





ARTICLE

Simple models for macro-parasite distributions in hosts

 Gonzalo Maximiliano Lopez^{*,1,2} and  Juan Pablo Aparicio^{1,3}

¹Instituto de Investigaciones en Energía no Convencional, Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional de Salta, Salta, Argentina.

²Departamento de Matemática, Facultad de Ciencias Exactas, Universidad Nacional de Salta, Salta, Argentina.

³Simon A. Levin Mathematical, Computational and Modeling Sciences Center, Arizona State University, Tempe-AZ, USA.

*Corresponding author. Email: gonzalo.maximiliano.lopez@gmail.com

(Received: June 9, 2022; Revised: August 1, 2022; Accepted: September 30, 2022; Published: June, 2023)

Abstract

The Negative binomial distribution is the most used distribution to model macro-parasite burden in hosts. However, reliable maximum likelihood parameter estimation from data is far from trivial. No closed formula is available and numerical estimation requires sophisticated methods. Using data from the literature, we show that simple alternatives to negative binomial, like zero-inflated geometric or hurdle geometric distributions, produce in some cases a better fit to data than the negative binomial distribution. We derived simple closed formulas for the maximum likelihood parameter estimators which constitutes a significant advantage of these distributions over the negative binomial distribution.

Keywords: Hurdle geometric distribution; Macroparasite, Maximum likelihood estimation; Negative binomial distribution; Zero-inflated geometric distribution;

1. Introduction

Macroparasites usually present over-dispersed distributions, that is, a few hosts harbor many parasites, while the rest of the hosts are practically free of parasites (see, for example, Crofton, 1971; Gourbière *et al.*, 2015; Seo *et al.*, 1979). The most used distribution to describe the distribution of parasites is the negative binomial (Bliss & Fisher, 1953; Shaw *et al.*, 1998) which provides an accurate description of the observations.

However, in many cases, the negative binomial distribution (or other similar distributions like the log-series and Polya-Eggenberger (Crofton, 1971; Seo *et al.*, 1979)) cannot account for the “excess” of zeros observed. A simple solution widely used is to consider zero-inflated distributions (Greene, 1994; Hall, 2000; Lambert, 1992).

The negative binomial distribution, denoted in the following by $NB(m, k)$, is a two parameters distribution, the mean burden of parasites in the host population m and the dispersion parameter k which is related to the degree of overdispersion of the distribution (Bliss & Fisher, 1953). Moreover, it can be shown that the limiting distribution of the $NB(m, k)$ distribution, as $k \rightarrow \infty$, is a Poisson distribution with parameter m , and $NB(m, k)$ distribution is a geometric distribution with parameter $\frac{1}{m+1}$ when $k = 1$.

One problem with the negative binomial distribution is parameter estimation from the observations. The method of moments estimation is simple but not always precise (Clark & Perry, 1989). Maximum likelihood estimation provides one of the best parameters estimation (Piegorisch, 1990) but for the negative binomial distribution there is no a closed formula for the parameter estimates in terms of the observations and maximum likelihood estimates should be obtained numerically which presents various complexities (Bandara *et al.*, 2019; Dai *et al.*, 2013).

In this article, we show simpler alternatives to the negative binomial distribution which describe the observations equally well, and in some cases provide a more precise description, but with the significant advantage that exist closed-form expressions for the maximum likelihood parameter estimators.

2. Zero-inflated (deflated) distributions and hurdle distributions

The distribution of parasites in hosts is often over-dispersed, that is, there tend to be a small number of hosts who harbor very large burdens while the majority of hosts harbor low burdens. It is colloquially said that parasite's distributions follow the 20-80 rule (Woolhouse *et al.*, 1997), that is, 20% of the individuals account for the 80% of the parasite burden. The negative binomial distribution offers such over-dispersed distribution, but in some cases, an excess of zeros is observed for which the distribution fail to account. A simple and widely used solution is to consider zero-inflated (in some cases zero-deflated, Dietz & Böhning (2000)) distribution and hurdle distribution (Min & Agresti, 2005; Welsh *et al.*, 1996).

2.1 Zero-inflated (deflated) distributions

A discrete random variable Y follows a zero-inflated (deflated) distribution if its probability mass function (pmf) is given by (Johnson *et al.*, 2005),

$$P(Y = y) = \begin{cases} \pi + (1 - \pi)p(y; \theta) & y = 0 \\ (1 - \pi)p(y; \theta) & y \geq 1 \end{cases} \quad (1)$$

where p is the pmf of a given base distribution and θ is a parameters vector. For example, θ may include the mean (μ) and variance (σ^2) of the base distribution. When $\pi < 0$ we have a zero-deflated distribution.

