REVUE DE STATISTIQUE APPLIQUÉE

ASTRID JOURDAN
CÉLESTIN C. KOKONENDJI

Surdispersion et modèle binomial négatif généralisé

Revue de statistique appliquée, tome 50, n° 3 (2002), p. 73-86
<http://www.numdam.org/item?id=RSA_2002__50_3_73_0>

Rev. Statistique Appliquée, 2002, **L** (3), 73-86

SURDISPERSION ET MODÈLE BINOMIAL NÉGATIF GÉNÉRALISÉ

Astrid JOURDAN* et Célestin C. KOKONENDJI*

* *Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques Appliquées, Avenue de l'Université - 64000 PAU, France.*

RÉSUMÉ

Le modèle binomial négatif ($\mathcal{BN}$) est souvent considéré comme le prototype des modèles surdispersés pour des données discrètes. A travers certaines propriétés, nous montrons que sa généralisation lagrangienne, appelée modèle binomial négatif généralisé ($\mathcal{BNG}$), lui offre une alternative dans ce domaine. En particulier, le $\mathcal{BNG}$ résulte aussi d'un mélange de Poisson par une loi absolument continue, il possède un degré de surdispersion plus important que le $\mathcal{BN}$, et il est aussi utilisable dans le cadre des modèles (non)linéaires généralisés. De plus, à l'aide de simulations et exemples, nous discutons divers points tels que l'unimodalité ou bien l'influence des paramètres sur la surdispersion.

Mots-clés : *Surdispersion, mélange de Poisson, modèles linéaires généralisés, modèle exponentiel de dispersion, fonction variance, binomial négatif généralisé.*

ABSTRACT

The negative binomial distribution is often considered in the literature as the prototype of overdispersed distributions for discrete data. Throught many properties, we show that its Lagrangian generalization called generalized negative binomial distribution offers to us an alternative in this way. In particular the generalized negative binomial distribution is a Poisson mixture, it is more overdispersed than the negative binomial distribution, and we can also use it for the generalized (non)linear models. Moreover, from some simulations and examples, we discuss about unimodality or the parameters effects on the overdispersion.

Keywords : *Overdispersion, Poisson mixture, generalized linear models, exponential dispersion model, variance function, generalized negative binomial distribution.*

1. Introduction

L'analyse statistique de certaines données discrètes présente une variabilité plus grande que la moyenne de l'échantillon. Ce phénomène dit de *surdispersion* a été largement et diversement étudié dans la littérature, en particulier, en relation avec la loi de Poisson. Si un modèle inadéquat est adopté en présence de la surdispersion, il peut y avoir perte d'efficacité pour les différentes statistiques en jeu. Cox (1983) a étudié en détail cet effet pour le modèle poissonnien.

Dès lors que la cause de la surdispersion est connue, un modèle paramétrique du type mélange de Poisson est souvent considéré (Breslow, 1984; McCullagh et Nelder, 1989). Le mélange de Poisson le plus utilisé est sans doute le modèle binomial négatif; lequel est obtenu avec une loi gamma (Greenwood et Yule, 1920). Cependant, d'autres mélanges aussi intéressants qu'utiles sont proposés par exemple dans Johnson *et al.* (1992, pages 328-335). Notons aussi que le modèle binomial négatif est une généralisation du modèle de Poisson pour la surdispersion dans le sens où, au lieu de modéliser une succession d'événements indépendants ayant une espérance constante, elle suppose qu'ils se produisent d'une manière contagieuse et/ou dans un milieu hétérogène. Ce modèle binomial négatif a bien sûr ses limites.

L'objet de cet article est de montrer qu'une généralisation sur les entiers de la loi binomiale négative, appelée *binomiale négative généralisée* (voir Jain et Consul, 1971) et notée $\mathcal{BNG}$, lui offre une alternative intéressante dans le domaine de la surdispersion. Pour cela, nous précisons tout d'abord les différentes définitions et propriétés intrinsèques de la $\mathcal{BNG}$, puis nous montrons enfin ses performances en matière de modèle surdispersé. Cette présentation générale du modèle $\mathcal{BNG}$ est illustrée par quelques graphiques et exemples, et insiste sur le parallèle qui existe avec le modèle binomial négatif.

