EX

Herman Persson

null

Abstract

Introduktion

The problem of drawing statistical conclusions regarding the size of an unknown population is known and very relevant in animal monitoring. When trying to fit a model it is necessary that the model assumptions to great extent are correct for the population estimation to be reliable and unbiased. If existing one-inflation in data is not taken into account when choosing a model, it can have major consequences for the analysis's conclusions. One-inflation in data simply means that for some reason extra ones occur in data compared to what is epected from the model, for example, as a consequence of behavioral change of observed individuals or as a post-collection error.

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 inventorying is repeated with five year intervals (one region each year and one year without inventory). During the inventorying hunters in the region are assigned test equipment to collect scat-samples. These samples are then sent to the NRM and used for genetic identification. NRM then build database of bear individuals in the region. The distribution for the number of times each bear was identified at the NRM can be seen in Figure 1. Note that the inventory of 2015 and 2020 was done in the same region. In Figure 1 we see that most bears were observed just a few number of times and that the most common number of observations is one, for all five years. Of course just because the most common number of observations per bear is 1 it does not mean that it exist inflation, but it is conceivable that genotyping errors could have occured when identifying the bears. This would lead to some samples being misidentified and observation data to include one-time observed non-existent bears.

The problem of one-inflation in data has been investigated before in reports such as REFERENSSAKEN and REFERENSSAKEN. Even tough many of the methods used in the previous work prove useful in our case there is an essential difference. As an example, in REFERENSSAKEN Böhning and Heijden develop methods to estimate the total number of drunk drivers in Britain based on distribution for number of times each individual was caught by the police. The data of drunk drivers is naturally zero truncated and since most drivers who get caught by the police change their behavior, the data also includes one inflation. Since the extra number of ones is generated as a consequence of behavioral change in some individuals the inflation value p directly corresponds to the expected number of ones generated by one individual. In the bear data one-inflation occur as consequence of genotyping errors. This means that each individual can give rise to as many extra ones as the number of times the individual was observed, since all its scat-samples with some probability p might get incorrectly identified. This means that expected number of extra ones generated by each individual is p multiplied by the expected number of observations. The other big difference is that in our case the base distribution is effected by p, where base distribution is the distribution of data when excluding extra ones. If the REFERENSSAKEN disregards the extra ones, the remaining data follow the base distribution as the individuals whose behavior does not change are not affected by p. If we in our case disregarded the extra ones, the remaining data will not follow the base distribution as the extra ones were created at the expense of the number of observations of the real bears.