If $G(z)$ is the probability generating function (pgf) of the base distribution, then the pgf of the corresponding zero-inflated (deflated) distribution is given by (ibid.)

$$F(z) = \pi + (1 - \pi)G(z) \quad (2)$$

From the pgf we may obtain the mean and the variance for the distribution

$$\begin{aligned} E(Y) &= (1 - \pi)\mu \\ \text{Var}(Y) &= (1 - \pi)\sigma^2 + \pi(1 - \pi)\mu^2 \end{aligned} \quad (3)$$

where μ , σ^2 are the mean and variance of the base distribution, respectively. The dispersion index, or variance-to-mean ratio, is given by $D = \frac{\text{Var}(Y)}{E(Y)}$ (Cox & Lewis, 1966), then D can be written as

$$D = \frac{\sigma^2}{\mu} + \pi\mu \quad (4)$$

where σ^2/μ is the dispersion index of the base distribution. Therefore, from (4), the zero-inflated distribution is more over-dispersed than the base distribution.

2.2 Hurdle distributions

Another common way to account for an excess of zeros is the hurdle distribution, where its pmf is given by (Johnson *et al.*, 2005)

$$P(Y = y) = \begin{cases} \omega & y = 0 \\ (1 - \omega) \frac{p(y; \theta)}{1 - p(0; \theta)} & y \geq 1 \end{cases} \quad (5)$$

and p, θ are defined as in section 2.1.

If $G(z)$ is the pgf for the base distribution, then the pgf for the hurdle distribution is (ibid.)

$$F(z) = \omega + (1 - \omega) \frac{G(z) - p(0; \theta)}{1 - p(0; \theta)} \quad (6)$$

and therefore, the mean and the variance of the hurdle distribution is,

$$\begin{aligned} E(Y) &= \alpha\mu \\ \text{Var}(Y) &= \alpha\sigma^2 + \alpha(1 - \alpha)\mu^2 \end{aligned} \quad (7)$$

where $\alpha = \frac{1 - \omega}{1 - p(0; \theta)}$, and μ, σ^2 are the mean and variance of the base distribution, respectively. Finally the dispersion index for the hurdle distribution is given by

$$D = \frac{\sigma^2}{\mu} + (1 - \alpha)\mu \quad (8)$$

If $\omega > p(0; \theta)$ the hurdle distribution is more over-dispersed than the base distribution.

3. Parameter estimation and Maximum likelihood

A simple, but in general inaccurate, way to fit the parameters of a distribution from a sample consist of the use of the method of moments. For example, the negative binomial distribution has two parameters that can be expressed in terms of the two first moments. While this method is quite simple, do not provide a reliable fit (Clark & Perry, 1989; Piegorsch, 1990). Maximum likelihood estimation (mle) is the most reliable method (Piegorsch, 1990), but it does not always produce closed-form expression, and the parameter estimations need to be obtained numerically, a task far from trivial for the negative binomial distribution (Bandara *et al.*, 2019; Bliss & Fisher, 1953; Dai *et al.*, 2013). In the following, we present the problem of maximum likelihood parameter estimation for zero-inflated and hurdle distributions.

3.1 Maximum likelihood estimation for zero-inflated (deflated) and hurdle distributions

3.1.1 Zero-inflated (deflated) distribution

As in section 2.1 we denote by θ to the vector of parameters of the base distribution, and therefore the set of parameters for the zero-inflated (deflated) distribution is (π, θ) .

Let y_i denote the i th observation, $i = 1, \dots, N$, where N denote the total number of observations in the sample and N_0 is the number of units with zero counts, then the log-likelihood function is given by

$$\begin{aligned} \ln L(\pi, \theta) &= \sum_{i=1}^N \ln P(Y = y_i) \\ &= N_0 \ln [\pi + (1 - \pi)p_0] + (N - N_0) \ln(1 - \pi) + \sum_{y_i \geq 1} \ln p_{y_i} \end{aligned} \tag{9}$$

where $p_{y_i} = p(y_i; \theta)$. Maximizing $\ell = \ln L$ for (π, θ) we obtain the following system of equations,

$$\begin{aligned} \frac{\partial \ell}{\partial \pi} &= \frac{N_0(1 - p_0)}{\pi + (1 - \pi)p_0} - \frac{N - N_0}{1 - \pi} = 0 \\ \frac{\partial \ell}{\partial \theta_i} &= \frac{N_0(1 - \pi) \frac{\partial p_0}{\partial \theta_i}}{\pi + (1 - \pi)p_0} + \sum_{y_i \geq 1} \frac{\frac{\partial p_{y_i}}{\partial \theta_i}}{p_{y_i}} = 0 \end{aligned} \tag{10}$$

