Monografías del Seminario Matemático García de Galdeano **33**, 237–244 (2006)

AN OVERVIEW OF PROBABILITY MODELS FOR STATISTICAL MODELLING OF COUNT DATA

S. Dossou-Gbété and D. Mizère

Abstract. The Poisson model is a benchmark model for the statistical analysis of the count data. Sometimes count data exhibit variation, refered to as overdispersion or underdispersion, resulting in the lack of fit of the Poisson model. The aim of this paper is to present an overview of potential families of discrete probability distributions that can provide alternative modelling framework for the statistical analysis of count data.

Keywords: Count data, Katz's model, exponential dispersion family, Poisson model, Poisson mixture model, probabilities ratio recursion, statistical modelling, weigthed Poisson model.

AMS classification: 60-xx

§1. Introduction

One of the crucial question in statistical analysis of count data is how to formulate an adequate probability model to describe observed variation of counts. The Poisson family of discrete distributions is used as a benchmark for statistical analysis of count data. This family made of distributions indexed by a positive parameter such the probability mass function is defined as $p(x,\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$, $\lambda \in]0,+\infty[$, $x \in \mathbb{N}$. It follows that this family of distributions is a natural exponential family with canonical parameter $\theta = \ln(\lambda)$ and cumulant function $\kappa(\theta) = \exp(\theta)$.

One of the important features of the Poisson family is that the variance-to-mean ratio, also called Fisher dispersion index, is equal to 1 whatecer the value of λ. Then, the Fisher dispersion index of a counts probability distribution is considered as a measure of its departure from Poisson model. Notice that the case where the variance-to-mean ratio equals to 1 characterises the Poisson family among the natural exponential family of discrete distributions. Overdispersion with respect to Poisson model (in short overdispersion) referes to the cases where there is evidence that the observed random variation is greater than the expected random variation under the Poison model. Otherwise underdispersion means that the expected variation is greater than the observed one. An other impotant feature of the Poisson family is the equality $\frac{1}{\lambda}\ln[p(0,\lambda)] + 1 = 0$ where $p(0,\lambda)$ is the probability of zero. The index $zi = \frac{1}{\lambda}\ln[p(0,\lambda)] + 1$, called zero inflation index, is also used as a measure of departure from Poisson model. Therefore, many goodness-of-fit procedures for Poisson distribution are built on the assessement of criteria based on one of these measures or both.

The reliance of the Poisson model on a single parameter results in a lack of flexibility in its application. The lack of fit of the Poisson model is a frequent issue in the count data analysis literature as a survey can show. This results in proposals of alternative statistical

analysis framework that take into account the knowledge on random mechanism underlying the occurences of the counted events. But one has to notice that more attention has been paid to overdispersion.

The Negative binomial distribution is one of the most widely used distributions when modelling count data that exhibit variation that Poisson distribution cannot explain. This distribution arises from various random scenarios ([12]) but it can be used only to model data that show overdispersion. Nevertheless some recent works focused on applications where count data exhibited underdispersed empirical distribution ([3, 27]).

This paper concentrates on providing an expository review of different approaches, old and recent, that are used to underpin the statistical modelling and analysis of count data. Cornerstones of these approaches include birth process modelling ([6, 28]), variance modelling as function of the mean assuming that the distribution belongs to an exponential dispersion family ([9, 17]), mixing Poisson distributions ([10, 28]), successive probabilities ratio modelling ([15, 23, 26, 5]) and modifying Poisson distribution by weighting ([21, 22, 18]).

§2. Modelling birth rate for birth processes

Count data are realizations of counting processes, as collection of numbers of events occuring in non overlaping periods of time of the same length. Assuming that events occur according to a random process that obeys the following rule: $P(N(t+h)=y+1 \mid N(t)=y)=\lambda h+o(h)$, $P(N(t+h)=y \mid N(t)=y)=1-\lambda h+o(h)$, where $N(u)$ denotes the number of events occurence during the time interval $]0,u[$ and λ is the conditional occurence rate (birth rate). Prior knowledge or assumption on this occurence process can be expressed throught λ by describing it as a random variable or as a non random function of total number of past occurences, parametric or nonparametric.

