Noname manuscript No.
(will be inserted by the editor)

On the Turing estimator in capture-recapture count data under the geometric distribution

Orasa Anan · Dankmar Böhning · Antonello Maruotti

Received: date / Accepted: date

Abstract We introduce an estimator for unknown population size in a capture-recapture framework where the count of identifications follows a geometric distribution. This can be thought of as a Poisson count adjusted for exponentially distributed heterogeneity. As a result, a new Turing-type estimator under the geometric distribution is obtained. This estimator can be used in many real life situations of capture-recapture, in which the geometric distribution is more appropriate than the Poisson. The proposed estimator shows a behavior comparable to the maximum likelihood one, on both simulated and real data. Its asymptotic variance is obtained by applying a conditional technique and its empirical behavior is investigated through a large-scale simulation study. Comparisons with other well-established estimators are provided. Empirical applications, in which the population size is known, are also included to further corroborate the simulation results.

Keywords Capture-Recapture data; Geometric distribution; Heterogeneity; Count data; Variance estimation

O. Anan
Department of Mathematics and Statistics, Faculty of Science, Thaksin University, Thailand
E-mail: aorasa@tsu.ac.th

D. Böhning
Southampton Statistical Sciences Research Institute and Mathematical Sciences, University of Southampton, UK
E-mail: D.A.Bohning@soton.ac.uk

A. Maruotti
Dipartimento di Giurisprudenza, Economia, Politica e Lingue Moderne, Libera Università Maria Ss. Assunta, Italy
E-mail: a.maruotti@lumsa.it

1 Introduction

Capture–recapture (CR) methods have become increasingly popular in the last decades and were adopted in a wide range of applications, focusing on the estimation of the size of hidden populations. To remark the importance of CR methods in data analysis, two books have been recently published by McRea and Morgan (2014) and Böhning et al. (2018), in which the CR methods are introduced and extensively discussed. CR analyses are based on the repeated sampling from a population and, consequently, on the use of recapture information to infer the number of uncaptured units. Throughout the paper, we consider the following CR setting. The target population is sampled over a certain number of capture occasions, and for each occasion, captured units are counted only once. Moreover, we consider a closed population, i.e. the unknown population size, is assumed to be constant (with no births/deaths during sampling stages), missclassification is not allowed and all units act independently. In capture-recapture analysis, the assumption of homogeneous catchability or identifiability across members of the target population is frequently in question. In these cases we speak of heterogeneity. The heterogeneity may influence capture probabilities, and failure to acknowledge this may lead to biased estimates of the unknown population size (Anan et al., 2017a; Farcomeni and Scacciatelli, 2013; Hwang and Huggins, 2005).

CR data are usually collected as follows. For each unit $i (i = 1, \dots, N)$ at occasion $t (t = 1, \dots, T)$, we record a binary indicator variable, say y_{it}, where $y_{it} = 1$ means that the i-th unit has been identified at the t-th occasion. It is further assumed that $y_i = \sum_{t=1}^{T} y_{it}$ is observed only if $y_i > 0$, that is if at least one $y_{it} > 0$ for $t = 1, \dots, T$. When $y_{i1} = y_{i2} = \cdots = y_{iT} = 0$, the i-th unit remains unobserved. The number of sampling occasions T may or may not be known a priori. In the following, we focus on the random variable X representing the distribution of the number of captures, i.e. X is a count variable.

To introduce heterogeneity, let us consider the pair (x, λ), where x is a realization of X, and $\lambda \geq 0$ is an unobserved realization of a non-negative random variable Λ, the parameter of the count data distribution, i.e. X depends on the parameter λ. It follows that the joint density $f(x, \lambda)$ can be written as $f(x \mid \lambda)g(\lambda)$, where $g(\lambda)$ is the marginal distribution of λ with respect to $f(x, \lambda)$. As we have not observed the value of λ, we consider the margin of $f(x \mid \lambda)g(\lambda)$ over λ leading to the mixture

$$\Pr(X = x) = \kappa_x = \int_0^{\infty} f(x \mid \lambda)g(\lambda)d\lambda, \tag{1}$$

for $x \in \{0, 1, 2, \dots\}$, the non-negative integers. In mixture model (1) we call $g(\lambda)$ the mixing distribution and $f(x \mid \lambda)$ the mixture kernel. In the following, we consider $f(x \mid \lambda) = \frac{\exp(-\lambda)\lambda^x}{x!}$, $\lambda \geq 0$, and $g(\lambda) = \frac{1}{\theta} \exp\left(-\frac{\lambda}{\theta}\right)$, $\theta > 0$, accounting for departures from assumptions implied by $f(x \mid \lambda)$, such that $\kappa_x = (1-p)^x p$, with $p = \frac{1}{1+\theta} \in (0, 1)$, i.e. κ_x follows a geometric distribution. Note that this result – mixing the Poisson with an exponential distribution

leads to the geometric – is a special case of mixing a Poisson with Gamma distribution which leads to a negative binomial (Fisher et al., 1943).

Of course, the binomial distribution can be also considered as the reference distribution to estimate population sizes. A justification of considering the Poisson distribution instead is as follows. Suppose the observational window consists of a large number of trapping occasions, each with the same positive capture probability θ. Then, let T being the number of possible identifications, using that $T\theta = \lambda$ remains constant when T becomes large, the binomial distribution converges to the Poisson distribution with parameter $T\theta = \lambda$. X is again the count of identifications per member of the target population. The only difference is that we do not know what could have been the largest possible count. The underlying assumption remains that identification occurs independently across occasions and with the same probability θ.

We are interested in using the geometric distribution in the capture-recapture setting. Let $X_1, \ldots, X_N$ be a sample from the geometric distribution. Here N is the size of the target population of interest and X_i is the count of identifications of the i-th member during the sampling period. The way identification occurs is determined by the application: it could be a live-trap, a hospital register, a police database, etc. In a recent paper, Coumans et al. (2017) considered estimating the numbers of homeless people in the Netherlands. In this case, X_i represents the number of nights stayed in a homeless shelter for homeless person i. However, not all members are identified, i.e. have a value of $X_i > 0$. Hence, we observe a zero-truncated sample $X_1, \ldots, X_n$, where we have, without loss of generality, $X_{n+1} = \cdots = X_N = 0$. We know the size n of the zero-truncated sample, but we do not know N, which needs to be estimated, and this is what this work is about. In the following we will use $f_x = \#\{x_i | x_i = x\}$, the frequency of counts exactly equal to x for $x = 0, 1, 2, \ldots$.

Whereas the Poisson distribution has been used frequently, we think that the geometric is more flexible in comparison with the former as it incorporates already some form of heterogeneity. Niwitpong et al. (2013) discuss various estimators for model (1) including a form of Mantel-Haenszel estimation.

In this work, we introduce a Turing-type estimator coping with heterogeneity, in the sense that the Poisson parameter of the conditional count-of-identifications distribution is mixed with an exponential density, leading to a geometric distribution. We argue that the geometric distribution is much better suitable for count distributions in the capture-recapture context as it can cope with simple forms of heterogeneity. We also derive its asymptotic variance, to have a measure of precision available. The Turing estimator has been used under several distributional assumptions on count data (Böhning et al., 2013; Hwang et al., 2015), in particular under the Poisson assumption where it is given by $\hat{N}_{Turing} = \frac{n}{1 - f_1 / \sum_{x=1}^{m} x f_x} = \frac{n}{1-\hat{p}_0}$, where m is the maximum number of observed counts. The benefits of Turing's estimator are that it is easy to calculate, its value can be obtained in a straightforward way, and there is no need for an iterative procedure. Nevertheless, under the Poisson assumption, it often underestimates the true target population size, though it has a better precision than the maximum likelihood estimator. ASK DANKMAR

FOR A REFERENCE We show in several case studies with known population size that the behavior of the geometric-based Turing estimator outperforms to other well-established estimators.

We investigate the empirical behavior of the introduced estimator by a large-scale simulation study with respect to several factors, such as the population size and the capture probabilities. To show the practical usefulness of this new estimator, we compare its performance to a few alternative estimators, widely used in the capture-recapture framework (see the simulation study in Section 3). Finally, we apply the proposal to several real datasets, often used as benchmarks in the capture-recapture framework and check the appropriateness of considering a geometric distribution to estimate the population size.

2 A Turing-type estimator under the geometric distribution

To generalise the Turing estimator under the geometric distribution, let us note that $\kappa_0 = p, \kappa_1 = (1-p)p$ and $E(X) = \frac{1-p}{p}$ and, accordingly,

$$\kappa_0 = p = \sqrt{p^2} = \sqrt{\frac{(1-p)p^2}{(1-p)}} = \sqrt{\frac{(1-p)p}{(1-p)/p}} = \sqrt{\frac{\kappa_1}{E(X)}}. \tag{2}$$

In practice, the κ_xs can be estimated by the relative frequencies so that

$$\widehat{\kappa}_0 = \sqrt{\frac{f_1/N}{S/N}} = \sqrt{\frac{f_1}{S}}, \tag{3}$$

where $S = \sum_{x=0}^{m} x f_x = \sum_{x=1}^{m} x f_x$. Hence, the resulting Turing estimator under the geometric distribution (TG) is given as

$$\widehat{N}_{TG} = \frac{n}{1 - \sqrt{\frac{f_1}{S}}}. \tag{4}$$

The form of the resulting TG estimator resembles its specification under the Poisson assumption. It uses the frequency f_1 of units observed only once, which is usually a large quantity. It also uses all information in the sample, by including S. This is in contrast with other well-established estimators which use only frequencies of ones and twos to estimate f_0.