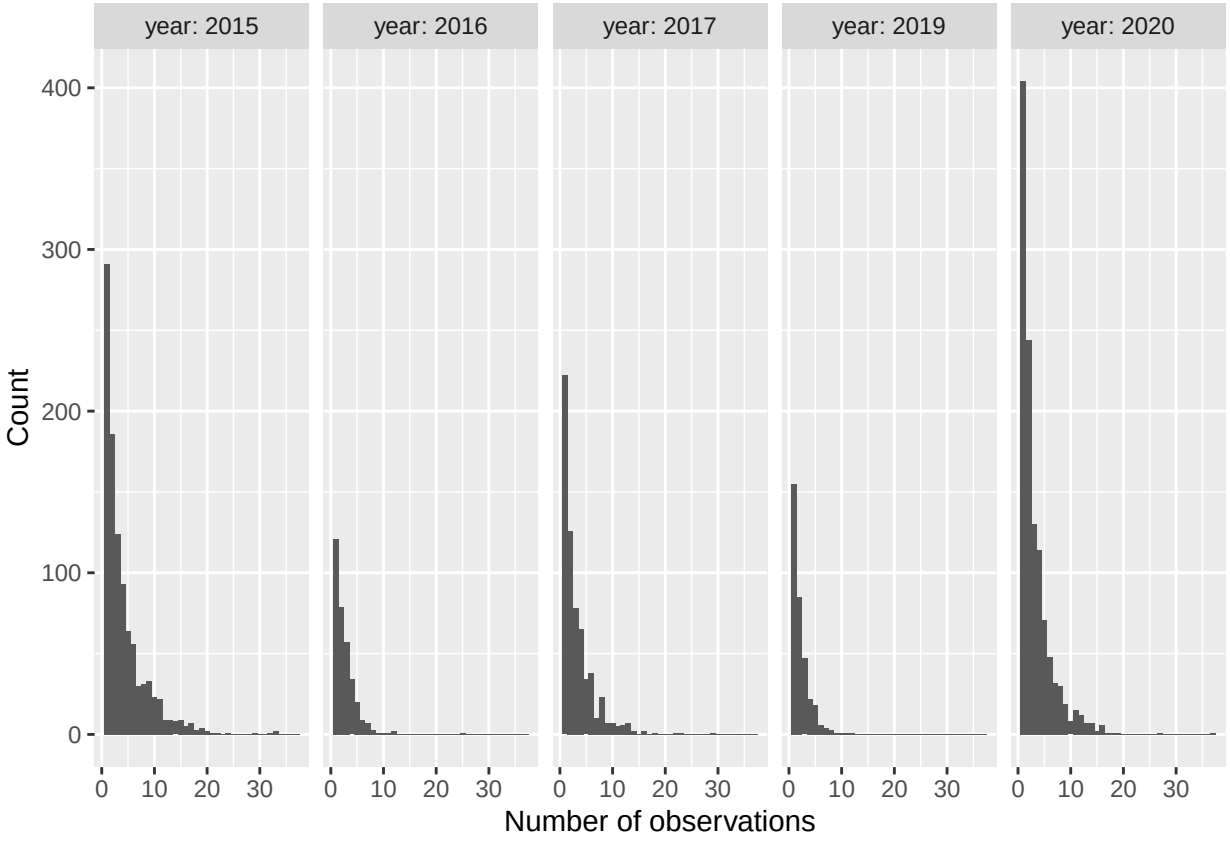

Figure 1: Histogram of number of observations

Method

We will assume that the bear population is closed during an ongoing inventory, ie that no individuals die, are born, move out or in. In this case the sample result will follow the vector $y = (n_1, n_2, ..., n_N)$, where n_i denotes the number of droppings from bear i and N is the size of the population. If then each collected scat-sample comes from a randomly selected bear, independent of the other bears, the vector y will follow a multinomial distribution with probabilities $p = (1/N, 1/N, ..., 1/N)$. Since the multinomial distribution is not desirable to work with, we approximate it with a poisson distribution REFERENCE. We then get that $n_i \sim Po(\lambda)$ with $\lambda = m/N$ where m is the total number of samples collected.

The poisson model

Suppose we have a population of size N from which we receive samples and that we are able to distinguish from which individual each sample is from. If there is no individual heterogeneity and the probability that a sample comes from one individual is the same for all individuals in the population, then the number of samples per individual of the observed individuals will be zero-truncated poisson distributed with probability mass function

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

Truncated distributions

Lets assume that for some reason a sample from one individual can be incorrectly identified as a sample from a not previously observed and non existent individual (ghost) with some probability p. For all p bigger than zero this would result in an increased number of observed ones, so called one-inflation, as well as a decreased number of expected observations per individual. To incorporate these conditions into we consider the new probability mass function of the zero-trunacted one-inflated poisson model (ZTOIP) REFERENSSAKEN. The distrubution for ZTOIP, denoted as p_{+1}, is

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

In the ZTOIP distrubution ω adjusts the extra mass at $x = 1$ and $\theta = \lambda(1-p)$ is adjusted density parameter with consideration to the reduced number of expected observation per individual. We can find ω by first making the observation

$$\begin{cases} E[f_1] = Np(1, \theta) + Np\lambda, \\ E[f_x] = Np(x, \theta), \quad x = 2, 3, ... \end{cases}$$

This means that for $x > 2$ we have

$$\begin{aligned} p_{+1}(x, \lambda, p) &= \frac{Np(x, \theta)}{Np\lambda + N\sum_{i=1}^{\infty} p(i, \theta)} = \frac{p(x, \theta)}{p\lambda + (1 - p(x, \theta))} \\ &= \frac{p(x, \theta)}{\big(1 - p(0, \theta)\big)\left(1 + \frac{p\lambda}{1 - p(0, \theta)}\right)} \\ &= \frac{p_+(x, \theta)}{1 + \frac{p\lambda}{1 - p(0, \theta)}}. \end{aligned} \tag{3}$$

