# 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

# Contents

## 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 e□ciency and bias.

*Some key words:*

# 1 Introduction and Background

For integer $N$, we consider a sample of counts $Y_1$, $Y_2$, ..., $Y_N$

$(1 - p)$, 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.

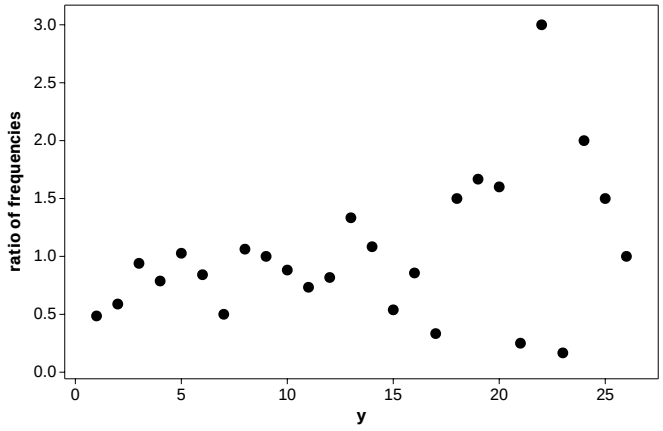


Figure 1: Ratio $f_{y+1}/f_y$ of neighboring frequencies for the data of the Scottish needle exchange program

We also see in Figure 2 that the geometric distribution provides a much better

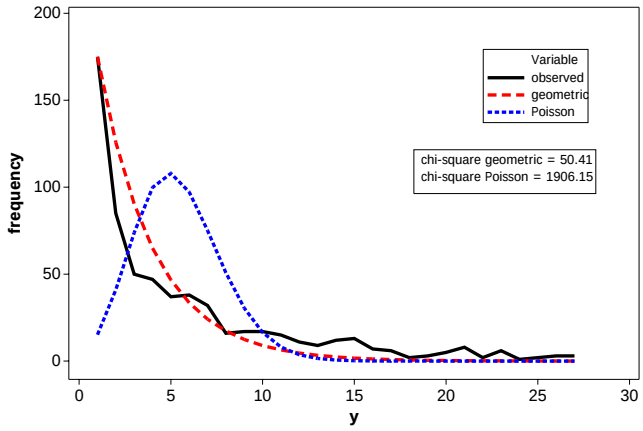


Figure 2: Observed frequencies with fitted frequencies under Poisson and geometric for the data of the Scottish needle exchange program

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 $S = \sum_{y=1} y f_y$. 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

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

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 $f_1 \log(q_1) + f_2 \log(q_2)$

# 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

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

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

# 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 bs 22a3(bs)l(si1TJ0h5(and)-28pmo2,mo 1h5

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 golofal.

[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

[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 pdf* $\log(p) = f_1 \log(q_1) + f_2 \log(q_2)$