Stockholms universitet

A zero-truncated one-inflated model with application to population monitoring

Herman Persson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2022:12
Matematisk statistik
Juni 2022

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2022:12**
http://www.math.su.se

A zero-truncated one-inflated model with application to population monitoring

Herman Persson*

June 2022

Abstract

When one-inflation in data arises due to samples from individuals being misidentified as samples from non-existent with probability p, a large bias arises in the population estimate if the inflation is not taken into account. This inflation causes a greater bias compared to previously analysed inflation where it arises due to individuals with some probability p_B succeed in deviating from being observed more than once (Böhning and Heijden (2019), Godwin (2017)). By using distributions that take into account that data contains one-inflation of the type miss identification, we can reduce the bias. With the same parameters in a base distribution and $p = p_B$, inflation that arises due to incorrect identification is always expected to give greater bias and variance on the population estimate than the corresponding behavior caused inflation population estimate. When we apply our models to brown bear data from the department of Environmental Research and Monitoring at the Swedish Museum of Natural History we see no evidence of one-inflation and note that further analysis regarding individual heterogeneity of the bears is required for a reliable population estimate.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: herman.olaspers@gmail.com. Supervisor: Martin Sköld.

Contents

Introduction **2**

Method **4**

Population estimation 4

Poisson model 5

Negative binomial model 7

Bear data 8

Simulations 8

Results **10**

Simulations 10

Application on bear data 12

Discussion **14**

Appendix **17**

(1) Derivation of ω 18

(2) Relative bias of mean population estimate 19

(3) Log-bias of NBin population estimates 20

(4) Histogram of $\hat{p}$ 21

(5) Comparison of base distribution NBin and Po 22

References **23**

Introduction

To draw conclusions about a populations size is an important problem in animal monitoring as well as other areas where the amount of information about the target group's behavior is limited. A common method is capture-recapture where a part the population is observed as well as identified and one with the help of the observed individuals attempt to estimate the total underlying population size. The problem can be summarized as wanting to estimate the size of the total population N, which is the sum of the number of observed individuals n and the number of unobserved individuals n_0. Since n is known from data, the problem is reduced to finding n_0. A common approach to yield inference about n_0 is to use the available data to estimate the parameters of the zero-truncated distribution which corresponds to the distribution of number of samples per individual. Using this approach there are several different methods for dealing with problems such as individual heterogeneity and contamination, e.g. Chao (1987) and Zelterman (1988), but as shown in Godwin and Böhning (2017), they fall very biased in the presence of one-inflation. One-inflation in data means that, as e.g. post-collection error or due to a behavioral change in the population, extra ones occur compared to the underlying distribution. The special case of one-inflation have been discussed in previous mathematical reports, methods and models for dealing with the problem have been developed. In Godwin and Böhning (2017) inflation occur as a consequence of behavioral change as some individuals change their behavior after being observed once and succeed in avoidance from further observations. In the report we see how one, with the probability of this behavioral change p_B, can adjust an underlying poisson distribution to include these extra ones. A more general variant of the method is developed in Böhning and Heijden (2019) where the one-inflated distribution for an arbitrary base probability density function (PDF) f with parameter $\theta = (\theta_1, ..., \theta_i)$ and corresponding zero-truncated PDF f_+ is shown to follow

$$\begin{cases} (1 - p_B) + p_B f_+(x, \theta) & \text{for } x = 1, \\ p_B f_+(x, \theta) & \text{for } x > 1. \end{cases} \tag{1}$$

In this report we will look at the previously untreated one-inflation which occur

when a sample from an individual with some probability p is incorrectly identified as a non-existent individual (ghost). Ghost inflation in data could occur due to genotyping errors or if photo identification is used. The case of ghost inflation differs from behavior inflation in a number of ways. One difference which is that the extra ones generated by ghost inflation does not represent real individuals, but in in behavior inflation they do. A second main difference is that the base distribution is affected by the inflation as the ghosts in data occurs at the expense of observations of real individuals with the same probability for all individuals, in contrast to behavior inflation where individuals who does not change their behavior is unaffected by the inflation. A third main difference is that the extra number of ones generated by individual i in ghost inflation is $p \cdot E[Y_i]$, where Y_i is the random variable which corresponds to the number of observations of said individual, where as in behavior inflation one individual only can contribute with one extra one. Because of the differences between ghost and behavioral inflation, the problems are obviously not equivalent, but as we will see later in this report, they are closely related.

The purpose of developing the models in this report is to apply them to data regarding the brown bear population in Sweden. Regions of Sweden which are inhabited by brown bears are divided into four parts which all have been monitored by the department of Environmental Research and Monitoring at the Swedish Museum of Natural History (NRM) since 2015. The bear population is surveyed in each region with five year intervals (one region each year and one year without survey). During the survey hunters in the region are asked to collect scat-samples. These samples are sent to the NRM, where samples are genotyped and stored in a database of observed individuals.

Method

We will assume that the population is closed during an ongoing inventory, which means that no individuals die, are born, move in or out. Our goal is to estimate the population size denoted N which then is a constant. We assume that data consists of one or more times observed individuals and the number of times each individual was observed. This means that our goal is reduced to estimating the number of unobserved individuals n_0, which is assumed to be unknown. One can summarize the most common method of estimating the population in this kind of setting (known as a capture recapture) as:

1. Fit observation data to its corresponding zero-truncated distribution f_+ to estimate its parameter/parameters θ
2. Estimate N and n_0 with the non-zero-truncated variant of the distribution f and the estimation of θ, denoted $\hat{\theta}$, from step 1.

The approach can be considered relatively simple, but problems may arise along the way, for example, for small sample sizes it can be difficult to get a reliable estimate of θ and it is not always obvious from which distribution data originates. Fitting data to an incorrect distribution can lead to devastating consequences as small differences in the base distribution can lead to large differences regarding inference about the population.