If we now combine (2) and (3) we get

$$\omega p_+(x,\theta) = \frac{p_+(x,\theta)}{1+\frac{p\lambda}{1-p(0,\theta)}} \qquad \Rightarrow \qquad \omega = \frac{1}{1+\frac{p\lambda}{1-p(0,\theta)}}.$$

If we denote the number of observations with value i as f_i and the highest observed value as m we get the likelihood function

$$L_{p_{+1}}(\lambda,p,x) = [(1-\omega)+p_+(1,\theta)]^{f_1}\prod_{i=2}^{m}[\omega p_+(x_i,\theta)]^{f_i}.$$

If we are not necessarily interested in the exact values of p and λ another useful way to look at data would be to ignore all the individuals observed only once and use a zero-one-truncated poisson model (ZOTP). By using this model we will increase the uncertainty of our estimates but avoid some bias which can occur when using the ZTOIP model. In the case of the OTPP model we have the probability density function

$$p_{++}(x,\theta) = \frac{p(x,\theta)}{1-p(0,\theta)-p(1,\theta)} = \frac{\theta^x}{(e^\theta-\theta-1)x!}, \qquad x=2,3,...$$

with the corresponding likelihood function

$$L_{p_{++}}(\theta,x) = \prod_{i=2}^{m} p_{++}(x_i,\theta)^{f_i}.$$

The final poisson based distribution we will look at is zero-truncated poisson model (ZTP). By using this distribution we will be able to see the effect on population estimates when existing inflation is not taken into account and to see how well our other models compare when there is no inflation. The ZTP probability mass function can be seen in (1) and its corresponding likelihood is

$$L_{p_+}(\theta,x) = \prod_{i=1}^{m} p_+(x_i,\theta)^{f_i}.$$

Negative binomial model

By expanding our poisson model to a negative binomial model we will be able to account for some variance introduced by individual heterogeneity. We can introduce individual variance for individual i by letting the parameter of its poisson distrubution be distributed according to a Gamma distribution, more precicely

$$\theta_i = \lambda_i(1-p) \sim \Gamma(k\lambda, \frac{k}{1-p}).$$

This means that $E[\theta_i] = \lambda(1-p)$ and $Var[\theta_i] = \lambda(1-p)^2/k$ and we are able to adjust for over dispersion by adjusting k. We know from e.g. REFERERENSSAKAN that

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

With some short boring clculations we then get

$$\theta_i \sim \Gamma(k\lambda, \frac{k}{1-p}) \quad \Rightarrow \quad Po(\theta_i) \overset{d}{=} NBin(k\lambda, \frac{k}{1-p+k}). \tag{4}$$

Notice that we can allow all $k > 0$ by using the extended negative binomial distribution which extends the binomial coefficient to all real-values by using the gamma function.

Truncated distributions

Let us denote the probability density function of our base distribution in (4) as $g(x, \lambda, k, p)$. Similar to the poisson model case we are going to construct three different distributions for estimating the total population, the zero-truncated one-inflated negative binomial (ZTOINB), zero-one-truncated negative binomial (ZOTNB) and the zero-truncated negative binomial (ZTNB). In the case of the last mentioned ZTNB distribution we assume that $p = 0$ and get the probability density function

$$g_+(x, \lambda, k, p \mid p = 0) = \frac{\frac{\Gamma(x+k\lambda)}{x!\Gamma(k\lambda)}\left(1 - \frac{k}{1+k}\right)^x \left(\frac{k}{1+k}\right)^{k\lambda}}{1 - \left(\frac{k}{1+k}\right)^{k\lambda}}.$$

In the ZOTNB model we get

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

In the case of the ZTOINB distribution we similarly to the ZTOIP model use our corresponding zero truncated distribution with the addition of ω to adjust the extra mass at 1. Hence we have the ZTOINB probability density function

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

In this case, similar to the poisson model, $\omega = 1/(1 + p\lambda/(1 - g(0, \lambda, k, p)))$.

