Population size estimation based upon zero-truncated, one-inflated and sparse count data

Estimating the number of dice snakes in Graz and flare stars in the Pleiades

Dankmar Böhning¹ · Herwig Friedl²

Accepted: 3 January 2021 / Published online: 30 January 2021 © The Author(s) 2021

Abstract

Estimating the size of a hard-to-count population is a challenging matter. In particular, when only few observations of the population to be estimated are available. The matter gets even more complex when one-inflation occurs. This situation is illustrated with the help of two examples: the size of a dice snake population in Graz (Austria) and the number of flare stars in the Pleiades. The paper discusses how one-inflation can be easily handled in likelihood approaches and also discusses how variances and confidence intervals can be obtained by means of a semi-parametric bootstrap. A Bayesian approach is mentioned as well and all approaches result in similar estimates of the hidden size of the population. Finally, a simulation study is provided which shows that the unconditional likelihood approach as well as the Bayesian approach using Jeffreys’ prior perform favorable.

Keywords Capture–recapture · Zero-truncation · One-inflation

The paper has been developed during an extended sabbatical visit of the first author to the Department of Statistics, Graz University of Technology during the summer term in 2019. The first author would like to express sincere thanks for all involved making this visit possible. Thanks also to Professor Sujit Sahu (University of Southampton) for the many discussions we had on Bayesian analysis. We are also grateful to two anonymous referees for providing valuable comments.

Dankmar Böhning
d.a.bohning@soton.ac.uk

Herwig Friedl
hfriedl@tugraz.at

¹ Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK

² Department of Statistics, Graz University of Technology, Graz, Austria
1 Introduction and motivation

The objective here is to determine the size $N$ of an elusive target population. To accomplish the purpose some mechanism (life trapping, register, surveillance system) is available which identifies a unit of the target population repeatedly. Hence, there is a count $X$ informing about the number of identifications of each unit in the target population. Furthermore, suppose a sample $X_1, X_2, \ldots, X_N$ of size $N$ is available which leads to the empirical count distribution as presented in Table 1.

There is, however, the well-known complication (Böning et al. 2018; McCrea and Morgan 2015) that any sample units with $X_i = 0$ would not be observed leading to a reduced observable sample $X_1, X_2, \ldots, X_n$.

In conclusion, we have that $f_0 = N - n$ is unknown. $f_0$ is also known as the dark or hidden figure and is the prime interest in this paper. In the following we illustrate the situation with two applications.

**Estimating the size of a dice snake population in Graz.** Tranninger and Friedl (2018) tried to estimate the size of a dice snake population in a closed area at the river Mur in Graz (Austria). The work was motivated by a resettlement project of the population due to the development of a water power plant in the vicinity of the living ground of the dire snakes. The major questions here was: how many dice snakes are there? We focus here on the year 2014 in which there were 31 capture occasions during the year. As above, $X$ denotes the identification count per dice snake. The empirical distribution of $X$ is provided in Table 3.

**The number of flare stars in the Pleiades.** Shortly after the appearance of two recent books on capture–recapture methods by McCrea and Morgan (2015); Böning et al. (2018), it was pointed out to the authors by Akopian (2018) that capture–recapture methods are also used in astro-physics to estimate the size of star clusters. Indeed, Ambartsumyan et al. (1970) published work where the number of stars in the Plaiades is estimated using capture–recapture techniques. The Pleiades is a star cluster about 444 light years away from planet Earth and consists of 100s of stars, only some of these are visible at certain times, the flare stars. In Table 4, we see the empirical distribution of $X$ representing the number of flares seen per star, for example, 123 stars were only seen once, 16 twice etc.

Both data situations have in common that there is perhaps sparsity exhibited by an abundance of unobserved zeros and relatively small non-zero counts. The second salient feature of both data sets is the occurrence of a relative large number of ones.

| Table 1 | Frequency distribution of count $X$ of repeated identifications |
|--------|---------------------------------------------------------------|
| $x$    | 0 | 1 | 2 | 3 | 4 | ... | Population size |
| $f_x$  | $f_0$ | $f_1$ | $f_2$ | $f_3$ | $f_4$ | ... | $N$ |
indicating potential one-inflation. One-inflation models have recently experienced some attention in capture–recapture modelling; see Böning and van der Heijden (2019), Böning et al. (2019), Farcomeni (2020), Godwin (2017, 2019), or Godwin and Böning (2017). In Böning and Ogden (2020) a more general investigation of inflation models is delivered and their close connection to truncation models established.

We see two major objectives for our paper:

1. With the motivation of the two case studies as background we are interested in raising the awareness of one-inflation in the capture–recapture context and its overestimation bias as consequence,
2. we are interested in illustrating some of the available approaches in estimating population size, with an emphasis on target populations that are small in size.