Population estimation

To estimate the population in a capture-recapture setting a variant of the Horvitz–Thompson estimator (Horvitz and Thompson (1952)) is commonly used. The modified Horvitz–Thompson estimator is $\hat{N}_{HT} = n/(1 - f(0, \hat{\theta}))$, where $\hat{\theta}$ is the estimated distribution parameter and n is the number of observed individuals. The estimator can be motivated by noting that $N = E[n]/(1 - f(0, \theta))$. The estimator is often suitable as it is asymptotically unbiased under the condition that $\hat{\theta}$ is unbiased. However, if one-inflation in data is not taken into account $\hat{N}_{HT}$ is strictly biased.

In Böhning and Heijden (2019) it is noted that in a similar way that $\hat{N}_{HT}$ can estimate the number of individuals that have been observed zero times, it can be used to estimate the number of individuals that have been observed once. Thus, if we suspect that the number of individuals observed once is inflated, we can filter the ones out of the data and try to estimate the correct number of individuals observed as

$$\frac{n - n_1}{1 - f(1, \hat{\theta})}. \tag{2}$$

Here n_1 denotes the total number of ones in data. If we now combine (2) with $\hat{N}_{HT}$, we get an estimator for the entire population where the number of ones and zeroes are estimated as per

$$\hat{N} = \frac{n - n_1}{1 - f(0, \hat{\theta}) - f(1, \hat{\theta})}.$$

In Böhning and Heijden (2019) the estimator is further developed to count the extra ones, but in our case this is not desired as the extra ones does not represent real individuals. The estimator $\hat{N}$ is thus useful for both our zero-truncated one-inflated and zero-one-truncated models, that we shall derive later in this report. Hence, $\hat{N}$ will be used to estimate the population for these models in this report.

Poisson model

There are many different ways to model the number of samples per individual. A common starting point in count data, where we will also begin, is to use a Poisson distributed number of samples per individual with no individual heterogeneity. Under this distribution, the number of samples from one individual is assumed to be independent of other individuals and the number of samples is assumed to be equally distributed for all individuals. We then have that the number of samples per individual is IID for all individuals according to $Y_i \sim \text{Po}(\lambda)$. However, since our observed data will not contain the number of individuals for which $Y_i = 0$, the observed data will follow the zero-truncated Poisson distribution (ZTP, also called positive Poisson). If we now introduce one-inflation by assuming that each samples from all individuals with some probability $p \in (0, 1)$ is incorrectly identified as a non-existent non-previously observed individual (ghost), then the number of observation per individual will be distributed according to $\text{Bin}(Y_i, 1 - p)$ and as we are unable to distinguish real individuals from ghost individuals, our collected data will include $Y_i - \text{Bin}(Y_i, 1 - p)$ ghost observations. Let us denote the PDF of the poisson distribution as $p(x, \lambda)$, the number of individuals observed k times as n_k and let

$$\mathrm{I}_a(b) = \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{if } a \neq b. \end{cases}$$

Then the distribution for the number of observations in the zero-truncated one-inflated poisson (ZTOIP) distribution can be found by first noting that

$$\begin{aligned} E[n_1] &= E[\mathrm{Bin}\Big(\sum_{i=1}^{N} \mathrm{I}_1(Y_i), 1-p\Big)] + E[\mathrm{Bin}\Big(\sum_{i=1}^{N} Y_i, p\Big)] \\ &= (1-p)E\Big[\sum_{i=1}^{N} \mathrm{I}_1(Y_i)\Big] + pE\Big[\sum_{i=1}^{N} Y_i\Big] \\ &= N(1-p)E[\mathrm{I}_1(Y_i)] + NpE[Y_i] \\ &= (1-p)Np(1,\lambda) + Np\lambda \\ &= Np(1, \lambda(1-p)) + Np\lambda, \\ E[n_k] &= Np(k, \lambda(1-p)), \qquad\qquad \text{for } k = 2, 3, ... \end{aligned}$$

With this result we get the one-inflated zero-truncated poisson PDF denoted as $p_{+1}(x, \lambda, p)$ with

$$\begin{aligned} p_{+1}(1, \lambda, p) &= \frac{E[n_1]}{\sum_{k=1}^{\infty} E[n_k]} \\ &= \frac{Np(1, \lambda(1-p)) + Np\lambda}{N[\sum_{k=1}^{\infty} p(k, \lambda(1-p))] + Np\lambda} \\ &= \frac{p(1, \lambda(1-p)) + p\lambda}{1 - p(0, \lambda(1-p)) + p\lambda}, \\ p_{+1}(k, \lambda, p) &= \frac{p(k, \lambda(1-p))}{1 - p(0, \lambda(1-p)) + p\lambda}, \qquad\qquad \text{for } k = 2, 3, ... \end{aligned}$$

Which as shown in Appendix (1) can be rewritten to resemble the distribution of previously analyzed inflation in i.a. Böhning and Heijden (2019) with the help of the zero-truncated poisson PDF p_+ and $\omega = 1/(1 + p\lambda/(1 - p(0, \lambda)))$ as

$$p_{+1}(x, \lambda, p) = \begin{cases} (1-\omega) + \omega p_+(x, \lambda(1-p)) & \text{for } x = 1, \\ \omega p_+(x, \lambda(1-p)) & \text{for } x = 2, 3, ... \end{cases}$$

