Tiziana Tuoto

Bayesian analysis of one–inflated models for population size estimation

Tiziana Tuoto[1,2]; Davide Di Cecco[2]; Andrea Tancredi[2]

[1] Italian National Institute of Statistics, Istat
[2] Sapienza University of Rome

Abstract
The treatment of one inflation in estimating the size of an elusive population has received an increasing attention from researchers in recent years. The inflation of counts "one", that is the presence of an excess of "one"s, larger than compatible with any distributional assumption, deserves specific attention due to its effect on usual estimators for the population size. One-inflation may lead to overestimation of the usual estimators when not accounted for, while in general, when the observed counting model presents other sources of heterogeneity and this is ignored by the estimation procedure, it may produce severe underestimation in the population size. In this paper we propose a Bayesian approach that considers one inflation in count data models for capture recapture. An application to real data for the estimate of the size of some illegal populations is used to illustrate the proposed methodology.

Keywords:
Capture-recapture; Illegal populations; Zero-truncated One–inflated Count Data Models; Bayesian Model Selection

1. Introduction

A popular methodology to estimate the size of hidden populations is the capture-recapture methods, originally used to estimate animal abundance. In capture–recapture, data can be collected at specific time points, therefore for each unit seen at least once there is a capture history. The data used in this work are a specific form of capture–recapture data. We collect data in continuous time hence we have only the total number of times a unit is captured. By means of an observational mechanism, i.e a register, a trapping system, a set of interviewers, a diagnostic devise, we can identify some population units and follow them in a specific time span. In this way, the observational mechanism provides us a list of population units with the count (i.e. the number of times) that they appear in the list. In the list we can observe individuals who are observed/captured 1, 2, 3, ..., times, however we cannot observe units not caught by the observational system. Hence, the list can be considered as incomplete and we want to estimate the hidden part of the population, i.e. the size of it not reported by our observational mechanism. Capture recapture data in this setting are usually called repeated counting data. To estimate the population size, one needs to model the counting process of observation/capturing.

In this framework, an increasing attention has been devoted to the inflation of counts "one", i.e. the presence of an excess of "one"s , larger than compatible with any distributional assumption, see Godwin and Böhning (2017), Godwin (2017), Godwin (2019), Böhning, Kaskasamkul, and van der Heijden, (2018), Böhning and van der Heijden (2019), Böhning and Friedl (2021).

One-inflation can occur for different reasons; for instance, we observe one-inflation when some units of the population cannot be captured anymore after the first capture. This may be the case of some wild animal populations, when the animals that experienced the capture once, find it so unpleasant that some of them develop the desire and/or ability to avoid subsequent captures. A similar reasoning can be applied also to human populations, particularly when the first capture consists of law enforcement, involves imprisonment or reveals an undesirable characteristic/behaviour. See Godwin and Böhning (2017) for a rich discussion on justifications and conditions for one-inflation in capture-recapture, also including an interpretation of one-inflation as limiting case of the so-called "*trap shy*" behavioural model.

One-inflation deserves specific attention due to its effect on usual estimators for population size. In general, when the observed counting model presents heterogeneity and this is ignored by the estimation procedure, it may produce severe underestimation in the population size. On the contrary, one-inflation may lead to overestimation of the usual estimators when not accounted for. This is true even for the well-known lower bound Chao estimator, which may severely overestimate the population size in the presence of one-inflation, as discussed in Chiu and Chao (2016) and in Böhning, Kaskasamkul, and van der Heijden, (2018).

In this paper we propose a Bayesian approach to count data models with one-inflation, for use in Horvitz–Thompson estimator of the population size. Under given conditions, the Bayesian approach results in population size estimates similar to the maximum likelihood ones, with the advantage of producing the credible intervals of the estimates as a by-product of the estimation procedure. Another advantage of the Bayesian approach is to allow incorporating previous knowledge on the hidden size of the population in the analysis, in the very favourable case in which this information is available to the analyst. Moreover, the Bayesian analysis allows testing the one-inflation assumption in a very natural way.

We apply our Bayesian proposals to real data for estimating the size of some illegal population, using also some popular dataset available from the literature on capture-recapture, where the issue of one-inflation has been recognised.

2. Methodology

According to the standard formulation, consider a closed population (no birth, death or migration) of size *N*. For each unit in the population, let *Y* be a random variable taking value *j=0, 1, 2, …* if the individual is observed/captured *j* times. We only observe the *n* individuals, such that *j>0*, with *n* ≤*N*.