Population estimation

For each one of our models we want to create an estimate of the total population N, which includes all the unobserved individuals. We denote the number of observed individuals (ghosts included) as n and estimate the total population in the zero-truncated models which ignore inflation with the Horvitz–Thompson estimator

$$\hat{N}_{h_+} = \frac{n}{1 - h(0)}.$$

Where h is our base distribution and h_+ is the corresponding zero-truncated distribution which ignores inflation. A similar estimator for our zero-one-truncated distributions will then be

$$\hat{N}_{h_{++}} = \frac{n - f_1}{1 - h(0) - h(1)},$$

where f_1 as before is the number of individuals observed once. Note that $\hat{N}_{h_{++}}$ is an unbiased estimator but that $\hat{N}_{h_+}$ only is unbiased for $p = 0$. In the case of the zero-truncated one-inflated models we use the estimator

$$\hat{N}_{f_{+1}} = \frac{n - f_1 + \hat{n}_1}{1 - f(0)} \tag{5}$$

where $\hat{n}_1$ is the estimated number of real individuals observed once, ie not including ghosts. In our zero-truncated one-inflted distrubutions the proportion of real individuals observed once is $\omega h_+(1, \theta)$ and the proportion of ghosts is $1 - \omega$. Our estimator for $\hat{n}_1$ is therefor

$$\hat{n}_1 = f_1 \left(\frac{h_{+1}(1) - \mathbb{P}(\text{"ghost"})}{h_{+1}(1)} \right) = f_1 \frac{\hat{\omega} h_+(1)}{(1 - \hat{\omega}) + \hat{\omega} h_+(1)}$$

which combined with (5) then gives us the estimator of $N_{h_{+1}}$ as

$$\hat{N}_{h_{+1}} = \frac{n - f_1 \left(1 - \frac{\hat{\omega} h_+(1)}{(1-\hat{\omega}) + \hat{\omega} h_+(1)} \right)}{1 - h(0)}.$$

Note that under the assumption that our estimated number of true individuals observed once is unbiased, $\hat{N}_{f_{+1}}$ is an unbiased estimator.

Simulation

By doing a simulation study we will be able to analyse the bias of our estimators for different values of our model parameters. All estimated parameters will be calculated with the maximum likelihood method, and since some of the likelihoods cant be solved algebraically all likelihoods will be maximized numerically.

Resultat

Simultaions

Positive poisson model

To begin we simulate data from the ZTOIP distribution and estimate the population with the ZTP model, which does not take into account that data is inflated. In Figure 2 we can see the percent error of the population estimate ($100 \cdot \hat{N}/N$) mean from 20 (ska öka) simulations repeated for three different values of λ, four different p as well as two different values of N.

One truncated and inflated positive poisson models

To compare the population estimators of the ZTOIP and ZOTP we simulate 1000 times for each combination of three different values for N, λ and p. For each simulation the underlying population size is estimated with both the ZTOIP population estimator $N_{p_{+1}}$ and the ZOTP population estimator $N_{p_{++}}$. The mean percent error of both our estimators can be seen in Figure 3.

To get an understanding of how the variance and bias differ between the ZTOIP and ZOTP population estimators we take a look at table 1. In the table we can see the 90 and 99% confidence intervals of our simulated population estimates as well as sample standard deviation.

