Brazilian Journal of Probability and Statistics
2010, Vol. 24, No. 1, 91–99
DOI: 10.1214/09-BJPS103

A generalized negative binomial distribution based on an extended Poisson process

Luis Ernesto Bueno Salasar, José Galvão Leite and Francisco Louzada-Neto

Universidade Federal de São Carlos

Abstract. In this article we propose a generalized negative binomial distribution, which is constructed based on an extended Poisson process (a generalization of the homogeneous Poisson process). This distribution is intended to model discrete data with presence of zero-inflation and over-dispersion. For a dataset on animal abundance which presents over-dispersion and a high frequency of zeros, a comparison between our extended distribution and other common distributions used for modeling this kind of data is addressed, supporting the fitting of the proposed model.

1 Introduction

Although the Poisson distribution is very usual for modeling discrete data, it may provide a poor fit in the presence of over-dispersion. Thus, in that case, the negative binomial (NB) distribution [e.g., Bliss and Fisher (1953), Greenwood and Yule (1920)] is a frequently used model. Besides over-dispersion, data may present a high frequency of zeros and neither the Poisson nor the negative binomial distributions achieve a good fit. So, we propose a generalization of the negative binomial distribution based on an extended Poisson process, hereafter GNB, which can handle data with both over-dispersion and high frequency of zeros. Although the inclusion of covariates is a very important issue for practical proposes it is out of the scope of this paper.

Several authors [e.g., Ridout, Hinde and Demétrio (2001), Yau, Wang and Lee (2003), Lewsey and Thomson (2004)] have considered zero inflated Poisson (ZIP) or zero inflated negative binomial (ZINB) model to handle data in presence of over-dispersion and high frequency of zeros. For these models, the underlying data generating process can be interpreted as a two-stage process. In the first stage, we choose from either a distribution that only generates zeros or a distribution that can generate any count (e.g., Poisson, negative binomial, etc.). In the second stage, an observation is generated from the chosen distribution. So, a zero count can arise from any of the two distributions. Although, this data-generating process may be adequate in many situations, the assumption that some observed units only provide zeros (structural zeros) may not be realistic. For instance, in insurance claim

Key words and phrases. Discrete probabilistic models, excess of zeros, extended Poisson process, negative binomial distribution, over-dispersion, probabilistic model.

Received December 2008; accepted April 2009.

frequency studies this assumption would mean that some policy holders cannot produce any accidents. However, the model proposed in this article does not rely on such assumptions. Indeed, the data-generation mechanism is a counting process, such that the probability of a new occurrence may depend on the accumulated number of occurrences. As we shall see later, the modeling idea considered here is similar to that of the hurdle models, which models separately the zero and nonzero occurrences, but in the usual instances with no covariates the model obtained is just a reparametrization of the zero-inflated one. A very good work comparing different modeling approaches to data with many zeros can be found in Ridout, Demétrio and Hinde (1998).

The remainder of this paper is organized as follow. In Section 2, we define the extended Poisson process and show how one can find the probability function for the GNB model. In Section 3, we present the expressions for the likelihood and log-likelihood function and comment on how maximum likelihood estimates are obtained. In Section 4, we consider a dataset on animal abundance and find the maximum likelihood estimates for the NB and GNB models. We also fit Poisson, ZIP and ZINB models and compare them by means of the AIC [Akaike (1973)] and BIC [Schwarz (1978)] criteria. Finally, in Section 5, we give some closing comments.

2 Probabilistic model

Faddy (1997) constructed discrete models based on an extended Poisson process allowing for over and under-dispersion relative to the Poisson distribution. A homogeneous continuous-time Markov chain $\{X(t); t \geq 0\}$ with state space $\mathbb{N} = \{0, 1, 2, \ldots\}$ is called an extended Poisson process (or a pure birth process) if the following conditions hold:

1. $X(0) = 0$;
2. if $s < t$, then $X(s) \leq X(t)$;
3. $P\{X(t+h) = n+1 | X(t) = n\} = \lambda(n)h + o(h)$;
4. $P\{X(t+h) > n+1 | X(t) = n\} = o(h)$;

for all $s, t \geq 0, h > 0, n \in \mathbb{N}$.

Interpreting t as time, the random variable $X(t)$ represents the number of occurrences of an event until the instant t. From conditions 3 and 4 above we can conclude that transition probabilities may depend on the current state n of the process, that is, the cumulative number of the event occurrences. Also, note that the homogeneous Poisson process is a particular case of this process when transition rates are constant.

