Sociedad de Estadística e Investigación Operativa
Test (2006) *Vol.* **15**, *No.* 2, pp. 423–432

A Bayesian Choice Between Poisson, Binomial and Negative Binomial Models

Jean-Yves Dauxois, Pierre Druilhet and Denys Pommeret*
Department of Statistics
ENSAI, France

Abstract

In this paper, we propose a Bayesian method for modelling count data by Poisson, binomial or negative binomial distributions. These three distributions have in common that the variance is, at most, a quadratic function of the mean. We use prior distributions on the variance function coefficients to consider simultaneously the three possible models and decide which one fits the data better. This approach sheds new light on the analysis of the Sibship data (Sokal and Rohlf, 1987). The Jeffreys-Lindley paradox is discussed through some illustrations.

Key Words: Jeffreys-Lindley paradox, natural exponential family, overdispersion, Sibship data, variance function.

AMS subject classification: 62C12.

1 Introduction

In count data modelling, the Poisson, the binomial or the negative binomial distributions are often used to fit the data and the problem of choosing between these three distribution families is a topic widely studied in the literature.

For instance, Greenwood and Yule (1920) found that Poisson distribution fitted poorly the number of accidents experienced by a group of workers whereas the negative binomial distribution fitted the data very well. Edwards and Gurland (1961) used the negative binomial and Poisson distributions for modelling accident data. Lenk (1999) proposed Poisson and

*Correspondence to: Denys Pommeret. ENSAI, Rue Blaise Pascal, BP 37203, 35175 Bruz Cedex, France. E-mail: pommeret@ensai.fr

Received: February 2003; Accepted: August 2004

negative binomial distributions for traffic accident counts (see also Gelfand and Dalal, 1990, for a similar study of accident data). Indeed, in such situations the presence of "non-Poisson" variation should suggest a negative binomial model. Similar problems of overdispersed models have been considered by Ramakrishnan and Meeter (1993) to analyse the data set from a pollution impact study (see also Dean and Lawless (1989) for a theoretical approach and Preece et al. (1988) for the study of Bortkewitsch's data). Note also that Poisson model can be used for "extra-binomial" variation as in Williams (1996) in the linear model area.

The important point to note here is that the Poisson, the binomial and the negative binomial distributions belong to one-parameter natural exponential families. Thus, they are characterized by their variance functions; that is, the relationships between their variances and their means. In our case, the variance functions are second-order polynomials. The purpose of this paper is to construct a Bayesian method of model choice based on prior distributions on the coefficients of the variance function. The method is quite automatic and one of its advantages appears when distributions from different families are very close. Roughly speaking, one can say that the Poisson model is on the borderline between the binomial and negative binomial models. In such a situation, p-values in Chi-squared goodness-of-fit tests are difficult to interpret and are sometimes misleading. Moreover, this property leads to an interesting version of the Jeffreys-Lindley paradox: when the prior distribution tends to a vague prior, the posterior distribution converges towards the prior distribution of each model and not towards a distribution that gives a mass 1 to the Poisson model.

In Section 2, we exhibit the common parameter in the variance function that distinguishes the three families of distributions. We then present the Bayesian method of model selection. In Section 3 a parallel with the Jeffreys-Lindley paradox is discussed. In Section 4, we present two applications. First of all, the overdispersed Sibship data set appears to be well fitted by a truncated binomial distribution. Next, our conclusion for the well-known Bortkewitsch data set agrees with the use of the negative binomial distribution, proposed by Preece et al. (1988).

2 A Bayesian choice

2.1 Natural exponential families and their variance functions

For all non negative mean parameter m, let us denote by $\mathcal{P}(m)$ the Poisson distribution and $\mathcal{B}(N, m)$ (resp. $\mathcal{NB}(N, m)$) the binomial (resp. negative binomial) distribution. One knows that all these models belong to the general framework of Natural Exponential Families (NEFs). Since different values of N in $\mathbb{N}$ give different NEFs in the binomial as well as in the negative binomial cases, the symbol $(B)_N$ (resp. $(NB)_N$) stands in the sequel for the binomial (resp. negative binomial) NEF with parameter N.