The ZTOIP PDF p_{+1} can be used to carry out the first step in estimating population size, i.e. to estimate the parameters of the underlying distribution, which in this case is assumed to be the Poisson distribution. However, as we will see later in this report, the ZTOIP model leads to bias in the estimation of p ($\hat{p}$)

when the real value is low. This in turn leads to a bias in the population estimate. One way to avoid this bias is by using a zero-one-truncated distribution, in this case the so-called zero-one-truncated Poisson (ZOTP) distribution. The PDF of the ZOTP distribution is

$$p_{++}(x,\lambda) = \frac{p(x,\lambda)}{1-p(0,\lambda)-p(1,\lambda)} = \frac{\lambda^x}{(e^\lambda-\lambda-1)x!}.$$

As we know from the derivation of p_{+1}, we must adjust the distribution parameter to take inflation into account. Hence, the parameter of p_{++} when one-inflation is present is $\lambda(1-p)$. Note that if we use the ZOTP distribution to estimate $\lambda(1-p)$, we do not need to know the exact value of either λ or p to estimate the size of the entire population as it is not necessary in $\hat{N}$. We will also use the ZTP distribution to estimate the population. When we use this distribution, we will ignore inflation in the data and look at the effect of in the population estimate. The density function for the ZTP distribution is

$$p_{+}(x,\lambda) = \frac{p(x,\lambda)}{1-p(0,\lambda)} = \frac{\lambda^x}{(e^\lambda-1)x!}.$$

Negative binomial model

All individuals being observed with the same probability is rarely the case. We are therefore very interested in introducing individual heterogeneity by allowing individual probability of observation. One way of introducing individual observation probability is by expanding our previous Poisson base distribution by letting the parameter for each individual be determined by a random variable. We choose to let the distribution parameter for individual i denoted as λ_i be distributed according to a Gamma distribution, more precisely $\lambda_i \sim \Gamma(k\lambda, k)$ where λ_i is a random variable, λ is the expected value. In this distribution we are able to adjust for over dispersion by adjusting k. We can also introduce inflation as $\lambda_i(1-p) \sim \Gamma(k\lambda, k/(1-p))$. In this distribution $E[\lambda_i(1-p)] = \lambda(1-p)$, $Var[\lambda_i(1-p)] = \lambda(1-p)^2/k$. Note that the distribution collapses to $\mathrm{Po}(\lambda_i(1-p))$ as $k \to \infty$. By using the Gamma distribution to determine λ_i we bring pleasant properties as we know from e.g. Wikipedia (2022) that for an arbitrary random variable X it holds that

$$X \sim \Gamma(r, \frac{p}{1-p}) \quad \Rightarrow \quad \mathrm{Po}(X) \overset{d}{=} \mathrm{NBin}(r,p).$$

Which when applied to our situation reveals that

$$\lambda(1-p) \sim \Gamma(k\lambda, \frac{k}{1-p}) \quad \Rightarrow \quad \mathrm{Po}(\lambda(1-p)) \overset{d}{=} \mathrm{NBin}(k\lambda, \frac{k}{1-p+k}). \quad (3)$$

This means that the number of observations per individual is negative binomial distributed. Note that we can allow all $k > 0$ by using the extended negative binomial distribution which extends the binomial coefficient to all real-values using the gamma function. Let us denote the PDF of our base distribution $\text{NBin}(k\lambda, k/(1-p+k))$ from (3) as $g(x, \lambda, k, p)$. Similarly to the poisson models we want to construct two different models for estimating the total population; zero-truncated one-inflated negative binomial (ZTOINB) and zero-one-truncated negative binomial (ZOTNB). In the ZOTNB model we get PDF

$$g_{++}(x, \lambda, k, p) = \frac{\frac{\Gamma(x+k\lambda)}{x!\Gamma(k\lambda)}\left(1 - \frac{k}{1-p+k}\right)^x}{\left(\frac{k}{1+k-p}\right)^{-k\lambda} - 1 - k\lambda\left(1 - \frac{k}{1+k-p}\right)}.$$

As shown in Appendix (1), in the case of the ZTOINB model the PDF can be constructed similarly to the ZTOIP model using the zero-truncated Gamma distribution g_+ with the addition of $\omega = 1/(1 + p\lambda/(1 - g(0, \lambda, k, p)))$ to adjust the extra mass at 1. Hence, we have the ZTOINB PDF

$$g_{+1}(x, \lambda, k, p) = \begin{cases} (1-\omega) + \omega g_+(x, \lambda, k, p) & \text{for } x = 1, \\ \omega g_+(x, \lambda, k, p) & \text{for } x = 2, 3, ... \end{cases}$$

Bear data

The distribution for the number of samples per individual each year can be seen in Figure 1. Note that the survey of 2015 and 2020 was done in the same region. In Figure 1 we see that most bears were observed only a few number of times and that the most common number of observations is one, for all five years. It is conceivable that some of these ones occur due to genotyping errors when identifying the bears which would lead to ghost inflation. An existing one-inflation in data will cause an overestimation of the population in data as it will increase the estimation of n_0. As we expect heterogeneity in the brown bear population we will investigate the suspected inflation by applying our negative binomial models to data which largely provides a good fit.