2. Modèle binomial négatif généralisé

Dans cette partie, nous revoyons les différentes définitions du modèle $\mathcal{BNG}$, ainsi que quelques interprétations probabilistes de ce modèle. Nous ferons ensuite le lien entre la $\mathcal{BNG}$ et d'autres familles de lois bien connues.

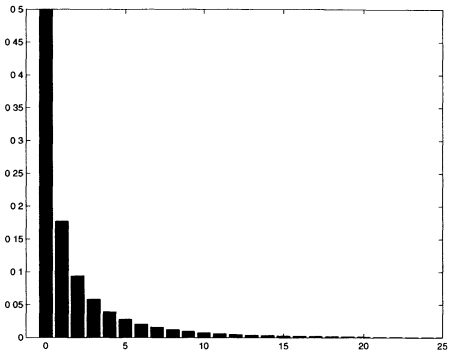
2.1. Définitions

Soit $a > 0$, $\lambda > 0$ et $0 < p < 1$ tels que $0 < a < (1-p)/p$. On dit qu'une variable aléatoire X à valeurs dans $\mathbb{N}$ suit une *loi binomiale négative généralisée* $\mathcal{BNG}(a, \lambda, p)$ de paramètres a, λ, p, si pour tout $x \in \mathbb{N}$ la probabilité de l'évenement $X = x$ est égale à

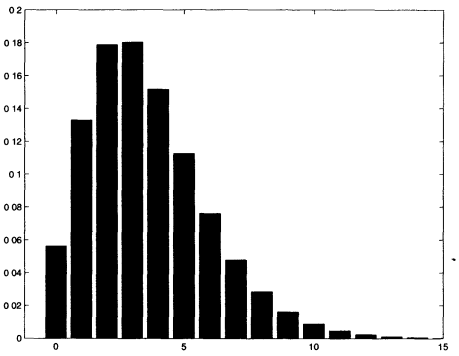
$$P(x; a, \lambda, p) = p^x (1-p)^{a(\lambda+x)} \frac{\lambda}{\lambda + x} \frac{\Gamma(a(\lambda + x) + x)}{\Gamma(a(\lambda + x))\,\Gamma(x+1)}, \qquad (1)$$

où Γ représente la fonction gamma. Cette définition est plus rigoureuse que celle trouvée habituellement dans la littérature (e.g., Jain et Consul, 1971; Johnson *et al.*, 1992; Famoye, 1997; Letac et Mora, 1990).

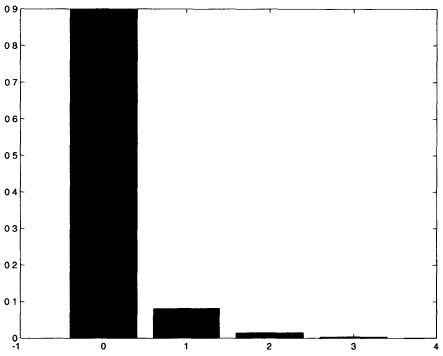
Les paramètres a, λ, p, sont tels que p est la probabilité de succès dans une épreuve dont la reparamétrisation $\theta = p(\theta)$ fournit le paramètre canonique de la famille exponentielle de dispersion associée (cf. § 2.3.2); λ est le paramètre d'échelle ou d'indice tel que $\sigma^2 = 1/\lambda$ se dit de dispersion (Jørgensen, 1997), il correspond à une puissance de produit de convolution; a est le paramètre d'indice de base d'après la généralisation lagrangienne de la binomiale négative (cf. § 2.1.1).



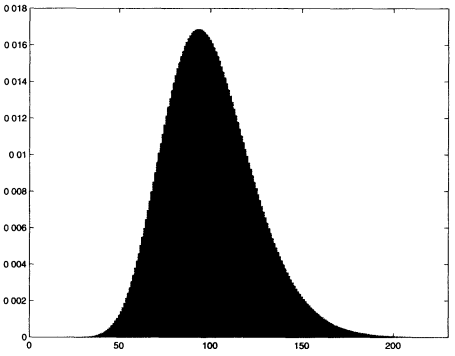
$p = 0.5$, $a = 0.5$ et $\lambda = 2$



$p = 0.25$, $a = 0.2$ et $\lambda = 50$



$p = 0.1$, $a = 1$ et $\lambda = 1$



$p = 1/3$, $a = 1$ et $\lambda = 100$

Figure 1. – Densité de probabilité de la $\mathcal{BNG}$ pour différentes valeurs de ses paramètres

