# *Capture–recapture estimation based upon the geometric distribution allowing for heterogeneity*

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

**Metrika**
International Journal for Theoretical and Applied Statistics

ISSN 0026-1335
Volume 76
Number 4

Metrika (2013) 76:495-519
DOI 10.1007/s00184-012-0401-0


Springer

Metrika (2013) 76:495–519
DOI 10.1007/s00184-012-0401-0

# Capture–recapture estimation based upon the geometric distribution allowing for heterogeneity

**Sa-aat Niwitpong · Dankmar Böhning · Peter G. M. van der Heijden · Heinz Holling**

Received: 25 October 2011 / Published online: 27 July 2012


**Abstract** Capture–Recapture methods aim to estimate the size of an elusive target population. Each member of the target population carries a count of identifications by some identifying mechanism—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

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

S. Niwitpong
Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology, North-Bangkok, Thailand
e-mail: snw@kmutnb.ac.th

D. Böhning (✉)
Southampton Statistical Sciences Research Institute and School of Mathematics,
University of Southampton, Southampton, UK
e-mail: d.a.bohning@soton.ac.uk

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

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

by an exponential density, the geometric distribution arises as the marginal. This note discusses population size estimation on the basis of the zero-truncated geometric (a geometric again itself). In addition, population heterogeneity is considered for the geometric. Chao's estimator is developed for the mixture of geometric distributions and provides 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). Another estimator based on a censored geometric likelihood is suggested which uses the entire sample information but is less affected by model misspecifications. 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.

**Keywords** Capture-recapture · Chao's estimator · Censored estimator · Censored likelihood · Estimation under model misspecification · Truncated likelihood

## 1 Introduction and background

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

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

where the mixture kernel $p(y|\lambda)$ comes from the Poisson family $p(y|\lambda) = Po(y|\lambda) = \exp(-\lambda)\lambda^y/y!$ and the mixing density $q(\lambda)$ is left unspecified. 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 1993; Wilson and Collins 1992 or Chao et al. 2001 for a review on the topic).

The sample of counts $Y_1, Y_2, \ldots, Y_N$ 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 $t = 1, \ldots, T$ and $T$ being potentially unknown and/or 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 $t = 1, \ldots, T$. Only if $y_{i1} = y_{i2} = \cdots = y_{iT} = 0$ 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. It is clear that,

**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 |

assuming independence and conditional on $T$, $Y_i$ has a binomial distribution with potentially unit-specific parameter $\pi_i$

$$\binom{T}{y_i}\pi_i^{y_i}(1-\pi_i)^{T-y_i} \approx \exp(-\lambda_i)\lambda_i^{y_i}/y_i!$$

which can be approximated by a Poisson with parameter $\lambda_i = T\pi_i$. Clearly, this approximation is most appropriate for small detection probabilities. If $\lambda_i$ is assumed to arise from a distribution with density $q(\lambda)$ the mixture (1) occurs.

*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. 2003; Roberts and Brewer 2006). One example area is family violence which is largely a hidden activity (Paluscia et al. 2010; Oosterlee et al. 2009). Another area of interest is determining the size of a population with addiction problems (Van Hest et al. 2008). Hay and Smit (2003) 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.

The model (1) is attractive since it incorporates population heterogeneity into the Poisson assumption. The general estimator available under this model is Chao's estimator (1987, 1989) with $\hat{N}_{\text{Chao}} = n + f_1^2/(2f_2)$ where $f_y$ is the frequency of count $y$ in the sample. However, $\hat{N}_{\text{Chao}}$ gives only an estimate of a lower bound for $N$ since it is based on the Cauchy-Schwarz inequality $[E(XY)]^2 \leq (EX^2)(EY^2)$ which produces for $X^2 = \exp(-\lambda)$ and $Y^2 = \exp(-\lambda)\lambda^2$ the result $g_1^2 \leq g_0 \times 2g_2$ from where Chao's estimator follows. If there is no variation in $\lambda$ then equality holds (and the lower bound becomes asymptotically sharp) and the more variation the stronger the underestimation.

The idea is to replace some of the (otherwise unspecified) heterogeneity density by a parametric density, potentially leaving some residual heterogeneity $q^*(\theta)$:

$$g_y = \int_0^\infty \left( \int_0^\infty p(y|\lambda)e(\lambda|\theta)d\lambda \right) q^*(\theta)d\theta,$$

where $e(\lambda|\theta) = (\frac{1}{\theta}\exp(-\frac{\lambda}{\theta})$ is the exponential density with parameter $\theta$. Under exponential mixing the integral can easily be solved so that for $y = 0, 1, \ldots$

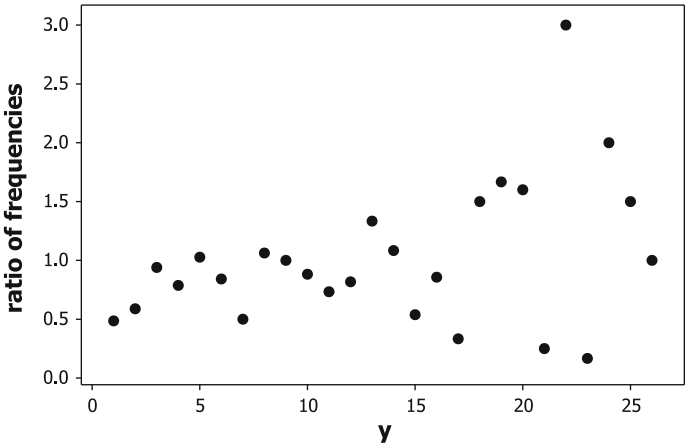


**Fig. 1** Ratio $f_{y+1}/f_y$ of neighboring frequencies for the data of the Scottish needle exchange program

$$k_y(p) = \int_0^\infty p(y|\lambda)e(\lambda|\theta)d\lambda = (1-p)^y p \tag{2}$$

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 often ignored for modelling count distributions. By incorporating some of the unobserved heterogeneity into the mixture kernel distribution it seems reasonable to expect an improvement in the lower bound based on this new form of mixture

$$g_y = \int_0^1 k_y(p)q^*(p)dp = \int_0^1 (1-p)^y p\, q^*(p)dp. \tag{3}$$

Indeed, the Cauchy-Schwarz inequality $[E(XY)]^2 \le (EX^2)(EY^2)$ produces for $X^2 = p$ and $Y^2 = p(1-p)^2$ the result $g_1^2 \le g_0 \times g_2$ from where the lower bound estimator $\hat{N}_C = n + f_1^2/f_2$ can be derived, clearly larger than the original Chao estimator $\hat{N}_{\text{Chao}} = n + f_1^2/(2f_2)$. Note that this difference stems from the fact that a geometric kernel is used in (3) intead of the Poisson kernel in (1).