The rest of the paper is organized as follows: a probabilistic class of models that is based on zero-truncation and one-inflation is introduced in Sect. 2. In Sect. 3 goodness-of-fits of the case studies with respect to various count distributions are provided and it is found that the relatively simple geometric model seems to show up the best fit. Thus Horvitz–Thompson estimates based on zero-truncated one-inflated models are discussed in Sect. 4. Unconditional profile likelihood estimators under a geometric and a one-inflated geometric model are derived in Sect. 5. Section 6 is dedicated the idea to estimate the population size under a Bayesian setting. The performance of all estimating techniques discussed so far is evaluated by means of a Monte Carlo simulation study in Sect. 7. Section 8 presents ideas on how a semi-parametric bootstrap algorithm can be applied in order to find variance estimates and confidence intervals. The paper concludes with a short discussion that

| Table 2 | Frequency distribution of count X of repeated identifications |
|---------|---------------------------------------------------------------|
| x       | 0 | 1 | 2 | 3 | 4 | ⋯ | Observed size |
| f_x     | − | f_1 | f_2 | f_3 | f_4 | ⋯ | n |

| Table 3 | Frequencies of the number of times dice snakes have been identified in the target area in 2014 |
|---------|---------------------------------------------------------------------------------------------------|
| Frequency of count of sightings per dice snake | f_0 | f_1 | f_2 | f_3 | f_4 | f_5 | n |
| − | 59 | 8 | 1 | 1 | 1 | 70 |

| Table 4 | Frequency distribution of the count (per star) of flares (Ambartsumyan et al. 1970) |
|---------|-------------------------------------------------------------------------------------|
| Frequency of count of flares | f_0 | f_1 | f_2 | f_3 | f_4 | f_5 | f_6 | f_9 | n |
| − | 123 | 16 | 2 | 1 | 1 | 1 | 1 | 145 |
is provided in Sect. 9. The analysis has been performed within the \texttt{R} environment and exploits various functions written by the authors that are available on request.

## 2 Modelling

For predicting \( f_0 \) some sort of modelling is unavoidable as the nonparametric estimates \( f_x, x = 1, \ldots, m \) carry no information for \( f_0 \). Hence, we need to find a base model for \( P(X = x) = b_x(\hat{\theta}) \) so that an estimate \( \hat{\theta} \) for \( \theta \) can be achieved. This leads to fitted probabilities \( b_x(\hat{\theta}) \) for \( x = 0, 1, \ldots, m \), where \( m \) denotes the largest number of identifications. In particular, we can use for \( x = 0 \) the Horvitz–Thompson-type estimator for estimating \( f_0 \)

\[
\hat{f}_0 = n \frac{b_0(\theta)}{1 - b_0(\theta)},
\]

from which, ultimately, the population size estimator \( \hat{N} = n + \hat{f}_0 \) follows. For justified inference, the valid specification of the model \( b_x(\theta) \) is crucial.

For both case studies mentioned in the previous section, we see a large number of counts of ones, the singletons. Hence, we are concerned about one-inflation, a situation where more counts of ones occur than compatible with the baseline model \( b_1(\theta) \) as this can lead to a highly inflated estimate of \( f_0 \) as the following example shows. See also Godwin (2017) for further illustrations of this point.

**A synthetic example.** The empirical distribution of 500 counts sampled from a Poisson distribution with parameter 1 and 500 extra-counts of 1 so that \( N = 1000 \) is shown in Table 5.

If we ignore the zeros and estimate \( \theta \) by means of zero-truncated maximum likelihood we find \( \hat{\theta} = 0.42344 \) and

\[
\hat{f}_0 = n \frac{\exp(-\hat{\theta})}{1 - \exp(-\hat{\theta})} = 1582,
\]

clearly overestimating \( f_0 \) almost by a factor of 10.

To accommodate one-inflation we need to include it into the modelling. Hence we will focus on one-inflation modelling

\[
p_x'(\theta) = \begin{cases} 
(1 - z) + z p_x(\theta), & x = 1 \\
z p_x(\theta), & x \neq 1 
\end{cases}
\]

where \( p_x(\theta) = b_x(\theta)/(1 - b_0(\theta)) \) is a zero-truncated base distribution and \( z \in [0, 1] \).

| \( f_0 \) | \( f_1 \) | \( f_2 \) | \( f_3 \) | \( f_4 \) | \( f_5 \) | \( n \) |
|---|---|---|---|---|---|---|
| 166 | 696 | 105 | 18 | 12 | 3 | 834 |
As also mentioned in the synthetic example above, for the Poisson case we generally set

\[ p_x(\theta) = \frac{\exp(-\theta) \, \theta^x}{1 - \exp(-\theta) \, x!} \]

in model (1).

The modelling is greatly simplified using the following general result from Böhning and van der Heijden (2019). Consider an arbitrary inflation point \( x_1 \) and an arbitrary count density \( p_x(\theta) \) with associated \( x_1 \)-inflation as

\[ p_x'(\theta) = \begin{cases} 
(1 - \alpha) + \alpha p_x(\theta), & x = x_1 \\
\alpha p_x(\theta), & x \neq x_1.
\end{cases} \]