2.1. Modelling birth rate as state-dependent non random function

The Poisson model is equivalent to a Poisson process, that is the conditional occurence rate λ is constant independently of occurence of events up to t. One can relax this constraint by allowing the conditional occurence rate λ to be state-dependent, that means a function of $N(t)$. This approach has been studied in [6] and improved later [7]. It was proved in [2], according to a conjecture in [6] that when λ is modelled as an increasing function of $N(t)$, the counting process results in a family of overdispersed counts distributions. Modelling λ as a decreasing function of $N(t)$ leads to a family of underdispersed counts distributions. Continuing on this way, a parametric framework was proposed in [27] by modelling the birth rate as $\lambda = a(b+N(t))^c$ with $a>0$, $b>0$ and $c \leq 1$. This class of models encompasses popular distributions models as the Poisson model when $c=0$ and the negative binomial models if $c=1$. One can readily show that the constraint $c>0$ gives overdispersed counts distributions while the constraint $c<0$, leads to underdispersed counts probability distributions. Recently a new class of models has been proposed in [8] to deal specifically with situation where there is evidence of underdispersion; it is based on the modelling of the birth rate as $\lambda = a(b+\exp(c\ln(N(t))-d))$.

2.2. Poisson mixtures: handling overdispersion by modelling birth rate as random variable

When data on hand shows a variance that significativly dominates the mean there is evidence that the assumption of constant birth rate is not realistic, since one should expect there is no significant difference between the variance and the mean. An alternative method to deal with this overdispersion is to consider a Poisson mixture models which results in modelling the birth rate λ as a random variable with support $]0,+\infty[$. By doing so one can take into account the heterogeneity in the occurence of events generated by the birth process underlying the counts. Assuming that the random variable λ is distributed according to the probability law H, the probability mass function of the sampling distribution of the counts has the form

$$p(x) = \int_0^{+\infty} \frac{1}{x!}\lambda^x e^{-\Lambda}\, dH(\lambda).$$

It is well-known that the specification of the mixing distribution is crucial in this approach. This may be done through a parametric modelling of the distribution as in [10, 28] or non-parametrically as suggested in [13]. Notice that some special mixing distributions models lead to well-known sampling distributions for the count. As an example, if the mixing distribution belongs to the Gamma family, the sampling distribution belongs to the negative binomial model. One of the drawback of the parametric modelling is that it is usually motivated by mathematical convenience (one can deal easily with the statistical inference) and computational issues.

§3. Exponential dispersion families and the variance modelling as function of the mean

Exponential dispersion families of discrete probability distributions have proved to be an appropriate distributional framework when dealing with generalized linear model (glm) [20] for count data regression analysis. Negative binomial family is an exponential dispersion family. A count probability distribution belongs to an exponential dispersion family if its probability mass function is of the form $p(x) = c(x,\phi)\exp(\theta x - \phi\kappa(\theta))$, $x \in \mathbb{N}$ ([14]) where $\theta \in \Theta \subset \mathbb{R}$ is the canonical parameter, $\phi \in \Phi \subset]0,+\infty[$ is the dispersion parameter and κ is the cumulant function. Let m and σ^2 denote respectively the mean and the variance of a discrete probability distribution belonging to an exponential dispersion family. One proves that $m = \phi d\kappa(\theta)/d\theta$ and $\sigma^2 = \phi d^2\kappa(\theta)/d\theta^2 = dm/d\theta$ and then one can write the variance as a function of the mean $\sigma^2 = \phi V(m/\phi)$, where V is named unit variance function.

It appears that modelling the variance σ^2 as a function of the mean may provide a gateway toward a probability distribution model building. An example of unit variance function is the function $V(\mu) = \mu + \mu^p$ with $p \geq 2$, considered by Hinde J. and Demétrio C. G. B in [9] in the framework of quasi-likelihood inference for fitting overdispersed count data. Kokonendji C. C. and al. proves in [16] that for the index p fixed in the interval $[2,+\infty[$, this function is the unit variance function of an exponential dispersion family. These probability distributions have support set $S = \mathbb{N} + p\mathbb{N}$ and then only integer values have to be considered for the index p when dealing with count data. $p = 2$ yields the negative binomial model, and $p = 3$, the strict arcsine model ([17]).