*Example* (*continued*) The geometric has the characteristic that $k_{y+1}/k_y = (1-p)$, in other words the ratio of neighboring geometric probabilities is constant. Frequently, it can be seen that an exponential mixing is more appropriate than a homogeneous Poisson. 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 Fig. 1. There appears to be evidence of a fairly constant pattern indicating a small amount of residual heterogeneity w.r.t. the geometric kernel.

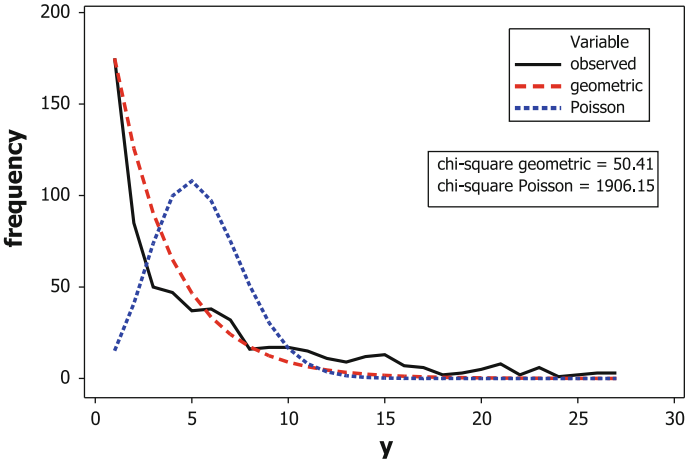


**Fig. 2** Observed frequencies with fitted frequencies under Poisson and geometric for the data of the Scottish needle exchange program

We also see in Fig. 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 as follows. In Sect. 2 we consider classical maximum likelihood estimation for the zero-truncated geometric including a form of Mantel-Haenszel estimation. In Sect. 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 Sect. 4 we develop an estimator that uses all available information but censors counts larger than 1. Finally, in Sect. 6 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 conditional maximum likelihood estimation under the Poisson-exponential mixture. For $y = 1, 2, \ldots$, let $k_y^+ = k_y/(1-p) = (1-p)^{y-1}p$ be the associated zero-truncated geometric. Then the log–likelihood, conditional upon $n$, is given as

$$\begin{aligned} \log L(p) &= \sum_{y=1}^{m} (y-1) f_y \log(1-p) + n \log(p) \\ &= S \log(1-p) + n(\log p - \log(1-p)), \end{aligned} \tag{4}$$

where $S = \sum_{y=1}^{m} y f_y$. It is easy to verify that (4) leads to the score–equation

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

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

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

Note that $\hat{N}_{ML}$ given above is usually a non-integer number which can be rounded for producing a sensible estimate of $N$.

A simple, alternative estimator arises as follows. Since $k_{y+1}/k_y = 1-p$ it is intuitively reasonable to consider a weighted estimator of the form $\left(\sum_{y=1}^{m-1} w_y f_{y+1}/f_y\right)/(\sum_{y=1}^{m-1} w_y)$. Any non-random choice of weights will give asymptotically unbiased estimators of $1-p$. Instead of searching for minimum variance estimators in this class, we consider the choice $w_y = f_y$ (a random weight) and get the Mantel-Haenszel estimator

$$1 - \hat{p}_{MH} = \frac{\sum_{y=1}^{m-1} f_{y+1}}{\sum_{y=1}^{m-1} f_y} = \frac{n-f_1}{n-f_m}, \tag{5}$$

which, with $\hat{N}_{MH} = n/(1-\hat{p}_{MH}) = n(n-f_m)/(n-f_1)$, is not only of a very simple form but also will avoid problems that might occur through zero frequencies in the ratios $f_{y+1}/f_y$ for the general weighted 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 in form of a density $q^*(p)$ on the parameter of the geometric as in (3).

The importance of the mixture (3) of geometric densities 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 (2005) for the discrete mixture model approach and Dorazio and Royle (2005) for the continuous mixture model approach) that a simple model $k_y(p)$ 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 2003, 2006; Holzmann et al. 2006). 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 2005, 2008). Hence, a renewed interest has re-occurred in the lower bound approach for population size estimation suggested by Chao (1987). By generalizing a moment inequality based upon the Cauchy-Schwarz inequality Mao (2007a,b, 2008a,b) developed in a series of papers a theory of lower bounds for the population size. This theory leads to a sequence of monotonically ordered lower bounds which include as a special case Chao's lower bound. 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.

We have previously derived Chao's estimator as $\hat{N}_C = n + f_1^2/f_2$ for a geometric mixture. It is interesting to see that a truncated, conditional 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. Recall that the geometric is given by $k_y(p) = (1-p)^y p$ for $y = 0, 1, 2, \ldots$. The associated binomially truncated geometric probabilities are

$$\pi_1 = \frac{(1-p)p}{(1-p)p + (1-p)^2 p} = 1/(2-p) \quad \text{and} \quad \pi_2 = (1-p)/(2-p).$$

This truncated sample leads to a binomial log-likelihood $f_1 \log(\pi_1) + f_2 \log(\pi_2)$ which is uniquely maximized for $\hat{\pi}_2 = 1 - \hat{\pi}_1 = f_2/(f_1 + f_2)$. Since $\pi_2 = (1-p)/(2-p)$ 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 $N_C = n + f_1^2/f_2$ for a geometric mixture follows. Note that the likelihood framework of a conditional binomial truncated likelihood 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

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

for $y = 1, 2, \ldots$. Then, if we consider all observations larger than 1 to be censored, $P(Y = 1) = k_1^+ = p$ and $P(Y > 1) = \sum_{y=2}^{\infty} k_y^+ = 1 - p$, 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 can expect that $\hat{N}_{Cen}$ and $\hat{N}_{MH}$ are close since typically $f_m$ will be small (often only equal to 1). Hence we won't consider $\hat{N}_{MH}$ any further in the following.