3.1.2 Hurdle distributions

The log-likelihood function for the hurdle distribution is given by

$$\begin{aligned} \ln L(\omega, \theta) &= \sum_{i=1}^N \ln P(Y = y_i) \\ &= N_0 \ln \omega + (N - N_0) \ln \left(\frac{1 - \omega}{1 - p_0} \right) + \sum_{y_i \geq 1} \ln p_{y_i} \end{aligned} \tag{11}$$

and therefore the parameter values estimates are obtained from the following system

$$\begin{aligned} \frac{\partial \ell}{\partial \omega} &= \frac{N_0}{\omega} - \frac{N - N_0}{1 - \omega} = 0 \\ \frac{\partial \ell}{\partial \theta_i} &= \frac{(N - N_0) \frac{\partial p_0}{\partial \theta_i}}{1 - p_0} + \sum_{y_i \geq 1} \frac{\frac{\partial p_{y_i}}{\partial \theta_i}}{p_{y_i}} = 0 \end{aligned} \tag{12}$$

4. Maximum likelihood for the negative binomial distribution and two simple alternatives.

The probability mass function for the negative binomial distribution is

$$P(Y = y) = \frac{\Gamma(y + k)}{\Gamma(y + 1)\Gamma(k)} q^k (1 - q)^y \tag{13}$$

where $0 \leq q \leq 1$ and $k > 0$.

Suppose we have data y_1, \dots, y_N sampled from a negative binomial distribution (13). Differentiating the log-likelihood function $\ell = \ln L$ partially and setting them equal to zero yields the following system of equations

$$\begin{aligned} \frac{\partial \ell}{\partial q} &= \frac{Nk}{q} - \frac{mN}{1-q} = 0 \\ \frac{\partial \ell}{\partial k} &= \left[\sum_i \psi(y_i + k) \right] - N\psi(k) + N \ln q = 0 \end{aligned} \tag{14}$$

where m is the mean of model (13) and $\psi(y) = \frac{\Gamma'(y)}{\Gamma(y)}$ is the digamma function. Substituting in the second equation $q = \frac{k}{m+k}$ (obtained from the first equation) we obtain

$$\left[\sum_i \psi(y_i + k) \right] - N\psi(k) + N \ln \left(\frac{k}{m+k} \right) = 0 \tag{15}$$

This equation cannot be solved for k in a closed form and must be solved numerically. Iterative techniques as Newton-Raphson method can be used, but this method may fail to find the mle value. Finding the mle value is a challenge, since we could not obtain the root or obtain more than one for the equation (15) (Clark & Perry, 1989; Dai *et al.*, 2013; Piegorsch, 1990; Saha & Paul, 2005; Willson *et al.*, 1984).

4.1 Zero-inflated geometric distribution

The geometric distribution is a special case of the negative binomial distribution for $k = 1$, and then, its pmf is given by

$$P(X = x) = q(1 - q)^x \tag{16}$$

where $x = 0, 1, 2, \dots$ and $0 \leq q \leq 1$. The mean is $\frac{1-q}{q}$ while the variance is $\frac{1-q}{q^2}$.

From (1) the corresponding zero-inflated (deflated) distribution is

$$P(Y = y) = \begin{cases} \pi + (1 - \pi)q & y = 0 \\ (1 - \pi)q(1 - q)^y & y \geq 1 \end{cases} \tag{17}$$

Mean and variance are given by (3),

$$\begin{aligned} E(Y) &= (1 - \pi) \frac{(1 - q)}{q} \\ \text{Var}(Y) &= (1 - \pi) \frac{(1 - q)}{q^2} [1 + \pi(1 - q)]. \end{aligned} \tag{18}$$

The dispersion index for the zero-inflated (deflated) geometric distribution is always greater than one, and therefore this distribution is always over-dispersed.