Theorem 1 *The TG estimator is asymptotically unbiased under the geometric distribution*

$$\lim_{N\to\infty} \frac{E(\widehat{N}_{TG})}{N} \to 1,$$

with $\hat{N}_{TG} > \hat{N}_{Turing}$.

Proof We have that $E(X) = E(S/N) = (1-p)/p$, $E(f_1) = Np(1-p)$ so that $\sqrt{\frac{E(f_1/N)}{E(S/N)}} = \sqrt{\frac{p(1-p)}{(1-p)/p}} = p = \kappa_0$ and $E(n/N) = (1-\kappa_0) = (1-p)$. Therefore,

$$E\left(\frac{\widehat{N}_{TG}}{N}\right) = E\left(\frac{\frac{n}{1-\sqrt{f_1/S}}}{N}\right) = E\left(\frac{n}{N}\frac{1}{1-\sqrt{f_1/S}}\right) \xrightarrow[N\to\infty]{} (1-p)\frac{1}{1-p} = 1.$$

This proves that the TG estimator is asymptotically unbiased under the geometric distribution. To show that $\hat{N}_{TG} > \hat{N}_{Turing}$, let us assume that $f_x > 0$ for some $x > 1$. The estimated probability of zero counts according to the (original) Turing estimator under the Poisson distribution is $\widehat{p}_{0,Turing} = \frac{f_1}{S}$; whereas the probability of zero counts according to the TG estimator is $\widehat{p}_{0,TG} = \sqrt{\frac{f_1}{S}}$. It is obvious that $f_1 < S$ where $S = \sum_{x=1}^{m} x f_x$, therefore $\frac{f_1}{S} < \sqrt{\frac{f_1}{S}}$. Then, we have that

$$\widehat{N}_{TG} = \frac{n}{1-\sqrt{\frac{f_1}{S}}} > \frac{n}{1-\frac{f_1}{S}} = \widehat{N}_{Turing}.$$

This property provides evidence of the importance of defining an estimator able to handle heterogeneity, as it is widely acknowledged that the Poisson-based Turing estimator is biased downward (Böhning et al., 2013) and a lower bound estimator in presence of heterogeneity (Puig and Kokonendji, 2018).

Following Böhning (2008), we derive the variance of the estimator by using a conditional technique mixed with the delta method.

Proposition 1 *The variance of the TG estimator is given as*

$$\widehat{Var}(\widehat{N}_{TG}) = \frac{n\sqrt{\frac{f_1}{S}}}{(1-\sqrt{\frac{f_1}{S}})^2} + n^2\left\{\frac{S+f_1}{4S^2\left(1-\sqrt{\frac{f_1}{S}}\right)^4}\right\}. \tag{5}$$

Proof The proof is given in the Appendix

3 Simulation study

A simulation study is undertaken to investigate the performance of the proposed estimator and its competitors. The count data sets were generated following the geometric distribution with a variety of parameters. That is $X \sim Geo(p)$ where $p = 0.1, 0.15, 0.2, 0.25, 0.3, 0.5$. The population size N is set to $N = 100, 250$ for small sizes , $N = 500, 1000$ for medium sizes, and $N = 5000, 10000$ for large sizes. Each data set is rearranged in the form of frequencies $f_0, f_1, f_2, f_3, ...f_m$, corresponding to the counts $0, 1, 2, 3, ..., m$. The frequency of zero counts f_0 was omitted before estimating population sizes $\widehat{N}$. The aim is to investigate the finite sample behavior of the proposed estimator

and to show how this may differ from other well-established estimators, known to work well under the geometric distribution. Moreover, we also look at how well we approximate the uncertainty surrounding the estimates.

To further remark the usefulness of the proposed estimator, we generate counts from a Negative Binomial model $\Gamma(x+\nu)/(\Gamma(\nu)+x!)p^{\nu}(1-p)^{x}$ for $x=0,1,\ldots...$, with $\nu=2,5$ and $p=0.1,0.5,0.7$ and estimate the population size using the proposed estimator.

3.1 Simulation results to investigate the performance of the estimator

To study the performance of the proposed estimator, 5000 samples were drawn from the geometric distribution for each combination of parameters. For each scenario the relative bias, relative variance and relative root mean squared error are computed. In the following, we show simulation study results, summarized in Figures 1-3 and, with a focus on the proposed estimator, in Table 1. The linear regression Conway-Maxwell-Poisson-based estimator (LCMP; $\widehat{N}_{LCMP}=n+f_1\exp(-\hat{\lambda})$; Anan et al., 2017a,b) , the maximum likelihood estimator under the geometric distribution (MLEGeo; $\widehat{N}_{MLEGeo}=\frac{n}{1-n/\sum x=1^{m}xf_x}$) and a non-parametric estimator based on Chao's lower bound under the geometric distribution (CG; $\widehat{N}_{CG}=n+\frac{f_1^2}{f_2}$) are compared with our Turing-based proposal (TG) and the extended Zelterman's estimator based on the zero-truncated geometric distribution (ZG; $\widehat{N}_{ZG}=\frac{nf_1}{f_2}$). Among these estimators, probably, the less-known one is that based on the Conway-Maxwell-Poisson distribution (Shmueli et al., 2005), whose probability distribution CMP(λ,ν) is given by

$$\kappa_x=\frac{\lambda^x}{(x!)^{\nu}}\frac{1}{z(\lambda,\nu)},\quad x=0,1,2,\ldots;\lambda>0;\nu\geq 0$$

where the normalizing constant

$$z(\lambda,\nu)=\sum_{j=0}^{\infty}\frac{\lambda^j}{(j!)^{\nu}}$$

is a generalization of well-known infinite sums. Anan et al. (2017a) introduced a population size estimator based on this distribution.

The relative bias ($RBias$), variance ($RVar$) and root mean squared error ($RRMSE$), calculated as

$$RBias(\widehat{N})=\frac{1}{N}\left[E(\widehat{N})-N\right],$$

$$RVar(\widehat{N})=\frac{1}{N^2}\left(\widehat{Var}(\widehat{N})\right),$$

and

$$RRMSE\{\widehat{N}\}=\frac{1}{N}\sqrt{Var\{\widehat{N}\}+\{bias(\widehat{N})\}^2}$$

where $bias(\widehat{N}) = E(\widehat{N}) - N$, are displayed.

Results are displayed and summarized in Figures 1, 2 and 3. All the considered estimators are asymptotically unbiased. All estimators provide a slight overestimation of population size for small population sizes and show a reduced bias as the population sizes increase. The TG estimator shows the most accurate behavior with the least bias on average. The relative bias of LCMP and TG estimators are very similar as p is small (i.e $p = 0.1, 0.15$). The MLEGeo provides the smallest variance under all settings as we expect, in line with the literature. Another interesting point is that the TG estimator has not only a small bias but also provides an estimated variance nearby the MLEGeo.

To summarize the performance of the different estimators, we use the relative root mean square error (see Figure 3). The simulation results show that TG and MLEGeo estimators are likely to be the best choices to estimate population size. The LCMP also shows a reasonable behavior, slightly worse, but comparable, than the TG and MLEGeo estimators. It happens because the Conway-Maxwell-Poisson distribution contains the geometric distribution as a special boundary case. It remains a reasonable estimator for a small value of p and, however, performs better than the CG estimator.

3.2 Variance approximation and confidence intervals

The aim of this section is to investigate the performance of the variance approximation introduced in equation (5). The full set of results of the simulation study, with the averaged estimates and standard errors of the new estimator, is provided in Table 1. Additionally, we consider the validity of the approximated variance estimator, used to derive confidence intervals, by investigating the ratio between approximated and sampled standard errors. Such an investigation shows that the performance of the approximated standard error of the TG estimator is more than reasonable, though it slightly underestimates the sample variability on average, but provides a satisfactory approximation for the true standard error, i.e. the ratio between the approximated and the sample standard errors is fairly close to one.