The transition probabilities are defined by

$$p_{i,n}(t) = P\{X(s+t) = n | X(s) = i\},$$

where $t > 0, s \geq 0, i, n \in \mathbb{N}$, can be expressed for the extended Poisson process, using the Chapman–Kolmogorov forward differential equations [Feller (1971)], as

$$p'_{i,i}(t) = -\lambda(i)\, p_{i,i}(t), \tag{2.1}$$

$$p'_{i,n}(t) = -\lambda(n)\, p_{i,n}(t) + \lambda(n-1)\, p_{i,n-1}(t), \qquad n > i, \tag{2.2}$$

with initial conditions $p_{i,i}(0) = 1$ and $p_{i,n}(0) = 0$ for $n > i$.

From equations (2.1) and (2.2), we can find the following recursive solution,

$$p_{i,i}(t) = e^{-\lambda(i)t}, \tag{2.3}$$

$$p_{i,n}(t) = \lambda(n-1)e^{-\lambda(n)t} \int_0^t e^{\lambda(n)s}\, p_{i,n-1}(s)\, ds, \qquad n > i. \tag{2.4}$$

Taking $i = 0$ on (2.3) and (2.4), we also obtain a recursive solution for the probabilities $p_n(t) = P\{X(t) = n\}$, $n = 0, 1, 2, \ldots$, which is given by

$$p_0(t) = e^{-\lambda(0)t}, \tag{2.5}$$

$$p_n(t) = \lambda(n-1)e^{-\lambda(n)t} \int_0^t e^{\lambda(n)s}\, p_{n-1}(s)\, ds, \qquad n > 0. \tag{2.6}$$

As shown by Faddy (1997), an interesting feature of probabilities $p_n(t)$ is that for any discrete probability distribution $(\pi_0, \pi_1, \pi_2, \ldots)$ on $\mathbb{N}$, a sequence $(\lambda(0), \lambda(1), \lambda(2), \ldots)$ of transition rates can be found such that for some fixed t, $p_n(t) = \pi_n$, for all $n \in \mathbb{N}$.

It is important to note that not all sequence of transition rates give rise to a honest process, that is, a process with $\sum_{n=0}^{\infty} p_n(t) = 1$, for all $t > 0$. A necessary and sufficient condition for this process to be honest is that the series $\sum_{j=0}^{\infty} (\lambda(j))^{-1}$ diverges [Feller (1971)].

Faddy (1997) also showed that if $\{X(t); t \geq 0\}$ is an extended Poisson process with transition rates

$$\lambda(n) = a(b+n), \tag{2.7}$$

with $a > 0$, $b > 0$, $n \in \mathbb{N}$, then the random variable $X(t)$ has NB distribution with parameters b and e^{-at}, that is,

$$P\{X(t) = n\} = f_{\mathrm{NB}}(n; b, e^{-at}) = \binom{b+n-1}{n} e^{-abt}(1 - e^{-at})^n, \tag{2.8}$$

where $f_{\mathrm{NB}}(n; r, p)$ is the probability distribution function at $n \in \mathbb{N}$ of a NB distribution with parameters r and p.

A generalization of this process is obtained by considering an extended Poisson process $\{Y(t); t \geq 0\}$ with transition rates given by

$$\lambda(n) = \begin{cases} \lambda_0, & \text{if } n = 0, \\ a(b+n), & \text{if } n \geq 1, \end{cases} \tag{2.9}$$

where $\lambda_0 > 0, a > 0$ and $b > -1$. Note that if $\lambda_0 = ab$ and $b > 0$, we have the same rates given in (2.7) and also, in this case, the probability distribution function of $Y(t)$ is the NB distribution given by (2.8). The proposed GNB model is defined as the probability distribution of $Y(t)$.

