N}_{ML} = 750$ whereas $\hat{N}_{Cen} = 887$ and $\hat{N}_C = 1007$ showing again the compromising character of the censored estimator between bias and efficiency. Note that the conventional estimator of Chao under Poisson heterogeneity is $\hat{N} = 827$ indicating that the classical Chao estimator is not flexible enough to cope with this form of heterogeneity.

5 Simulation Study

To illustrate the performance of the estimators a simulation study was undertaken. Since we show in the appendix that, under geometric homogeneity, all estimators are asymptotically unbiased, the focus of the simulation will be on scenarios where the model is misspecified.

5.1 Design

A number of scenarios were investigated. Initially, the case was considered that the geometric density is the true model. This is the situation under which all estimators were derived. Secondly, a contamination model $(1-\alpha)g_y(p)+\alpha g_y(q)$ was considered with $\alpha = 0.1$ (small amount of contamination) and with $\alpha = 0.5$ (large amount of contamination). We also study as a continuous heterogeneity distribution the beta-distribution with density

$$b(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1},$$

so that sampling arises from the marginal

$$\int_0^1 g_y(p)\ b(p|\alpha,\beta)\ dp.$$

The forms of the beta-density we have considered are provided in Figure 3.

5.2 Results

Table 2 and Table 3 presents the results in terms of mean, standard error of estimate and root mean squared error for the maximum likelihood estimator, Chao's lower bound estimator adapted to to the geometric case, and the proposed censored estimator. We are not presenting any results for the Mantel-Haenszel estimator since they are almost identical to the censored case. Table 2 provides

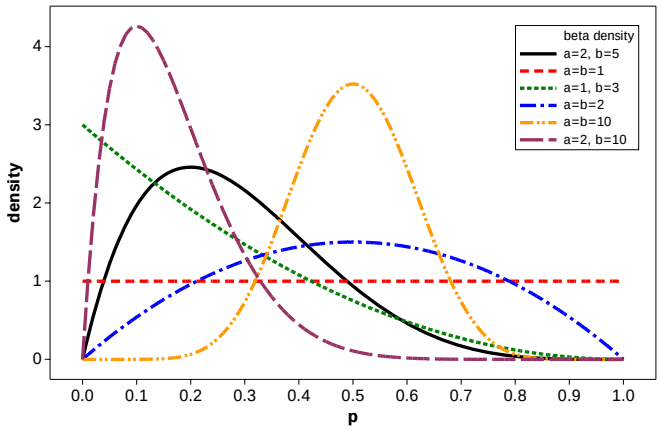


Figure 3: Some beta-densities characterized by parameters α and β to model heterogeneity in the parameter p of the geometric

results for $N = 1000$ whereas Table 3 shows results for $N = 100$. We summarize a few major results:

- under geometric homogeneity all three estimators are asymptotically unbiased (this is also proved in the appendix as Theorem 2, so that the simulation part referring to this situation (populations 1-4) serves only as illustration,
- the efficiency of the censored estimator ranges typically between 80%–90% whereas Chao's estimator varies between 40%–50% in its efficiency,
- for cases of mild heterogeneity, such as for populations 5–12, 15, 16, 21 and 22, the censored estimator behaves well. It has still a small bias and its variance is close to the variance of the maximum likelihood estimator,

- for cases of stronger heterogeneity, such as populations 13, 14 and 17–20, the bias is reasonably small (except populations 17 and 19) and well balanced by a small standard error,
- if focus is on achieving an estimator with small bias, then the choice should be Chao's estimator which has smallest bias for all populations with heterogeneity.

In summary, the simulation study confirms and provides evidence for the hypothesis that the censored estimator is a reasonable compromise between maximum likelihood estimation and Chao's lower bound estimator.