Then the likelihood and log-likelihood functions are

\[
L(\theta, x|x) = [(1 - \alpha) + \alpha p_1(\theta)]^{f_1} \prod_{x \neq x_1} [\alpha p_x(\theta)]^{f_x}, \\
\log L(\theta, x|x) = f_1 \log[1 - \alpha + \alpha p_1(\theta)] + \sum_{x \neq x_1} f_x \log p_x(\theta) + (n - f_1) \log \alpha,
\]

respectively, where \( p_1(\theta) = p_{x_1}(\theta) \), \( f_1 = f_{x_1} \), and \( n \) is the sample size. Therefore the profile log-likelihood in \( \theta \) is

\[
\log PL(\theta|x) = \sup_x \log L(\theta, x|x)
\]

and

\[
\hat{\alpha} = \frac{1 - f_1/n}{1 - p_1(\theta)}
\]

maximizes (2) for fixed \( \theta \). It follows that

\[
1 - \hat{\alpha} + \hat{\alpha} p_1(\theta) = 1 - \frac{1 - f_1/n}{1 - p_1(\theta)} + \frac{1 - f_1/n}{1 - p_1(\theta)} p_1(\theta) = f_1/n
\]

and the profile log-likelihood (3) becomes

\[
\log PL(\theta|x) = f_1 \log[1 - \hat{\alpha} + \hat{\alpha} p_1(\theta)] + \sum_{x \neq x_1} f_x \log p_x(\theta) + (n - f_1) \log \hat{\alpha}
\]
\[
= f_1 \log(f_1/n) + (n - f_1) \log \frac{1 - f_1/n}{1 - p_1(\theta)} + \sum_{x \neq x_1} f_x \log p_x(\theta)
\]
\[
= f_1 \log(f_1/n) + (n - f_1) \log(1 - f_1/n) + \sum_{x \neq x_1} f_x \log \left( \frac{p_x(\theta)}{1 - p_1(\theta)} \right)
\]

as \( \sum f_x = n - f_1 \). Thus, this \( x_1 \)-inflated profile log-likelihood equals the \( x_1 \)-truncated log-likelihood.
\[
\sum_{x \neq x_1} f_x \log \left( \frac{p_x(\theta)}{1 - p_1(\theta)} \right)
\]

plus

\[f_1 \log(f_1/n) + (n - f_1) \log(1 - f_1/n),\]

which is independent of \(\theta\). This result implies that \(x_1\)-inflation models can be simply fitted by \(x_1\)-truncated models.

To diagnose \(x_1\)-inflation we may fit the \(x_1\)-truncated log-likelihood

\[
\log T_1(\hat{\theta}) = \sum_{x \neq x_1} f_x \log \left( \frac{p_x(\hat{\theta})}{1 - p_1(\hat{\theta})} \right),
\]

construct the fitted \(x_1\)-inflated profile log-likelihood

\[
\log PL_1(\hat{\theta}|x) = f_1 \log(f_1/n) + (n - f_1) \log(1 - f_1/n) + \log T_1(\hat{\theta}),
\]

and finally form the likelihood ratio statistic

\[\lambda = 2 \log(PL_1(\hat{\theta}|x)/L_0(\hat{\theta}|x))\]

where

\[
\log L_0(\hat{\theta}|x) = \sum_x f_x \log p_x(\hat{\theta})
\]

is the non-inflated log-likelihood using all data.

We apply these ideas to zero-truncated distributions. For an arbitrary count density \(b_x(\theta)\), the base density, consider the associated zero-truncated count density

\[p_x(\theta) = \frac{b_x(\theta)}{1 - b_0(\theta)}, \quad x = 1, 2, \ldots\]

According to the previous result, for the one-inflated density we can restrict inference on the zero-one-truncated density

\[p_x^{++}(\theta) = \frac{b_x(\theta)}{1 - b_0(\theta) - b_1(\theta)}, \quad x = 2, 3, \ldots,\]

which then provides the one-inflated, zero-truncated density.

### 3 Finding the base distributions in the case studies

An important issue is the choice of the base distribution in the case studies. Graphical analysis using ratio plotting has been previously suggested; see Böning et al. (2013, 2018) or Böning (2016). However, these techniques require large samples sizes to avoid misleading conclusions, and in the cases discussed here we have clearly small sizes. Hence we base our analysis on likelihood methods including information criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). To cope with small samples we specifically use the modified version of the AIC in which the penalty for the
model complexity, say $2k$, is further increased by the factor $1 + (k + 1)/(n - k - 1)$ (see also McCrea and Morgan 2015).

Table 6 provides a comparative analysis including the Poisson and geometric distribution as well as the negative-binomial distribution. For both data situations the best model is the geometric model since it shows up the smallest $\text{AIC}_c$ and BIC values.

For completeness, we mention the probability mass function of the negative-binomial with $\theta = (\mu, \delta)$:

$$b_x(\theta) = \frac{\Gamma(x + 1/\delta)}{\Gamma(x + 1)\Gamma(1/\delta)} \left(\frac{1/\delta}{\mu + 1/\delta}\right)^{1/\delta}\left(\frac{\mu}{\mu + 1/\delta}\right)^x, \quad x = 0, 1, 2, \ldots$$

using the mean parameterization, so that $E(X) = \mu$ and $\text{Var}(X) = (1 + \delta \mu)\mu$, where $\mu > 0$ is the mean and $\delta > 0$ is the dispersion parameter. This yields the geometric distribution for $\delta = 1$ and the Poisson as the limiting case when $\delta \to 0$. Table 6 gives evidence for both case studies that the geometric is a reasonable distribution here.