It is well known that, for all these NEFs, writing the variance V as a function of the mean m, yields

$$V(m) = a\, m^2 + m, \tag{2.1}$$

where the value of a characterizes the NEF. Thus, a null value for a relates to the Poisson NEF, a negative one to the binomial $(B)_{-1/a}$ NEF and a positive one to the negative binomial $(NB)_{1/a}$ NEF. Then the problem of model choice reduces to the estimation of the sign of a.

2.2 The method

We assume that, given a, $X_1, \cdots, X_q$ are independent and identically distributed with distribution in the Poisson NEF if $a = 0$, in a binomial $(B)_{-1/a}$ NEF if $a \in -1/\mathbb{N}^*$ or in a negative binomial $(NB)_{1/a}$ NEF if $a \in 1/\mathbb{N}^*$.

A Bayesian choice of the type of NEFs among the three ones under consideration consists in comparing the three posterior probabilities $P(a = 0|X = x)$, $P(a > 0|X = x)$ and $P(a < 0|X = x)$, where $X = (X_1, \cdots, X_q)$. Then we can choose the type of model corresponding to the highest posterior probability. We can also conclude that different types of models explain data equally well when their associated posterior probabilities are close.

If, rather than selecting a type of model between three types, one is interested in choosing the best model between an infinite set of models, then a Bayesian choice of model consists in choosing the value of a, which maximizes its posterior probabilities.

2.3 Prior and posterior distributions on a

In this paper, we consider two simple prior distributions for a, denoted by Π and Π^*, allocating equal weights to the three families of distributions. The first prior distribution is derived from Poisson distributions so that it gives a non-null weight for all possible values of a. It is defined, for all n in $\mathbb{N}^*$, by

$$\begin{aligned}
\Pi_{\alpha,\beta}(a=0) &= \frac{1}{3} \\
\Pi_{\alpha,\beta}(a=1/n) &= \frac{1}{3}\, e^{-\alpha}\frac{\alpha^{n-1}}{(n-1)!} \\
\Pi_{\alpha,\beta}(a=-1/n) &= \frac{1}{3}\, e^{-\beta}\frac{\beta^{n-n_0}}{(n-n_0)!} I\{n \geq n_0\}
\end{aligned}$$

where α and β are positive hyperparameters and $n_0 = \max_{i=1,\ldots,q} X_i$.

The second prior distribution is simply a mixture of uniform distributions on truncated supports of a. It is defined by

$$\Pi_K^*(a) = \begin{cases} \frac{1}{3} & \text{if} \quad a = 0 \\ \frac{1}{3\,K} & \text{if} \quad \frac{1}{a} \in \{1, ..., K\} \\ \frac{1}{3\,K} & \text{if} \quad -\frac{1}{a} \in \{n_0, ..., n_0 + K - 1\} \end{cases}$$

where $K \in \mathbb{N}^*$ is an hyperparameter.

Using Bayes formula

$$P(a = z|X = x) \quad = \quad C(x)\, p(X = x|a = z)\, \pi(a = z),$$

where π is the prior distribution on a and $C(x)$ is a constant of normalization, one can easily obtain the following posterior probabilities:

$$\begin{aligned}
P(a > 0|X = x) &= \sum_{n\in\mathbb{N}^*} P(a = 1/n|X = x) \\
P(a < 0|X = x) &= \sum_{n\in\mathbb{N}^*} P(a = -1/n|X = x).
\end{aligned}$$

When the mean parameter is unknown but the sample size q is large enough, one can replace m by its empirical mean $\bar{X} = \sum_{i=1}^{q} X_i$, as we will

do for the two real data sets studied in Section 4. For small sample size (and still with m unknown), it may be of interest to consider also a prior distribution on the mean parameter m. This can easily be done following the modelling detailed e.g. in Robert (2001).