Simulations

All parameter estimates are calculated via the maximum likelihood method. Our zero-one-truncated and zero-truncated one-inflated models will both use $\hat{N}$ to estimate the population and our zero-truncated models, which does not take into account that data includes inflation, will use $\hat{N}_{HT}$. Since the likelihood functions of some of our models are unable to be maximized algebraically all

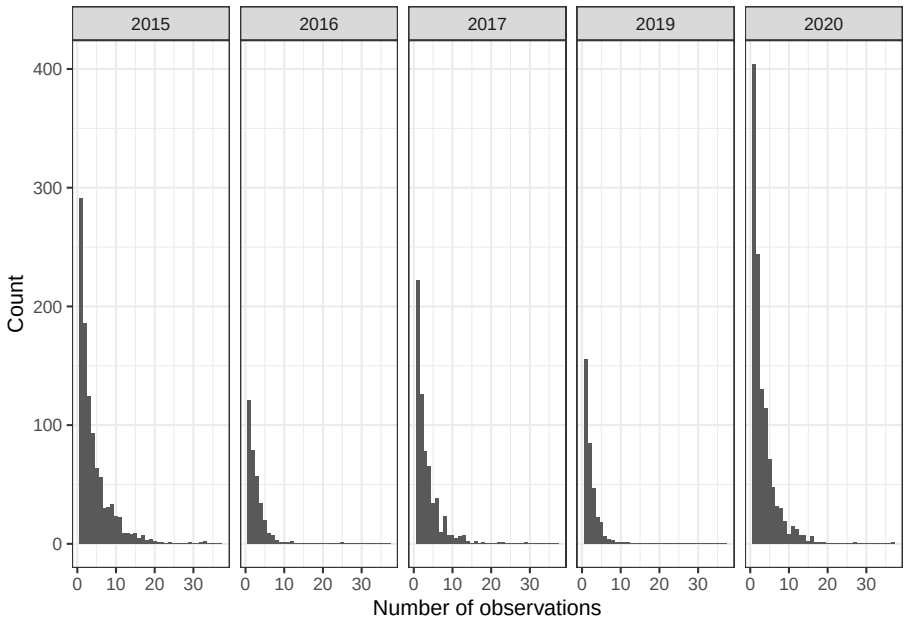


Figure 1: Histogram of number of samples per bear

likelihoods will be maximized numerically for the sake of comparability. An important note is the fact that the negative binomial model is rather unstable when using numerical maximization and therefore some non-existent trends might be visible for even very large simulations. A link to all the code used in the creation the simulations as well as a complete summary of the simulation results can be found in the Appendix.

Results

Since we want to see how well our different models estimate the population, a natural starting point is to use simulation study. We will start by looking at the consequences of not taking ghost inflation into account to then move on to our models that include inflation.

Simulations

To begin our analysis we simulate data from the ZTOIP distribution and estimate the population with the regular ZTP model using $\hat{N}_{HT}$, which does not take into account that data is inflated. Results of the simulation are shown in Figure 2. The mean percent error of the population estimate ($100 \cdot \hat{N}/N$) for each combination of λ, p and N is calculated from 100 simulations. A comparison with the bias in the case of behavior inflation from Godwin and Böhning (2017) can be seen in Table 1.

Table 1: Percentage bias of $\bar{N}_{ZTP}$ for diffrent inflation types as N = 500

p/p_B	λ	% bias of $\bar{N}_{ZTP}$	Type
0.1	1	6.2	Behavior
0.3	1	23.2	Behavior
0.1	2	3.0	Behavior
0.3	2	12.0	Behavior
0.1	1	27.6	Ghosts
0.3	1	121.4	Ghosts
0.1	2	32.6	Ghosts
0.3	2	137.9	Ghosts

To see how well our ZTOIP and ZOTP models are able to estimate the population we simulate data from the ZTOIP distribution 1000 times for each combination of three different values for N, λ and p. For each simulation the base

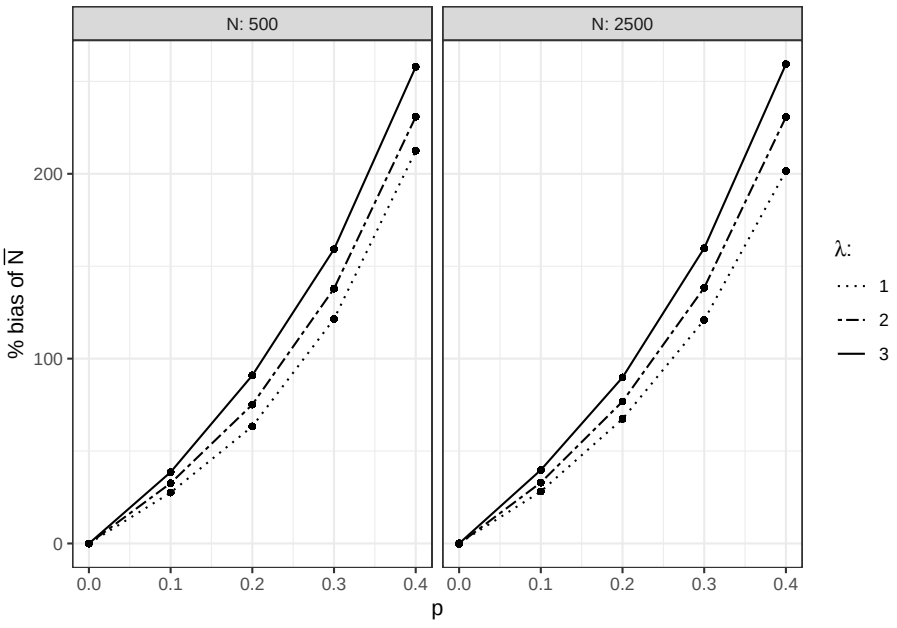


Figure 2: ZTP, relative bias of mean population estimate

distribution parameters are estimated using the ZTOIP, ZOTP and ZTP models and the population size is estimated with $\hat{N}_{HT}$ for ZTP and $\hat{N}$ for ZTOIP and ZOTP. To get an understanding of how the variance and bias differ between the ZTOIP and ZOTP population estimators we observe Table 2. In the table, together with the mean percent error, we see the root-mean-square error (RMSE) as well as 90% and 99% confidence intervals. The bounds of our two-sided confidence intervals are chosen from our population estimates and contains 90% and 99%, respectively, of the estimates.