However, the question arises if there is any evidence of one-inflation as the mere existence of many ones does not necessarily mean that there is one-inflation. Table 7 provides a diagnostic analysis of one-inflation. Note that we are testing here $H_0 : \alpha = 1$ vs. $H_1 : \alpha < 1$, so that the null-hypothesis is in the boundary of the alternative hypothesis and non-standard inference applies. In this case, the asymptotic distribution of the likelihood ratio test statistic $2\log(\lambda)$ is a $\chi^2$-distribution, namely

$$\frac{1}{2} \chi^2_0 + \frac{1}{2} \chi^2_1,$$

where $\chi^2_k$ is the $\chi^2$-distribution with $k$ degrees of freedom (Self and Liang 1987). $\chi^2_0$ is the singular distribution putting all its mass at 0. In practice, this means that conventional $\chi^2$-values need to be halved. For example, for the dice snake data we have a value of the likelihood ratio statistic of 3.0 which would give a conventional $p$ value of 0.084 under a $\chi^2$-distribution with 1 df. Halving this leads to the correct $p$ value of 0.042. For the Pleiades data we have a clear indication of one-inflation, whereas this is borderline for the dice snake data. As the overestimation effect of the

| Case study | Distribution | 0/1 log-lik | $\text{AIC}_c$ | BIC |
|------------|--------------|-------------|----------------|-----|
| Dice snakes | Poisson      | −41.856     | 87.890         | 96.208 |
|            | Geometric    | −41.480     | 87.140         | 91.458 |
|            | NB           | −41.485     | 89.324         | 95.706 |
|            | NB dispersion: 0.9999 (0.9995–1.0005) |          |                |     |
| Flare stars | Poisson      | −93.226     | 190.536        | 196.405 |
|            | Geometric    | −89.250     | 182.585        | 188.454 |
|            | NB           | −88.335     | 182.840        | 191.600 |
|            | Log-NB dispersion: 11.16 (−85.66 to 107.98) |          |                |     |

Table 6  Comparative distributional analysis for the two case studies based on the zero-truncated-one-inflated log-likelihood
population size is severe when ignoring one-inflation, we will use the one-inflation model when estimating population size.

### 4 Horvitz–Thompson estimation

The conventional Horvitz–Thompson estimator

\[ \hat{f}_0 = n \frac{b_0(\theta)}{1 - b_0(\theta)} \]  

has the property \( E(\hat{f}_0) = Np_0(\theta) \), if there is no inflation. The estimator (5) needs to be modified here as \( n \) contains the one-inflated part. This leads to

\[ \hat{f}_0 = (n - f_1) \frac{b_0(\theta)}{1 - b_0(\theta) - b_1(\theta)}. \]

which again has the property \( E(\hat{f}_0) = Np_0(\theta) \) and, ultimately, we can define the modified Horvitz–Thompson estimator

\[ \hat{N} = n + (n - f_1) \frac{b_0(\theta)}{1 - b_0(\theta) - b_1(\theta)}, \]  

which is unbiased in the sense that \( E(\hat{N}) = N \).

As \( \theta \) is unknown, a plug-in estimate is used based on the 0–1-truncated geometric as evidenced in the previous analysis. In Table 8 we see the estimated population sizes for the two case studies. The conventional Horvitz–Thompson estimator (cHTE) uses the 0-truncated geometric distribution whereas the modified Horvitz–Thompson estimator (mHTE) uses the 0–1-truncated geometric as described above.

### 5 Unconditional maximum likelihood estimation

So far we maximized the conditional (zero-truncated) likelihood of the observed counts. In the following we discuss the general sampling mechanism that has generated these observations. In particular we will discuss the unconditional maximum likelihood approach which has a long history in capture–recapture modelling. Chao and Bunge (2002) give a nice discussion on the conditional and unconditional approach and how they connect. Sanathanan (1972, 1977) provides a comprehensive analysis of their statistical properties. The unconditional likelihood

| Case study   | 0/1 log-lik | 0 log-lik | 2 log i. (p value) |
|-------------|-------------|-----------|-------------------|
| Dice snakes | -41.480     | -42.975   | 3.000 (0.042)     |
| Flare stars | -89.250     | -96.584   | 14.668 (0.000)    |
approach leads naturally, as we will see below, to a profile likelihood in \( N \). The latter suggests also a generic way of constructing confidence intervals as outlined in Venzon and Moolgavkar (1988). Cormack (1992) provides an application to capture–recapture settings as well as do Lebreton et al. (1992).

Let \( m \) denote the largest number of sightings, then the joint pmf of the sample is a multinomial model defined on the counts \( 0, 1, \ldots, m \) coming from the population of size \( N \). Since we only observe the counts of \( 1, \ldots, m \), the conditional model used is a zero-truncated multinomial for the \( n \) observed counts. This conditioning process is described by a binomial variable that is responsible for splitting the population into an observed part (of size \( n \)) and an unobserved part (of size \( N - n = f_0 \)). Together we have

\[
\text{multinom}(b_0(\theta), b_1(\theta), \ldots, b_m(\theta)|N) = \text{multinom}\left( \frac{b_1(\theta)}{1 - b_0(\theta)}, \ldots, \frac{b_m(\theta)}{1 - b_0(\theta)} | n \right) \\
\times \text{binom}(1 - b_0(\theta)|N),
\]

or equivalently

\[
\frac{N!}{f_0! f_1! \cdots f_m!} \prod_{x=0}^{m} b_x(\theta)^{f_x} = \frac{n!}{f_1! \cdots f_m!} \prod_{x=1}^{m} \left( \frac{b_x(\theta)}{1 - b_0(\theta)} \right)^{f_x} \\
\times \frac{N!}{f_0! n!} b_0(\theta)^{f_0} (1 - b_0(\theta))^n,
\]

which allows now to check the validity of this factorization.