4.1.1 Maximum likelihood estimation

The system of likelihood equations, according to (10), is given by

$$\begin{aligned} \frac{\partial \ell}{\partial \pi} &= \frac{N_0(1 - q)}{\pi + (1 - \pi)q} - \frac{N - N_0}{1 - \pi} = 0 \\ \frac{\partial \ell}{\partial q} &= \frac{N_0(1 - \pi)}{\pi + (1 - \pi)q} + \frac{N - N_0}{q} - \frac{mN}{q} = 0 \end{aligned} \tag{19}$$

Therefore, the maximum likelihood estimators for π and q are

$$\begin{aligned} \hat{\pi} &= \frac{mN_0 - N + N_0}{mN - N + N_0} \\ \hat{q} &= \frac{N - N_0}{mN} \end{aligned} \tag{20}$$

which is expressed in terms of the observed sample mean m , the sample size N and the number of zeros in the sample N_0 .

4.2 Hurdle geometric distribution

The hurdle distribution for the geometric distribution is given by

$$P(Y = \gamma) = \begin{cases} \omega & \gamma = 0 \\ (1 - \omega) \frac{q^{(1-q)^\gamma}}{1-q} & \gamma \geq 1 \end{cases} \tag{21}$$

Mean and variance are obtained from (7)

$$\begin{aligned} E(Y) &= \frac{(1 - \omega)}{q} \\ \text{Var}(Y) &= \frac{(1 - \omega)}{q^2} [1 + (\omega - q)] \end{aligned} \tag{22}$$

The dispersion index for the hurdle geometric distribution is always greater than one, and therefore this distribution is always over-dispersed.

4.2.1 Maximum likelihood estimation

According to (12) the system of likelihood equations is given by

$$\begin{aligned} \frac{\partial \ell}{\partial \omega} &= \frac{N_0}{\omega} - \frac{N - N_0}{1 - \omega} = 0 \\ \frac{\partial \ell}{\partial q} &= (N - N_0) \left(\frac{1}{1 - q} + \frac{1}{q} \right) - \frac{mN}{1 - q} = 0 \end{aligned} \tag{23}$$

therefore, the maximum likelihood estimators for ω and q are given by

$$\begin{aligned} \hat{\omega} &= \frac{N_0}{N} \\ \hat{q} &= \frac{N - N_0}{Nm} \end{aligned} \tag{24}$$

Like in the previous case, maximum likelihood estimators are expressed in terms of the observed sample mean m , the sample size N and the number of zeros in the sample N_0 .

Note that by the reparameterization $\omega = \pi + (1 - \pi)q$, the probability mass functions of the zero-inflated geometric and hurdle geometric models coincide. On the other hand the pair $\left(\frac{\hat{\omega} - \hat{q}}{1 - \hat{q}}, \hat{q} \right)$ is a maximum of the log-likelihood function (9). Therefore we can obtain that the corresponding Akaike's information criterion values of both models will be the same.

Due to these facts, the fit of the data by both models will be the same.

5. Applications

5.1 Study of frequency distribution of *Ascaris lumbricoides* infection

The nematode *Ascaris lumbricoides* is one of the most common intestinal parasites of humans. It is prevalent in tropical and temperate populations where poverty and lack of sanitation is common (Pullan & Brooker, 2012).

Burden of infestation is computed using the number of parasites in each host of the sample. At the present time, this type of study is no longer conducted and we will use the results from Seo *et al.* (1979). He studied six rural populations in Korea where an endemic situation was observed.

The samples obtained from each rural population were denoted by A, B, C, D, E and F respectively. These samples are over-dispersed (see Table 1) and can be fitted by a negative binomial distribution (see Figure 1).