To obtain the probability distribution of $Y(t)$, consider a random variable T_0 denoting the holding time of the process at state 0, that is, the time of the first occurrence of the event. The time T_0 is exponentially distributed with parameter λ_0, since $P\{T_0 \leq t\} = 1 - P\{Y(t) = 0\}$ and from (2.5) and (2.9) we have

$$P\{Y(t) = 0\} = e^{-\lambda_0 t}. \tag{2.10}$$

Hence, the probability distribution of $Y(t)$ can be obtained as

$$\begin{aligned} P\{Y(t) = n\} &= \int_0^\infty P\{Y(t) = n | T_0 = \tau\} f_{T_0}(\tau)\, d\tau \\ &= \int_0^\infty P\{Y(t) = n | T_0 = \tau\} \lambda_0 e^{-\lambda_0 \tau}\, d\tau, \qquad n \in \mathbb{N}. \end{aligned} \tag{2.11}$$

For $n \geq 1$, the conditional probability in (2.11) is given by

$$\begin{aligned} &P\{Y(t) = n | T_0 = \tau\} \\ &\quad = \begin{cases} 0, & \text{if } \tau > t, \\ P\{Y(t) = n | [Y(s) = 0, \forall s < \tau], Y(\tau) = 1\}, & \text{if } \tau \leq t, \end{cases} \end{aligned}$$

which, from the Markovian property [Cox and Miller (1965)], implies that

$$P\{Y(t) = n | T_0 = \tau\} = \begin{cases} 0, & \text{if } \tau > t, \\ p_{1,n}(t - \tau), & \text{if } \tau \leq t. \end{cases} \tag{2.12}$$

Replacing (2.12) in (2.11), it follows that, for $n \geq 1$,

$$P\{Y(t) = n\} = \int_0^t p_{1,n}(t - \tau) \lambda(0) e^{-\lambda(0)\tau}\, d\tau. \tag{2.13}$$

From (2.1) and (2.2) with $i = 1$, we have the following differential equations

$$\begin{aligned} p'_{1,1}(x) &= -\lambda(1) p_{1,1}(x), \\ p'_{1,n}(x) &= -\lambda(n) p_{1,n}(x) + \lambda(n-1) p_{1,n-1}(x), \qquad n \geq 2. \end{aligned}$$

Now, considering the transformations

$$\begin{aligned} g_n(x) &= p_{1,n+1}(x), \qquad n = 0, 1, 2, \ldots, \\ \xi(n) &= \lambda(n+1), \qquad n = 0, 1, 2, \ldots, \end{aligned}$$

we obtain the following system of differential equations for the functions $\{g_n(x), n = 0, 1, \ldots\}$,

$$g'_0(x) = -\xi(0) g_0(x), \tag{2.14}$$

$$g'_n(x) = -\xi(n) g_n(x) + \xi(n-1) g_{n-1}(x), \qquad n \geq 1, \tag{2.15}$$

where $\xi(n) = \lambda(n+1) = a(b+n+1)$, for $n \geq 0$. The functions $\{g_n(x), n = 0, 1, \ldots\}$ can be thought as the solution of the Chapman–Kolmogorov differential equations with transition rates $\xi(n)$, $n \geq 0$, that are linearly increasing. So, from the previous commented result proved by Faddy (1997), the solution of the system of differential equations (2.14) and (2.15) is given by

$$g_n(x) = f_{\mathrm{NB}}(n; b+1, e^{-ax}), \qquad n = 0, 1, 2, \ldots,$$

which implies that,

$$\begin{aligned} p_{1,n}(x) &= g_{n-1}(x) \\ &= f_{\mathrm{NB}}(n-1; b+1, e^{-ax}) \\ &= \binom{b+n-1}{n-1} (e^{-ax})^{b+1} (1-e^{-ax})^{n-1} \qquad \text{for } n \geq 1. \end{aligned} \tag{2.16}$$

Replacing (2.16) in (2.13), for $n \geq 1$, we have

$$\begin{aligned} &P\{Y(t) = n\} \\ &\quad = \int_0^t \binom{b+n-1}{n-1} \left[e^{-a(t-\tau)}\right]^{b+1} \left[1 - e^{-a(t-\tau)}\right]^{n-1} \lambda_0 e^{-\lambda_0 \tau}\, d\tau \\ &\quad = \binom{b+n-1}{n-1} \lambda_0 \int_0^t e^{-a(b+1)(t-\tau)} e^{-\lambda_0 \tau} \left[1 - e^{-a(t-\tau)}\right]^{n-1} d\tau, \end{aligned}$$

and substituting $x = t - \tau$ in the last integral, follows that

$$\begin{aligned} &P\{Y(t) = n\} \\ &\quad = \binom{b+n-1}{n-1} \lambda_0 e^{-\lambda_0 t} \int_0^t e^{[\lambda_0 - a(b+1)]x} [1 - e^{-ax}]^{n-1}\, dx \\ &\quad = \binom{b+n-1}{n-1} \lambda_0 e^{-\lambda_0 t} \int_0^t (e^{-ax})^{b+1-\lambda_0/a} [1 - e^{-ax}]^{n-1}\, dx. \end{aligned} \tag{2.17}$$