Remarque. – Il semble, au vu d'arguments empiriques, que la loi $\mathcal{BNG}$ soit unimodale. De plus, en établissant le rapport $P(x+1; a, \lambda, p)/P(x; a, \lambda, p)$, on peut montrer que la condition

$$1 - a\lambda p(1-p)^a > 0$$

est nécessaire pour que le mode soit en 0.

2.1.1. Généralisation lagrangienne

Une manière de montrer les différentes liaisons entre les lois binomiale négative ($\mathcal{BN}$) et binomiale négative généralisée ($\mathcal{BNG}$) est d'utiliser leurs fonctions génératrices des moments. En effet, si une variable aléatoire X suit une $\mathcal{BN}(\lambda, p)$, de probabilité individuelle

$$\mathrm{P}(X = x) = p^x (1-p)^\lambda \frac{\Gamma(\lambda + x)}{\Gamma(\lambda)\,\Gamma(x+1)},$$

alors sa fonction génératrice est donnée par

$$G_X(z) = \mathbb{E}(z^X) = \left(\frac{1-pz}{1-p}\right)^{-\lambda}.$$

Alors que si une variable aléatoire Y suit une $\mathcal{BNG}(a, \lambda, p)$, alors sa fonction génératrice se met sous la forme

$$G_Y(z) = G_X(h(z)),$$

où h est une fonction vérifiant :

$$h(0) = 0 \text{ et } h(\omega) = \omega G_{X'}(h(\omega)),$$

$G_{X'}$ étant la fonction génératrice d'une variable X' de loi $\mathcal{BN}(a, p)$.

L'existence et le développement en série entière de la fonction implicite h, est garantie par la formule d'inversion de Lagrange (Dieudonné, 1971), d'où la qualification de "généralisation lagrangienne" de la binomiale négative par Jain et Consul (1971). La terminologie "généralisée" dans binomiale négative généralisée peut donc être considérée comme abusive dans le sens où elle se rapporte uniquement à la généralisation lagrangienne alors qu'il existe d'autres généralisations de la binomiale négative (Gupta, 1977; Kokonendji, 1994). Letac et Mora (1990) ont proposé l'appelation "loi de Takács", car cette loi $\mathcal{BNG}(a, \lambda, p)$ est apparue dans l'article de L. Takács (1962) en théorie des files d'attente.

2.1.2. Cas particuliers

Habituellement, la loi $\mathcal{BNG}(a, \lambda, p)$ est définie à partir du produit $a\lambda$ positif (Jain et Consul, 1971; Johnson *et al.,* 1992; Famoye, 1997) avec comme unique cas pour a négatif : $a = -1$. Ce qui permet d'englober la loi binomiale $\mathcal{B}(-\lambda, p)$ de paramètres $-\lambda$ et p, où λ est un entier négatif. Pour a fixé, on retrouve d'autres lois de probabilité classiques, par exemple la loi binomiale négative avec $a = 1$ et en remplaçant dans (1) $\lambda + x$ par λ pour tout $x \in \mathbb{N}$, ou encore la loi binomiale inverse de Yanagimoto (1989) avec uniquement $a = 1$ dans (1). Enfin, lorsque a tend vers l'infini, on obtient la loi de Poisson généralisée (Consul, 1989) qui est une généralisation lagrangienne de la loi de Poisson classique.

2.2. Interprétations probabilistes

Nous proposons ici deux types d'interprétations probabilistes de la binomiale négative généralisée en liaison une fois de plus avec la binomiale négative.

2.2.1. Par une chaîne de Markov

Rappelons qu'une suite $(X_n)_{n\geqslant 0}$ de variables aléatoires est appelée chaîne de Markov sur $(\Omega, \mathcal{F}, \mathrm{P})$, si pour tout n, la loi de X_{n+1} sachant $X_0 = x_0, \cdots, X_n = x_n$ ne dépend que de x_n.