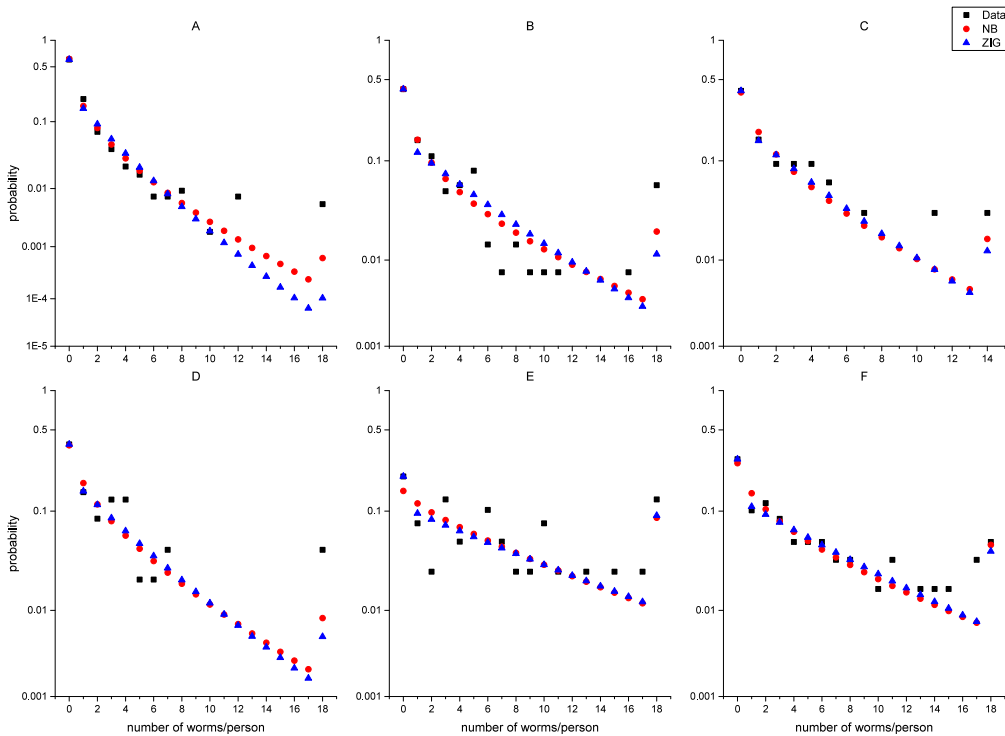


Figure 1. Fitting the parasite counts data (black) by NB (red) and ZIG (blue) distribution for Seo *et al.* (*ibid.*) data set. Except in the first case, the simple zero-inflated geometric distribution fit the data as well as the negative binomial distribution (see Table 1).

In Figure 1 we show the observed (black) and expected values of the fitted models (negative binomial and zero-inflated geometric). In addition, Table 1 includes the maximum likelihood estimations, the chi-squared statistics for goodness-of-fit (D’Agostino, 2017), their corresponding *p*-values, and Akaike’s information criterion (AIC) (Konishi & Kitagawa, 2008). As we see, the zero-inflated geometric distribution fit the data as well as the negative binomial distribution in most cases. The AIC results show that the negative binomial and zero-inflated geometric models produce a similar fit. Indeed, the fit in Figure 1 improves the results obtained by negative binomial distribution in samples E and F. On the other hand, using the AIC, the geometric zero-inflated model showed the best performance in samples C, E and F. Hence, the zero-inflated geometric distribution, $ZIG(\pi, q)$, is a suitable candidate model to fit such data.

Table 1. Parameters of NB and ZIG distributions calculated from observed Seo *et al.* (1979) data and results of chi-squared test and AIC

Theoretical distribution	Calculated parameters	Samples					
		A (n = 540)	B (n = 136)	C (n = 32)	D (n = 47)	E (n = 39)	F (n = 59)
NB	<i>m</i>	1.0167	2.8235	2.3125	2.5106	6.6410	4.6102
	<i>k</i>	0.3546	0.4240	0.5761	0.5893	0.8726	0.6193
	chi-squared statistic	50.0660	25.8883	6.8092	17.4885	15.6914	11.2927
	<i>p</i> -value	<0.0001	0.0556	0.8699	0.3547	0.4747	0.7911
	AIC	1438.0235	574.5621	131.7761	198.1410	235.0388	310.0239
ZIG	π	0.3653	0.2687	0.2011	0.1819	0.0971	0.1581
	<i>q</i>	0.3843	0.2057	0.2568	0.2458	0.1197	0.1544
	chi-squared statistic	213.1743	43.2591	6.8197	22.2336	14.4552	10.6473
	<i>p</i> -value	<0.0001	0.0003	0.8693	0.1358	0.5648	0.8307
	AIC	1458.0335	579.2921	131.5360	198.1585	233.3277	308.9315
	df	16	16	12	16	16	16
	dispersion index	4.7973	7.5077	4.6992	5.9204	5.8447	6.5972

5.2 Parasite distribution in crabs

Crofton (Crofton, 1971) contributed significantly to the study of parasite distributions in hosts. In his works he observed that over-dispersion is one of the main characteristics of parasite-host distributions.