Let $p_j = Prob(Y = j)$ denote the probability of a unit being captured j times, and n_j denote the number of individuals observed *j* times, that is n_j is the frequency of the count *j*.
Our interest is to estimate the unknown $n_0 = N - n$ units remain unobserved, actually the units observed in zero counts n_0, and consequently *N*, on the basis of some model for the observed n_j.

The Horvitz-Thompson estimator arises from the sum of both the unobserved and the observed cases *n*, by the solution of the estimating equation

$$N = Np_0 + N(1 - p_0) = n/(1 - p_0)$$

where $N(1 - p_0)$ is the expected number of cases identified by the capture mechanism, estimated by the observed cases *n*. In the Horvitz–Thompson estimator, the probability of

being observed *0* times, p_0, needs to be estimated by an appropriate model that describe the observed zero-truncated counts.

Note that Bayesian inference for the population size *N* is straightforward under standard models for *Y*. Suppose for example that *Y* is Poisson distributed with mean *λ*. Let $y = (y_1, \dots, y_n)$ be the observed number of captures for the n observed units and let *s* be their sum $s = \sum_{i=1}^{n} y_i$
The likelihood function for $\theta = (\lambda, N)$ is given by

$$f(y, n|\, \lambda, N) = \binom{N}{n} exp\{-\lambda\,(N-n)\}\,\lambda^n\, exp\{-\lambda\, s\}$$

Assuming a priori $p(N) \propto 1/N$ and $\lambda \sim Gamma\,(a_l, b_l)$ and setting $n_0 = N - n$ the posterior for(λ, n_0) is given by

$$p(n, \lambda\,|y, n) \;\propto\; \binom{n_0 + n}{n} exp\{-\lambda\, n_0\}\,\lambda^{n+a_l}\, exp\{-\lambda\,(s + b_l)\}\frac{1}{n_0 + n}$$

which can be easily simulated via a Gibbs sampler. In fact, the conditional distribution of $\lambda\,|n_0, y, n$ is a Gamma with parameters $n + a_l$, $s + b_l$ and the conditional distribution $n_0\,|\lambda, y, n$ is a Negative Binomial with size parameters equal to *n* and probability $exp\{-\lambda\,\}$.

To include the one-inflation in our model, we assume that in our population a specific behavioural mechanism is acting. That is, an individual that without that mechanism would face multiple captures, now has a positive probability ω of being captured just once.

Let *Y* denote the observed number of captures for a unit, and *Y** the latent value we would observe without the behavioural mechanism. The two variables are linked by means of the following infinite transition matrix

P=

1	0	0	0	0	…
0	1	0	0	0	…
0	ω	1- ω	0	0	…
0	ω	0	1- ω	0	…
0	ω	0	0	1- ω	…
…	…	…	…	…	…

where the $(k, j) - th$ element represents the conditional probability $P(Y = j|Y^* = k)$. Note that the first row comprises the probabilities $\{p_{0j}\}_{j=0,1,\dots}$ and the first column the probabilities $\{p_{k0}\}_{k=0,1,\dots}$.

Let $f(k|\theta) = P(Y^* = k|\theta)$ be the probability distribution, depending on some parameter *θ*, of the number of captures without the behaviour effect, and let *F(θ)* denote the associated c.d.f. Then, the resulting distribution for *Y* is the one-inflated model defined as follows:

$$P(Y = j) = \begin{cases} f(0|\theta) & if\ j = 0 \\ (1-\omega)f(1|\theta) + \omega(1 - f(0|\theta)) & if\ j = 1 \\ (1-\omega)f(j|\theta) & if\ j > 1 \end{cases}$$

The conditional distribution of *Y** when *Y=j* is concentrated on *j* when $j \neq 1$, while, when *j=1*, we have:

$$P(Y^* = k \mid Y = 1) = \begin{cases} 0 & if\ k = 0 \\ \dfrac{f(1|\theta)}{f(1|\theta) + \omega(1 - F(1|\theta)} & if\ k = 1 \\ \dfrac{\omega f(k|\theta)}{f(1|\theta) + \omega(1 - F(1|\theta)} & if\ k > 1 \end{cases}$$

This setting allows for different specifications for our count data *Y**, i.e. $f(\theta)$ can be a Poisson density, as in Godwin and Böhning (2017), or alternatively *Y** can follow a Negative Binomial distribution, as in Godwin (2107). In both cases, a Gibbs sampler can be outlined for simulating from the conditional distributions of $\omega, Y^*, and\ \theta$.