Table 2: Performance measures for the MLE, Chao and Censored estimator in the case $N = 1000$

		$E(\hat{N})$			SE			RMSE		
Population	model	MLE	Chao	Cens	MLE	Chao	Cens	MLE	Chao	Cens
		homogeneity: geometric $G(p)$								
1	$p = 0.1$	1000.08	1002.18	1000.00	11.24	26.79	15.48	11.24	26.88	15.48
2	$p = 0.3$	1000.84	1003.36	1000.70	24.68	51.85	32.20	24.69	51.96	32.21
3	$p = 0.5$	1003.43	1007.75	1003.80	45.05	82.10	54.80	45.18	82.46	54.94
4	$p = 0.7$	1007.71	1014.25	1007.43	91.62	140.08	104.17	91.95	140.80	104.43
		heterogeneity: $0.9G(0.1) + 0.1G(q)$								
5	$q = 0.2$	994.16	1002.45	999.17	11.64	28.14	16.19	13.03	28.24	16.21
6	$q = 0.3$	984.01	1001.00	995.37	11.91	28.84	16.34	19.93	28.85	16.98
7	$q = 0.4$	972.67	999.07	989.51	12.36	31.07	17.48	29.99	31.09	20.38
8	$q = 0.5$	961.37	994.70	981.54	12.64	32.56	17.82	40.58	32.99	25.66
		heterogeneity: $0.9G(0.3) + 0.1G(q)$								
9	$q = 0.6$	970.51	993.78	983.17	25.51	54.36	33.45	38.99	54.71	37.45
10	$q = 0.7$	954.79	982.01	969.03	25.62	55.45	33.82	51.96	58.29	45.85
		heterogeneity: $0.5G(0.1) + 0.5G(q)$								
11	$q = 0.2$	978.25	1002.88	997.11	14.05	34.31	19.78	25.89	34.43	19.99
12	$q = 0.3$	931.69	1000.57	984.99	15.96	41.08	23.04	70.14	41.09	27.49
13	$q = 0.4$	875.11	995.93	961.76	16.86	47.89	25.53	126.02	48.07	45.97
14	$q = 0.5$	814.46	986.49	925.38	16.97	55.44	27.42	186.30	57.06	79.49
		heterogeneity: $0.5G(0.3) + 0.5G(q)$								
15	$q = 0.4$	984.22	1001.51	994.31	28.31	58.83	36.67	32.41	58.85	37.11
16	$q = 0.5$	939.43	994.12	974.73	30.28	64.34	39.76	67.71	64.61	47.11
		heterogeneity: $\int_0^1 G(p)b(p\|\alpha,\beta)dp$								
17	$\alpha = 1, \beta = 1$	536.50	840.04	750.79	22.71	68.41	32.92	464.05	173.97	251.37
18	$\alpha = 1, \beta = 3$	775.34	977.90	937.46	17.86	46.95	24.60	225.36	51.89	67.20
19	$\alpha = 2, \beta = 2$	669.82	905.43	834.08	31.40	73.71	39.91	331.66	119.90	170.65
20	$\alpha = 2, \beta = 5$	835.15	986.08	952.95	21.62	51.87	28.49	166.25	53.70	55.00
21	$\alpha = 2, \beta = 10$	910.45	999.53	985.34	15.04	35.88	20.09	90.80	35.88	24.87
22	$\alpha = \beta = 10$	909.89	980.79	955.18	40.33	80.29	51.06	98.72	82.55	67.93

Table 3: Performance measures for the MLE, Chao and Censored estimator in the case $N = 100$