## 5 Standard errors of estimates

It is important to have measures of precisions available for the developed estimators. In the following we summarize the variance estimators for the three population size estimators, namely the conditional likelihood based estimator $\hat{N}_{ML} = n/(1 - n/S)$, the censored estimator $\hat{N}_{Cen} = n/(1 - f_1/n)$ and Chao's estimator for geometric mixtures $\hat{N}_C = n + f_1^2/f_2$. We have derived (for full details see Appendix 2) the following variance estimates

$$\widehat{var}(\hat{N}_{ML}) = \frac{S^2 n^2}{(S-n)^3} \tag{6}$$

$$\widehat{var}(\hat{N}_{Cen}) = \frac{f_1}{(1 - f_1/n)^2}\frac{2n - f_1}{n - f_1} \tag{7}$$

$$\widehat{var}(\hat{N}_C) = \frac{f_1^4}{f_2^3} + \frac{4f_1^3}{f_2^2} + \frac{f_1^2}{f_2}. \tag{8}$$

*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, using that $n = 647$ and $S = 7034$ we find that $\hat{N}_{ML} = 750$ (727 – 773) whereas $\hat{N}_{Cen} = 887$ (832–942) and $\hat{N}_C = 1007$ (871–1,144) showing, at least for this example, the compromising character of the censored estimator between bias and efficiency. Note that the conventional estimator of Chao under general Poisson heterogeneity is $\hat{N} = 827$ indicating the bias reduction potential of the new Chao estimator upon the classical one.

## 6 Simulation study

To illustrate the performance of the estimators a simulation study was undertaken. Since we show in the Appendix 1 that, under geometric homogeneity, all estimators are asymptotically unbiased, the focus of the simulation will be on scenarios where the model is misspecified.

### 6.1 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 $(1-\alpha)k_y(p)+\alpha k_y(q)$ 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

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

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

### 6.2 Results

Tables 2 and 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 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 results for $N = 1{,}000$ 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 1 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 and 90 % whereas Chao's estimator varies between 40 and 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,

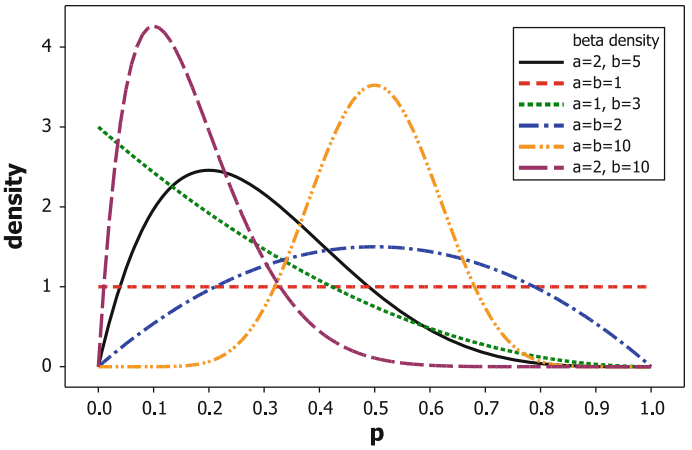


**Fig. 3** Some beta-densities characterized by parameters $\alpha$ and $\beta$ to model heterogeneity in the parameter $p$ of the geometric

- 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.

We have also investigated with the simulation study how well the variance estimators (6)–(8) approximate the true variance. The results are presented in Table 4. The approximations, expressed in the ratio $= E[\widehat{\text{Var}}(\hat{N})]/\text{Var}(\hat{N})$, work reasonably well, under homogeneity even in the small population size case $N = 100$ (Table 5).

- For $N = 1{,}000$ (Table 4) and the homogeneity case we have that $|\text{ratio} - 1| \leq 0.05$ for all three estimators. For $N = 100$ (Table 5) and the homogeneity case we have that $|\text{ratio} - 1| \leq 0.18$ for the MLE and the censored estimator, and $|\text{ratio} - 1| \leq 0.55$ for Chao's estimator.
- It is interesting that for the considered heterogeneity cases for $N = 1{,}000$ (Table 4) we also have good approximations for Chao's estimator with $|\text{ratio} - 1| \leq 0.13$ as well as for the censored estimator with $|\text{ratio} - 1| \leq 0.15$. Here, however, the variance estimator for the MLE breaks down completely (see populations 13–14 and 17–21).
- The main message is that variance estimators for the censored and Chao's estimator will work for most scenarios under reasonable population size ($N \geq 1{,}000$), whereas the variance estimator for the MLE will only work under homogeneity and can become entirely unsatisfactory in certain situations (populations 17 and 18).

**Table 2** Performance measures for the MLE, Chao and Censored estimator in the case $N = 1{,}000$

| Population | Model | $E(\hat{N})$ | | | SE | | | RMSE | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MLE | Chao | Cens | MLE | Chao | Cens | MLE | Chao | Cens |
| | | Homogeneity: geometric $G(p)$ | | | | | | | | |
| 1 | $p = 0.1$ | 1,000.08 | 1,002.18 | 1,000.00 | 11.24 | 26.79 | 15.48 | 11.24 | 26.88 | 15.48 |
| 2 | $p = 0.3$ | 1,000.84 | 1,003.36 | 1,000.70 | 24.68 | 51.85 | 32.20 | 24.69 | 51.96 | 32.21 |
| 3 | $p = 0.5$ | 1,003.43 | 1,007.75 | 1,003.80 | 45.05 | 82.10 | 54.80 | 45.18 | 82.46 | 54.94 |
| 4 | $p = 0.7$ | 1,007.71 | 1,014.25 | 1,007.43 | 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 | 1,002.45 | 999.17 | 11.64 | 28.14 | 16.19 | 13.03 | 28.24 | 16.21 |
| 6 | $q = 0.3$ | 984.01 | 1,001.00 | 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 | 1,002.88 | 997.11 | 14.05 | 34.31 | 19.78 | 25.89 | 34.43 | 19.99 |
| 12 | $q = 0.3$ | 931.69 | 1,000.57 | 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 | 1,001.51 | 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 |