Soient W_0 une variable aléatoire de loi $\mathcal{BN}(\lambda, p)$ et $(W_n)_{n\geqslant 0}$ une chaîne de Markov sur $\mathbb{N}$ de fonction de transition

$$\mathbb{E}(z^{W_{n+1}}|W_n) \;=\; (\mathbb{E}(z^Y))^{W_n}$$

où Y suit une loi $\mathcal{BN}(a, p)$. Puisqu'il existe un certain rang à partir duquel W_n est nul p.s. (e.g., Harris, 1963), on peut définir la variable aléatoire

$$Z = \sum_{n\geqslant 0} W_n,$$

et montrer qu'elle suit une loi $\mathcal{BNG}(a, \lambda, p)$. Voir Kokonendji (1994, Théorème 4.5) pour une démonstration plus générale des lois lagrangiennes.

La variable aléatoire Z représente les descendances de l'individu W_0 en théorie de processus de branchement. Par exemple, elle peut être utilisée pour modéliser la progéniture d'un individu asexué.

2.2.2. Par des temps de visite d'une marche aléatoire

Soit $(X_k)_{k\geqslant 0}$ une suite de variables aléatoires indépendantes de même loi $\mathcal{BN}(a, p)$. On définit alors la suite $(S_n)_{n\geqslant 0}$ par $S_0 = 0$ et $S_n = \sum_{k=1}^{n}(1 - X_k)$. Notons pour tout $k \geqslant 0$,

$$T_k = \inf\{n > 0 \;;\; S_n = k\}$$

le temps de premier passage de S_n en k, et $\mathcal{T}_k$ l'image de T_k par $x \mapsto x - k$. Alors on a :

$$\mathcal{T}_1 \text{ suit une loi } \mathcal{BNG}(a, 1, p) \text{ et } \mathcal{T}_\lambda \text{ suit une loi } \mathcal{BNG}(a, \lambda, p),\ \forall \lambda > 0,$$

où $\mathcal{BNG}(a, \lambda, p)$ est la convolée λ fois de la loi $\mathcal{BNG}(a, 1, p)$ de $\mathcal{T}_1$. Pour plus de précisions sur ce résultat, on pourra consulter Letac et Mora (1990) ou Kokonendji (1994). En particulier, on pourra trouver une extension au temps de passage en 0 (lié à $\mathcal{T}_1$) dans Kokonendji (2001).

Ainsi, à partir d'une suite de lois $\mathcal{BN}$ on obtient un temps de premier passage de loi $\mathcal{BNG}$. Rappelons que le temps de passage obtenu à partir d'une suite de lois gaussiennes suit la loi gaussienne-*inverse*. Par analogie, une $\mathcal{BNG}$ peut être appelée $\mathcal{BN}$-*inverse*. C'est dans cet esprit que Yanagimoto (1989) a présenté le cas $a = 1$ (cf. § 2.1.2) comme $\mathcal{B}$-*inverse*.

2.3. Liens avec d'autres familles de lois

Dans les paragraphes précédents, nous nous sommes attachés à établir un lien entre les lois $\mathcal{BN}$ et $\mathcal{BNG}$. Il existe cependant d'autres familles de lois auxquelles peut se rattacher la $\mathcal{BNG}$ comme nous allons le voir maintenant.

2.3.1. Poisson pondérée

La $\mathcal{BNG}$ peut être considérée comme un cas particulier de la loi de Poisson pondérée surdispersée de Castillo *et al.* (1998). En fait, la $\mathcal{BNG}$ est vue dans Hassairi (1992) comme une action du groupe linéaire sur une loi de Poisson en terme de sa fonction variance. Cette action sur la densité de Poisson permet d'obtenir une expression de la densité de la $\mathcal{BNG}$ similaire à celle de la Poisson pondérée de Castillo *et al.* (1998) avec un paramétrage adéquat.