		$E(\hat{N})$			SE			RMSE		
Population	model	MLE	Chao	Cens	MLE	Chao	Cens	MLE	Chao	Cens
		homogeneity: geometric $G(p)$								
1	$p = 0.1$	100.14	102.92	100.09	3.53	12.04	4.92	3.53	12.39	4.92
2	$p = 0.3$	100.55	103.99	100.55	8.00	19.86	10.48	8.02	20.25	10.49
3	$p = 0.5$	101.91	106.77	101.76	14.75	36.88	18.31	14.87	37.50	18.39
4	$p = 0.7$	110.52	120.15	110.10	41.74	72.18	45.04	43.04	74.94	46.16
		heterogeneity: $0.9G(0.1) + 0.1G(q)$								
5	$q = 0.2$	99.51	102.89	99.97	3.71	11.91	5.12	3.74	12.26	5.12
6	$q = 0.3$	98.51	103.03	99.70	3.82	12.61	5.34	4.10	12.97	5.35
7	$q = 0.4$	97.38	102.67	99.03	3.83	13.44	5.45	4.63	13.70	5.53
8	$q = 0.5$	96.12	102.44	98.08	3.94	15.09	5.61	5.53	15.29	5.93
		heterogeneity: $0.9G(0.3) + 0.1G(q)$								
9	$q = 0.6$	97.65	103.43	98.97	8.15	20.84	10.86	8.48	21.12	10.91
10	$q = 0.7$	96.04	102.59	97.55	8.19	21.43	11.06	9.09	21.59	11.32
		heterogeneity: $0.5G(0.1) + 0.5G(q)$								
11	$q = 0.2$	97.96	103.29	99.77	4.42	13.57	6.16	4.87	13.91	6.16
12	$q = 0.3$	93.35	103.11	98.66	5.14	15.49	7.32	8.40	15.80	7.44
13	$q = 0.4$	87.69	103.29	96.44	5.37	8.26	8.31	13.42	19.79	9.04
14	$q = 0.5$	81.63	103.44	92.78	5.51	23.77	8.83	19.17	24.01	11.40
		heterogeneity: $0.5G(0.3) + 0.5G(q)$								
15	$q = 0.4$	99.33	104.21	100.37	9.19	21.89	11.96	9.21	22.29	11.96
16	$q = 0.5$	95.09	104.98	98.74	10.05	25.29	13.17	11.19	25.78	13.23
		heterogeneity: $\int_0^1 G(p)b(p\|\alpha,\beta)dp$								
17	$\alpha = 1, \beta = 1$	56.07	91.37	75.74	6.62	32.13	10.95	44.42	33.26	26.61
18	$\alpha = 1, \beta = 3$	79.03	102.14	94.01	5.06	20.28	7.79	21.56	20.40	9.82
19	$\alpha = 2, \beta = 2$	69.08	97.37	84.39	9.30	30.97	13.49	32.28	31.09	20.62
20	$\alpha = 2, \beta = 5$	84.54	102.57	95.86	6.49	19.56	9.07	16.76	19.73	9.97
21	$\alpha = 2, \beta = 10$	91.50	103.00	98.59	4.60	15.48	6.53	9.66	15.77	6.68
22	$\alpha = \beta = 10$	92.86	105.15	97.45	13.15	34.13	16.79	14.96	34.52	16.98

6 Discussion

We have tried in section 5 to compare the suggested estimators by means of a simulation study. There is one problem which arises in any comparison involving biased estimators. Recall that we are considering in the simulation study tow types of misspecified models: in one model the geometric parameter is sampled from a tow-component mixture and in the other model it sampled from a beta-distribution. Under these two models all three estimators are asymptotically biased. Whereas with increasing sample size the bias stabilizes and persists, the standard error decreases. Hence, with increasing sample size, the mean squared error will be dominated by the bias and the evaluation, if done solely on the basis of the mean squared error, will ultimately favor the estimator with the smallest bias. This point is best illustrated using the example in Table 4 where we consider the ratio $\hat{N}/N$. It is clear that from Table 4 that asymptotically Chao's estimator will perform best, since it has the smallest asymptotic bias and the standard error (of $\hat{N}/N$, not of $\hat{N}$) converging to zero.

Table 4: Mean and standard error of $\hat{N}/N$ for increasing N for the geometric parameter p coming from a 2-component mixture giving equal weight to $p = 0.3$ and $q = 0.5$

	$E(\hat{N}/N)$			$SE(\hat{N}/N)$		
N	MLE	Chao	Cen	MLE	Chao	Cen
100	0.95	1.04	0.98	0.10	0.24	0.13
1,000	0.94	0.99	0.97	0.03	0.06	0.04
10,000	0.94	0.99	0.97	0.01	0.02	0.01

As a consequence, one should either limit oneself to realistic values of the population size if using the mean squared error (as we have done here) or, for asymptotic considerations, choose a performance measure different from the MSE.