Since \( f_1, \ldots, f_m \) are fixed given the observed counts, the relevant part of the unconditional likelihood is

\[
L(f_0, \theta|f_1, \ldots, f_m) = \frac{N!}{f_0!} \prod_{x=0}^{m} b_x(\theta)^{f_x}.
\]

Therefore, we have to maximize the unconditional log-likelihood function

\[
\ell(f_0, \theta|f_1, \ldots, f_m) = \sum_{x=0}^{m} f_x \log b_x(\theta) + \log(N!/f_0!). \quad (7)
\]

For a given value of \( f_0 \), the \( \theta \)-score function is

\[
|\theta| \quad \text{Case study} \quad \hat{N}
\]

| \( n \) | mHTE | cHTE |
|---|---|---|
| **Dice snakes** | 70 | 127 | 358 |
| **Flare stars** | 145 | 205 | 671 |

\[\text{Springer}\]
\[
\frac{\partial}{\partial \theta} \ell(f_0, \theta | \cdot) = \sum_{x=0}^{m} f_x \frac{\partial b_x(\theta)}{b_x(\theta)}. 
\]

If we specify the base distribution to be the geometric, i.e. \( b_x(\theta) = \theta(1 - \theta)^{x} \), then

\[
\frac{\partial b_x(\theta)}{b_x(\theta)} = \frac{(1 - \theta) - x\theta}{\theta(1 - \theta)}
\]

and the maximum likelihood estimator becomes

\[
\hat{\theta} = \frac{1}{1 + \frac{1}{N} \sum_{x=1}^{m} xf_x}.
\]

This estimator depends on the value of \( N \) and thus on the unknown \( f_0 \). We propose to evaluate the profile log-likelihood \( \ell(f_0, \hat{\theta} | \cdot) \) for a grid of \( f_0 \) values to find the maximizer \( \hat{f}_0 \). This is shown in Fig. 1.

Since \( \hat{f}_0 = 286 \) with 95% profile confidence interval (159, 527) for \( f_0 \), the total size of the population is estimated to be 356 snakes, which seems to be a reasonable number. This unconditional estimate can now be compared to the respective conditional estimate \( \hat{N}_c = 358 \) given in Table 8.

We also apply this model to estimate the size of the flare stars. From the right panel of Fig. 1 we get \( \hat{f}_0 = 523 \) with 95% profile confidence interval (353, 782). The respective estimate of the population size \( \hat{N} = 668 \) is again slightly smaller compared with \( \hat{N}_c = 671 \) from Table 8.

Under an arbitrary one-inflated count model the respective unconditional log-likelihood function (7) is

\[
\ell(f_0, \theta, \alpha | f_1, \ldots, f_m) = f_1 \log(1 - \alpha + \alpha b_1(\theta)) + \sum_{x \neq 1} f_x \log[\alpha b_x(\theta)] + \log(N! / f_0!)
\]

for \( x = 0, 2, \ldots, m \). Since for any fixed value of \( f_0 \) the derivation of the maximizer of this function w.r.t. \( \alpha \) is the analogue to finding the maximizer (4) of the respective conditional log-likelihood (2), we immediately have

\[
\hat{\alpha} = \frac{1 - f_1/N}{1 - b_1(\theta)}
\]

and define the profile log-likelihood function as

\[
\log PL(\theta, f_0 | f_1, \ldots, f_m) = f_1 \log \frac{f_1}{N} + (N - f_1) \log \left(1 - \frac{f_1}{N}\right)
\]

\[
+ \sum_{x \neq 1} f_x \log \frac{b_x(\theta)}{1 - b_1(\theta)} + \log(N! / f_0!)
\]

Notice that in this unconditional log-likelihood the total population size \( N \) takes
over the role of the observed sample size $n$ in its conditional version and the sum
also includes the additional term for $x = 0$.

Under the one-inflated geometric situation the relevant term depending on $\theta$ becomes

$$
\sum_{x \neq 1} f_x \log \frac{x(\theta)}{1 - x(\theta)} = \sum_{x \neq 1} f_x \log \frac{\theta(1 - \theta)^x}{1 - \theta(1 - \theta)}
$$

$$
= \log \frac{\theta}{1 - \theta(1 - \theta)} \sum_{x \neq 1} f_x + \log(1 - \theta) \sum_{x \neq 1} f_x x
$$

where $x = 0, 2, \ldots, m$. With

$$
N(-1) = \sum_{x \neq 1} f_x \quad \text{and} \quad S(-1) = \sum_{x \neq 1} f_x x = \sum_{x=2}^m f_x x
$$

the above profile log-likelihood simplifies to

$$
\log PL(\theta, f_0|f_1, \ldots, f_m) = f_1 \log \frac{f_1}{N} + N(-1) \left( \log \left( 1 - \frac{f_1}{N} \right) + \log \frac{\theta}{1 - \theta(1 - \theta)} \right)
$$

$$
+ S(-1) \log(1 - \theta) + \log(N! / f_0!)
$$

with corresponding $\theta$-score function

$$
\frac{\partial}{\partial \theta} \log PL(\theta, f_0|f_1, \ldots, f_m) = N(-1) \left( \frac{1}{\theta} + \frac{1 - 2\theta}{1 - \theta(1 - \theta)} \right) - S(-1) \frac{1}{1 - \theta}.
$$

Since $N(-1)$ is a sum over all frequencies except $f_1$, this score function actually depends on both, $\theta$ and the unobserved $f_0$. Thus, it is natural to find the maximizer of this profile log-likelihood using a grid of $f_0$ values and maximize the corresponding likelihood function in $\theta$ conditional on each $f_0$ value.