As we expand our model and introduce individual heterogeneity in the negative binomial distribution the effects of the population estimate as λ, N and p varies are unaltered (Appendix (3)). Therefore, we will focus on the consequences on our population estimates as k varies. In Table 3 the mean percent error, RMSE and 90% and 99% confidence intervals can be seen for the ZOTNB population estimate of 1000 simulations repeated for different combinations of λ and k.

Table 2: ZTOIP, ZOTP and ZTP, confidence intervals and RMSE of pupulation estimate as $N = 500$ and $\lambda = 2$

Type	p	$\bar{N}$	90% CI	99% CI	RMSE
ZOTP	0.00	500	[450, 552]	[430, 583]	31
ZOTP	0.05	502	[446, 563]	[420, 584]	35
ZOTP	0.10	503	[445, 568]	[418, 595]	38
ZTOIP	0.00	493	[450, 526]	[430, 538]	24
ZTOIP	0.05	502	[446, 563]	[420, 584]	35
ZTOIP	0.10	503	[445, 568]	[418, 595]	38
ZTP	0.00	501	[482, 519]	[475, 525]	11
ZTP	0.05	578	[550, 604]	[542, 616]	80
ZTP	0.10	666	[627, 704]	[612, 721]	168

Table 3: ZOTNB, confidence intervals and RMSE of population estimate as $N = 500$ and $p = 0.1$

λ	k	$\bar{N}$	90% CI	99% CI	RMSE
1	0.2	115418	[142, 718583]	[118, 993213]	268513
1	0.6	83239	[193, 687465]	[153, 1035121]	249703
1	1.0	86196	[204, 705441]	[172, 1070459]	266781
1	2.0	62574	[241, 567445]	[212, 1144568]	235195
2	0.2	56004	[275, 468107]	[248, 1554165]	253450
2	0.6	4904	[342, 1060]	[311, 2342]	78061
2	1.0	2153	[369, 837]	[336, 1331]	50749
2	2.0	532	[381, 776]	[354, 1109]	190
3	0.2	3721	[354, 1026]	[332, 3346]	78915
3	0.6	513	[412, 656]	[391, 801]	84
3	1.0	509	[427, 626]	[404, 717]	63
3	2.0	504	[444, 583]	[428, 635]	43

Application on bear data

Before we apply our models to bear data it is important that we get an understanding of how well bear data fits a negative binomial model. We do this by creating QQ-plots where we compare the distribution of our bear data, as seen in Figure 1, with the theoretical quantiles of a zero-truncated negative binomial (ZTNB) distribution with parameters estimated by the ZOTNB model. The result can be seen in Figure 3. It is important to notice that the data which consists of all years combined (All) does not meet the model assumptions to the same degree as one year separately since some bears have been observed

during two different surveys and therefore have a much higher probability of being observed.

In Table 4 the estimated underlying distribution parameters from the ZTOINB distribution are shown together with the estimated population size by all of our three negative binomial models.

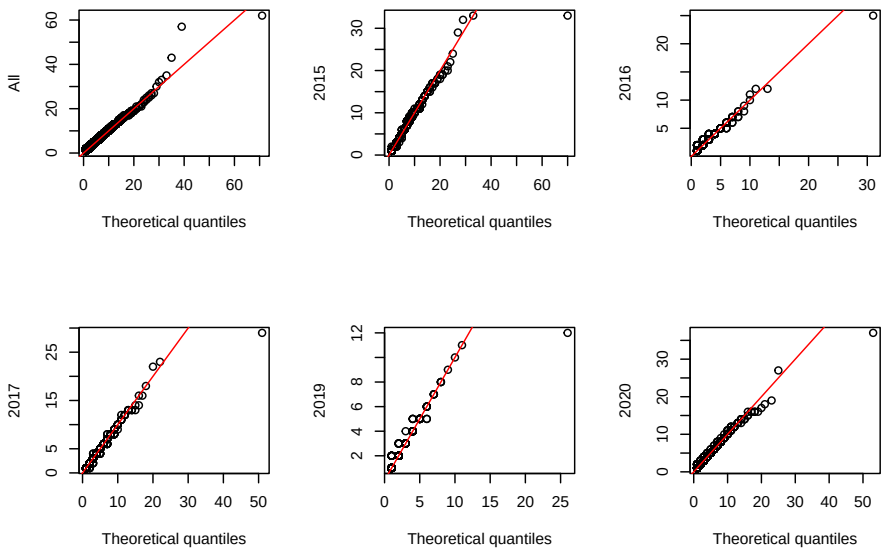


Figure 3: QQ-plot, bear data and theoretical NBin quantiles

Table 4: Population and samples per bear distribution paramter estimates

Year	n	$\hat{\lambda}_{g+1}$	$\hat{k}_{g+1}$	$\hat{p}_{g+1}$	$\hat{N}_{g+1}$	$\hat{N}_{g++}$	$\hat{N}_{g+}$
All	3076	1.821	0.189	0.000	6659	11484	6564
2015	1016	2.512	0.197	0.000	1722	1732	1722
2016	336	1.722	0.656	0.000	526	700	518
2017	636	1.818	0.280	0.004	1168	1168	1231
2019	344	1.158	0.711	0.000	672	829	668
2020	1154	1.641	0.265	0.000	2350	2713	2338

Discussion

In Godwin and Böhning (2017) where the inflation parameter p_B directly corresponds to the expected number of extra ones generated by one individual, the estimation of the population for the regular ZTP model is positively biased. From Table 1 where the bias of the ZTP model in our simulations is compared to the bias in Godwin and Böhning (2017) we can see that the two different types of inflation result in much different bias, more precisely the bias of our estimator when ghost inflation is present is much greater than for behavioral inflation. In the case of behavioral inflation the bias reduces as λ increase, which is not the case for ghost inflation where the opposite is true. The difference can be explained as in behavioral inflation, when λ increase, the expected number of extra ones generated by the inflation does not change, but $\hat{\lambda}$ will increase which makes $\hat{n}_0$ and the positive bias decrease. In the case of ghost inflation an increased λ also means an increased number of extra ones which will lead to negative bias in $\hat{\lambda}$. This negative effect dominates the positive effect on the bias previously described which causes $\hat{n_0}$ to increase and the positive bias in $\hat{N}$ increase. As can be seen in Figure 2 the relative bias of the estimator grows exponentially as p increase and the lower the value of λ the more does an increased population size reduce the bias. In Godwin and Böhning (2017) a similar result for behavioral inflation can be observed as an increased population size reduce the relative bias more for lower values of λ and the relative bias also grows exponentially as p increase.