Simulation studies are an important tool to evaluate a series of estimators. However, they also have their limitations since they can only mirror a reality envisioned in the design of the study with natural restrictions in complexity. Hence it is of interest to study the proposed estimators in data sets where the population size is known in advance. Borchers *et al.* (2004) report the following capture–recapture experiment in St. Andrews. $N = 250$ groups of golf tees were placed in a survey region of $1,680\ m^2$. They were then surveyed by eight different students of the University of St. Andrews and $n = 162$ were identified. Typically, an unknown number of golf tees would be missed, but here we know that exactly 88 golf tees remained missed. The data are provided in Table 5.

Table 5: Frequency of recovery counts in golf-tees experiment (true $N = 250$) with associated estimators of N

y	1	2	3	4	5	6	7	8
f_y	46	28	21	13	23	14	6	11

estimator of N				
geometric			Poisson	
MLE	Chao	Cens	Chao	Turing
230	226	238	200	177

The estimators under geometric sampling are fairly similar and close to the true number $N = 250$. Note that Chao's estimator (adjusting for heterogeneity) is close to the maximum likelihood estimator indicating that the exponential mixing is coping well with any heterogeneity in the data. We have also computed two estimators under Poisson sampling: the Chao estimator $n + f_1^2/(2f_2)$ and the Turing estimator $n/(1 - f_1/S)$, both being too small and also different from each other. This means that there is residual heterogeneity under Poisson sampling which evidently the geometric estimators can pick up and adjust for.

The geometric (and mixtures of geometrics to adjust for heterogeneity) appears to be an interesting alternative to the Poisson (and mixtures thereof).

We have presented two estimators, Chao's estimator and the censored estimator, which appear to work well under geometric heterogeneity. Frequently, the geometric provides a better initial fit than then Poisson and hence can be expected to cope with some of the potentially available heterogeneity. It is also technically easy to deal with. However, diagnostic devices such as the suggested ratio plot $y \to f_{y+1}/f_y$ or goodness-of-fit measures should also be used to check for the appropriateness of the approach.

References

[1] Borchers, D.L., Buckland, S.T., and Zucchini, W. (2004). *Estimating Animal Abundance. Closed Populations.* London, Springer.

[2] Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association* **88**, 364–373.

[3] Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**, 783–791.

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

[5] Chao A., Tsay P.K., Lin S.H, Shau W.Y, Chao D.Y. (2001). Tutorial in Biostatistics: The applications of capture-recapture models to epidemiological data. *Statistics in Medicine* **20**, 3123–3157.

[6] Dorazio, R.M. and Royle, J.A. (2005). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics* **59**, 351–364.

[7] Hay, G. and Smit, F. (2003). Estimating the number of drug injectors from needle exchange data. *Addiction Research and Theory*, **11** 235–243.

[8] Van Hest, N.A.H., De Vries, G., Smit, F., Grant, A.D., and Richardus, J.H. (2008). Estimating the coverage of Tuberculosis screening among drug users and homeless persons with truncated models. *Epidemiology and Infection* **136**, 14–22..

[9] Van der Heijden, P. G. M., 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.

[10] Holzmann, H., Munk, A., and Zucchini, W. (2003). On identifiability in capture-recapture models. *Biometrics* **62**, 934–939.

[11] Link, W.A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* **59**, 1123–1130.

[12] Link, W.A. (2003). Response to a paper by Holzmann, Munk and Zucchini. *Biometrics* **62**, 936–939.

[13] Oosterlee, A., Vink, R.M., Smit, F. (2009). Prevalence of family violence in adults and children: estimates using the capture–recapture method. *European Journal of Public Health* **19**, 586-591.

[14] Paluscia, V.J., Wirtz, S.J., and Covington, T.M. (2010). Using capture-recapture methods to better ascertain the incidence of fatal child maltreatment. *Child Abuse & Neglect* **34**, 396402.

[15] Pledger, S. A. (2005). The performance of mixture models in heterogeneous closed population capture-recapture. *Biometrics* **61**, 868–876.

[16] Roberts, J.M. and Brewer, D.D. (2006). Estimating the prevalence of male clients of prostitute women in Vancouver with a simple capture-recapture method. *Journal of the Royal Statistical Society (Series A)* **169**, 745–756.

[17] Wang, J.-P. and Lindsay, B.G. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association* **100**, 942–959.