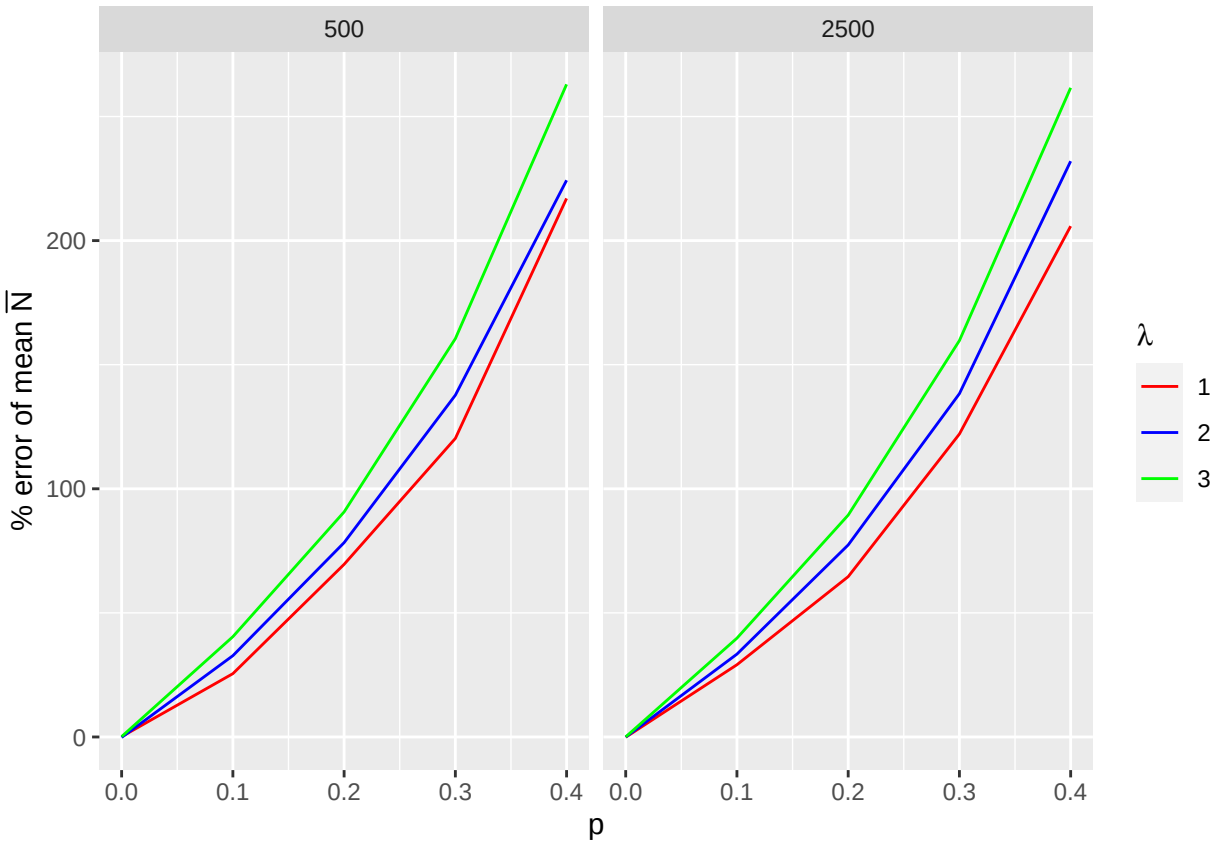

Figure 2: ZTP, percent error of mean population estimate

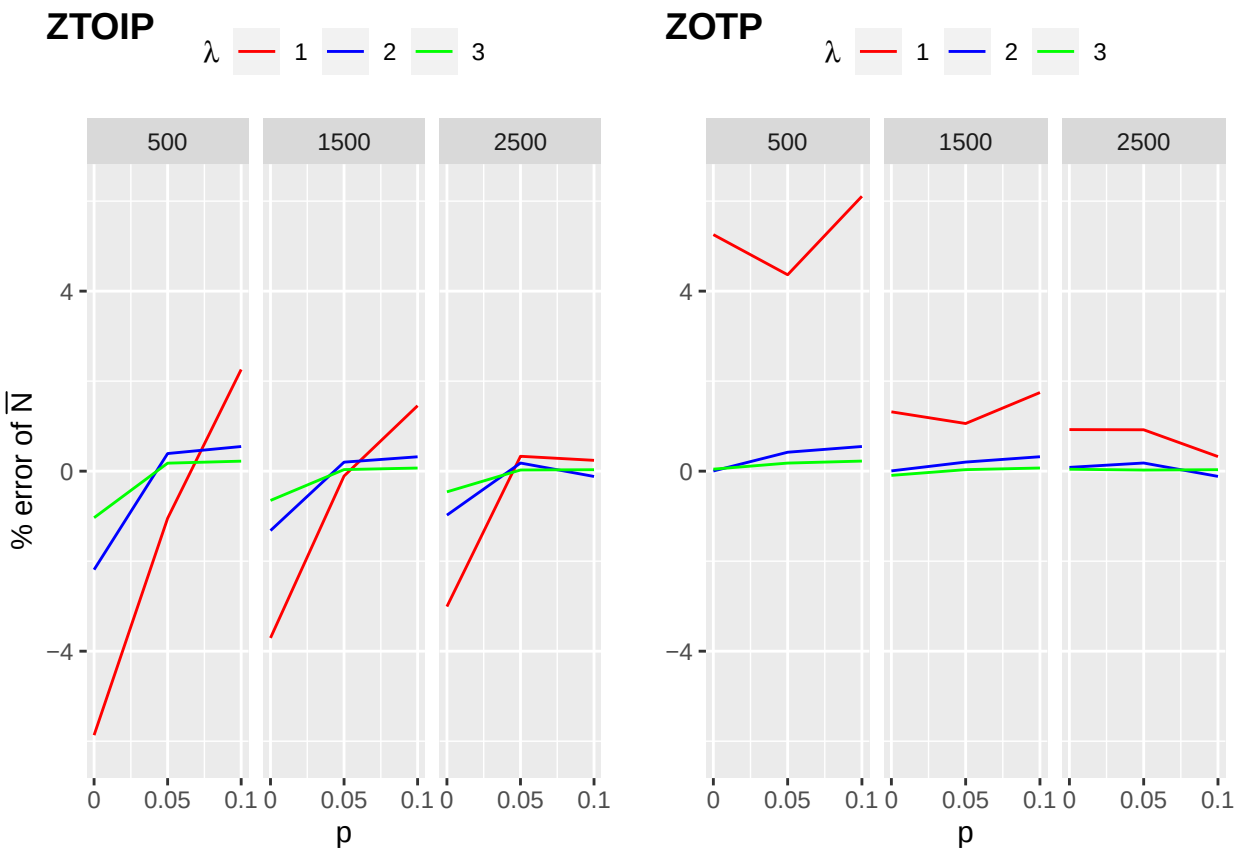

Figure 3: ZTOIP and ZOTP, percent error of mean population estimate

Table 1: ZTOIP and ZOTP, confidence intervals and standard error of pupulation estimate

p	N	λ	Type	$\bar{N}$	90% CI	99% CI	σ
0.00	500	2	ZOTP	500.017	[450, 552]	[430, 583]	30.990
0.00	500	2	ZTOIP	489.055	[450, 516]	[430, 523]	19.774
0.00	500	2	ZTP	500.655	[482, 519]	[475, 525]	10.970
0.05	500	2	ZOTP	502.104	[446, 563]	[420, 584]	34.820
0.05	500	2	ZTOIP	501.959	[446, 563]	[420, 583]	34.506
0.05	500	2	ZTP	578.105	[550, 604]	[542, 616]	16.646
0.10	500	2	ZOTP	502.741	[445, 568]	[418, 595]	37.696
0.10	500	2	ZTOIP	502.738	[445, 568]	[418, 595]	37.698
0.10	500	2	ZTP	666.115	[627, 704]	[612, 721]	23.900

Negative binomial models

Since all our conclusions regarding the population estimate for our three different model types and various values of p, λ and N still apply in the negative binomial models (see figur Appendix), we will only look at the effect of individual heterogeneity in the negative binomial model. To do this we simulate from the ZTOINB model using four different values of k. For each simulation we estimate the population with $\hat{N}_{g++}$. Mean, confidence intervals as well as sample standard deviation can be seen in table 2.