**Table 2** continued

| Population | Model | $E(\hat{N})$ | | | SE | | | RMSE | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MLE | Chao | Cens | MLE | Chao | Cens | MLE | Chao | Cens |
| | | 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 $N = 100$

| Population | Model | $E(\hat{N})$ | | | SE | | | RMSE | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 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 |

## 7 Discussion

We have tried in Sect. 6 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 two types of misspecified models: in one model the geometric parameter is sampled from a two-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,

**Table 4** Comparison of the estimated and true variance for $N = 1{,}000$; ratio $= E[\widehat{\text{Var}}(\hat{N})]/\text{Var}(\hat{N})$

| Population | Model | MLE | | | Chao | | | Cens | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\text{Var}(\hat{N})$ | $E[\widehat{Var}(\hat{N})]$ | Ratio | $\text{Var}(\hat{N})$ | $E[\widehat{Var}(\hat{N})]$ | Ratio | $\text{Var}(\hat{N})$ | $E[\widehat{Var}(\hat{N})]$ | Ratio |
| | | Homogeneity: geometric $G(p)$ | | | | | | | | |
| 1 | $p = 0.1$ | 123.26 | 123.76 | 1.00 | 705.62 | 721.05 | 1.02 | 232.51 | 235.62 | 1.01 |
| 2 | $p = 0.3$ | 609.54 | 615.12 | 1.01 | 2,790.50 | 2,749.66 | 0.99 | 1,060.96 | 1,049.70 | 0.99 |
| 3 | $p = 0.5$ | 2,007.47 | 2,030.66 | 1.01 | 6,819.54 | 6,831.98 | 1.00 | 3,039.06 | 3,056.02 | 1.01 |
| 4 | $p = 0.7$ | 8,047.58 | 8,173.58 | 1.02 | 19,052.23 | 19,494.19 | 1.02 | 10,475.03 | 10,653.87 | 1.02 |
| | | Heterogeneity: $0.9G(0.1) + 0.1G(q)$ | | | | | | | | |
| 5 | $q = 0.2$ | 137.35 | 130.34 | 0.95 | 788.12 | 801.08 | 1.02 | 259.40 | 259.56 | 1.00 |
| 6 | $q = 0.3$ | 143.54 | 130.35 | 0.91 | 865.09 | 895.07 | 1.03 | 281.12 | 278.67 | 0.99 |
| 7 | $q = 0.4$ | 147.07 | 128.57 | 0.87 | 964.57 | 980.21 | 1.02 | 295.62 | 290.01 | 0.98 |
| 8 | $q = 0.5$ | 151.46 | 126.17 | 0.83 | 1,033.37 | 1,045.32 | 1.01 | 311.32 | 294.46 | 0.95 |
| | | Heterogeneity: $0.9G(0.3) + 0.1G(q)$ | | | | | | | | |
| 9 | $q = 0.6$ | 654.45 | 634.14 | 0.97 | 3,020.02 | 3,069.75 | 1.02 | 1,132.78 | 1,133.05 | 1.00 |
| 10 | $q = 0.7$ | 650.33 | 619.27 | 0.95 | 3,082.37 | 3,114.01 | 1.01 | 1,134.13 | 1,119.96 | 0.99 |
| | | Heterogeneity: $0.5G(0.1) + 0.5G(q)$ | | | | | | | | |
| 11 | $q = 0.2$ | 194.73 | 169.66 | 0.87 | 1,095.71 | 1,136.89 | 1.04 | 370.60 | 373.74 | 1.01 |
| 12 | $q = 0.3$ | 251.64 | 178.91 | 0.71 | 1,636.34 | 1,669.58 | 1.02 | 523.58 | 509.32 | 0.97 |
| 13 | $q = 0.4$ | 278.48 | 170.61 | 0.61 | 2,229.63 | 2,290.92 | 1.03 | 636.89 | 620.95 | 0.97 |
| 14 | $q = 0.5$ | 292.26 | 154.83 | 0.53 | 3,063.04 | 3,070.84 | 1.00 | 756.90 | 692.74 | 0.92 |
| | | Heterogeneity: $0.5G(0.3) + 0.5G(q)$ | | | | | | | | |
| 15 | $q = 0.4$ | 795.55 | 768.30 | 0.97 | 3,405.26 | 3,469.17 | 1.02 | 1,335.92 | 1,342.37 | 1.00 |
| 16 | $q = 0.5$ | 940.59 | 829.75 | 0.88 | 4,263.91 | 4,317.39 | 1.01 | 1,630.62 | 1,599.80 | 0.98 |

**Table 4** continued

| Population | Model | MLE | | | Chao | | | Cens | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\text{Var}(\hat{N})$ | $E[\widehat{Var}(\hat{N})]$ | Ratio | $\text{Var}(\hat{N})$ | $E[\widehat{Var}(\hat{N})]$ | Ratio | $\text{Var}(\hat{N})$ | $E[\widehat{Var}(\hat{N})]$ | Ratio |
| | | Heterogeneity: $\int G(p)b(p\|\alpha, \beta)dp$ | | | | | | | | |
| 17 | $\alpha = 1, \beta = 1$ | 547.09 | 41.94 | 0.08 | 4,825.95 | 4,657.66 | 0.97 | 1,121.04 | 948.52 | 0.85 |
| 18 | $\alpha = 1, \beta = 3$ | 292.38 | 26.94 | 0.09 | 2,354.52 | 2,234.09 | 0.95 | 582.50 | 532.07 | 0.91 |
| 19 | $\alpha = 2, \beta = 2$ | 976.90 | 314.02 | 0.32 | 5,663.43 | 5,549.54 | 0.98 | 1,595.60 | 1,510.14 | 0.95 |
| 20 | $\alpha = 2, \beta = 5$ | 454.35 | 164.84 | 0.36 | 2,301.42 | 2,597.45 | 1.13 | 714.04 | 748.76 | 1.05 |
| 21 | $\alpha = 2, \beta = 10$ | 219.82 | 90.86 | 0.41 | 1,339.52 | 1,342.58 | 1.00 | 411.52 | 392.67 | 0.95 |
| 22 | $\alpha = \beta = 10$ | 1,550.90 | 1,373.10 | 0.89 | 6,432.23 | 6,556.47 | 1.02 | 2,574.00 | 2,581.51 | 1.00 |