3 The Jeffreys-Lindley paradox

The use of improper vague prior, as for instance $\pi(a = 1/n) = 1$, is always delicate in hypothesis testing (see DeGroot, 1973). Moreover, in our case the series $\sum_{i\in\mathbb{N}^*} p(x|a = 1/n)$ and $\sum_{i\in\mathbb{N}^*} p(x|a = -1/n)$ do not converge. This lead to an indeterminacy in the posterior probabilities.

An alternative is then to use a sequence of probabilities that tends to a vague prior. For instance, we may examine the limit of the posterior probabilities when α and β tends to $+\infty$ for the prior distribution $\Pi_{\alpha,\beta}$ or when $K \mapsto +\infty$ for the uniform prior Π^*_K. However, it is well known that limiting arguments are not valid in hypothesis settings, especially when there is a simple hypothesis (here "$a = 0$"). This usually lead to the well-known Jeffreys-Lindley paradox (see for instance Robert, 2001). This paradox is characterized by the fact that, when the prior distribution tends to a vague prior, the limit of the posterior probability does not depend on the data and, in many cases, gives a mass 1 to the simple null hypothesis. In our problem, we observe the first part of the paradox, mainly because the likelihood $p(x|a)$ converges to a limit ℓ when a tends to 0, or equivalently when n or $-n$ tends to $+\infty$. However, because ℓ is not 0 but $p(x|0)$, the limits of the posterior probabilities are equal to the prior probabilities. This result can easily be generalized by the following proposition.

Proposition 3.1. *Let* $\{\pi_\gamma; \gamma \in \mathbb{R}$ *or* $\mathbb{N}\}$ *a family of prior probabilities such that,* $\forall\gamma$, $\pi_\gamma(a = 0)$, $\pi_\gamma(a > 0)$ *and* $\pi_\gamma(a < 0)$ *do not depend on* γ *and such that*

(a) $\pi_\gamma(\{1, \frac{1}{2}, ..., \frac{1}{A}\}) \overset{\gamma\to\infty}{\longrightarrow} 0$, $\forall A \in \mathbb{N}^*$,

(b) $\pi_\gamma(\{-1, -\frac{1}{2}, ..., -\frac{1}{A}\}) \overset{\gamma\to\infty}{\longrightarrow} 0$, $\forall A \in \mathbb{N}^*$.

Then:

- $P(a = 0|x) \overset{\gamma\to\infty}{\longrightarrow} \pi(a = 0)$

- $P(a > 0|x) \overset{\gamma\to\infty}{\longrightarrow} \pi(a > 0)$
- $P(a < 0|x) \overset{\gamma\to\infty}{\longrightarrow} \pi(a < 0)$

Proof. The limit of $p(x|a)$ is $p(x|0)$ when a tends to 0 (it corresponds to the standard convergence of binomial or negative-binomial distributions towards Poisson distribution). From (a) and (b), we have

$$\begin{aligned}
&\lim_{\gamma\to+\infty} \sum_{n\in\mathbb{N}^*} \frac{1}{\pi(a>0)}\, p(x|a=1/n)\, \pi_\gamma(a=1/n) \\
&= \lim_{\gamma\to+\infty} \sum_{n\in\mathbb{N}^*} \frac{1}{\pi(a<0)}\, p(x|a=-1/n)\, \pi_\gamma(a=-1/n) \\
&= p(x|a=0),
\end{aligned}$$

which completes the proof. □

We now illustrate this paradox by some Monte-Carlo studies. We simulate data sets of size 100. We consider three cases: the data are drawn respectively from a Poisson distribution of mean 10, a binomial distribution of mean 10 and $n = 20$ or a negative binomial of mean 16 and $n = 20$. The prior distribution chosen is Π_K^*, for several values of K. Note that $\Pi_K^*(a = 0) = \Pi_K^*(a < 0) = \Pi_K^*(a > 0) = 1/3$.

Table 1: *Simulation results with* $X_i|m \sim \mathcal{P}(10)$.