For the snake data $\hat{f}_0 = 45$ maximizes the profile likelihood as shown in the left panel of Fig. 2. The respective population size estimate $\hat{N} = 115$ is therefore rather small. The reason for this surprising result might be the fairly wide 95% profile confidence interval (7, 399), reflecting the large variance of the estimator $\hat{f}_0$. 
The situation is similar with the flare stars data shown in the right panel of Fig. 2. Since $f_0 = 53$ with 95% profile confidence interval (17, 165), the estimate $\hat{N} = 198$ is fairly small compared to the results under the previously considered models.

6 Bayesian analysis

We use the geometric density in the form

$$\frac{1}{\exp(z) + 1} \frac{\exp(z)^x}{(\exp(z) + 1)^x} = \theta(1 - \theta)^x$$

with $\theta = 1/(\exp(z) + 1)$ and for $x = 0, 1, \ldots$. Now, as we have 0–1-truncated data the associated 0–1-truncated density is

$$\frac{1}{\exp(z) + 1} \frac{\exp(z)^{x-2}}{(\exp(z) + 1)^{x-2}} = \theta(1 - \theta)^{x-2}$$

which is the form we use for the analysis. A non-informative normal prior (mean zero and standard deviation 100) is used for $\alpha$ and the Bayesian analysis is implemented using 10 parallel Markov chains, each run to produce a sample size of 10,000 after 2500 burn-in iterations. A nonparametric density estimate (Epanechnikov kernel with optimal bandwidth) for the posterior distribution of $\alpha$ is given in Fig. 3 for the dice snake data (left panel) and the flare stars data (right panel). To find the posterior distribution of $N$ we use (6) for the geometric pmf, i.e. the transformation

$$\hat{N} = n + (n - f_1) \frac{\theta}{1 - \theta - \theta(1 - \theta)}.$$

Note that the latter is a monotone increasing function in the interval (0, 1). The associated values for the median as well as for the 95% HPD credible interval are given in Table 9.

As an alternative we might consider a Bayesian analysis using Jeffreys’ invariance prior. This leads here to a prior distribution proportional to $1/\sqrt{1 - \theta}$.
which corresponds to an improper beta prior. The posterior distribution is proportional to

$$h^{n-1} (1 - h)^{s-1/2},$$

which corresponds to a beta distribution with parameters $a = n$ and $b = s + 1/2$ where $s = \sum_i (x_i - 2)$. The corresponding values for median, 0.025-quantile, and 0.975-quantile of this posterior are provided in Table 10.

### 7 Monte Carlo simulation

To allow a comparison on the performance of all suggested estimators we are including some simulation work to provide a more in-depth study of these suggestions. We have considered the following population sizes for $N$: 50, 100, 200, 500, 1000. We did not consider any sizes larger than 1000 as this deemed not appropriate for our setting and also most differences between estimators can be expected for smaller sizes. We have chosen the geometric distribution as baseline distribution with parameter values $\theta = 0.3, 0.4, 0.5$. We studied three settings: no,
10% and 30% one-inflation. N units were sampled under the respective setting and zero-counts removed. Then six estimators were considered: the modified Chao estimator (1) discussed in Böhning et al. (2019), the estimator based on the conditional likelihood with no one-inflation (2), the estimator based on the conditional likelihood with one-inflation modelled (3), the estimator based on the unconditional likelihood with no one-inflation (4), the estimator based on the unconditional likelihood with one-inflation modelled (5) and the Bayes estimator using Jeffreys’ prior (6).

We like to provide results as relative bias and relative standard deviation defined as

\[ rb = \left( \frac{1}{R} \sum_{r=1}^{R} \frac{\hat{N}_r - N}{N} \right) \]  

and

\[ rsd = \sqrt{\frac{1}{R} \sum_{r=1}^{R} \left( \frac{\hat{N}_r - N}{\hat{N}} \right)^2} \]

respectively. Here \( R \) is the number of replications and \( \hat{N}_r \) is the estimate of interest in the \( r \)–th simulation run while \( \hat{N} \) is the mean estimate of interest over all simulation runs. These relative definition forms are required to allow comparisons across different population sizes and also to permit meaningful asymptotic statements. Note that the usual relationship \( rb^2 + rsd^2 = rmse \) holds where

\[ rmse = \frac{1}{R} \sum_{r=1}^{R} \left( \frac{\hat{N}_r - N}{N} \right)^2 / N^2 \]

is the relative mean squared error.