**Table 5** Comparison of the estimated and true variance for $N = 100$; ratio $= E[\widehat{\text{Var}}(\hat{N})]/\text{Var}(\hat{N})$

| Population | Model | MLE | | | Chao | | | Cens | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\text{Var}(\hat{N})$ | $E[\widehat{Var}(\hat{N})]$ | Ratio | $\text{Var}(\hat{N})$ | $E[\widehat{Var}(\hat{N})]$ | Ratio | $\text{Var}(\hat{N})$ | $E[\widehat{Var}(\hat{N})]$ | Ratio |
| | | Homogeneity: geometric $G(p)$ | | | | | | | | |
| 1 | $p = 0.1$ | 12.07 | 12.59 | 1.04 | 146.67 | 199.80 | 1.36 | 23.18 | 24.30 | 1.05 |
| 2 | $p = 0.3$ | 61.19 | 64.30 | 1.05 | 367.72 | 418.77 | 1.14 | 105.67 | 111.64 | 1.06 |
| 3 | $p = 0.5$ | 223.22 | 230.90 | 1.03 | 999.14 | 1,145.09 | 1.15 | 326.67 | 352.57 | 1.08 |
| 4 | $p = 0.7$ | 1,548.66 | 1,747.41 | 1.13 | 5,794.14 | 8,696.11 | 1.50 | 1,947.34 | 2,303.63 | 1.18 |
| | | Heterogeneity: $0.9G(0.1) + 0.1G(q)$ | | | | | | | | |
| 5 | $q = 0.2$ | 13.75 | 13.39 | 0.97 | 160.58 | 205.37 | 1.28 | 26.30 | 26.30 | 1.00 |
| 6 | $q = 0.3$ | 14.78 | 13.39 | 0.91 | 186.19 | 239.47 | 1.29 | 28.66 | 27.77 | 0.97 |
| 7 | $q = 0.4$ | 14.96 | 13.20 | 0.88 | 203.14 | 269.24 | 1.33 | 30.10 | 29.01 | 0.96 |
| 8 | $q = 0.5$ | 14.74 | 13.06 | 0.89 | 226.91 | 301.68 | 1.33 | 30.58 | 29.77 | 0.97 |
| | | Heterogeneity: $0.9G(0.3) + 0.1G(q)$ | | | | | | | | |
| 9 | $q = 0.6$ | 68.41 | 68.85 | 1.01 | 438.02 | 479.42 | 1.09 | 115.76 | 116.82 | 1.01 |
| 10 | $q = 0.7$ | 68.46 | 67.56 | 0.99 | 436.23 | 490.13 | 1.12 | 114.33 | 116.04 | 1.01 |
| | | Heterogeneity: $0.5G(0.1) + 0.5G(q)$ | | | | | | | | |
| 11 | $q = 0.2$ | 19.83 | 17.37 | 0.88 | 188.07 | 227.44 | 1.21 | 38.44 | 38.94 | 1.01 |
| 12 | $q = 0.3$ | 25.17 | 18.43 | 0.73 | 245.09 | 293.09 | 1.20 | 51.90 | 52.81 | 1.02 |
| 13 | $q = 0.4$ | 29.00 | 17.58 | 0.61 | 367.51 | 436.89 | 1.19 | 65.87 | 64.93 | 0.99 |
| 14 | $q = 0.5$ | 29.60 | 16.02 | 0.54 | 520.98 | 618.84 | 1.19 | 77.04 | 71.91 | 0.93 |
| | | Heterogeneity: $0.5G(0.3) + 0.5G(q)$ | | | | | | | | |
| 15 | $q = 0.4$ | 83.62 | 81.63 | 0.98 | 477.21 | 533.59 | 1.12 | 139.26 | 144.75 | 1.04 |
| 16 | $q = 0.5$ | 99.56 | 89.41 | 0.90 | 624.85 | 682.50 | 1.09 | 174.24 | 174.38 | 1.00 |

**Table 5** continued

| Population | Model | MLE | | | Chao | | | Cens | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\text{Var}(\hat{N})$ | $E[\widehat{Var}(\hat{N})]$ | Ratio | $\text{Var}(\hat{N})$ | $E[\widehat{Var}(\hat{N})]$ | Ratio | $\text{Var}(\hat{N})$ | $E[\widehat{Var}(\hat{N})]$ | Ratio |
| | | Heterogeneity: $\int G(p)b(p\|\alpha,\beta)dp$ | | | | | | | | |
| 17 | $\alpha = 1, \beta = 1$ | 44.35 | 8.70 | 0.20 | 1096.11 | 1405.61 | 1.28 | 117.51 | 103.43 | 0.88 |
| 18 | $\alpha = 1, \beta = 3$ | 25.96 | 4.59 | 0.18 | 412.89 | 475.83 | 1.15 | 61.55 | 55.36 | 0.90 |
| 19 | $\alpha = 2, \beta = 2$ | 88.28 | 39.34 | 0.45 | 965.50 | 1,143.82 | 1.18 | 173.05 | 168.17 | 0.97 |
| 20 | $\alpha = 2, \beta = 5$ | 40.36 | 18.70 | 0.46 | 406.61 | 466.35 | 1.15 | 81.49 | 78.64 | 0.97 |
| 21 | $\alpha = 2, \beta = 10$ | 21.53 | 9.99 | 0.46 | 236.70 | 284.62 | 1.20 | 41.09 | 40.96 | 1.00 |
| 22 | $\alpha = \beta = 10$ | 170.73 | 156.31 | 0.92 | 979.87 | 1,118.63 | 1.14 | 278.76 | 294.01 | 1.05 |

**Table 6** 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 $q = 0.5$

| $N$ | $E(\hat{N}/N)$ | | | $SE(\hat{N}/N)$ | | |
|---|---|---|---|---|---|---|
| | 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 |

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 6 where we consider the ratio $\hat{N}/N$. It is clear that from Table 6 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.

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.