§4. Probabilities ratio recursion approach

A family p_ϑ of count probability distributions $\{p(x,\vartheta),\, x\in\mathbb{N}\}$ is completly specified by the recursion formula

$$\begin{cases} p(0;\vartheta)\neq 0 & x=0,\\ \dfrac{p(x;\vartheta)}{p(x-1;\vartheta)} = f(x,\vartheta) & x\geq 1, \end{cases}$$

where f is a specified function of $x\in\mathbb{N}$ and ϑ is a vector of numeric parameters. Several types of functions have been studied resulting in classes of discrete probability distributions that include Poisson family as a special case. Among them are Katz's recursion and its various extensions ([5, 23, 26, 29]).

4.1. Basic Katz's recursion

The basic Katz's recursion [15] is defined as follows: $f(x,\alpha,\lambda)=\alpha+\lambda/x$, for $x\geq 1$. When $\alpha\in\,]0,1[$ and $\lambda>0$, this ratio yields the negative binomial model with α the probability of failure and $\phi=1+\lambda/\alpha$ the dispersion parameter; when $\alpha\in\,]0,1[$ and $\lambda=0$, it yields the geometric model with α the probability of failure, when $\alpha=0$ and $\lambda>0$, it yields the Poisson model of parameter λ and when $\alpha<0$, it yields the binomial model $B(n,p)$ with $p=-\alpha/(1-\alpha)$ the probability of success and $-\lambda/\alpha=n+1$, $n\in\mathbb{N}^*$. Notice that Katz's recursion leads to Poisson distribution or two parameters models for counts probability distribution as well. Several extensions of Katz's recursion have been proposed to enlarge the hierachy of models available for modelling and analysing count data.

4.2. Lerch family

The Lerch family has been introduced by in [19] as revisiting Good distribution, motivated by the modelling of non zero counts data arising in ecology. The proposal results in a successive probability ratio model of the following form:

$$f(x,\alpha,\beta,\nu)=\alpha\left(1-\frac{1}{\beta+x}\right)^{\nu},$$

where $\alpha\in\,]0,1[$, $\beta>0$ and $\nu\neq 0$. This models proves its ability to fit data for which the observed variation could be significatively greater than the mean, less than the mean or equal to the mean. Further statistical and probabilistics properties of this model is studied in [1] where the support of the counts distribution is taken as the set of the non negative integer $\mathbb{N}$.

4.3. Revival of Conway-Maxwell class

The Conway-Maxwell class of discrete probability distributions is defined by modelling the consecutive probabilites ratios as follows:

$$f(x,\alpha,\lambda,\nu)=\alpha+\frac{\lambda}{x^{\nu}},\quad x\in\mathbb{N}^{\star},$$

where α, λ, ν are numeric parameters such that $\alpha \in]0,1[$, $\lambda \in \mathbb{R}$ and $\nu > 0$. Shmueli et al. studied in [26] the statistical and the probabilitics properties of the distributions arising from this model when $\alpha = 0$ and $\lambda > 0$. This family of discrete distributions, named Conway-Maxwell-Poisson distributions (in short COM-Poisson), is discussed as a two-parameter extention of Poisson family that generalizes well-known family (Poisson, Benouilli, Geometric). Its also leads to the generalisation of distributions derived these families as Binomial and Negative Binomial. Further results were obtained in [5] where the general case of this model is studied as extention of Conway-Maxwell distributions. This class of distributions proves to encompass probability distributions for which the variance-to-mean ratio is greater than 1, less than 1 or equal to 1. Then the family have more flexibility to fit count data generated by various random mechanisms.

§5. Modifying Poisson distributions by weighting

Modifying Poisson distribution by weighting is an other approach for the building of discrete probability distributions that can account for overdispersion and under-dispersion. One motivation of this approach is that the count x could be recorded with a probability proportionally to some function $w(x)$ (cf. [12, 22]). Then the count x is the realization of a sampling distribution called the weighted version of the Poisson distribution and its probability mass function is defined as

$$P_w(x,\lambda) = \frac{w(x)}{E_\lambda[w]}\frac{\lambda^x}{x!}e^{-\lambda}, \quad x \in \mathbb{N}, \quad \lambda > 0,$$

where

$$E_\lambda[w] = e^{-\lambda}\sum_{s=0}^{+\infty} w(x)\frac{\lambda^s}{s!} < \infty$$