Finally, considering $y = e^{-ax}$ in the integral of (2.17), it follows, for $n \geq 1$, that

$$\begin{aligned} &P\{Y(t) = n\} \\ &\quad = \binom{b+n-1}{n-1} \frac{\lambda_0 e^{-\lambda_0 t}}{a} \int_{e^{-at}}^1 y^{b-\lambda_0/a} [1-y]^{n-1}\, dy. \end{aligned} \tag{2.18}$$

Hence, the probability distribution function of the GNB model is determined by (2.10) and (2.18). Also, note that if $\lambda_0 = ab$ and $b > 0$ in (2.10) and (2.18), we obtain the probability distribution function of the NB model given in (2.8).

3 Estimation and inference

Considering $Y_1, Y_2, \ldots, Y_n$ a random sample of $Y(1)$, whose distribution is given by (2.10) and (2.18) with $t = 1$, the likelihood function is calculated as

$$\begin{aligned}
L(\lambda_0, a, b) &= \prod_{i=1}^{n} P\{Y_i = y_i\} \\
&= \prod_{i:y_i=0} P\{Y_i = y_i\} \prod_{i:y_i>0} P\{Y_i = y_i\} \\
&= \prod_{i:y_i=0} e^{-\lambda_0} \prod_{i:y_i>0} \binom{b+y_i-1}{y_i-1} \frac{\lambda_0 e^{-\lambda_0}}{a} \int_{e^{-a}}^{1} x^{b-\lambda_0/a}[1-x]^{y_i-1}\,dx \\
&= \frac{e^{-n_0\lambda_0}\lambda_0^{(n-n_0)}e^{-(n-n_0)\lambda_0}}{a^{n-n_0}} \\
&\quad \times \prod_{i:y_i>0} \binom{b+y_i-1}{y_i-1} \int_{e^{-a}}^{1} x^{b-\lambda_0/a}[1-x]^{y_i-1}\,dx \\
&= \frac{\lambda_0^{(n-n_0)}e^{-n\lambda_0}}{a^{n-n_0}} \prod_{i:y_i>0} \binom{b+y_i-1}{y_i-1} \int_{e^{-a}}^{1} x^{b-\lambda_0/a}[1-x]^{y_i-1}\,dx,
\end{aligned}$$

where $y_1, \ldots, y_n$ are the observed values of $Y(1)$, $n_0 = \sum_{j=1}^{n} I_{\{0\}}(y_j)$ is the number of zeros in the sample, $\lambda_0 > 0$, $a > 0$, and $b > -1$.

Thus, the log-likelihood function is obtained as

$$\begin{aligned}
l(\lambda_0, a, b) &= -n\lambda_0 + (n-n_0)\big(\log(\lambda_0) - \log(a)\big) \\
&\quad + \sum_{i:y_i>0} \log\left\{ \binom{b+y_i-1}{y_i-1} \int_{e^{-a}}^{1} x^{b-\lambda_0/a}(1-x)^{y_i-1}\,dx \right\}.
\end{aligned}$$

To improve the performance of the maximization algorithm used to obtain the maximum likelihood estimates, we consider the following reparametrization: $l_0 = \log(\lambda_0), l_a = \log(a), l_b = \log(b+1)$, which has no constraints.

Hence, the log-likelihood function for the new parametrization is expressed as

$$\begin{aligned}
&l(l_0, l_a, l_b) \\
&\quad = -ne^{l_0} + (n-n_0)(l_0 - l_a) \\
&\qquad + \sum_{i:y_i>0} \log\left\{ \binom{e^{l_b}+y_i-2}{y_i-1} \int_{\exp(-e^{l_a})}^{1} x^{e^{l_b}-e^{l_0-l_a}-1}(1-x)^{y_i-1}\,dx \right\}.
\end{aligned} \tag{3.1}$$

For a given dataset, maximum likelihood estimates can be obtained by direct maximization of the expression (3.1), since no major simplifications are possible.

For model comparison, we shall consider the AIC and BIC criteria, which are defined, respectively, by $-2\log(L(\widehat{\theta})) + 2p$ and $-2\log(L(\widehat{\theta})) + p\log(n)$, where $\widehat{\theta}$ is the maximum likelihood estimate, p is the number of parameters estimated, and n is the sample size.