Analyzing data from Hynes & Nicholas (1963) of six samples (A to F) of the parasitic infection of the crustacean *Gammarus pulex*, by the parasite acanthocephalan *Polymorphus minutus*, Crofton showed that the negative binomial distribution provides a good fit to the data.

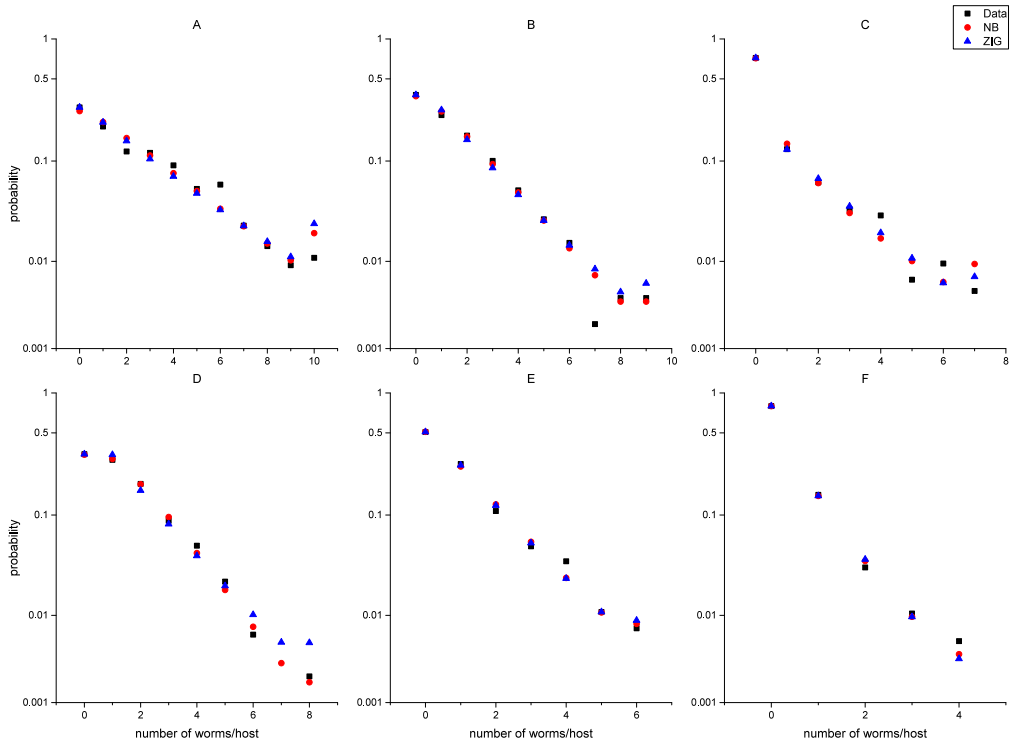


Figure 2. Fitting the parasite counts data (black) by NB (red) and ZIG (blue) distribution for data in parasite in crabs (Crofton, 1971). Zero-inflated (deflated) geometric distribution fit the data as well or better than the negative binomial distribution (see Table 2).

In Figure 2 we compare the fit to the data analyzed by Crofton using the negative binomial distribution and the zero-inflated (deflated) geometric distribution. In most cases our simple proposal provides a better fit than the negative binomial distribution.

Based on the chi-square goodness-of-fit test and the AIC reported in Table 2 we conclude that the zero-inflated (deflated) geometric distribution provides a fit of the data as good as the negative binomial distribution.

Table 2. Parameters of NB and ZIG distributions calculated from observed of parasite in crabs data (Crofton, 1971) and results of chi-squared test and AIC

Theoretical distribution	Calculated parameters	Samples					
		A (n = 549)	B (n = 509)	C (n = 633)	D (n = 486)	E (n = 276)	F (n = 191)
NB	<i>m</i>	2.2732	1.4165	0.6003	1.3189	0.8913	0.2670
	<i>k</i>	1.2564	1.5837	0.2974	3.0544	1.2679	0.6069
	chi-squared statistic	20.6558	3.1086	10.5075	2.9993	2.3843	0.2776
	<i>p</i> -value	0.0081	0.8748	0.0621	0.8089	0.6655	0.8704
	AIC	2211.4460	1662.1623	1279.5742	1506.5751	724.9286	252.4139
ZIG	π	-0.0256	-0.1304	0.4875	-0.3313	-0.1020	0.2195
	<i>q</i>	0.3109	0.4438	0.4605	0.5023	0.5528	0.7451
	chi-squared statistic	23.4185	6.0467	6.5825	8.6706	2.1793	0.4542
	<i>p</i> -value	0.0029	0.5343	0.2536	0.1930	0.7028	0.7969
	AIC	2215.4026	1665.9254	1274.8293	1514.0280	724.8346	252.4989
	df	8	7	5	6	4	2
	dispersion index	2.4235	1.8340	2.5774	1.4125	1.6757	1.4464