3. Result

In this section, we apply the proposed Bayesian model to a selection of popular case-studies in capture recapture literature.
The following real cases are considered:

1. the number of prostitutes in Vancouver, presented in Rossmo and Routledge (1990);
2. opiate users in Rotterdam, already analysed by Cruyff and van der Heijden (2008);
3. heroin users in Bangkok, from Böhning et al. (2004) Viwatwongkasem et al (2008).

The evidence of one inflation in these data sets have been largely discussed already in Godwin and Böhining (2017) and Godwin (2017), where results from maximum likelihood analysis under Poisson and Negative Binomial distribution, respectively, can be found and compared with the results presented in this section.

Table 1 shows the results from our Bayesian models for the selected case studies, reporting the posterior modes and credible intervals of the population sizes, *N*, as well as the posterior medians of the model parameters. In the table, we indicate with λ the parameter of the Poisson distribution. For the Negative Binomial distribution, we adopt the size and probability parameterization, indicated with r and p, respectively. The one-inflated models are indicated with the suffix 'OI'. In table 1, we show also the results from a Bayesian analysis assuming the Poisson distribution and ignoring the one-inflation, to show the over-estimation occurring when one-inflation is ignored.

Table 1. The posterior mode and credible intervals for the population size *N*, posterior median for ω and model parameters, for some popular real cases analysed

Prostitutes in Vancouver		N	HPD(N)	ω	λ	r	p
	Poisson	1237	1178 - 1301		1.253		
	OIP	1016	980 - 1057	0.439	2.037		
	OINB	1304	1100 - 2174	0.274		1.947	0.604
Opiate users in Rotterdam		N	HPD(N)	ω	λ	r	p
	Poisson	2929	2832 - 3038		1.174		
	OIP	2500	2418 - 2587	0.336	1.663		
	OINB	3769	3140 - 5533	0.086		1.397	0.618
Heroin users in Bangkok		N	HPD(N)	ω	λ	r	p
	Poisson	9453	9427 - 9477		4.134		
	OIP	9364	9349 - 9380	0.207	5.004		
	OINB	10865	10629 - 11111	0.055		1.612	0.300

4. Discussion and Conclusion

In this paper we propose a Bayesian approach to count data models with one-inflation, for use in Horvitz–Thompson estimator of the population size. The one-inflation can introduce severe overestimation in traditional capture-recapture estimator, if not properly accounted for. Our Bayesian approach results in population size estimates similar to the maximum likelihood ones, under non informative prior distribution, with the advantage of producing the credible intervals of the estimates as a by-product of the estimation procedure. Another advantage of the Bayesian approach is to allow incorporating previous knowledge on the hidden size of the population in the analysis, in the very favourable case in which this information is available to the analyst. The role of prior distributions is topic for further investigation.

We apply our Bayesian proposals to real data for estimating the size of some illegal population, using also some popular dataset available from the literature on capture-recapture, where the issue of one-inflation has been recognised. The properties of our model are highlighted also by a simulation study, not presented in these proceedings for the sake of brevity.

A possible extension of this work aims at incorporating, in a natural way, a Bayesian test for the one-inflation assumption.

References

1. Böhning, D., Suppawattanabodee, B., and Kusolvisitkul, W. and Viwatwongkasem, C. (2004). Estimating the number of drug users in Bangkok 2001: A capture-recapture approach using repeated entries in one list. European journal of epidemiology, 19, 12, 1075-1083
2. Böhning, D. and Friedl, H. (2021). Population size estimation based upon zero-truncated, one-inflated and sparse count data. Statistical Methods & Applications, 1-21
3. Böhning, D. and Kaskasamkul, P. and van der Heijden, PGM. (2018). A modification of Chao's lower bound estimator in the case of one-inflation. Metrika, 82, 3, 361-384
4. Böhning, D. and 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. Annals of Applied Statistics,13, 1198-1211
5. Cruyff, M. J. and van der Heijden, P. G. (2008). Point and interval estimation of the population size using a zero-truncated negative binomial regression model. Biometrical Journal, 50, 1035-1050
6. Godwin, R.T. (2019). The one-inflated positive Poisson mixture model for use in population size estimation. Biometrical Journal, 61, 6, 1541-1556
7. Godwin, R.T. and Böhning, D. (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, 425-448
8. Rossmo, D., and Routledge, R. (1990). Estimating the size of criminal populations. Journal of quantitative criminology, 6 (3), 293–314
9. Viwatwongkasem, C., Kuhnert, R. and Satitvipawee, P. (2008). A comparison of population size estimators under the truncated count model with and without allowance for contaminations, Biometrical Journal: Journal of Mathematical Methods in Biosciences, 50, 6, 1006-1021