We further look at 95% confidence intervals for the proposed estimator of the population size N. We compare our proposal with other estimators often used to deal with population size estimation, derived under the Poisson and the geometric distributions. An 95% approximate confidence interval of population size N is constructed under the symmetric normal approximation as:

$$95\%CI = (\widehat{N}_L, \widehat{N}_U) = (\widehat{N} \pm Z_{.975} S.E(\widehat{N})), \tag{6}$$

where, for the proposed Turing-based estimator, $S.E(\widehat{N})$ is appoximated by the square root of (5). Their performances are quantified using the coverage probability (Cov) and the average length (AL) defined as follows:

$$Cov = \frac{\sum_{t=1}^{T} A_{(t)}}{T} \times 100, \tag{7}$$

where $A_{(t)}$ equal to 1 if the true value N is in the target confidence interval, and 0 otherwise; and T is the number of replicates. The confidence intervals are produced based upon the assumption of asymptotic normality that might affect the coverage probability for the small population sizes. It can be seen from Table 2 and Figure 4, in the Supplementary Material, that the coverage probabilities from almost all estimators show lower level than the nominal confidence interval level when the population sizes are $N = 100$ and the converge to the nominal level increases with increasing N. Overall, the simulation results suggest that the MLEGeo estimator provides the best performance with respect to coverage probability of confidence intervals. As can be seen, the coverage probability of the MLEGeo is close to the nominal level on average and provides the shortest length. The proposed estimator TG estimator has a satisfying and comparable behavior, even for small population sizes. The CG estimator requires large population sizes to provide a good performance of confidence interval at 95%. Finally, the LCMP estimator shows strong an anti-conservative behavior with lower coverage than the nominal confidence levels for the small population sizes. Additionally, it tends to show higher coverage than the nominal levels when the population sizes increases, as a consequence of the large standard errors.

Table 1 Simulation results: population size estimates $E(\widehat{N})$, approximated standard errors $E[\widehat{SE}(\widehat{N})]$ and *true* standard errors $SE(\widehat{N})$

N	TG estimator					
	$E(\widehat{N})$	$SE(\widehat{N})$	$E[\widehat{SE}(\widehat{N})]$	$E(\widehat{N})$	$SE(\widehat{N})$	$E[\widehat{SE}(\widehat{N})]$
	$Geo(0.10)$			$Geo(0.15)$		
100	100.6	4.01	3.96	100.5	5.13	5.02
250	250.6	6.12	6.13	250.6	8.11	7.82
500	500.5	8.78	8.60	500.5	11.42	10.99
1,000	1,000.4	12.37	12.11	1,000.9	15.87	15.51
5,000	5,001.1	27.15	27.02	5,001.4	36.13	34.59
10,000	10,001.5	39.51	38.19	10,001.3	49.22	48.89
	$Geo(0.20)$			$Geo(0.25)$		
100	100.6	6.17	6.08	100.8	7.43	7.20
250	250.6	9.74	9.49	250.9	11.49	11.21
500	500.7	13.82	13.33	500.9	16.44	15.74
1,000	1,001.3	19.34	18.82	1,001.4	23.04	22.23
5,000	5,002.9	41.13	41.97	5,000.2	51.07	49.55
10,000	10,000.9	62.37	59.30	10,000.8	71.79	70.07
	$Geo(0.30)$			$Geo(0.50)$		
100	101.2	8.81	8.44	102.7	16.21	16.16
250	251.0	13.76	13.07	252.4	24.52	24.35
500	500.9	19.47	18.37	502.1	34.01	33.97
1,000	1,001.2	26.72	25.90	1,001.1	46.83	47.67
5,000	4,999.5	60.13	57.76	5,005.9	103.73	106.35
10,000	10,000.9	83.75	81.68	10,000.8	150.10	150.08

To draw further conclusions on the usefulness of the proposed Turing-type estimator, we discuss results based on simulated data from a misspecified model, i.e. data are drawn from the Negative Binomial distribution (see Table

Table 2 Comparing the performance of confidence intervals of five estimators when data are generated from the geometric distribution

N	Coverage probability (%)					Average length				
	LCMP	MLEGeo	TG	ZG	CG	LCMP	MLEGeo	TG	ZG	CG
	$Geo(0.1)$									
100	91.6	94.4	92.9	91.4	88.5	21.2	14.8	15.5	241.3	40.9
250	94.6	94.7	94.2	93.6	92.9	34.8	22.4	24.0	327.8	55.7
500	94.8	94.5	93.6	94.8	93.8	49.8	31.2	33.7	446.1	75.3
1,000	95.3	94.5	94.1	94.8	94.3	71.0	43.9	47.5	613.5	103.4
5,000	96.8	94.8	94.9	95.3	95.3	161.9	97.6	105.9	1,349.9	227.4
10,000	96.3	94.4	94.0	94.5	95.0	230.1	137.8	149.7	1,907.9	321.2
	$Geo(0.15)$									
100	91.8	94.2	93.0	91.2	89.5	28.9	18.9	19.7	189.9	47.3
250	93.7	94.2	93.5	94.0	92.8	47.2	28.9	30.6	274.0	68.2
500	94.1	94.7	93.8	94.5	93.4	67.6	40.4	43.1	373.8	92.8
1,000	95.5	94.2	93.9	94.8	94.4	96.9	56.8	60.8	522.4	130.0
5,000	96.2	94.3	93.8	94.7	94.4	221.1	126.5	135.6	1,154.9	287.0
10,000	96.9	95.1	94.8	94.7	95.1	314.2	178.7	191.7	1,630.1	404.9
	$Geo(0.20)$									
100	92.0	94.5	93.6	91.6	90.8	37.5	23.0	23.8	169.5	55.1
250	94.2	94.9	93.4	94.5	93.4	61.0	35.4	37.2	245.5	80.2
500	94.7	94.6	93.5	94.2	93.8	87.4	49.5	52.3	339.6	110.6
1,000	95.6	95.0	93.9	94.7	94.4	124.9	69.7	73.7	474.1	154.4
5,000	97.6	96.0	96.2	94.0	96.0	283.8	155.1	164.5	1,053.7	342.9
10,000	96.4	94.6	93.5	95.0	94.5	404.0	219.2	232.5	1,483.1	482.1
	$Geo(0.25)$									
100	91.6	94.8	93.3	92.8	91.4	47.2	27.4	28.2	158.2	63.2
250	94.0	95.0	94.0	94.4	94.0	76.4	42.2	43.9	231.0	92.3
500	94.1	94.9	93.7	94.4	94.0	108.8	59.0	61.7	319.3	127.2
1,000	95.6	94.7	93.8	94.0	94.5	156.6	83.1	87.2	448.3	178.8
5,000	96.3	95.2	94.0	94.9	94.9	355.0	184.9	194.2	993.3	395.6
10,000	96.9	95.5	94.7	94.9	95.2	504.6	261.4	274.7	1,402.6	558.9
	$Geo(0.3)$									
100	91.3	95.0	93.2	91.9	91.2	61.4	32.3	33.1	152.7	71.8
250	93.4	94.6	93.5	93.7	93.3	94.4	49.4	51.2	224.4	105.3
500	94.2	94.5	92.6	94.7	94.1	134.8	69.2	72.0	309.1	144.9
1,000	95.5	95.2	93.6	94.4	94.5	192.8	97.5	101.5	433.6	203.1
5,000	96.3	95.0	94.0	95.2	95.2	437.7	217.0	226.4	961.6	450.0
10,000	96.7	95.4	94.5	95.2	95.3	620.6	306.9	320.2	1,357.7	635.4
	$Geo(0.50)$									
100	89.5	95.5	94.6	92.8	92.3	148.8	59.8	63.4	165.9	117.9
250	92.5	94.8	94.5	94.3	94.2	219.5	90.1	95.4	235.9	167.1
500	94.2	94.5	94.6	94.9	94.7	308.9	125.7	133.1	324.1	229.4
1,000	94.0	95.3	95.0	94.9	94.7	433.4	176.5	186.9	450.7	318.8
5,000	97.2	97.4	96.6	95.0	95.4	969.0	393.1	416.8	1,004.7	710.5
10,000	96.7	94.7	94.9	95.3	95.2	1,380.2	554.8	588.3	1,414.1	1,000.0

3). A crucial role is played by the probability of success in each trial, which drives the percentage of zeros generated in the data. To a lower value of p, i.e. to lower percentage of zeros, corresponds a better performance of the estimator, which overestimates the true population sizes, though still reasonably, for higher values of p and proportions of zeros. This is also somehow expected. As shown in Böhning (2015), even sampling under the Negative Binomial

distribution may lead to severe bias in the population size estimate, due to a boundary problem.

Table 3 Simulation results: population size estimates $E(\widehat{N})$, approximated standard errors $E[\widehat{SE}(\widehat{N})]$ and *true* standard errors $SE(\widehat{N})$. Data are generated under the Negative Binomial model.