Probabilities	$K=5$	$K=10$	$K=50$	$K=100$	$K=1000$	$K=10000$
$P(a=0\|X=x)$	1	0.99	0.54	0.43	0.34	0.33
$P(a<0\|X=x)$	0	0.01	0.44	0.49	0.39	0.34
$P(a>0\|X=x)$	0	0.00	0.03	0.08	0.27	0.32

Table 2: *Simulation results* $X_i|n, m \sim \mathcal{NB}(20, 10)$.

Distribution	$K=5$	$K=10$	$K=20$	$K=100$	$K=1000$	$K=10000$
$P(a=0\|X=x)$	1	0.14	0.02	0.06	0.13	0.32
$P(a<0\|X=x)$	0	0.00	0.00	0.00	0.06	0.29
$P(a>0\|X=x)$	0	0.85	0.98	0.93	0.80	0.40

Table 3: *Simulation results* $X_i|n,m \sim \mathcal{B}(20, 40/3)$.

Distribution	$K = 10$	$K = 500$	$K = 1000$	$K = 10000$	$K = 100000$
$P(a = 0\|X = x)$	0	0.01	0.02	0.14	0.29
$P(a < 0\|X = x)$	1	0.99	0.97	0.73	0.41
$P(a > 0\|X = x)$	0	0.00	0.01	0.13	0.29

4 Illustrations

4.1 Sibship data

The data in Table 4 are taken from Sokal and Rohlf (1987). They consist of frequencies of males in 6115 sibship of size 12 in Saxony (1876-85).

Table 4: *Sibship data*

Males	0	1	2	3	4	5	6	7	8	9	10	11	12
Observed	3	24	104	286	670	1033	1343	1112	829	478	181	45	7

The nature of the data set suggests the use of a binomial NEF $(B)_{12}$ as the model. However, Gelfand and Dalal (1990) proved that there is a significant overdispersion; that is, the estimated variance significantly exceeds the theoretical one. As suggested by formula (2.1), such overdispersed data could be fitted better by binomial model with parameter $N > 12$, Poisson model or negative binomial one.

We first select the best type of NEF using the method described in Section 2.2 with prior distribution $\Pi_{\alpha,\beta}$. The hyperparameter β has been arbitrarily fixed equal to α, since we have observed that it has no effect on the posterior probability.

We observe that the binomial type of NEFs is the most likely one for this data set. Its posterior probability is always very close to 1, except for extreme values of α. In that case we are in the presence of the Jeffreys-Lindley paradox.

The model that has the highest posterior probability is the binomial one with parameter $a = -1/14$. The distribution which would best fit the data set is the binomial $B(14, 6.23)$, where 6.23 is the empirical mean for these data. Of course, the probability to observe 13 or 14 males among 12 children is null ! So we consider a truncated binomial distribution on $\{0, \cdots, 12\}$.

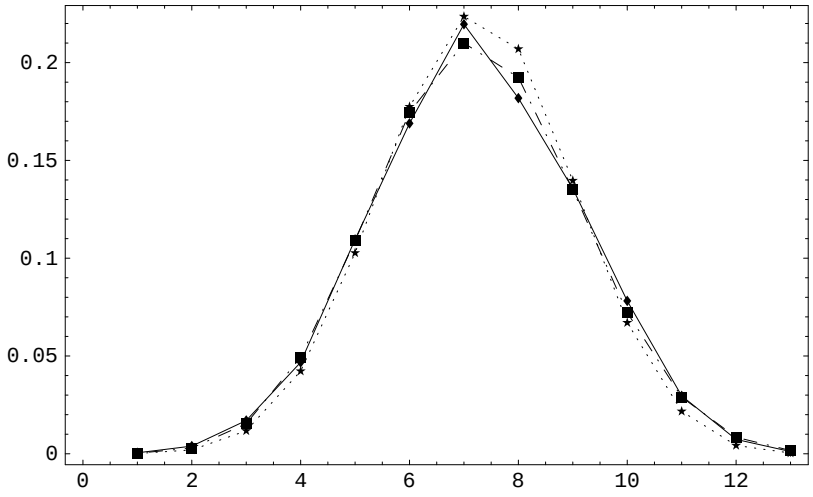

Figure 1: *Frequencies for Sibship data (♦), binomial with parameter 12 (★) and truncated binomial with parameter 14 (■)*

As can be seen in Figure 1, the truncated distribution fits quite well the sibship data set. Its related Chi-squared distance is equal to 18.75 although it is equal to 110.5 for the initial binomial $B(12, 6.23)$. The Chi-squared test of fit gives a p-value equal to 10^{-6} for the $B(12, 6.23)$ distribution and 0.12 for the truncated one. Thus, such a result shows that overdispersed data could be fitted by truncated distributions.