As both the ZTOIP and ZOTP models in our modeling for ghost inflation use akin population estimators one could suspect their result to be similar. However, as can be seen in Table 2, their population estimates differ, especially for low values of p. Both of the models suffer positive bias as a consequence of high uncertainty for low values of λ and N (see Appendix (2)), but for low values of p the positive bias in the ZTOIP model is dominated by a negative bias. The negative bias can be explained as the ZTOIP model will estimate $p > 0$ in about half of the cases where $p = 0$ due to variance in the number of ones in data. The negative bias decrease as N increases since data then converge towards theoretical distribution and the relative variance in the number of ones decrease. One possible way to remove this bias would be to expand the model to allow deflation, i.e. $\mid p \mid < 1$, as this would create a symmetry in $\hat{p}$ for low values of p

which is currently missing (see Appendix (4)). In Godwin and Böhning (2017) deflation is allowed and as a result the ZTOIP population estimate does not suffer from negative bias. However, the method used is not directly applicable to our models and another approach is necessary in the case of ghost inflation. Due to the negative bias of the ZTOIP estimator, the ZOTP estimator is a better alternative for low values of p, and since we can see in Table 2 that ZTOIP and ZOTP converges towards each other for higher values of p, ZOTP looks to be the better estimator between the two. Its however important to mind the high bias of the ZOTP population estimator when λ is small. If we compare our ghost inflation estimator ZOTP to the behavior inflation estimator ZTOIP in Godwin and Böhning (2017) we see that the our estimate is much more sensitive to low population sizes, low values of λ and high inflation parameters and is always expected to giver higher bias and error. The difference is expected as the ZOTP estimator does not take into account the number of ones and ghost inflation means a greater number of ones and zeroes compared to behavioral inflation. This also implies that if one were to find an estimator in ghost inflation which allows deflation it is expected to give greater bias and error compared to the corresponding estimator in behavioral inflation.

As we expand our model to $NBin(k\lambda, k/(1-p+k))$ and introduce k to adjust for over dispersion we get a heavier tail on our base distribution compared to a poisson distribution with the same λ and p. As a result the variance of our population estimations increase. As can be observed from Table 3 the ZOTNB confidence intervals and RMSE is strictly greater than for the ZOTP estimator using the same λ and p in Table 2. The result is expected due to the increased variance of the negative binomial model. As can be seen in Appendix (3) and the full data of our ZTOINB simulations, for low values of λ and high values of k the ZTOINB model provide better estimates and should definitely be considered when choosing a model. A better understanding of the difference between the poisson and the negative binomial distributions can be obtained by studying the difference between the two distributions shown in Appendix (5). In Godwin (2017) the effect of behavioral inflation in the negative binomial model is examined and for $\lambda = 2$, $N = 500$ and $p = 0.1$ the ZTOINB population estimate when behavioral deflation is present (which allows deflation) has less bias and variance compared to our ZOTNB estimator for ghost inflation. The difference is explained analogously to that in the poisson model.

When estimating the bear population and distribution parameters for the number of observations, we can from Table 4 see that the inflation for all year except 2017 is estimated to be zero. Assuming that the negative binomial distribution provides a good fit with data this result go against the hypothesis that bear data include inflation. Based on QQ-plots in Figure 3, data seem to fit the negative binomial distribution quite well for lower values, however the distribution fits poorly on data for higher values. This is not be overlooked as the negative binomial distribution is very flexible and existing discrepancy is alarming. Therefore, to say that the data fits the distribution well would not be true. The problem of finding the correct model has proven to be very difficult in capture-recapture

and from Link (2003) we know that it is impossible to distinguish among reasonable models of heterogeneity (without additional information) even though they can yield very unlike inferences about population size. This makes the problem of finding a better fit for bear data very difficult and more information should be taken into account for reliable estimates.

It is possible that there often exists an understanding of how a possible inflation value p can arise and be distributed. In the case of genotyping errors this is especially true since there in many cases is possible to get a good understanding of what the probability error can be based on the method used (Pompanon et al. (2005)). If we prior to our study have an understanding of our inflation parameter p, priori distributions can be of use. In Tuoto, Di Cecco, and Tancredi (2022) the use of priori in behavior inflation is introduced and similar work would be interesting to apply to the case of ghost inflation.

A summary from our analysis is that ghost inflation compared with behavioral inflation always leads to more uncertain population estimates when $p = p_B$ and the same base distribution is used. As mentioned previously an improvement of zero-truncated one-inflated estimates can be done by expanding the model to include deflation and such a model would in most scenarios be ideal. However, this improvement does not mean that population estimate in ghost inflation will be as good as in behavior inflation as the expected bias and error still will be higher using the same base distribution and when $p = p_B$.

Appendix

All the code used in the production of this report can be found at https://github.com/herm4np/A-zero-truncated-one-inflated-model-with-application-to-population-monitoring.