The results for \( R = 1000 \) are presented visually in Figs. 4, 5 and 6 and show a clear picture. Estimators (2) and (4) show high overestimation bias under one-inflation, all other estimators behave reasonable in all settings with respect to bias and appear to be asymptotically unbiased. The modified Chao estimator shows larger variance than the conditional and unconditional estimators as well as the Bayes estimator. However, it should be kept in mind that the modified Chao estimator does allow heterogeneity under the geometric sampling distribution. In summary, the unconditional and Bayes estimator seem to perform best among the considered estimators.

8 Variance and bootstrap

For finding a variance and confidence interval estimate of the population size estimate under the zero-truncated one-inflated model we use the bootstrap approach. The conventional, nonparametric bootstrap works as follows:
1. Draw a sample of size \(N\) from the observed distribution defined by the relative frequencies \(f_0/N, f_1/N, \ldots, f_m/N\).

2. Derive \(\hat{\theta}\) and \(\hat{N}\) for the bootstrap sample in 1).

3. Repeat step 1. and 2. \(B\) times, leading to a sample of estimates \(\hat{N}^{(1)}, \ldots, \hat{N}^{(B)}\).

4. Calculate the bootstrap standard error as
Fig. 5 Relative bias (upper panel) and relative standard deviation (lower panel) of estimators of $N$ for the setting with 10% 1-inflation: 1 = modified Chao, 2 = CMLE no 1-inflation, 3 = CMLE under 1-inflation, 4 = UMLE no 1-inflation, 5 = UMLE under 1-inflation, 6 = Bayes with Jeffreys’ prior under 1-inflation

$$SE^* = \sqrt{\frac{1}{B} \sum_{b=1}^{B} (\hat{N}^{(b)} - \overline{N}^*)^2},$$

where $\overline{N}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{N}^{(b)}$. 
The problem with this bootstrap algorithm is that neither $f_0$ nor $N$ are known. This has been acknowledged in the capture–recapture community for some time. Norris and Pollock (1996) suggest three bootstrap methods to account for the uncertainty involved in estimating $N$. One method bases the bootstrap only on the observed sample size $n$ and estimates the remaining uncertainty analytically. Another method uses a bootstrap based on a complete modelling approach. The third method they
suggest is also used here. This method is also discussed favorably in Anan et al. (2017).

We call this method a *semi-parametric* bootstrap and it can be described as follows:

1. Draw a sample of size $||\hat{N}||$ from the observed distribution defined by the relative frequencies $\hat{f}_0/\hat{N}, \hat{f}_1/\hat{N}, \ldots, \hat{f}_m/\hat{N}$. Here $||x||$ denotes the rounding of $x$ to the nearest integer.
2. Derive $\hat{\theta}$ and $\hat{N}$ for the bootstrap sample in 1).
3. Repeat step 1) and 2) $B$ times, leading to a sample of estimates $\hat{N}^{(1)}, \ldots, \hat{N}^{(B)}$.
4. Calculate the bootstrap standard error as

$$SE^* = \sqrt{\text{median} \{ R^{(b)} \mid b = 1, \ldots, B \}},$$

where $R^{(b)} = (\hat{N}^{(b)} - \overline{N}^*)^2$ for $b = 1, \ldots, B$ and now with $\overline{N}^* = \text{median} \{ \hat{N}^{(b)} \mid b = 1, \ldots, B \}$.

We call this bootstrap *semi-parametric* as it is *non-parametric* conditional on $||\hat{N}||$ and *parametric* as it uses the estimated model to find $\hat{N}$. Note that we have chosen a robust estimator for the mean and for the variance.

We now apply this bootstrap procedure to all estimators studied in the simulation work of the previous section. These are the modified Chao estimator (1) discussed in Böhning et al. (2019), the estimator based on the conditional likelihood with no one-inflation (2), the estimator based on the conditional likelihood with one-inflation modelled (3), the estimator based on the unconditional likelihood with no one-inflation (4), the estimator based on the unconditional likelihood with one-inflation (5), and the Bayes estimator (6).