Another issue is whether uncertainty evaluation should be based upon estimating $N$ or predicting $f_0$ as a referee pointed out. This seems equivalent at first glance since $N = n + f_0$ and, hence $\hat{N} = \hat{f_0} + n$, where $\hat{f_0}$ is the predicted value for the unobserved random variable $f_0$. However, the issue is how $n$ should be treated and this can be done in two ways. One way is, and this is most current practice, to consider $n$ as random. Then the variance of $\hat{N}$ has two sources of error, the one coming from the random variable $n$, the other from predicting $f_0$. The second way is to treat $n$ as fixed and hence there is only one source of error variance $Var(\hat{N}) = Var(\hat{f_0})$. This conditional inference seems more appropriate for capture-recapture problems. Suppose one is interested in finding the errors in a software system. A capture-recapture experiment finds 3 errors. Then the only interest is in how many more errors are hidden in the system and what is the random error attached to it. Whether the 3 observed errors are random or not is irrelevant for the prediction question. Hence we tend to agree with the second view, but still have done the performance assessment in the first way since it is commonly done and widely accepted.

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 7. 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

**Table 7** Frequency of recovery counts in golf-tees experiment (true $N = 250$) with associated estimators of $N$ (95 % CI)

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

Estimator of $N$ (95 % CI)

| Geometric | | | Poisson | |
|---|---|---|---|---|
| MLE | Chao | Cens | Chao | Turing |
| 230(207–253) | 238(183–292) | 226(198–255) | 200(180–241) | 177(170–190) |

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. Note also that none of the two has confidence intervals including the true population size. This means that there is residual heterogeneity under Poisson sampling which evidently the geometric estimators can pick up and adjust for.

The geometric is a an exponential mixture of Poisson densities and, hence, the set of geometric mixtures is a subset of the class of Poisson mixture distributions. Indeed, it is a proper subset, since no non-trivial Poisson distribution can be expressed as a mixture of geometric distributions. From this perspective, working with geometric mixtures is more restrictive than working with Poisson mixtures and leads to different inferences as can be seen, in particular, for Chao's estimator which becomes an improved lower bound when using the assumption of a mixture of geometric distribution for $Y$.

However, let us consider comparing the *homogeneous* Poisson with the *homogeneous* geometric distribution taking the geometric simply as another one-parameter count distribution. In this comparison, we feel that the geometric provides a more flexible model than the Poisson because of the fact that the geometric is a Poisson-exponential mixture, as pointed out above. On the other hand, the Poisson is not a special case of the geometric. Only if the parameter of the exponential mixing distribution becomes small, meaning mean and variance become small, the geometric becomes close to a Poisson. Hence, for situations with small detection probabilities we can expect the geometric to be provide better fits. To illustrate we consider data discussed previously in Van der Heijden et al. (2003) on the illegal possession of firearms in the Netherlands based on a police registration system. According to this $f_1 = 2561$ were caught once possessing a firearm, $f_2 = 72$ were caught twice and $f_3 = 5$ were caught 3 times. The $\chi^2-$goodness-of-fit statistic is 7.31 the Poisson and $\chi^2-$goodness-of-fit statistic is 3.42 for the geometric. Note that only the second value is non-significant ($P$-value 0.064 with 1 df). It is clear that this advantage of the geometric might be lost in situations with increased detection probabilities. But even here we see often a better fit of the geometric distribution in comparison to the Poisson. This was case in the example discussed previously (see Fig. 2), but it is also the case in a capture-recapture study on ant species mentioned in Mao (2008b). Clearly, this long-tailed distribution is fitted a lot better by the geometric than by the Poisson (see Table 8). The associated population size estimates (with 95 % CIs) are $\hat{N}_{ML} = 228$

**Table 8** Observed and fitted frequency distribution for ant species data discussed in Mao (2008b)

| $y$ | $f_y$ | Poisson | Geometric | $y$ | $f_y$ | Poisson | Geometric |
|---|---|---|---|---|---|---|---|
| 1 | 50 | 25.16 | 28.75 | 16 | 2 | 0.00 | 2.59 |
| 2 | 29 | 41.07 | 24.49 | 17 | 2 | 0.00 | 2.21 |
| 3 | 24 | 44.69 | 20.86 | 18 | 2 | 0.00 | 1.88 |
| 4 | 13 | 36.48 | 17.77 | 19 | 3 | 0.00 | 1.60 |
| 5 | 6 | 23.82 | 15.14 | 20 | 5 | 0.00 | 1.36 |
| 6 | 9 | 12.96 | 12.89 | 21 | 1 | 0.00 | 1.16 |
| 7 | 3 | 6.05 | 10.98 | 22 | 1 | 0.00 | 0.99 |
| 8 | 4 | 2.47 | 9.35 | 23 | 3 | 0.00 | 0.84 |
| 9 | 1 | 0.90 | 7.97 | 24 | 1 | 0.00 | 0.72 |
| 10 | 7 | 0.29 | 6.79 | 25 | 1 | 0.00 | 0.61 |
| 11 | 6 | 0.09 | 5.78 | 26 | 4 | 0.00 | 0.52 |
| 12 | 2 | 0.02 | 4.92 | 27 | 1 | 0.00 | 0.44 |
| 13 | 6 | 0.01 | 4.19 | 28 | 1 | 0.00 | 0.38 |
| 14 | 5 | 0.00 | 3.57 | 29 | 1 | 0.00 | 0.32 |
| 15 | 1 | 0.00 | 3.04 | | | | |

(214–241), $\hat{N}_{Cen} = 261$ (233–290) and $\hat{N}_C = 280$ (220–340) which are well in the range of estimators given in Mao (2008b).

As mentioned previously Mao (2007a,b, 2008a,b) constructs a series of lower bounds that allow improving upon the Chao's lower bound, in fact, a sharpest lower bound can be derived. However, the choice between these bounds may be not easy. As Mao (p. 132) Mao (2008b) emphasizes: *A higher-order lower bound seems desirable, but it may have a larger estimation bias and variance.* $\cdots$ *It is a difficult problem to select one estimator from the sequence* $\cdots$. We point out that the lower bound approach also depends on the choice of the mixing kernel (Poisson vs. geometric). In this context it is interesting to note that for the second example discussed in Mao (2007b) (the ESTs data) the geometric gives a much better fit than the Poisson which is also supported in the associated ratio plots. This supports the likewise importance of the choice of the mixing kernel in (1).

The geometric (and mixtures of geometric distributions) 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, ultimately diagnostic devices such as the suggested ratio plot $y \rightarrow f_{y+1}/f_y$ or goodness-of-fit measures should also be used to check for the appropriateness of the approach.