6. Discussion and Conclusions

The Negative binomial distribution is widely used to describe parasite burden in populations because this distribution often provides a good fit of the observations which can be improved by the corresponding zero-inflated (deflated) distribution (Crofton, 1971; Seo *et al.*, 1979).

However, parameters estimation is far from trivial, and maximum likelihood estimates must be found always numerically. Simple numerical methods (as the Newton method, Atkinson (2008)) are not easy to implement as there is not a closed expression for the derivative of the Gamma function and generally fail if the starting value is not chosen appropriately (Bandara *et al.*, 2019).

The Negative binomial distribution may also fail to fit the zero counts (Crofton, 1971). This issue may be overcome using the zero-inflated negative binomial distribution, however, parameters estimation by maximum likelihood is even more complex and also the AIC criteria penalize models with a larger number of parameters.

However zero-inflated geometric or hurdle geometric models are simpler alternatives to the negative binomial distribution, for which we found simple formulas for the maximum likelihood parameter estimates.

In the examples analyzed in this work zero-inflated geometric or hurdle geometric models present, in most cases, a similar fit to the data than the negative binomial model, and in some cases these models significantly improve the fit.

The AIC results show that the models are similar performance in both datasets.

However the major advantage of these models is not a little improvement in the data fitting but the fact that these models are simpler and with maximum likelihood parameter estimates easier to obtain than the negative binomial model.

For most of the sample considered, observations in the tail of the distribution may look too far from the expected values. However this fact does not necessarily indicates the need to consider other distributions. Since the samples are small, rare events in the tail, when observed acts like outliers given too much weight to this rare observation. The simulations presented in the Appendix are consistent with this explanation.

Acknowledgments

This work was partially supported by grant CIUNSA 2018-2467. JPA is a member of the CONICET. GML is a postdoctoral fellow of CONICET. We would like to thank reviewers and editors for their comments.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Atkinson, K. *An introduction to numerical analysis* (John Wiley & Sons, 2008).
2. Bandara, U., Gill, R. & Mitra, R. On computing maximum likelihood estimates for the negative binomial distribution. *Statistics & Probability Letters* **148**, 54–58 (2019).
3. Bliss, C & Fisher, R. Fitting the negative binomial distribution to biological data. *Biometrics* **9**, 176–200 (1953).
4. Clark, S. & Perry, J. Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics* **45**, 309–316 (1989).
5. Cox, D. R. & Lewis, P. A. *The statistical analysis of series of events* (Springer, 1966).
6. Crofton, H. A quantitative approach to parasitism. *Parasitology* **62**, 179–193 (1971).
7. D'Agostino, R. B. *Goodness-of-fit-techniques* (Routledge, 2017).
8. Dai, H., Bao, Y. & Bao, M. Maximum likelihood estimate for the dispersion parameter of the negative binomial distribution. *Statistics & Probability Letters* **83**, 21–27 (2013).
9. Dietz, E. & Böhning, D. On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics & Data Analysis* **34**, 441–459 (2000).
10. Gourbière, S., Morand, S. & Waxman, D. Fundamental factors determining the nature of parasite aggregation in hosts. *PloS one* **10**, e0116893 (2015).
11. Greene, W. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. *NYU Working Paper No. EC-94-10* (1994).
12. Hall, D. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56**, 1030–1039 (2000).
13. Hynes, H & Nicholas, W. The importance of the acanthocephalan *Polymorphus minutus* as a parasite of domestic ducks in the United Kingdom. *Journal of Helminthology* **37**, 185–198 (1963).
14. Johnson, N., Kemp, A. & Kotz, S. *Univariate discrete distributions* (John Wiley & Sons, 2005).
15. Konishi, S & Kitagawa, G. *Information criteria and statistical modeling* (Springer, 2008).
16. Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14 (1992).
17. Min, Y. & Agresti, A. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling* **5**, 1–19 (2005).
18. Piegorsch, W. Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics* **46**, 863–867 (1990).
19. Pullan, R. & Brooker, S. The global limits and population at risk of soil-transmitted helminth infections in 2010. *Parasites & vectors* **5**, 1–14 (2012).