(1) Derivation of ω

From the derivation of the ZTOIP distribution in the method chapter we find that

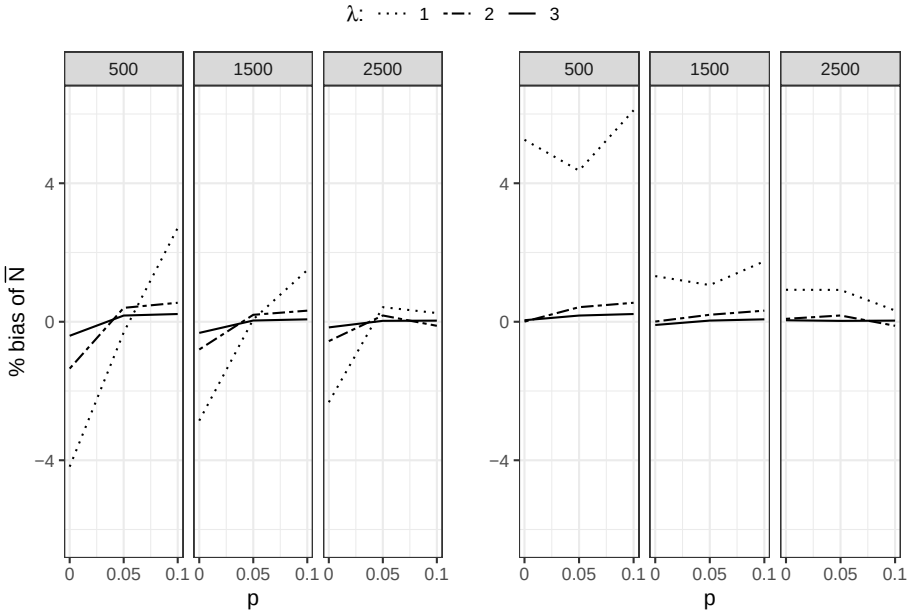
$$p_{+1}(1,\lambda,p) = \frac{E[n_1]}{\sum_{k=1}^{\infty} E[n_k]} = \frac{p(1,\lambda(1-p)) + p\lambda}{1 - p(0,\lambda(1-p)) + p\lambda},$$

which for an arbitrary base PDF f translates to

$$\begin{aligned}
f_{+1}(1) &= \frac{pE[Y_i] + (1-p)f(1)}{pE[Y_i] + \sum_{k=1}^{\infty}(1-p)f(k)} \\
&= \frac{pE[Y_i] + f(1,p)}{pE[Y_i] + (1 - f(0,p))} \\
&= \frac{pE[Y_i]}{1 - f(0,p) + pE[Y_i]} + \frac{f(1,p)}{1 - f(0,p) + pE[Y_i]} \\
&= 1 - \frac{1 - f(0,p) + pE[Y_i]}{1 - f(0,p) + pE[Y_i]} + \frac{pE[Y_i]}{1 - f(0,p) + pE[Y_i]} + \frac{\frac{f(1,p)}{1-f(0,p)}}{\frac{1-f(0,p)+pE[Y_i]}{1-f(0,p)}} \\
&= 1 - \frac{1 - f(0,p) + pE[Y_i]}{1 - f(0,p)} + \frac{f_+(1,p)}{1 - \frac{pE[Y_i]}{1-f(0,p)}} \\
&= \left(1 - \frac{1}{1 - \frac{pE[Y_i]}{1-f(0,p)}}\right) + \frac{1}{1 - \frac{pE[Y_i]}{1-f(0,p)}} f_+(1,p) \\
&= (1-\omega) + \omega f_+(1,p) \\
\Rightarrow \omega &= \frac{1}{1 - \frac{pE[Y_i]}{1-f(0,p)}}.
\end{aligned}$$

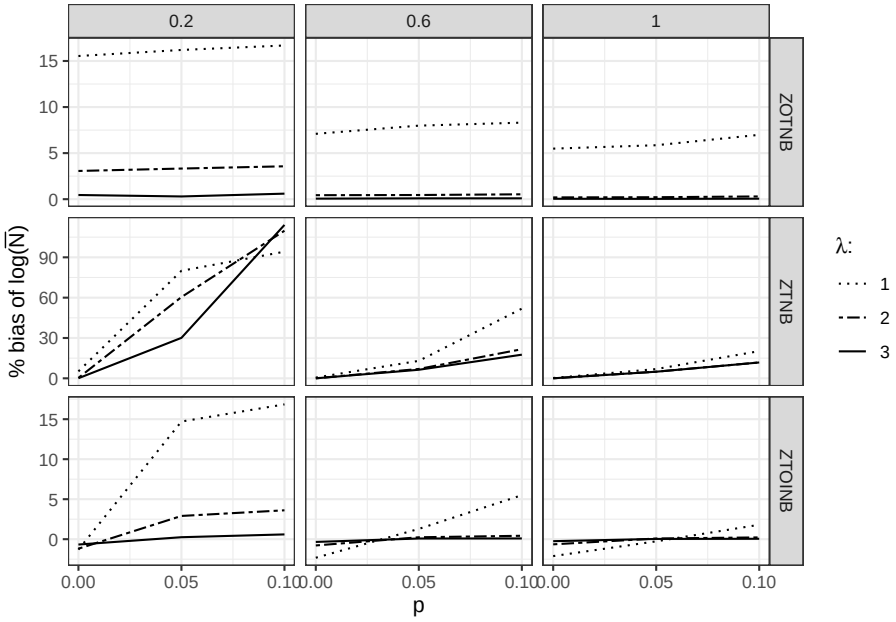
(2) Relative bias of mean population estimate

Faceted with respect to N. ZTOIP (left) and ZOTP (right).