N	TG estimator							
	$E(\widehat{N})$	$SE(\widehat{N})$	$E[\widehat{SE}(\widehat{N})]$	Average % zeros	$E(\widehat{N})$	$SE(\widehat{N})$	$E[\widehat{SE}(\widehat{N})]$	Average % zeros
	$NB(0.1,2)$				$NB(0.1,5)$			
100	101.6	1.6	2.3	1.8	100.5	0.7	1.5	1.0
500	510.9	3.7	4.9	1.0	502.7	1.3	2.6	<1.0
5,000	5111.6	47.6	51.7	1.0	5014.1	4.6	6.6	<1.0
	$NB(0.5,2)$				$NB(0.5,5)$			
100	116.6	10.0	10.6	25.0	110.3	3.6	4.9	3.0
500	580.1	21.5	23.4	25.0	553.5	7.9	10.9	3.0
5,000	5800.2	67.7	73.7	25.0	5536.0	24.9	34.6	3.0
	$NB(0.7,2)$				$NB(0.7,5)$			
100	126.8	23.2	24.9	48.9	126.9	9.2	10.8	16.8
500	618.2	47.6	51.7	49.0	553.5	7.9	10.9	16.8
5,000	6159.1	148.1	161.4	49.5	6331.6	63.5	75.7	16.8

4 Applications

In this section, we firstly consider two widely analysed datasets to show the appropriateness and importance of our proposal in empirical data analyses and then provide further examples to better understand the behavior of the proposed estimator. We compare our estimator with other well established ones, based on homogeneous and heterogeneous Poisson models and on the geometric model. Turing's estimator and the maximum likelihood estimator under a Poisson model, with parameter λ, are considered as estimators in the homogeneous case. Estimators for heterogeneous populations as Zelterman's estimator (Zelterman, 1988) and Chao's lower bound estimator (Chao, 1987, 1989; Chao and Colwell, 2017) are considered as well. The maximum likelihood estimator under a geometric distribution and the LCMP estimator, with parameters λ and ν, are included as potential competitors to our estimator, as well as Chao's estimator under the geometric distribution. We also report an estimate of the variance associated with the population size estimate based on the so-called *imputed* bootstrap (SE_{boot}; Anan et al., 2017b). In our setting, the imputed bootstrap is based on $\widehat{N}$ and the corresponding estimate $\hat{f}_0$ of f_0. We draw 1000 samples containing $\widehat{N}$ observation from a Multinomial distribution with parameters $\widehat{N}$ and $\{\hat{f}_0/\widehat{N}, \hat{f}_1/\widehat{N}, \ldots, \hat{f}_m/\widehat{N}\}$. For each bootstrapped sample we estimate $\widehat{N}$ and then compute the variance over all 1000 samples.

4.1 Golf tees data

In a field experiment, $N = 250$ groups of golf tees were placed in a survey region, either exposed above the surrounding grass or hidden by it. They were surveyed by the 1999 statistics honor class at the University of St Andrews (Scotland), see Borchers et al. (2004). A total of $n = 162$ groups of tees were observed, but a (potentially unknown) number is missed and needs to be estimated. The corresponding frequency distribution is given by $(f_0, \ldots, f_8) = (88, 46, 28, 21, 13, 23, 14, 6, 11)$. This toy example is very useful for comparing the performance of several estimators as the true value of the population size is known. In the following, we compare several estimators based on the Poisson or other distributions, accounting for heterogeneity. The population size estimators under the geometric distribution (i.e., TG, MLEGeo, LCMP and CG) provide the population size estimates close to the true number $N = 250$, confirming that the homogeneity assumption of the capture probabilities, as in the Poisson distribution, is unreliable. In detail, the proposed TG estimator and the MLEGeo estimator show a negligible difference in terms of both the estimate of the population size and variability. Comparing these results with those from other estimators, the TG and MLEGeo are the best for estimating the number of golf tees. In detail, with respect to the other estimators, it is no surprise for the original Turing and MLEPoi to underestimate the population size, and their confidence intervals do not even cover the true value. $\hat{N}_{Turing}$, $\hat{N}_{MLEPoisson}$ and $\hat{N}_{Chao}$ are lower bound estimates for mixed (heterogeneous) Poisson distributions, where the geometric distribution is a special case (Puig and Kokonendji, 2018). Although the Zelterman estimator provides an estimated population size closer to the true value, the standard errors are very large, leading to the a wide confidence interval at 95%. The LCMP estimator might be the alternative choice for estimating the number of golf tees giving a slight bias.

4.2 Bowel cancer data

Over several years, from 1984 onwards, about 50,000 subjects were screened for bowel cancer at St Vincent's Hospital in Sydney (Australia), see Lloyd and Frommer (2004). The screening procedure was based on a sequence of binary diagnostic tests, self-administered on T = 6 successive days. Since no screening test is 100% accurate, replications of the diagnostic test over a number of days may help identify most cases. On each of the six occasions, the presence of blood in feces has been recorded. People with six negative tests were not further assessed and it remains unknown which disease status they have, while people with at least one positive test had their true disease status determined by physical examination, sigmoidoscopy, and colonoscopy. The aim is to estimate how many (say f_0) cancer patients have been missed by adopting this screening procedure. Lloyd and Frommer (2004) mention that 122 patients with confirmed cancer status were screened again using the identical screening

Table 4 Golf tees data: Estimated population size, standard errors, confidence intervals and lengths of confidence interval from eight different estimators.

Model	$\widehat{N}$	$\widehat{SE}(\widehat{N})$	$95\%CI$	Length	SE_{boot}
Homogeneous Poisson					
Turing	177	4.58	(168 – 186)	18	4.85
MLE Poisson ($\widehat{\lambda} = 3.23$)	169	2.83	(163 – 175)	12	3.07
Heterogeneous Poisson					
Chao	200	13.09	(174 – 226)	52	15.33
Zelterman	231	29.90	(171 – 289)	118	33.61
Geometric					
MLEGeo ($\widehat{p} = 0.3$)	230	11.75	(207 – 253)	46	11.12
TG	228	12.03	(204 – 252)	48	12.24
LCMP	223	33.09	(159 – 288)	129	14.50
($\widehat{\lambda} = 0.77$ and $\widehat{\nu} = 0$)					
Contaminated Geometric					
CG	238	27.86	(183 – 293)	110	30.33

procedure. We will focus on this secondary distribution as f_0 is known there, with $(f_0, \ldots, f_6) = (22; 8; 12; 16; 21; 12; 31)$.
From Table 5, it is clear that ignoring any heterogeneity source leads to an underestimation of the population size. The Turing and MLEPoi estimators fail to recover the unknown f_0. The situation does not improve even if we consider estimators that relax the Poisson assumption. Zelterman's and Chao's estimators do not show a satisfying behavior. The true value is not covered by confidence intervals of any of these estimators. The proposed TG estimator clearly outperforms its competitors with $\hat{f}_0 = 17$ and $\hat{N} = 117$, and both the MLEGeo and the LCMP estimators have reasonable behaviors.

Table 5 Bowel cancer data: Estimated population size, standard errors, confidence intervals and lengths of confidence interval from eight different estimators.

Model	$\widehat{N}$	$\widehat{SE}(\widehat{N})$	$95\%CI$	Length	SE_{boot}
Homogeneous Poisson					
Turing	102	1.59	(99 – 105)	6	1.62
MLE Poisson ($\widehat{\lambda} = 4.02$)	102	1.42	(99 – 105)	6	1.48
Heterogeneous Poisson					
Chao	103	2.47	(98 – 108)	10	3.17
Zelterman	105	7.91	(90 – 120)	30	13.89
Geometric					
MLEGeo ($\widehat{p} = 0.24$)	132	7.51	(117 – 147)	30	6.73
TG	117	5.64	(106 – 128)	22	5.59
LCMP	107	7.75	(92 – 123)	31	3.83
($\widehat{\lambda} = 1.32$ and $\widehat{\nu} = 0$)					
Contaminated Geometric					
CG	105	4.68	(96 – 114)	18	5.40

4.3 Other examples where the number of zeros is known

The geometric distribution works well for the two illustrative datasets considered above. A more extended range of examples is provided in the following to support the view that the geometric is useful generally, though it may lead to *non-optimal* estimates when the Poisson assumption is tenable.

We are going to check the performance of our estimators with four real data sets where the numbers of zeros are known. The first two were analysed in Böhning and Schön (2005) in order to check the performance of the estimators introduced there. The last two (number of dicentrics) were analysed in Puig and Barquinero (2011) using rth-order Hermite distributions. The data sets are as follows:

1. Daily numbers of deaths in 1989 of women with brain vessel disease in West Berlin: $(f_0, \ldots, f_{14}) = (1, 4, 15, 31, 39, 55, 54, 49, 47, 31, 16, 9, 8, 4, 3)$.
2. Weekly number of packs of a product that were purchased within the previous 7 days in 456 stores. Each frequency is the number of stores that sold exactly x packages: $(f_0, \ldots, f_{20}) = (102, 54, 49, 62, 44, 25, 26, 15, 15, 10, 10, 10, 10, 3, 3, 5, 5, 4, 1, 2, 1)$.
3. Number of dicentric chromosomes after the exposure of a radiation dose of 0.405 Gy. Each frequency is the number of cells having exactly x dicentric chromosomes: $(f_0, \ldots, f_4) = (437, 66, 15, 1, 1)$.
4. Number of dicentric chromosomes after the exposure of a radiation dose of 0.600 Gy. The frequencies are as follows:$(f_0, \ldots, f_4) = (473, 119, 34, 3, 2)$.

All the results given by our estimators are shown in Table 6.