is the normalization constant and $w(x)$ the weighting function. The choice of this weighting function is an important step of this method. The weighting function could be specified through a parametric models or non parametrically. One should have in mind that it is not an easy task to choose a parametric model for the weighting function. Very often the choice is guided by mathematical convenience and computational issues. A well-known weighted Poisson distribution is the size-biased Poisson distribution ([21]) which corresponds to the weighting function $w(x) = x$. The class of Poisson distributions modified by weighting with the functions $w(x,\alpha,r) = (x+\alpha)^r$ has been studied in [4]. More recently, Kokonendji C.C. and Mizère D. studied in [18] general properties of these classes of distributions in the perspective of fitting count data that exhibit over-dispersion or under dispersion. It is proved that a Poisson distribution modified by weighting with a function $w(x)$ is over-dispersed (resp. under-dispersed) if the function $\lambda \longmapsto E_\lambda[w]$ is log-convex (resp. log-concave). Furthermore the weighted Poisson distribution is over-dispersed (resp. under-dispersed) if the weighting function w is log convex (resp. log-concave) as function of counts x.

The concept of the dual weighted Poisson distributions was introduced in [18] in order to prove that the class of the distributions obtained by weighting Poisson distributions is flexible in the sense it encompasses distributions that can account for over-dispersion as well as under dispersion.

Definition 1. Two weighting function $w_1(x)$ and $w_2(x)$ leads to a dual pair of weighted Poisson distributions if they satisfy the following condition: $w_1(x)\,w_2(x) = 1$, $\forall x \in \mathbb{N}$.

The following result holds [18]:

Proposition 1.

(i) Let's consider the Poisson distribution with mean λ and a weighting function w such

$$\lim_{y\to+\infty} \frac{w(x-1)}{xw(x)} = \frac{c_0}{\lambda}$$

for some $c_0 \in]0, 1[$. Then the weighted version of this Poisson distribution obtained by weighting by w admits a dual distribution.

(ii) Let's consider a dual pair of two weighted Poisson distributions with weighting functions w_i, $i = 1, 2$. If one of the weighting functions w_i is log-convex (or log concave) then this pair is constituted of an over-dispersed and an under-dispersed counts distribution.

As an example, a dual pair of weighted Poisson distributions is obtained by weighting a Poisson distribution with the functions $w_1(x) = (x+\alpha)^r$ and $w_2(x) = (y+\alpha)^{-r}$ where $\alpha > 0$, $r \geq 0$.

§6. Concluding comments and miscellaneous

The Poisson distribution and the Negative binomial distribution are the most widely used discrete probability distributions for the analysis of count data. But survey can show they demonstrate limitations since in many applications the shape of the empirical distribution of data is significantly different of the expected shape from Poisson and Negative Binomial models. Although our presentation is limited to the framework of parametric models its does not cover all of the modelling proposal that can be encountered in the huge literature on count data analysis. By modelling the successive probability ratio, Pestana D. D. and Velosa S. F. have constructed a wide class of discrete probability distributions indexed with three parameters that encompasses Poisson stopped sums and Geometric stopped sums ([23]). Puig J. and co-authors studied count distributions modelled by two parameters, the mean and the Fisher dispersion index, for the analysis of over-dispersed count data [24, 25]. However, the present paper proves that it is possible to obtain flexible parametric models that can account for over-dispersion, under-dispersion as well as equidispersion, depending on the value of some model parameters. Such discrete probability models can be used for applications in various domains as ecology, linguistics, information sciences, statistical physics, etc.

References

[1] AKSENOV, S. V., AND SAVAGEAU, M. A. Some properties of the Lerch family of discrete distributions. Preprint, Elsevier Science, arXiv:math.PR/054485v1, (2005).

[2] BALL, F. G. A note on variation in birth processes. *Math. Scientist 20* (1995), 50–55.

[3] BOSCH, R. J., AND RYAN, L. M. Generalized Poisson distributions arising from Markov processes. *Statistics and Probability Letters 39* (1998), 205–212.

[4] CASTILLO, J. D., AND PEREZ-CASANY, M. Weighted Poisson distributions for over-dispersion and under-dispersion situations. *Ann. Inst. Statist. Math. 50*, 3 (1998), 567–585.

[5] DOSSOU-GBÉTÉ, S., AND MIZÈRE, D. Modelling count data: a probability ratio recursion approach. In preparation.

[6] FADDY, M. J. On variation in Poisson processes. *Math. Scientist 19* (1994), 47–51.

[7] FADDY, M. J. Extending Poisson process modelling and analysis of count data. *Biometrical Journal 39*, 4 (1997), 431–440.