4.2 Bortkewitsch's data

The data in Table 5 are taken from von Bortkewitsch (1898). They consist of frequencies of Prussian soldiers killed by horse-kicks.

Table 5: *Bortkewisch's data set.*

Number of deaths	0	1	2	3	4	5+
Observed freq.	144	91	32	11	2	0

Although these counts are historically associated to the Poisson distribution, Preece et al. (1988) showed that the negative binomial distribution may be derived as a model for these data. Applying our Bayesian method

with prior distribution, $\Pi_{\alpha,\beta}$, we obtain the results listed in Table 6. We have observed that the parameter α has no effect on the conclusion and then we have fixed arbitrarily α equal to β here. In Table 7 are listed the results with Π_K^* as prior distribution.

Table 6: *Posterior probabilities for Bortkewisch's data set with* $\Pi_{\alpha,\beta}$ *prior* .

β	10^{-7}	0.1	1	5	10	20	50	500
$P(a=0\|X=x)$	0.998	0.98	0.75	0.38	0.33	0.33	0.33	0.333
$P(a<0\|X=x)$	0.001	0.00	0.02	0.08	0.14	0.21	0.28	0.328
$P(a>0\|X=x)$	0.000	0.02	0.23	0.54	0.53	0.46	0.39	0.339

Table 7: *Posterior probabilities for Bortkewisch's data set with* Π_K^* *prior* .

K	1	10	50	100	500
$P(a=0\|X=x)$	0.998	0.42	0.35	0.34	0.33
$P(a<0\|X=x)$	0.002	0.08	0.21	0.25	0.30
$P(a>0\|X=x)$	0.000	0.50	0.44	0.41	0.36

One can see that for small values of β and K the Poisson model is chosen. For larger values the method suggests a negative binomial model even if its posterior probabilities are close those of the Poisson model. For very large values we observe the Jeffreys-Lindley paradox.

Acknowledgements

The authors thank the editor and a referee for their comments and suggestions which have helped improve the original manuscript.

References

Dean, C. and Lawless, J. F. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association*, 84:467–472.

DeGroot, M. H. (1973). Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association*, 68:966–969.

Edwards, C. B. and Gurland, J. (1961). A class of distributions applicable to accidents. *Journal of the American Statistical Association*, 56:503–517.

Gelfand, A. E. and Dalal, S. R. (1990). A note on overdispersed exponential families. *Biometrika*, 77:55–64.

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.

Lenk, P. J. (1999). Bayesian inference for semiparametric regression using a Fourier representation. *Journal of the Royal Statistical Society. Series B*, 61:863–879.

Preece, D. A., Ross, G. J. S., and Kirby, S. P. J. (1988). Bortkewitsch's horse-kicks and the generalised linear model. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 37:313–318.

Ramakrishnan, V. and Meeter, D. (1993). Negative binomial cross-tabulations, with applications to abundance data. *Biometrics*, 49:195–207.

Robert, C. P. (2001). *The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer-Verlag, New York, 2nd ed.

Sokal, R. R. and Rohlf, F. J. (1987). *Introduction to Biostatistics*. W. H. Freeman & Co, New York, 2nd ed.

von Bortkewitsch, L. (1898). *Das Gesetz der Kleinen Zahlen*. B.G. Teubner, Leipzig.

Williams, D. A. (1996). Overdispersion in logistic-linear models. In B. Morgan, ed., *Statistics in Toxicology*. Clarendon Press, Oxford.