There is clear evidence that all the Poisson-based estimators provide a general good performance for the first data set (Brain vessel). The similarity of all the estimates could be a sign of an underlying Poisson distribution, as already noticed by Böhning and Schön (2005). The proposed estimator overestimates the true population size, driven by some heterogeneity that is not present in the data. The number of packs data is included here to show that our estimator outperforms its Poisson-based counterpart, which does not provide a value close to the true number of zeros, and performs similarly compared to the MLEGeo estimator. Poor results were obtained by Böhning and Schön (2005) and Puig and Kokonendji (2018). Both works highlighted the presence of heterogeneity (mainly due to zero-inflation) which is difficult to capture and to model properly. It seems that our proposal is, instead, able to capture this data feature. The last two data sets come from an experiment where the counts of chromosome aberrations (dicentrics) can be modelled by a physical mechanism leading to compound-Poisson distributions Puig and Barquinero (2011). In both examples, the geometric-based estimators provide much better results than the Poisson-based ones, capturing the heterogeneity in the data.

Table 6 Other examples where the number of zeros is known: Estimated population size and standard errors.

Model	Brain Vessel (N=366) $\widehat{N}$	$\widehat{SE}(\widehat{N})$	SE_{boot}	# packs (N = 456) $\widehat{N}$	$\widehat{SE}(\widehat{N})$	SE_{boot}
Homogeneous Poisson						
Turing	366	0.85	1.11	365	3.63	3.87
MLE Poisson	366	0.80	0.97	357	1.51	1.96
Heterogeneous Poisson						
Chao	366	0.91	1.22	384	9.33	11.11
Zelterman	367	0.96	9.55	423	30.82	32.72
Geometric						
MLEGeo	433	9.77	9.04	439	11.48	11.69
TG	381	4.12	5.71	428	11.25	11.67
LCMP	366	0.96	1.13	408	13.20	15.40
Contaminated Geometric						
CG	368	4.79	1.87	414	19.85	20.76
Model	Dic 0.405 Gy (N = 520) $\widehat{N}$	$\widehat{SE}(\widehat{N})$	SE_{boot}	Dic. 0.600 Gy (N = 631) $\widehat{N}$	$\widehat{SE}(\widehat{N})$	SE_{boot}
Homogeneous Poisson						
Turing	231	62.45	42.68	379	59.47	47.57
MLE Poisson	229	44.02	38.04	381	46.96	45.75
Heterogeneous Poisson						
Chao	229	49.34	64.03	367	49.63	59.07
Zelterman	227	55.05	69.06	363	56.65	61.85
Geometric						
MLEGeo	427	95.58	79.42	701	103.31	95.01
TG	416	137.78	86.32	669	133.10	96.48
LCMP	383	106.38	111.38	462	152.38	165.05
Contaminated Geometric						
CG	373	104.99	123.99	574	106.53	116.06

4.4 The ratio-plot under the geometric distribution

As we extensively discussed throughout the main text, the geometric distribution is potentially a suitable candidate for count of cases distribution as it catches naturally some heterogeneity present in the population. Hence, in some sense the geometric distribution should be preferred to the Poisson distribution. Nevertheless, the geometric distribution needs to be investigated to see if it is appropriate for our real data example.

In Böhning et al. (2013) a diagnostic tool was suggested to investigate a count dataset for a specific distribution. This diagnostic tool, called the *ratio plot*, is built on the observation that the ratios of neighboring probabilities are constant. The ratio plot is then given by $r_x = \frac{\kappa_{x+1}}{\kappa_x} = 1-p$, for $x = 0, 1, \ldots, m$, and p being the geometric event parameter. Note that these ratios are not dependent on whether untruncated or truncated distributions are considered. A natural estimate of r_x occurs when replacing the unknown probabilities by

the estimate f_x/N

$$\hat{r}_x = \frac{f_{x+1}}{f_x}$$

as the unknown N cancels out.

An obvious question relates to the fact that we could have used another popular distribution for modelling count data, as e.g. the Poisson model. The Poisson distribution is given as $\kappa_x = exp(-\lambda)\lambda^x/x!$ so that $r_x = (x+1)\kappa_{x+1}/\kappa_x = \lambda$ and we expect that $\hat{r}_x = (x+1)f_{x+1}/f_x$ shows a horizontal line pattern.

Figure 5 shows both ratio plots in comparison for both empirical data considered in Sections 4.1 and 4.2 and there is clear evidence that there is a positive trend in the Poisson ratio plots. Consequently, we argue here that the geometric distribution, whose ratio-plot is an almost constant line, is more appropriate in this case study.

Whereas the ratio plot focuses on the idea whether an empirical, nonparametric estimate of the ratio would follow a straight line, a major difficulty with interpreting the ratio plots is the qualitative judgment on constancy across the count range. Böhning and Punyapornwithaya (2018) developed the idea to construct a diagnostic device, namely the ratio-plot under the null hypothesis, that shows the observed ratio within limits expected if the data would follow a geometric distribution, so that we are able to examine more easily if the observed ratios lie in the specified geometric-defined region. This can be achieved by considering the 95% point-wise error bars. Figure 6 shows the error bars for the (log)-ratio-plots for the golf tees and bowel cancer data and supports the use of the geometric distribution in these case studies.

To provide further evidence of the results discussed in Section 4.3, we show the ratio-plots under the null hypothesis for those data as well (see Figure 7). The geometric distribution looks appropriate for three out of four datasets, while it does not seem appropriate for the brain vessel one, confirming the results previously discussed. To conclude, this graphical device could preliminary used to asses the adequacy of the geometric distribution, avoiding the use of geometric-based estimators if not appropriate.

Appendix: Proof of Proposition 1

According to the conditional technique, we have

$$Var(\hat{N}_{TG}) = Var_n\left\{E(\widehat{N}_{TG}|n)\right\} + E_n\left\{Var(\widehat{N}_{TG}|n)\right\}. \tag{8}$$

Starting from the first term on the right hand side of (8), the delta method we have $E(\widehat{N}_{TG}|n) \approx \frac{n}{1-\kappa_0}$ and, accordingly,

$$Var_n\left\{E(\widehat{N}_{TG}|n)\right\} \approx Var_n\left\{\frac{n}{1-\kappa_0}\right\} = \frac{1}{(1-\kappa_0)^2}Var(n) = \frac{N(1-\kappa_0)\kappa_0}{(1-\kappa_0)^2} \tag{9}$$

Since $E(n) = N(1-\kappa_0)$ and $\widehat{\kappa}_{0(TG)} = \sqrt{\frac{f_1}{S}}$, the variance in (9) can be estimated as:

$$\widehat{Var}_n\left\{E(\widehat{N}_{TG}|n)\right\} = \frac{n\sqrt{\frac{f_1}{S}}}{\left(1-\sqrt{\frac{f_1}{S}}\right)^2}.$$

Additionally,

$$Var(\widehat{N}_{TG}|n) = Var\left(\frac{n}{1-\sqrt{\frac{f_1}{S}}}|n\right) = n^2 Var\left(\frac{1}{1-\sqrt{\frac{f_1}{S}}}\right).$$

We know that $Var\left(\frac{1}{1-\sqrt{\frac{f_1}{S}}}\right)$ can be approximated by the delta-method. Hence, let $y = \frac{f_1}{S}$ and we take $h(y) = \frac{1}{1-\sqrt{y}}$. Then,

$$h'(y) = -(1-y^{1/2})^{-2}\left(-\frac{1}{2}y^{-1/2}\right) = \frac{1}{2\sqrt{y}(1-\sqrt{y})^2}.$$

Furthermore,

$$Var\left(\frac{1}{1-\sqrt{\frac{f_1}{S}}}|n\right) \approx \left(\frac{1}{2\sqrt{y}(1-\sqrt{y})^2}\right)^2 Var\left(\frac{f_1}{S}\right) = \left(\frac{1}{4\frac{f_1}{S}(1-\sqrt{\frac{f_1}{S}})^4}\right) Var\left(\frac{f_1}{S}\right).$$

As next step, using the conditional variance technique to estimate $Var\left(\frac{f_1}{S}\right)$, we have that

$$Var\left(\frac{f_1}{S}\right) = Var_{f_1}\left\{E\left(\frac{f_1}{S}\right)|f_1\right\} + E_{f_1}\left\{Var\left(\frac{f_1}{S}|f_1\right)\right\}. \tag{10}$$

With the approximation $E\left(\frac{f_1}{S}|f_1\right) = f_1 E(\frac{1}{S}) \approx \frac{f_1}{S}$, we have that

$$\begin{aligned} Var_{f_1}\left\{E\left(\frac{f_1}{S}|f_1\right)\right\} &\approx Var_{f_1}\left(\frac{f_1}{S}\right) = \frac{1}{S^2}Var(f_1) = \frac{1}{S^2}Np_1(1-p_1) \\ &= \frac{1}{S^2}\left(N\frac{f_1}{N}(1-\frac{f_1}{N})\right) = \frac{f_1}{S^2}\left(1-\frac{f_1}{N}\right). \end{aligned} \tag{11}$$