[8] FADDY, M. J., AND BOSCH, R. J. Likelihood modelling and analysis of data under-dispersed relative to Poisson distribution. *Biometrics 57* (2001), 620–624.

[9] HINDE, J., AND DEMÉTRIO, C. G. B. Over-dispersion: Models and Estimation. *Sao Paulo: ABE*, 1998.

[10] HOUGAARD, P., LEE, M-L. AND WHITMORE, G. A. Analysis of over-dispersed count data by mixtures of Poisson variable and Poisson processes. *Biometrics 53* (1997), 1255–1238.

[11] JANSAKUL, N. AND HINDE, J. P. Linear mean-variance negative binomial models for analysis of orange tissue culture. *Songklanakaria J. Sci. Technol. 26*, 5 (2004), 683–696.

[12] JOHNSON, N. L., KOTZ, S., AND KEMP, A. W. *Univariate Discrete Distributions*. John Wiley, Ed., 1992.

[13] JONGBLOED, G., AND KOOLE, G. Managing uncertainty in call centres using Poisson mixtures. *Applied Stochastic Models in Business and Industry 17*, 4 (2001), 307–318.

[14] JORGENSEN, B. *Theory of dispersion models*. Chapman & Hall, 1997.

[15] KATZ, L. Unified of a broad class of discrete probability distributions. In *Distribution Symposium*, G. P. Patil, Ed. Michigan State University, 1965, pp. 175–182.

[16] KOKONENDJI, C. C., DOSSOU-GBÉTÉ, S., AND DEMÉTRIO, C. G. B. Some discrete exponential dispersion models: Poisson-Tweedie and Hinde-Demétrio classes. *Sort 28*, 2 (2004), 201–214.

[17] KOKONENDJI, C. C., AND KHOUDAR, M. On strict arcsine distribution. *Communications in Statistics, Theory and Methods 33*, 5 (2004), 993–1006.

[18] KOKONENDJI, C. C., AND MIZÈRE, D. Over-dispersion and under-dispersion characterization of weighted Poisson distributions. Prépublication, LMA, Université de Pau, n. 0523, 2005.

[19] KULASEKERA, K. B., AND TONKYN, D. W. A new discrete distribution, with applications to survival, dispersal and dispersion. *Communications in Statistics and Simulation 21* (1992), 499–518.

[20] MCCULLAGH, P., AND NELDER, J. A. *Generalized Linear Models*. Chapman and Hall, 1989.

[21] PATIL, G. P., AND RAO, C. R. Weighted distributions and size biased sampling with application to wildlife population and human families. *Biometrics 34* (1978), 179–189.

[22] PATIL, G.P. Weighted distributions. In *Encyclopedia of Environmetrics*, vol. 4 , A. H. El-Shaarawi and W.W. Piegorrsch, Eds. 2002, pp. 2369–2377.

[23] PESTANA, D. D., AND VELOSA, S. F. Extensions of Katz-Panjer families of discrete distributions. *Revstat-Statistical Journal 2*, 2 (2004), 146–162.

[24] PUIG, P. Characterizing additively closed discrete models by a property of their maximum likelihood estimators, with an application to generalized Hermite distributions. *JASA 98* (2003), 687–692.

[25] PUIG, P., AND VALERO, J. Count data distributions: some characterisation with applications. *JASA 101*, 473 (2006), 332–340.

[26] SHMUELI, G., MINKA, TH. P., KADANE, J. B., BORLES, S., AND BATWRIGHT, P. A useful distribution for fitting discrete data: revival of the Conway-Maywell-Poisson distribution. *Appl. Statist. 54*, 1 (2005), 127–142.

[27] TOSCAS, P. J., AND FADDY, M. J. Likelihood-based analysis of longitudinal count data using a generalized Poisson model. *Statistical Modelling 3* (2003), 99–108.

[28] WALHIN, J. F., AND PARIS, J. A general family of over-dispersed probability laws. *Belgian Actuarial Bulletin 2*, 1 (2002), 1–8.

[29] WINKELMANN, R. *Econometric Analysis of Count Data*, fourth edition. Springer Verlag, 2003.

Simplice Dossou-Gbété
Laboratoire de Mathématiques Appliquées
CNRS UMR 5142 - Université de Pau et des Pays de l'Adour
IPRA, B.P. 1155, 64013 Pau Cedex, France

Dominique Mizère
Université Marien Ngouabi
BP. 69, Brazzaville, Congo.