| Case study | n  | Estimator       | $\hat{N}$ | BTmean | BTmed | BTsd | BTse (8) | 95% BTciv |
|------------|----|-----------------|----------|--------|-------|------|---------|-----------|
| Dice snakes| 70 | Mod. Chao (1)   | 126      | 201.5  | 129.0 | 198.6| 47.0    | 72–735    |
|            |    | Geometric (2)   | 358      | 383.9  | 362.2 | 120.7| 68.9    | 218–682   |
|            |    | 1inf-geom (3)   | 127      | 213.8  | 129.4 | 319.6| 33.5    | 79–948    |
|            |    | Marg. geom (4)  | 356      | 380.5  | 357.0 | 118.7| 67.5    | 220–680   |
|            |    | Marg. 1inf. geom (5) | 115    | 175.9  | 117.0 | 179.7| 27.0    | 76–958    |
|            |    | Bayes (6)       | 122      | 182.8  | 123.8 | 205.7| 30.4    | 78–713    |
| Flare stars| 145| Mod. Chao (1)   | 425      | 749.8  | 439.0 | 883.7| 200.0   | 169–3552  |
|            |    | Geometric (2)   | 671      | 716.3  | 684.5 | 185.3| 108.5   | 455–1160  |
|            |    | 1inf-geom (3)   | 205      | 297.1  | 209.8 | 597.0| 30.8    | 160–809   |
|            |    | Marg. geom (4)  | 668      | 713.0  | 683.0 | 181.9| 107.0   | 450–1157  |
|            |    | Marg. 1inf. geom (5) | 198    | 256.9  | 203.0 | 196.5| 28.0    | 158–776   |
|            |    | Bayes (6)       | 203      | 278.2  | 207.2 | 442.1| 29.5    | 159–766   |
inflation modelled (5) and the Bayes estimator using Jeffreys’ prior (6). The results of the bootstrap procedure are provided in Table 11. Due to the small sample size, the confidence intervals are rather wide. The upper interval end provides for both case studies valuable information on an upper bound for the hidden population units. Due to the sparsity of the data the bootstrap samples generate occasionally very large population size estimates. Typical, we would expect the bootstrap mean to be close to the population size estimate. However, this is not the case due to the occasional occurrence of spurious large size estimates. The bootstrap median does get close to the sample population size estimate. This aspect is of interest in practice as we might want to check if the bootstrap median is in agreement with the population size estimate for the given sample as this could indicate that the latter is spurious. It can be expected that also the conventional bootstrap standard deviation experiences a similar inflation and estimating the true variation by means of (8) is likely more useful. Note that the ranking of estimators according to BTse (8) is in line with the results of the simulation study. We can ignore the two estimators under no inflation as these are using a wrong model which contributes to their large variance in this case. Under the remaining estimators, Chao’s modified estimator has by far the largest standard error. Jeffreys’ Bayes estimator and the conditional under one-inflation are on par whereas the unconditional under one-inflation seems to perform best.

9 Discussion and conclusion

It is widely known that parameter heterogeneity which is not accounted for in the modelling can lead to severe bias in the estimation of population size. However, it is mostly assumed that the bias occurs in a form of underestimation. IWGDMF (1995) provide a generic argument for this fact. In particular, this is justified in zero-truncated count models as any heterogeneity which can be modelled as a mixture of parametric densities leads to an underestimation bias if the mixture is ignored and only a homogeneous model is fitted (Böhning and Schön 2005, van der Heijden et al. 2003). Here we have seen that in the case of one-inflation heterogeneity serious overestimation of population size can occur. This is particularly disturbing if Chao’s lower bound estimator (Chao 1987, 1989) is used which seemingly provides a lower bound estimate for population size whereas under one-inflation the opposite is true.

We have seen that under sparsity modelling of the remaining counts (after truncating inflated counts) is crucial for the predictive value. Of course, having a well-fitting model for the observed data does not automatically apply it is also a good fit for the unobserved part as the model might not be valid for this part. This assumption needs to be made and it is untestable given the data constellation for this paper. The inclusion of covariates (if available) will always help to improve the fit of the model and increase the likelihood of valid predictions of population size. This aspect needs to be investigated in future research.

All estimation methods provide similar results. The modified Chao estimator is least favourable as its standard error is relatively large when compared to the others,
but has the benefit of avoiding distributional assumptions. The unconditional and the Bayesian approach both seem to perform better than the conditional one. Most important is, and this cannot be emphasized enough, that one-inflation is not ignored, as, if it is, it leads not only to large bias in the estimate but also inflates standard errors considerably.

Clearly, due to the small sample sizes, confidence intervals based on profile log-likelihoods are rather wide, but we like to notice, however, that standard errors of estimators based on the appropriate one-inflated model are remarkably smaller than those ignoring one-inflation (see Table 11).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Akopian A (2018) Personal communication
Ambartsumyan VA, Mirzoyan LV, Parsamyan ES, Chavushyan OS, Erastova LK (1970) Flare stars in the Pleiades. Astrofizika 6(1):7–30
Anan O, Böhning D, Maruotti A (2017) Uncertainty estimation in heterogeneous capture–recapture count data. J Stat Comput Simul 87(10):2094–2114
Böhning D (2016) Ratio plot and ratio regression with applications to social and medical sciences. Stat Sci 31:205–218
Böhning D, Schön D (2005) Nonparametric maximum likelihood estimation of population size based on the counting distribution. J R Stat Soc Ser C 54:721–737
Böhning D, van der Heijden PGM (2019) The identity of the zero-truncated, one-inflated likelihood and the zero-one-truncated likelihood for general count densities with an application to drink-driving in Britain. Ann Appl Stat 13:1198–1211
Böhning D, Ogden HE (2020) General inflation models for count data. Metrika. https://doi.org/10.1007/s00184-020-00786-y
Böhning D, Baksh MF, Lerdauansri R, Gallagher J (2013) The use of the ratio-plot in capture–recapture estimation. J Comput Graph Stat 22:133–155
Böhning D, van der Heijden PGM, Bunge J (2018) Capture–recapture methods for the social and medical sciences. Chapman & Hall/CRC, Boca Raton
Böhning D, Kaskasamkul P, van der Heijden PGM (2019) A modification of Chao’s lower bound estimator in the case of one-inflation. Metrika 82:361–384
Borchers DL, Buckland ST, Zucchini W (2004) Estimating animal abundance. Closed populations. Springer, London
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, Bunge J (2002) Estimating the number of species in a stochastic abundance model. Biometrics 58:531–539
Cormack RM (1992) Interval estimation for mark-recapture studies of closed populations. Biometrics 48:567–576
Population size estimation based upon zero-truncated...