2.3.2. Famille exponentielle de dispersion

Rappelons d'abord brièvement quelques définitions et propriétés des familles de lois à structure exponentielle. Nous renvoyons le lecteur aux livres de Jørgensen (1997) et de Kotz *et al.* (2000, Chapitre 54) pour plus de détails. Soit $\theta \in \Theta$ et $\lambda \in \Lambda$, où Θ est généralement un intervalle d'intérieur ($\mathrm{int}\Theta$) non vide de $\mathbb{R}$ et Λ un sous ensemble de $]0, +\infty[$ contenant $\{1, 2, ...\}$. On dit qu'une variable aléatoire X suit une loi *exponentielle de dispersion* de paramètres θ et λ (notée $X \sim ED(\theta, \lambda)$) si sa densité par rapport à une mesure de référence peut se mettre sous la forme :

$$c(x; \lambda) \exp\{\theta x - \lambda\, \kappa(\theta)\}\,, \quad x \in \mathbb{R}. \tag{2}$$

La *famille exponentielle de dispersion* (FED) associée est alors l'ensemble des probabilités $ED(\theta, \lambda)$ avec $\theta \in \Theta$ et $\lambda \in \Lambda$.

Pour $\lambda > 0$ fixé, la FED est une *famille exponentielle naturelle* (FEN). Ainsi, les paramètres θ et λ sont respectivement dits canonique et d'indice, et satisfont la formule de convolution

$$ED(\theta, \lambda_1) * ED(\theta, \lambda_2) = ED(\theta, \lambda_1 + \lambda_2).$$

D'où, la famille est stable par convolution et donc $\{1, 2, ...\} \subseteq \Lambda$ avec $\Lambda =]0, +\infty[$ pour $ED(\theta, \lambda)$ indéfiniment divisible.

La fonction κ dans (2) est telle que, si μ est cette mesure de référence (positive et σ-finie) alors $\kappa(\theta) = \log \int e^{\theta x} c(x; 1)\, d\mu(x)$. Ainsi, κ est strictement convexe sur $\mathrm{int}\Theta$ et, pour une variable aléatoire $X \sim ED(\theta, \lambda)$, on a :

$$\mathbb{E}(X) = \lambda\, \kappa'(\theta) \quad \text{et} \quad \mathrm{Var}(X) = \lambda\, \kappa''(\theta), \tag{3}$$

où $\kappa'(\theta)$ et $\kappa''(\theta)$ sont respectivement les dérivées première et seconde de κ au point θ. À partir de (3) avec $\lambda = 1$, la fonction V définie sur le domaine $M = \kappa'(\mathrm{int}\Theta)$ telle que

$$\kappa''(\theta) = V\{\kappa'(\theta)\} \tag{4}$$

est appelé *fonction variance unité*. Tout comme pour les FEN, elle caractérise la FED. De nombreuses propriétés ont été établies dans la littérature et, dans la plupart des cas, la fonction variance unité présente une expression plus simple que celle de la densité. Les fonctions variance unités les plus connues sont les polynômes de degré

inférieur ou égal à 3 (Morris, 1982; Letac et Mora, 1990) dont un résumé est donné par le Tableau 1 ci-dessous.

Tableau 1. – Fonctions variance unités cubiques

Nom du type (symbole)	$V(v)$
Normale ($\mathcal{N}$)	1
Poisson ($\mathcal{P}$)	v
Binomiale ($\mathcal{B}$)	$v(1-v)$
Binomiale négative ($\mathcal{BN}$)	$v(1+v)$
Gamma ($\mathcal{G}$)	v^2
Cosinus hyperbolique ($\mathcal{CH}$)	v^2+1
Gaussienne inverse ($\mathcal{GI}$)	v^3
Ressel-Kendall ($\mathcal{RK}$)	$v^2(1+v)$
Poisson généralisée ($\mathcal{PG}$)	$v(1+v)^2$
Binomiale négative généralisée ($\mathcal{BNG}$)	$v(1+v)\,(1+v(a+1)/a)$
Arcsinus stricte ($\mathcal{AS}$)	$v(1+v^2)$
Arcsinus large ($\mathcal{AL}$)	$v\left(1+2v/a+v^2(1+a^2)/a^2\right)$

Notons enfin que les FED sont utilisées pour modéliser l'erreur dans les modèles (non)linéaires généralisés (McCullagh et Nelder, 1989; Wei, 1998). Les reparamétrisations $v=\kappa'(\theta)$ et $\sigma^2=1/\lambda$ permettent d'écrire la FED comme suit : $\{ED(v,\sigma^2); v\in M, \sigma^{-2}\in\Lambda\}$. Remarquons qu'une FED n'est évidemment pas une FEN à deux paramètres.