[18] Wang, J.-P. and Lindsay, B.G. (2008). An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Statistical Methodology* **5**, 30–45.

[19] Wilson, R.M. and Collins, M.F. (1992). Capture-recapture estimation with samples of size one using frequency data. *Biometrika* **79**, 543–553.

Appendix: Proof of Theorems

Theorem 1 *a) Let* $\log L(p) = f_1 \log(q_1) + f_2 \log(q_2)$ *with* $q_1 = 1/(2-p)$ *and* $q_2 = (1-p)/(2-p)$ *being the geometric probabilities truncated to counts of ones and twos. Then* $\log L(p)$ *is maximized for* $\hat{p} = (f_1 - f_2)/f_1$.

b) $E(f_0|f_1, f_2; \hat{p}) = f_1^2/f_2$, *for* $\hat{p} = (f_1 - f_2)/f_1$.

Proof. For the first part, it is clear that $f_1 \log(q_1) + f_2 \log(q_2)$ is maximal for $\hat{q}_1 = f_1/(f_1 + f_2) = 1/(2 - \hat{p})$, which is attained for $\hat{p} = (f_1 - f_2)/f_1$. For the second part, we see that with $e_y = E(f_y|f_1, f_2; p) = g_y(p)N$ we have the following:

$$e_y = g_y(p)N = g_y(p)(e_0 + f_1 + f_2 + \sum_{j=3}^{\infty} e_j)$$

so that

$$e_0 + e_3^+ = [1 - g_1(p) - g_2(p)](e_0 + e_3^+) + [1 - g_1(p) - g_2(p)](f_1 + f_2)$$

with $e_3^+ = \sum_{j=3}^{\infty} e_j$. Hence

$$e_0 + e_3^+ = \frac{1 - g_1(p) - g_2(p)}{g_1(p) + g_2(p)}(f_1 + f_2)$$

and

$$e_0 = g_0(p)(f_1 + f_2 + e_0 + e_3^+) \quad = g_0(p)(f_1 + f_2)[1 + \tfrac{1-g_1(p)-g_2(p)}{g_1(p)+g_2(p)}]$$
$$= \tfrac{g_0(p)}{g_1(p)+g_2(p)}(f_1 + f_2) = \tfrac{f_1+f_2}{(1-p)(2-p)}.$$

Plugging in the maximum likelihood estimate $\hat{p} = (f_1 - f_2)/f_1$ for p yields

$$\frac{f_1 + f_2}{(1-\hat{p})(2-\hat{p})} = \frac{f_1 + f_2}{\frac{f_2}{f_1}\frac{f_1+f_2}{f_1}} = f_1^2/f_2,$$

the desired result. □

Theorem 2 *Let $g_y(p) = (1-p)^y p$ for $y = 0, 1, ..$ and $p \in (0, 1)$. Then,*

$$\lim_{N\to\infty} \frac{E(\hat{N})}{N} = 1$$

for $\hat{N} = \hat{N}_{ML}, \hat{N}_C,$ or $\hat{N}_{Cen}$.

Proof. Let $\hat{N} = \hat{N}_{ML} = n/(1 - n/S)$. Note that $E(n) = Np$ and $E(S/N) = (1-p)/p$ so that

$$\frac{E(n/(1-n/S))}{N} \xrightarrow[N\to\infty]{} \frac{p}{1 - \frac{p}{p/(1-p)}} = 1.$$

Let $\hat{N} = \hat{N}_C = n + f_1^2/f_2$. Note that $E(f_1) = Np(1-p)$ and $E(f_2) = Np(1-p)^2$ so that

$$\frac{E(n + f_1^2/f_2)}{N} \xrightarrow[N\to\infty]{} (1-p) + \frac{p^2(1-p)^2}{p(1-p)^2} = 1.$$

Finally, let $\hat{N} = \hat{N}_{Cen} = \frac{n}{1-f_1/n}$. Using the above we have

$$\frac{E\left(\frac{n}{1-f_1/n}\right)}{N} \xrightarrow[N\to\infty]{} \frac{1-p}{1 - \frac{(1-p)p}{(1-p)}} = 1,$$

which ends the proof. □