4 Data analysis

As an example, we consider the count data of *Macoma liliana*, a small clam, observed in four sites. Faddy (1997) investigated the counts of site A, that presents over-dispersion and a high frequency of zeros. For these data, the maximum likelihood and the standard deviation estimates (in brackets) for the NB and GNB models were obtained by direct maximization of (3.1). The calculations involved were carried out using the R statistical software [R Development Core Team (2007)] and are presented in Table 1.

Table 2 displays a comparison between observed and expected frequencies for Poisson (PO), NB, ZIP, ZINB and GNB models as well as the AIC and BIC criterion values for each model.

Table 1 *Maximum likelihood estimates*

	NB	GNB
l_0		0.4007 (0.1931)
l_a	0.8318 (0.1311)	−0.8068 (0.9675)
l_b	0.6223 (0.1310)	3.3058 (1.1784)

Table 2 *Observed and expected frequency distribution*

Counts	Observed	PO	NB	ZIP	ZINB	GNB
0	9	0.02	5.50	9.00	9.00	8.99
1	2	0.14	4.27	0.01	0.72	1.26
2	1	0.53	3.58	0.07	1.21	1.37
3	2	1.36	3.07	0.24	1.65	1.50
4	1	2.62	2.67	0.60	1.99	1.62
5	1	4.05	2.34	1.19	2.21	1.75
6	0	5.21	2.05	1.98	2.32	1.86
7	4	5.75	1.81	2.82	2.33	1.96
8	1	5.56	1.60	3.51	2.26	2.03
9	4	4.77	1.42	3.89	2.14	2.06
10	1	3.68	1.26	3.88	1.97	2.04
11	3	2.59	1.12	3.51	1.79	1.98
≥12	11	3.72	9.29	9.28	10.41	11.57
AIC		386.08	252.25	273.20	241.49	238.75
BIC		387.77	255.63	276.58	246.56	243.81

We observe that the dataset presents a high frequency of zeros that cannot be accommodated neither by the PO nor by the NB model, but the ZIP, ZINB and GNB models achieve a good fit at this count (see Table 2). Strong evidence in favor of the GNB model when compared to the NB model is observed when carrying out a likelihood ratio test of the hypothesis $H_0 : \lambda_0 = ab$, that provides a p-value equal to 8.22×10^{-5}. Also, both selection criterion (AIC and BIC) lead to the choice of the GNB model in comparison to the other models.

5 Final remarks

We argue that the generalized distribution GNB is useful to handle data with presence of over-dispersion and a high frequency of zeros, which cannot be accommodated, for instance, by the usual NB distribution.

For the count data considered in Section 4, we observed a better fit of the GNB model in comparison to the PO and NB in the light of observed and expected frequencies. Also, both selection criteria (AIC and BIC) indicate a better fitting of the proposed model when compared to the other usual models. This fact can be explained by the over-dispersion and the large number of zeros in the sample.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* 267–281. Akadémia Kiadó, Budapest. MR0483125

Bliss, C. I. and Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics* **9** 176–200. MR0055625

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Wiley, New York. MR0192521

Faddy, M. J. (1997). Extended Poisson process modelling and analysis of count data. *Biometrical Journal* **39** 431–440.

Feller, W. (1971). *An Introduction to Probability Theory and Applications*, Vol. I, 3rd ed. Wiley, New York. MR0228020

Greenwood, M. and Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happening with special reference of multiple attacks of disease or repeated accidents. *Journal of the Royal Statistical Society* **83** 255–279.

Lewsey, J. D. and Thomson, W. M. (2004). The utility of the zero-inflated Poisson and zero-inflated negative binomial models: A case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. *Community Dentistry and Oral Epidemiology* **32** 183–189.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at http://www.R-project.org.

Ridout, M., Demétrio, C. G. B. and Hinde, J. (1998). Models for count data with many zeros. In *Proceedings of the XIXth International Biometric Conference* 179–192. Cape Town, South Africa.

Ridout, M., Hinde, J. and Demétrio, C. G. B. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* **57** 219–223. MR1833310

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6** 461–464. MR0468014

Yau, K. K. W., Wang, K. and Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal* **45** 437–452. MR1984622

L. E. B. Salasar
J. G. Leite
F. Louzada-Neto
Departamento de Estatística
Universidade Federal de São Carlos
CP 676, ZIP 13565-905
São Carlos, SP
Brazil
E-mails: luis.salasar@gmail.com
leite@ufscar.br
dfln@ufscar.br