Again, estimating $E_{f_1}\left\{Var\left(\frac{f_1}{S}|f_1\right)\right\}$ by $Var\left(\frac{f_1}{S}|f_1\right)$ we have that

$$E_{f_1}\left\{Var\left(\frac{f_1}{S}|f_1\right)\right\} \approx Var\left(\frac{f_1}{S}|f_1\right) = f_1^2 Var\left(\frac{1}{S}\right)$$

Using the delta method, we achieve that

$$Var\left(\frac{1}{S}\right) \approx \frac{1}{S^4}Var(N\bar{X}) = \frac{1}{S^4}N^2 Var(\bar{X}) = \frac{1}{S^4}N^2\frac{Var(X)}{N}.$$

Since $X \sim Geo(p)$ we have that $E(X) = \frac{1-p}{p}$ and $Var(X) = \frac{1-p}{p^2}$.

$$Var\left(\frac{1}{S}\right) \approx \frac{1}{S^4} N^2 \frac{\left(\frac{1-p}{p^2}\right)}{N} = \frac{1}{S^4} N^2 \frac{\left(\frac{E(X)}{p}\right)}{N} = \frac{1}{S^4} N^2 \frac{\left(\frac{E(S/N)}{p}\right)}{N} \approx \frac{1}{pS^3}.$$

Let us note that

$$E\left(\frac{S}{N}\right) = \frac{1-p}{p}; \quad \frac{S}{N} \approx \frac{1-p}{p} \quad \text{or} \quad p(S+N) \approx N \quad \text{or} \quad \frac{1}{p} \approx \frac{S+N}{N} \tag{12}$$

Hence,

$$\widehat{Var}\left(\frac{f_1}{S}|f_1\right) = \frac{f_1^2}{S^3}\left(\frac{S+N}{N}\right).$$

Substituting (11) and (12) into (10), this leads to

$$\begin{aligned}\widehat{Var}\left(\frac{f_1}{S}\right) &= \frac{1}{S^2}\left\{f_1\left(1-\frac{f_1}{N}\right)\right\} + \frac{f_1^2}{S^3}\left(\frac{S+N}{N}\right) = \frac{f_1}{S^2}\left\{\frac{N+f_1}{N} + \frac{f_1}{S}\left(\frac{S+N}{N}\right)\right\} \\ &= \frac{f_1}{S^2}\left\{\frac{NS - Sf_1 + f_1S + f_1N}{NS}\right\} = \frac{f_1}{S^2}\left\{\frac{N(S+f_1)}{NS}\right\} = \frac{f_1S + f_1^2}{S^3}.\end{aligned}$$

We have that

$$\begin{aligned}\widehat{Var}\left(\frac{1}{1-\sqrt{\frac{f_1}{S}}}\right) &= \left\{\frac{1}{\frac{4f_1}{S}\left(1-\sqrt{\frac{f_1}{S}}\right)^4}\right\}\left\{\frac{f_1S+f_1^2}{S^3}\right\} = \widehat{Var}\left(\frac{1}{1-\sqrt{\frac{f_1}{S}}}\right) \\ &= \left\{\frac{S}{4f_1\left(1-\sqrt{\frac{f_1}{S}}\right)^4}\right\}\left\{\frac{f_1S+f_1^2}{S^3}\right\} = \frac{Sf_1+f_1^2}{4f_1S^2\left(1-\sqrt{\frac{f_1}{S}}\right)^4} \\ &= \frac{S+f_1}{4S^2\left(1-\sqrt{\frac{f_1}{S}}\right)^4}.\end{aligned}$$

Acknowledgments

This work is developed under the PRIN2015 supported-project "Environmental processes and human activities: capturing their interactions via statistical methods (EPHASTAT)" funded by MIUR (Italian Ministry of Education, University and Scientific Research). Antonello Maruotti is grateful to the "Centro di Ateneo per la Ricerca e l'Internalizzazione" (LUMSA) for the financial support.

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest

References

Anan, O., Böhning, D., Maruotti, A. (2017a). Population size estimation and heterogeneity in capture–recapture data: a linear regression estimator based on the Conway–Maxwell–Poisson distribution. *Statistical Methods and Applications* **26**, 49–79.

Anan, O., Böhning, D., Maruotti, A. (2017b). Uncertainty estimation in heterogeneous capture–recapture count data. *Journal of Statistical Computation and Simulation* **87**, 2094–2114.

Böhning, D., Schön, D. (2005). Nonparametric maximum likelihood estimation of population size based on the counting distribution. *Journal of the Royal Statistical Society: Series C* **54**, 721–737.

Böhning, D. (2008). A simple variance formula for population size estimators by conditioning. *Statistical Methodology* **5**, 410–423.

Böhning, D., Baksh, M.F., Lerdsuwansri, R., Gallagher, J. (2013). Use of the ratio plot in capture–recapture estimation. *Journal of Computational and Graphical Statistics* **22**, 135–155.

Böhning, D. (2015). Power series mixtures and the ratio plot with applications to zero-truncated count distribution modelling. *METRON* **73**, 201–216.

Böhning, D., Punyapornwithaya, V. (2018). The geometric distribution, the ratio plot under the null and the burden of Dengue Fever in Chiang Mai province. In *Capture-Recapture Methods for the Social and Medical Sciences*, Böhning, D., Bunge, J. and van der Heijden, P.G.M. (eds.), p. 55–60, CRC Press.

Böhning, D., van der Heijden, P.G.M., Bunge, J. (2018). *Capture-recapture Methods for the Social and Medical Sciences*. CRC Press.

Borchers, D.L., Buckland, S.T., Zucchini, W. (2004) *Estimating Animal Abundance – Closed Populations.*, London: Springer.

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., Colwell, R.K. (2017). Thirty years of progeny from Chao's inequality: estimating and comparing richness with incidence data and incomplete sampling. *SORT: statistics and operations research transactions* **41**, 3–54.

Coumans, A. M., Cruyff, M., Van der Heijden, P. G. M., Wolf, J. and Schmeets, H. (2017). Estimating Homelessness in the Netherlands Using a Capture-Recapture Approach. *Social Indicators Research* **130**,189–212.

Farcomeni, A., Scacciatelli, D. (2013). Heterogeneity and behavioural response in continuous time capture-recapture, with application to street cannabis use in Italy. *Annals of Applied Statistics* **7**, 2293–2314.

Fisher, R.A., Corbet, A.S and Williams, C.B. (1943). The relation between the number of species and the number of individuals in a random sample from one animal population. *Journal of Animal Ecology* **12**, 42–58.

Hwang, W.H., Huggins, R. (2005). An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data. *Biometrika* **92**, 229–233.

Hwang, W-H., Lin, C-W., Shen, T-J. (2015). Good–Turing frequency estimation in a finite population. *Biometrical Journal* **57**, 321–339.

Lloyd, C.J. and Frommer, D. (2004). Regression based estimation of the false negative fraction when multiple negatives are unverified. *Journal of the Royal Statistical Society, Series C* 53: 619–631.

McRea, R.S., Morgan, B.J.T (2014). *Analysis of capture-recapture data.* CRC Press.

Niwitpong, S.A., Böhning, D., van der Heijden, P.G., Holling, H. (2013). Capture–recapture estimation based upon the geometric distribution allowing for heterogeneity. *Metrika* **76**, 495–519.

Puig, P., Barquinero, J. F. (2011). An application of compound poisson modelling to biological dosimetry. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **467**, 897–910.

Puig, P., Kokonendji, C.C. (2018). Non-parametric Estimation of the Number of Zeros in Truncated Count Distributions. *Scandinavian Journal of Statistics* **45**, 347–365.

Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. Journal of the Royal Statistical Society: Series C 54: 127–142.

Zelterman, D. (1988) Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *Journal of statistical planning and inference* **18**, 225–237.