**Acknowledgments** The authors would like to thank the Editor and two anonymous referees for their very helpful comments which considerably improved the paper. We also would like to thank Jeerapa Sappakitkamjorn (Department of Applied Statistics, King Mongkut's University of Technology North–Bangkok) for her great support in finalizing the simulation study.

## 8 Appendix 1: Proof of theorems

**Theorem 1** *Let* $k_y(p) = (1-p)^y p$ *for* $y = 0, 1, \cdots$ *and* $p \in (0, 1)$.

*(a) Let* $\log L(p) = f_1 \log(\pi_1) + f_2 \log(\pi_2)$ *with* $\pi_1 = 1/(2-p)$ *and* $\pi_2 = (1-p)/(2-p)$ *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 $f_1 \log(\pi_1) + f_2 \log(\pi_2)$ is maximal for $\hat{\pi}_1 = f_1/(f_1 + f_2) = 1/(2 - \hat{p})$, which is attained for $\hat{p} = (f_1 - f_2)/f_1$. For the second part, we see that with $e_y = E(f_y|f_1, f_2; p) = k_y(p)N$ we have the following:

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

so that

$$e_0 + e_3^+ = [1 - k_1(p) - k_2(p)](e_0 + e_3^+) + [1 - k_1(p) - k_2(p)](f_1 + f_2)$$

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

$$e_0 + e_3^+ = \frac{1 - k_1(p) - k_2(p)}{k_1(p) + k_2(p)}(f_1 + f_2)$$

and

$$\begin{aligned} e_0 = k_0(p)(f_1 + f_2 + e_0 + e_3^+) &= k_0(p)(f_1 + f_2)\left[1 + \frac{1 - k_1(p) - k_2(p)}{k_1(p) + k_2(p)}\right] \\ &= \frac{k_0(p)}{k_1(p) + k_2(p)}(f_1 + f_2) = \frac{f_1 + f_2}{(1-p)(2-p)}. \end{aligned}$$

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* $k_y(p) = (1-p)^y p$ *for* $y = 0, 1, \cdots$ *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. □

## 9 Appendix 2: Standard errors

Let $\hat{N}$ be the estimator of the population size $N$ of interest, the latter being a fixed but unknown quantity. Also, let the random quantity $n$ be the observed number of units. We will make use of the result

$$\mathrm{Var}(\hat{N}) = E_n\{\mathrm{Var}(\hat{N}|n)\} + \mathrm{Var}_n\{E(\hat{N}|n)\}, \tag{9}$$

where $\hat{N}|n$ refers to the distribution of $\hat{N}$ conditional upon $n$ and $E_n(.)$ and $\mathrm{Var}_n(.)$ refer to the first and second (central) moment w.r.t. the distribution of $n$. For more details see Böhning (2008).

### 9.1 Maximum likelihood estimator

We consider the maximum likelihood estimator $\hat{p}_{ML} = n/S$ and the associated population size estimator $\hat{N} = \hat{N}_{ML} = n/(1-n/S)$. We start with the second term in (9) and have that $E(\hat{N}|n) \approx n/(1-p)$, approximately, so that

$$\mathrm{Var}_n[n/(1-p)] = \frac{1}{(1-p)^2} Np(1-p).$$

Note that $N(1-p)$ can be estimated by $n$ and $p$ by the maximum likelihood estimator $n/S$, so that the variance estimator $\frac{Sn^2}{(S-n)^2}$ arises.

For the first term in (9), we use the $\delta$-method to determine $\mathrm{Var}(\hat{N}|n)$ as

$$\frac{n^2}{(1-p)^4}\mathrm{Var}_n(\hat{p}_{ML})$$

and, using the Fisher information for $p$, we can determine $\mathrm{Var}_n(\hat{p}_{ML})$ as

$$\mathrm{Var}_n(\hat{p}_{ML}) \approx \frac{n(S-n)}{S^3}.$$

The expected value $E_n\{\mathrm{Var}(\hat{N}|n)\}$ is then replaced by its moment estimate $\mathrm{Var}(\hat{N}|n)$ to achieve the total variance

$$\frac{Sn^2}{(S-n)^2} + \frac{n^2}{(1-n/s)^4}\frac{n(S-n)}{S^3} = \frac{S^2n^2}{(S-n)^3} \tag{10}$$

### 9.2 Censored estimator

We consider the censored estimator $\hat{p}_{Cen} = f_1/n$ and the associated population size estimator $\hat{N} = \hat{N}_{Cen} = n/(1-f_1/n)$. We have $E(\hat{N}|n) \approx n/(1-p)$, approximately, so that, as before, $\mathrm{Var}_n n/(1-p) = \frac{1}{(1-p)^2}Np(1-p)$, which can be estimated as $\frac{f_1}{(1-f_1/n)^2}$ by replacing $N(1-p)$ by $n$ and $p$ by $f_1/n$.

For the first term in (9), $\mathrm{Var}(\hat{N}|n)$, using the $\delta-$method once more we achieve the approximation

$$\mathrm{Var}\left(\frac{n}{1-f_1/n}|n\right) \approx \frac{n^2}{(1-f_1/n)^4}\mathrm{Var}\left(\frac{f_1}{n}|n\right),$$

from where the variance estimator $\frac{f_1(1-f_1/n)}{(1-f_1/n)^4} = \frac{f_1}{(1-f_1/n)^3}$ arises. In total, taking both variance terms into account, we achieve the variance estimator

$$\frac{f_1}{(1-f_1/n)^2} + \frac{f_1}{(1-f_1/n)^3} = \frac{f_1}{(1-f_1/n)^2}\frac{2n-f_1}{n-f_1} \tag{11}$$

### 9.3 Chao's estimator

Finally, we consider the Chao-type estimator $\hat{N} = \hat{N}_C = n + f_1^2/f_2$. Note that it differs from the original Chao-estimator $n + f_1^2/(2f_2)$ for which a variance estimator is provided in Chao (1987). If we would be only interested in a variance estimator of $\hat{f}_0$ we could simply multiply the Chao-variance-estimator by a factor of 4. However, interest is usually in the population size estimator $\hat{N}$ for which this simple adjustment is not valid. Hence we provide a full analysis in the following, again using the conditioning technique (9).