20. Saha, K. & Paul, S. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* **61**, 179–185 (2005).
21. Seo, B, Cho, S. & Chai, J. Frequency distribution of *Ascaris lumbricoides* in rural Koreans with special reference on the effect of changing endemicity. *Korean J Parasitol* **17**, 105–113 (1979).
22. Shaw, D., Grenfell, B. & Dobson, A. Patterns of macroparasite aggregation in wildlife host populations. *Parasitology* **117**, 597–610 (1998).
23. Welsh, A., Cunningham, R., Donnelly, C. & Lindenmayer, D. Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling* **88**, 297–308. ISSN: 0304-3800 (1996).
24. Willson, L., Folks, J & Young, J. Multistage estimation compared with fixed-sample-size estimation of the negative binomial parameter k. *Biometrics* **40**, 109–117 (1984).
25. Woolhouse, M. *et al.* Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proceedings of the National Academy of Sciences* **94**, 338–342 (1997).

Appendix

As we show in Figure 1, some observations are not close to the predicted value obtained from the best fitted negative binomial distribution. Here we show that even for the case of the two largest samples ($n = 540$ and $n = 136$, see Table 1) the observations are compatible with a negative binomial distribution. In each case, we produced 1000 simulated samples of sizes $n = 540$ and $n = 136$, drawn from the best fitted negative binomial.

The simulation results are shown in Figure 3 where the simulated samples are in black dots and the Seo *et al.* (1979) samples in red diamonds.

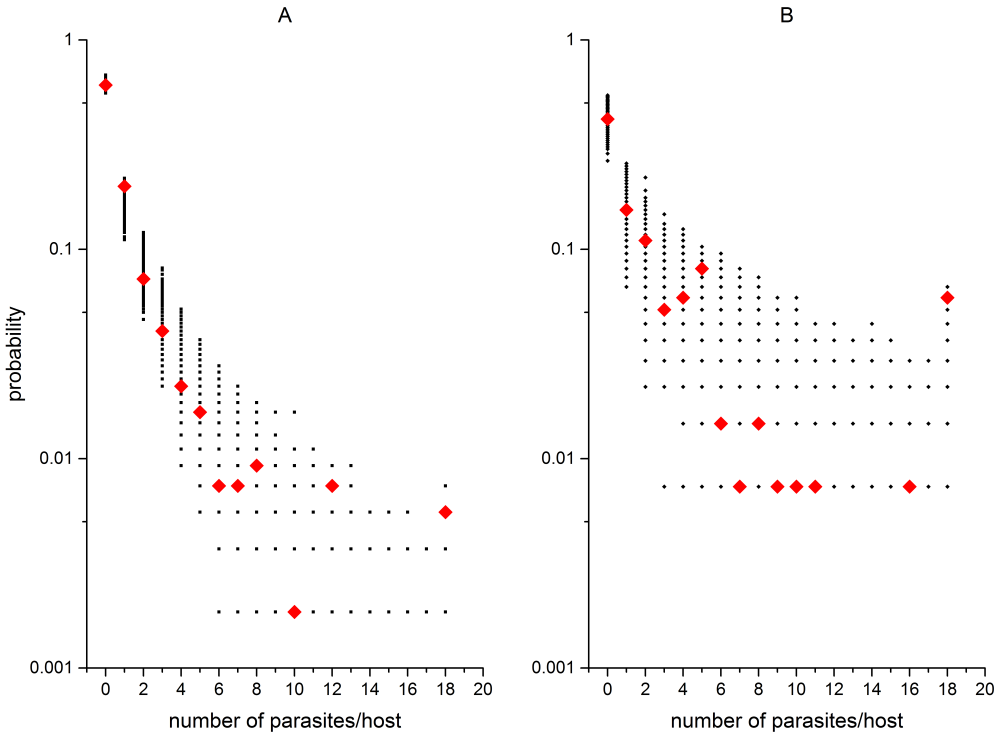


Figure 3. A and B samples of Seo *et al.* (ibid.) in red diamonds and the simulated samples in black dots.