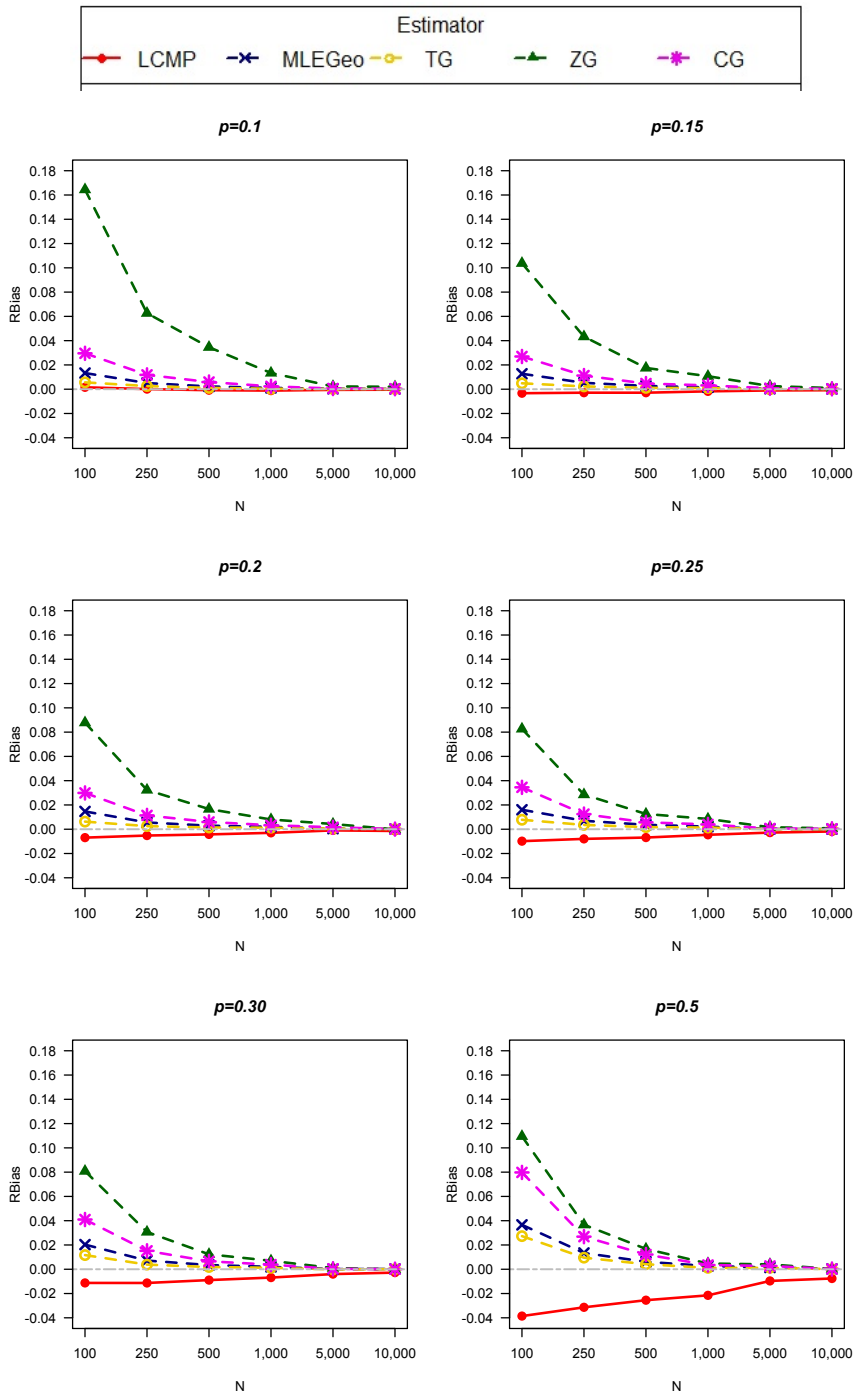

Fig. 1 Relative bias of six estimators with different parameters following the geometric distribution

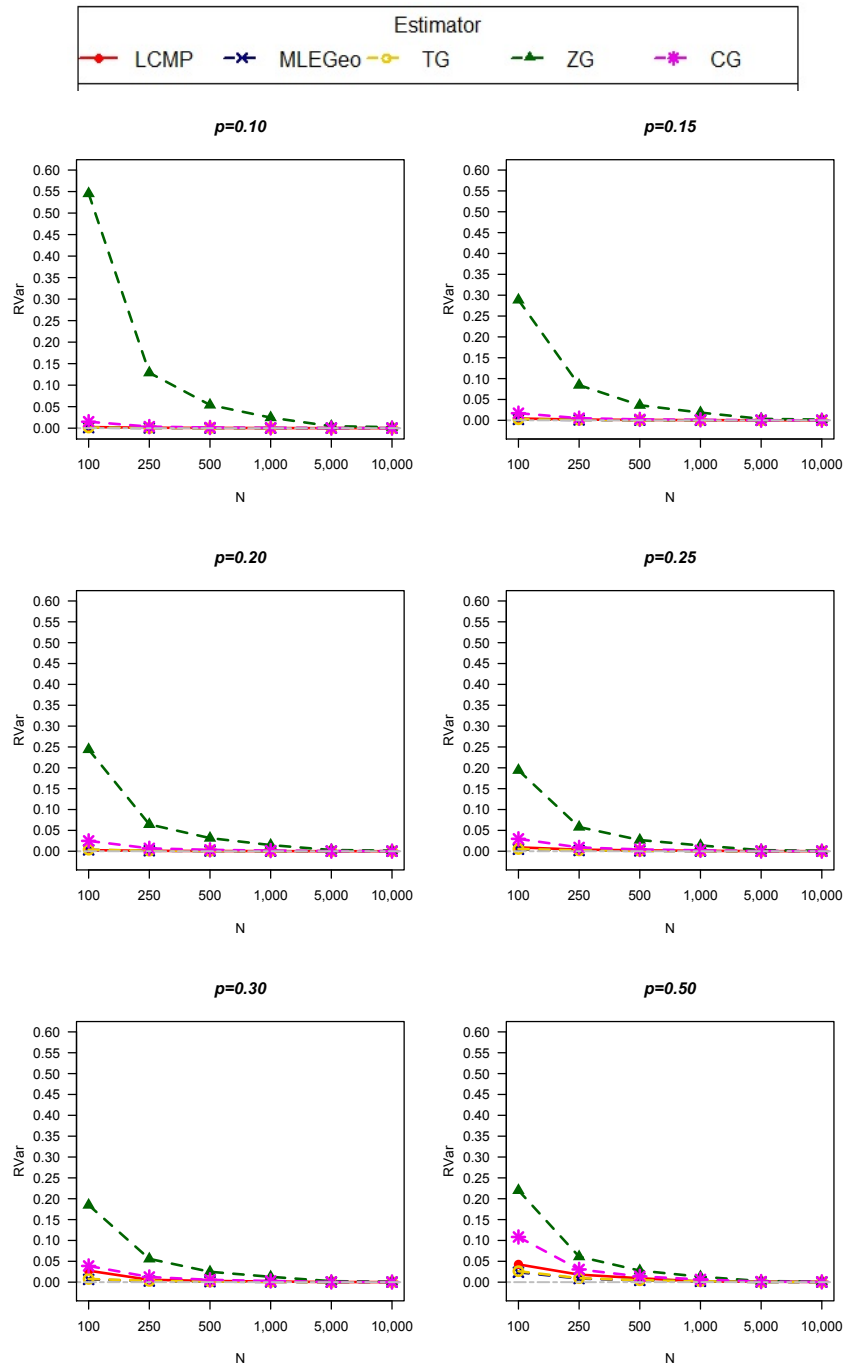

Fig. 2 Relative variance of six estimators with different parameters following the geometric distribution

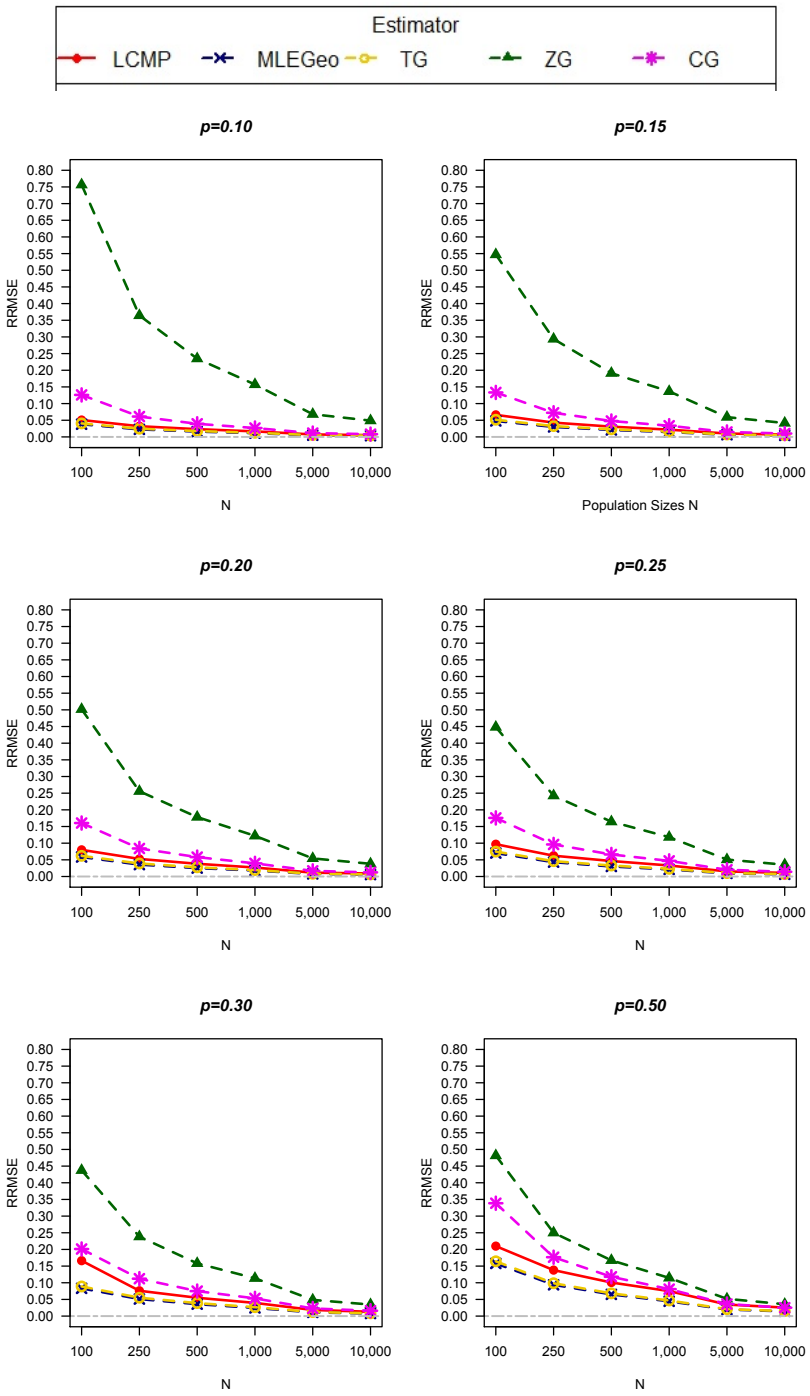

Fig. 3 Relative root mean square error of six estimators with different parameters following the geometric distribution