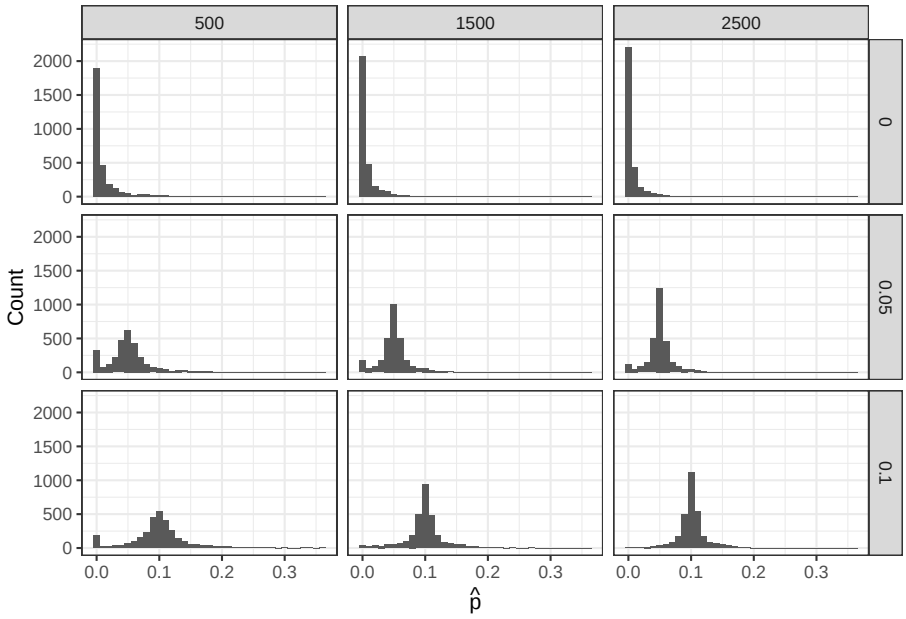
(3) Log-bias of NBin population estimates

Faceted horizontally with respect to k. Log for improved visibility.



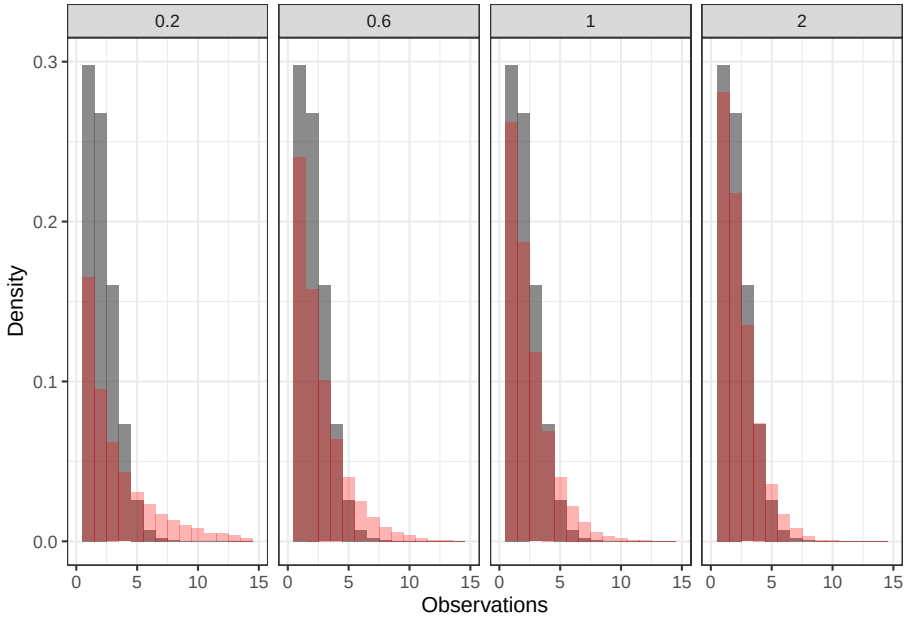
(4) Histogram of $\hat{p}$

Estimated by ZTOIP. Faceted horizontally with respect to N and vertically to p. $\lambda \in \{1, 2, 3\}$.



(5) Comparison of base distribution NBin and Po

Poisson distribution shown in grey, extra ones excluded, $\lambda = 2$ and $p = 0.1$. Faceted horizontally with respect to k.



References

Böhning, Dankmar, and Peter GM van der Heijden. 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." *The Annals of Applied Statistics* 13 (2): 1198–1211.

Chao, Anne. 1987. "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability." *Biometrics*, 783–91.

Godwin, Ryan T. 2017. "One-Inflation and Unobserved Heterogeneity in Population Size Estimation." *Biometrical Journal* 59 (1): 79–93.

Godwin, Ryan T, and Dankmar Böhning. 2017. "Estimation of the Population Size by Using the One-Inflated Positive Poisson Model." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 66 (2): 425–48.

Horvitz, Daniel G, and Donovan J Thompson. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47 (260): 663–85.

Link, William A. 2003. "Nonidentifiability of Population Size from Capture-Recapture Data with Heterogeneous Detection Probabilities." *Biometrics* 59 (4): 1123–30.

Pompanon, François, Aurélie Bonin, Eva Bellemain, and Pierre Taberlet. 2005. "Genotyping Errors: Causes, Consequences and Solutions." *Nature Reviews Genetics* 6 (11): 847–59.

Tuoto, Tiziana, Davide Di Cecco, and Andrea Tancredi. 2022. "Bayesian Analysis of One-Inflated Models for Elusive Population Size Estimation." *Biometrical Journal*.

Wikipedia. 2022. "Negative binomial distribution — Wikipedia, the Free Encyclopedia." http://en.wikipedia.org/w/index.php?title=Negative%20binomial%20distribution&oldid=1085162622.

Zelterman, Daniel. 1988. "Robust Estimation in Truncated Discrete Distributions with Application to Capture-Recapture Experiments." *Journal of Statistical Planning and Inference* 18 (2): 225–37.