We have $E(\hat{N}|n) = E(n + \frac{f_1^2}{f_2}) \approx n + (g_1^+)^2 n/g_2^+ = n(1 + (g_1^+)^2/g_2^+)$, approximately. Recall that $g_y^+ = g_y/(1 - g_0)$ for $y = 1, 2, \ldots$. (Note that $E(\hat{N}|n)$ refers to the conditional count distribution $g_1^+, g_2^+, \ldots$ which is estimated by $f_1/n, f_2/n, \ldots$. Hence $\mathrm{Var}_n\{n(1 + (g_1^+)^2/g_2^+)\} = (1 + (g_1^+)^2/g_2^+)^2 N g_0(1 - g_0)$ which can be estimated as follows. $Ng_0$ can be estimated as $\hat{f}_0 = f_1^2/f_2$ and $(1 - g_0)$ as $1 - f_1^2/(\hat{N} f_2) = \frac{f_2 n}{f_2 n + f_1^2}$, so that in total the estimate $(1 + \frac{f_1^2}{f_2 n})^2 \frac{f_1^2 n}{f_2 n + f_1^2}$ arises, which we can simplify as

$$\left(1 + \frac{f_1^2}{f_2 n}\right)^2 \frac{f_1^2 n}{f_2 n + f_1^2} = f_1^2/f_2 + f_1^4/(f_2^2 n). \tag{12}$$

For the first term in (9), $\mathrm{Var}(\hat{N}|n)$, using the bivariate $\delta-$method, we achieve the approximation

$$\mathrm{Var}(\hat{N}|n) \approx \nabla\phi_0(f_1, f_2)^T cov(f_1, f_2) \nabla\phi_0(f_1, f_2)$$

where $\phi_0(f_1, f_2) = f_1^2/f_2$ and $\nabla\phi_0(f_1, f_2)$ is the two-vector of partial derivatives with respect to $f_1$ and $f_2$:

$$\nabla\phi_0(f_1, f_2)^T = (2f_1/f_2, -f_1^2/f_2^2).$$

The covariance matrix, conditional on $n$, is $cov(f_1, f_2) = n(\mathrm{dia}(\mathbf{g}^+) - \mathbf{g}^+\mathbf{g}^{+T})$, where $\mathbf{g}^+$ is the two-vector of probabilities, conditional on $n$, for observing a one or a two, respectively. Also, $\mathrm{dia}(\mathbf{g}^+)$ is the diagonal $2 \times 2$ matrix with $g_1^+$ and $g_2^+$ on the main diagonal. This matrix is estimated by

$$\begin{pmatrix} f_1 - f_1^2/n & -f_1 f_2/n \\ -f_1 f_2/n & f_2 - f_2^2/n \end{pmatrix}.$$

Hence we find for

$$\nabla\phi_0(f_1, f_2)^T \widehat{cov}(f_1, f_2) \nabla\phi_0(f_1, f_2) = \frac{4f_1^3}{f_2^2} + \frac{f_1^4}{f_2^3} - \frac{f_1^4}{f_2^2 n}. \tag{13}$$

Ultimately, taking (12) and (13) together, we achieve the variance estimator for $\hat{N} = \hat{N}_C = n + f_1^2/f_2$ as

$$\frac{f_1^4}{f_2^3} + \frac{4f_1^3}{f_2^2} + \frac{f_1^2}{f_2}, \tag{14}$$

being of remarkably simple form.

## References

Böhning D (2008) A simple variance formula for population size estimators by conditioning. Stat Methodol 5:410–423

Borchers DL, Buckland ST, Zucchini W (2004) Estimating animal abundance. Closed populations. Springer, London

Bunge J, Fitzpatrick M (1993) Estimating the number of species: a review. J Am Stat Assoc 88:364–373

Chao A (1987) Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43:783–791

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

Chao A, Tsay PK, Lin SH, Shau WY, Chao DY (2001) Tutorial in biostatistics: The applications of capture-recapture models to epidemiological data. Stat Med 20:3123–3157

Dorazio RM, Royle JA (2005) Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics 59:351–364

Hay G, Smit F (2003) Estimating the number of drug injectors from needle exchange data. Addict Res Theory 11:235–243

Holzmann H, Munk A, Zucchini W (2006) On identifiability in capture-recapture models. Biometrics 62:934–939

Link WA (2003) Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. Biometrics 59:1123–1130

Link WA (2006) Response to a paper by Holzmann, Munk and Zucchini. Biometrics 62:936–939

Mao CX (2007a) Estimating population sizes for capture-recapture sampling with binomial mixtures. Comput Stat Data Anal 51:5211–5219

Mao CX (2007b) Estimating the number of classes. Ann Stat 35:917–930

Mao CX (2008a) On the nonidentifiability of population sizes. Biometrics 64:977–981

Mao CX (2008b) Lower bounds to the population size when capture probabilities vary over individuals. Aust N Z J Stat 50:125–134

Oosterlee A, Vink RM, Smit F (2009) Prevalence of family violence in adults and children: estimates using the capture-recapture method. Eur J Public Health 19:586–591

Paluscia VJ, Wirtz SJ, Covington TM (2010) Using capture-recapture methods to better ascertain the incidence of fatal child maltreatment. Child Abuse Neglect 34:396–402

Pledger SA (2005) The performance of mixture models in heterogeneous closed population capture-recapture. Biometrics 61:868–876

Roberts JM, Brewer DD (2006) Estimating the prevalence of male clients of prostitute women in Vancouver with a simple capture-recapture method. J R Stat Soc Ser A 169:745–756

Van der Heijden PGM, Cruyff M, van Houwelingen HC (2003) Estimating the size of a criminal population from police records using the truncated poisson regression model. Stat Neerlandica 57:1–16

Van Hest NAH, De Vries G, Smit F, Grant AD, Richardus JH (2008) Estimating the coverage of Tuberculosis screening among drug users and homeless persons with truncated models. Epidemiol Infect 136:14–22

Wang J-P, Lindsay BG (2005) A penalized nonparametric maximum likelihood approach to species richness estimation. J Am Stat Assoc 100:942–959

Wang J-P, Lindsay BG (2008) An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. Stat Methodol 5:30–45

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