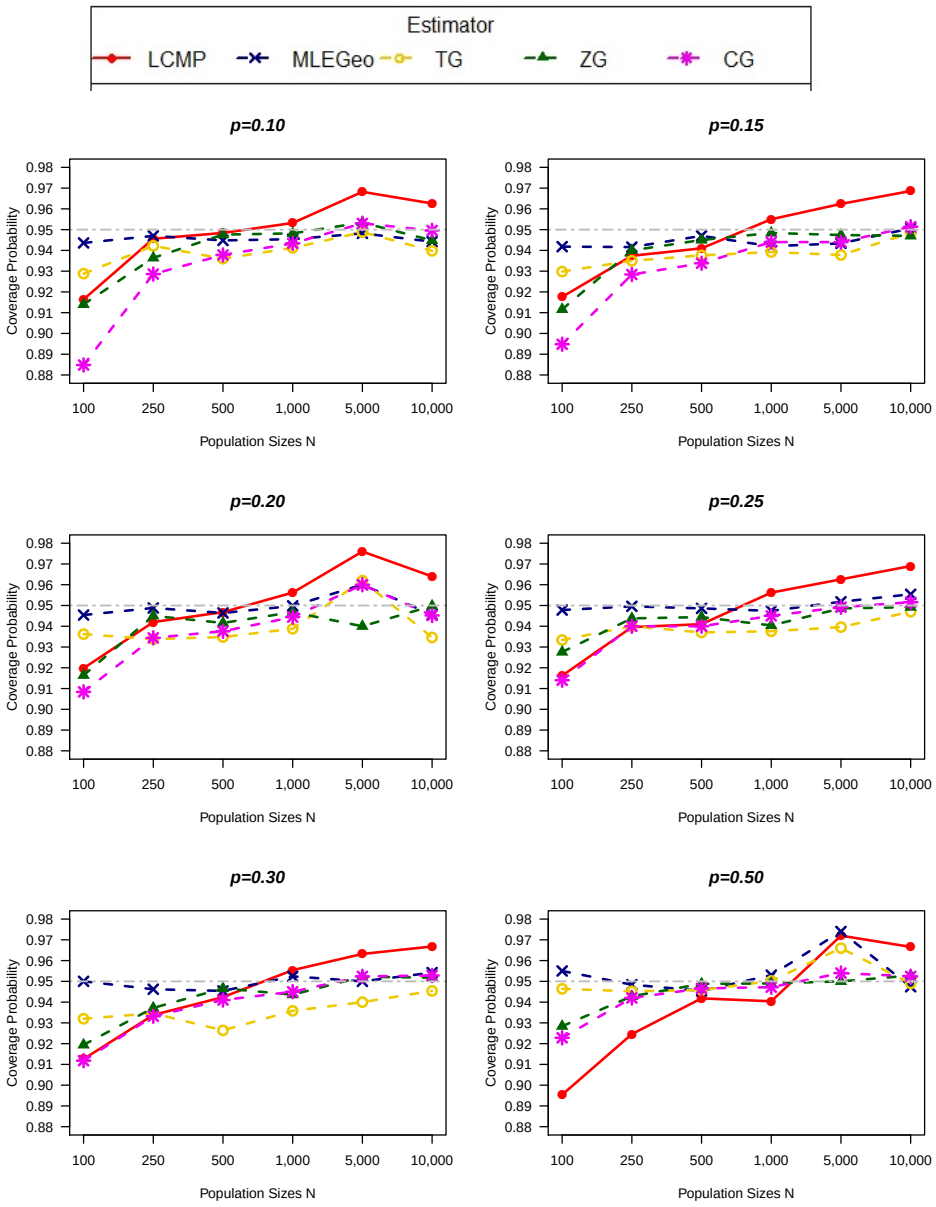

Fig. 4 Coverage Probabilities of 95% confidence interval when data is generated from the geometric distribution

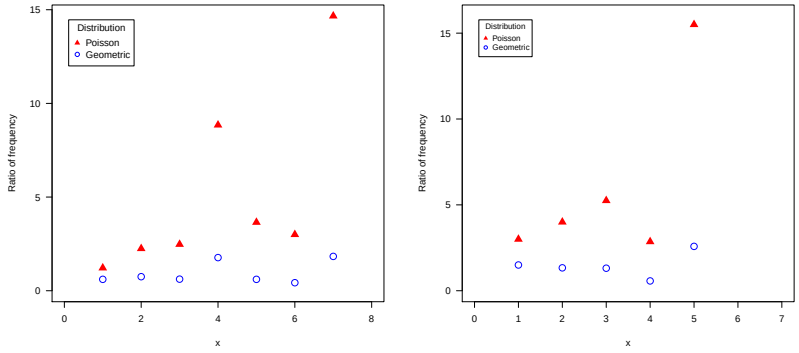

Fig. 5 Poisson and geometric ratio plots. Left panel: Golf tees data. Right panel: Bowel Cancer data.

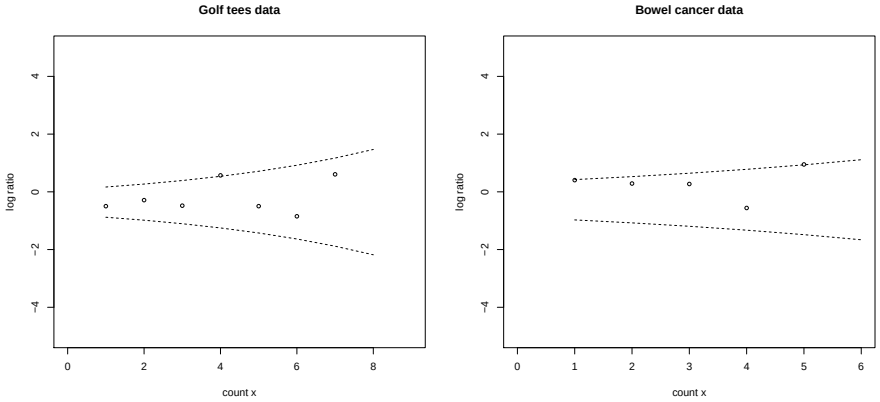

Fig. 6 Ratio plots for a geometric distribution under the null hypothesis with the 95% point-wise error bars (see Böhning and Punyapornwithaya (2018) for further details)

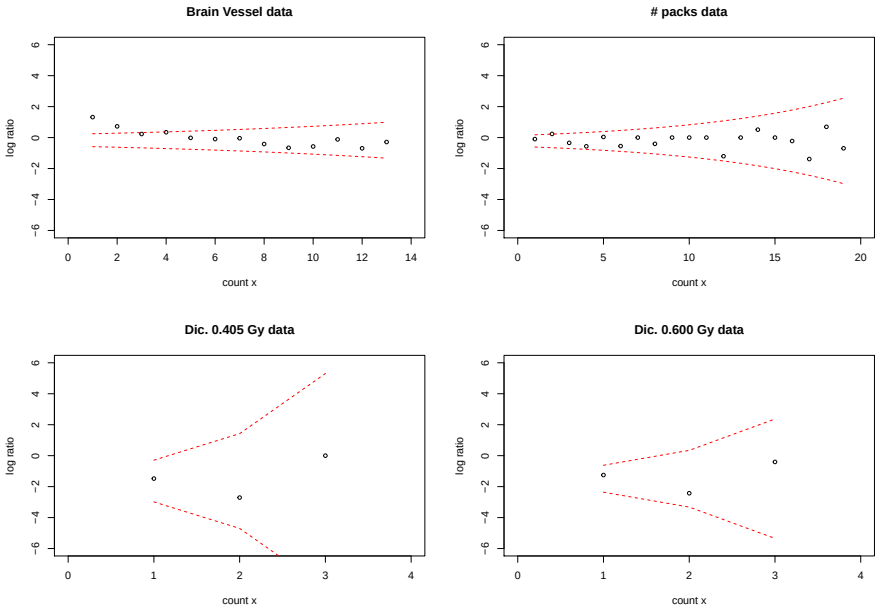

Fig. 7 Ratio plots for a geometric distribution under the null hypothesis with the 95% point-wise error bars (see Böhning and Punyapornwithaya (2018) for further details)