Table 2: ZOTNB, confidence intervals and standard error of pupulation estimate

p	N	λ	k	$\bar{N}$	90% CI	99% CI	σ
0.1	2500	1	0.2	332640.597	[1107, 2789185]	[937, 4690045]	984844.570
0.1	2500	1	0.6	73808.500	[1412, 12170]	[1197, 2983149]	554239.953
0.1	2500	1	1.0	34487.377	[1535, 8149]	[1344, 150163]	372776.129
0.1	2500	1	2.0	16866.201	[1553, 5956]	[1376, 29879]	220505.959
0.1	2500	2	0.2	5489.282	[1756, 4475]	[1601, 6921]	88056.102
0.1	2500	2	0.6	2557.703	[2078, 3253]	[1909, 3726]	376.021
0.1	2500	2	1.0	2531.200	[2136, 3061]	[2024, 3349]	286.846
0.1	2500	2	2.0	2516.536	[2181, 2926]	[2083, 3174]	231.209
0.1	2500	3	0.2	2540.567	[2091, 3107]	[1973, 3731]	343.296
0.1	2500	3	0.6	2508.969	[2290, 2769]	[2208, 2919]	150.405
0.1	2500	3	1.0	2500.881	[2316, 2713]	[2264, 2822]	117.549
0.1	2500	3	2.0	2505.279	[2366, 2671]	[2297, 2752]	96.212

Application on bear data

Before we look at whether there is any inflation in the bear population or try to estimate its size, 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 for our bear data with the theoretical quantiles of a ZTNB model. The result can be seen in Figure 4.

In table 3 we can see population estimates from our three different negative binomial models as well as parameter estimates of the ZTOINB model.

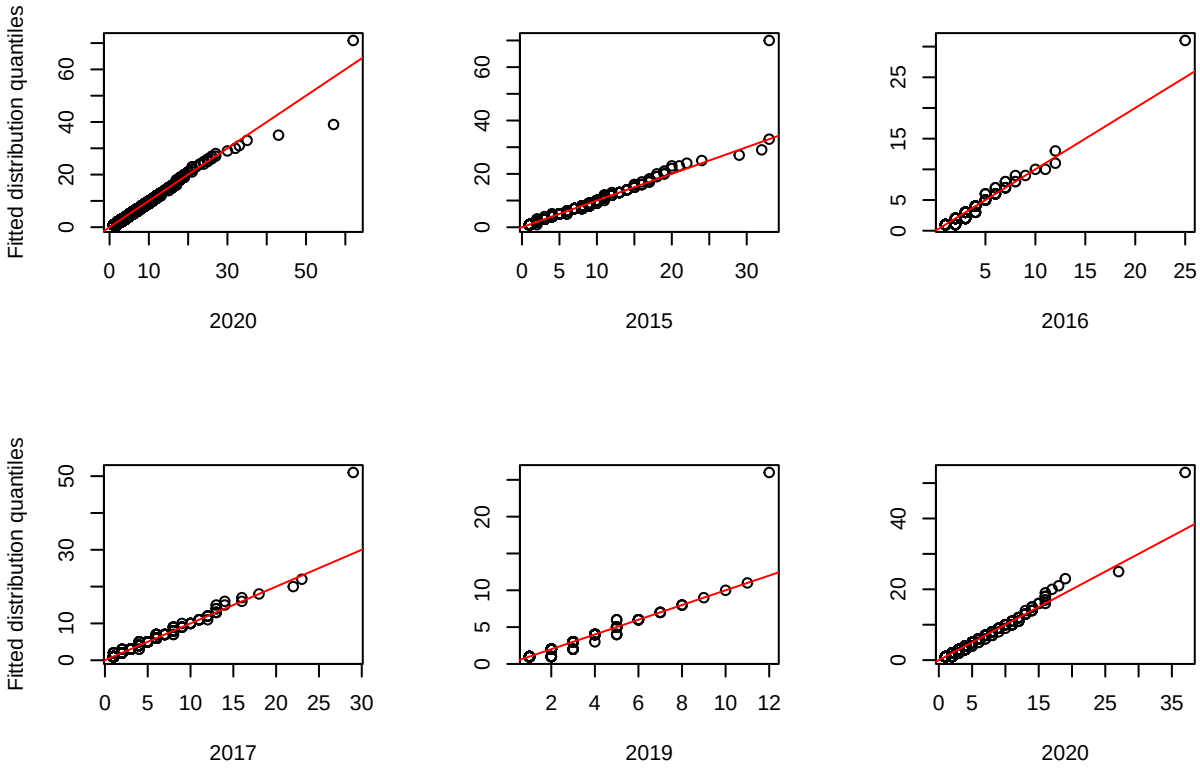

Figure 4: QQ-plot, number of observations data against theoretical NBin quantiles

Table 3: ZOTNB, confidence intervals and standard error of pupulation estimate

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	6564	11484	6564
2015	1016	2.512	0.197	0.000	1722	1732	1722
2016	336	1.722	0.656	0.000	518	700	518
2017	636	1.818	0.280	0.004	1168	1168	1231
2019	344	1.158	0.711	0.000	668	829	668
2020	1154	1.641	0.265	0.000	2338	2713	2338

Diskussion

In our first result from Figure 2 we see that for both our population sizes the bias of $\hat{N}_{p_+}$ grows exponentially as p increases. We notice that the effect does not depend on the population size. As expected, the bias grows quicker for greater values of λ, this is of course since the extra ones then have a larger impact on the estimated underlying distribution parameter $\hat{\lambda}$. Its interesting to see that for an inflation as small as 0.05 we can already observe a bias of over 10% for even our smallest lambda.

In Figure 3 we can observe a great bias of the ZTOIP model for small values of p. The bias for $p = 0$ occurs since data in about half of the cases will contain more ones than expected, based on the remaining data. In these cases the ZTOIP likelihood maximum will be maximized for some $p > 0$, which then leads to a lower population estimate. It is possible that this bias can be eliminated by expanding the model to allow negative bias values. One way of introducing negative bias can be seen in PPSIZRERFefeERENS where the link $\omega = \theta/(1 + \theta)$ is used, however the approach would need to be adapted to be applicable in our case.

We can see that the negtive bias for low p values in Figure 3 ZTOIP estimate decreases as the expected number of ones decrease, i.e for higher values of λ. The bias also decreases for greater population sizes as the maximum likelihood estimator is asymptoticly unbiased. The positive bias that can be seen in Figure 3 ZOTP estimates arises as a consequence of high variance and also exists in the ZTOIP estimates, but is overshadowed by the bias previously described. Its worth noting that a combination of the ZOTP and ZTOIP model which first uses the zero-one truncated model to estimate the λ parameter and then estimates p can be used. This combined model is in most cases preferable to the ZTOIP since the model does not lead to negative bias for low p values.

In Table 1 we see how the ZTOIP and ZOTP differ for lower values of p but converge towards each other for higher values and are largely equal for $p = 0.1$. We can also see that ZTP has the lowest variance for all values of p.

In Table 2, we see how the variance for our estimator increases when we introduce different amounts of individual heterogeneity through k. We note that low values of k in combination with lower values of *lambda* result in high uncertainty and bias in the population estimates, but as both or one of the parameters increases in value, the population estimate stabilize.

From QQ-plots in Figure 4, we see that the data, largely, seems to fit well with the model with the only major difference being that the collected data appears to have a slightly thinner tail compared to the theoretical quantiles. The good fit is expected as the negative binomial model is very flexible.

From the inflation parameter estimations $\hat{p}_{g_{+1}}$ in Table 3 we see no sign that bear data contain one-inflation as only one of the years contains more one-observations than expected. Even if inflation does not seem to exist, we can estimate the population with our different models and get an estimate of the size of the population in all regions for the different years. Since inflation does not seem to exist in data, $\hat{N}_{g_+}$ is the best estimator. Note, however, that these population estimates are very uncertain due to the observed variances and bias in in table 2 and as model assumptions are to some extent not met and a more complicated model

accounting for more individual heterogeneity is required for reliable estimates. The problem of individual heterogeneity in capture-recapture can be read more about in WILLIAM A LINK.

It is possible that there is often an understanding of how a possible inflation value p can arise and be distributed, especially in in the case of genotyping errors since there in many cases is possible to get a good understanding of what the probability of such an error can be based on the method used POMPANON2005. If we have a prior to our study have an understanding of our inflation parameter p, priori distributions can be of use. In The Bayesian analysis of on a the problem of one-inflation is examined from a Bayesian Inference perspective and the methods used can be applied to our work in this report after possible adjustment to work with our different type of inflation.

Appendix

Referenser