Ainsi, pour seulement a fixé, la loi $\mathcal{BNG}(a,\lambda,p)$ est une loi exponentielle de dispersion de densité de la forme (2), indexée ici par a :

$$c_a(x,\lambda)\exp\left\{\theta x-\lambda\kappa_a(\theta)\right\}. \tag{5}$$

En effet, l'équivalence entre les définitions (1) et (5) est donnée par la reparamétrisation de θ par

$$p=h(e^{\theta}), \tag{6}$$

avec $\theta=\log[p(1-p)^a]$, $\kappa_a(\theta)=-a\log(1-p)$ d'où l'on déduit que $h^{-1}(p)=p(1-p)^a$ et $h(w)=w[1-h(w)]^{-a}$ où $h(w)$ est la fonction (implicite) d'inversion de Lagrange (e.g., Kokonendji, 1999).

Une conséquence importante tirée de cette appartenance est que la famille des lois $\mathcal{BNG}(a,\lambda,p)$, pour a fixé, est caractérisée par sa fonction variance sur $]0,+\infty[$

$$V_\lambda(m)=m\left(1+\frac{m}{\lambda}\right)\left(1+\frac{m}{\lambda}\frac{(a+1)}{a}\right), \tag{7}$$

m étant l'espérance mathématique de la loi $\mathcal{BNG}(a,\lambda,p)$, qui est égale à

$$m=\frac{a\lambda p}{1-p(a+1)} \tag{8}$$

et $V_\lambda(m)$ la variance associée qui est aussi égale à $\lambda p(1-p)/[1-p(a+1)]^3$.

3. Surdispersion

Le phénomène de surdispersion dans le cadre de la FED se traduit le plus simplement par

$$V_\lambda(m) > m \tag{9}$$

pout tout $m > 0$, où V_λ est la fonction variance de la famille.

La condition de surdispersion (9) est trivialement vérifiée par la fonction variance (7) de la $\mathcal{BNG}$ quels que soient les paramètres a, λ et p. On peut noter que lorsque le paramètre λ tend vers 0 dans l'expression de la fonction variance (7), alors la surdispersion augmente et *a contrario*, lorsque ce paramètre tend vers l'infini alors la fonction variance tend vers celle d'une loi de Poisson. De plus, sans perte de généralité on fixe $\lambda = 1$, on peut établir les inégalités suivantes sur les fonctions variance unités : pour tout $m > 0$,

$$V_{\mathcal{BNG}}(m) > V_{\mathcal{PG}}(m) > V_{\mathcal{BN}}(m) > V_{\mathcal{P}}(m), \tag{10}$$

où $V_{\mathcal{BNG}}$ (resp. $V_{\mathcal{PG}}$, $V_{\mathcal{BN}}$ et $V_{\mathcal{P}}$) est la fonction variance unité de la $\mathcal{BNG}$ (resp. $\mathcal{PG}$, $\mathcal{BN}$ et $\mathcal{P}$) donnée par le Tableau 1.

Puisque, pour $\lambda = 1$ dans (8), m est une fonction croissante de a et de p, alors à m constant, si a augmente, p diminue. On remarque que plus a est petit plus la surdispersion est grande et inversement lorsque a augmente alors la surdispersion diminue avec comme cas limite la fonction variance unité de la loi de Poisson généralisée (Figure 2). Donc quels que soient les paramètres a, λ et p, la loi $\mathcal{BNG}$ est toujours plus surdispersée que la $\mathcal{PG}$ et *a fortiori* la $\mathcal{BN}$, sauf dans le cas de très petites moyennes.

3.1. Surdispersion et mélange de Poisson

Dans cette partie, au lieu de montrer uniquement le cas $\mathcal{BNG}$ comme une loi résultante d'un mélange de Poisson par une autre loi, nous donnons cette interprétation de modèles surdispersés pour toutes les lois exponentielles de dispersion dont la binomiale négative, la Poisson généralisée et bien sûr la $\mathcal{BNG}$ en font partie. Commençons tout d'abord par rappeler brièvement ce qu'est un mélange de Poisson.

La densité de la loi de Poisson (classique) de paramètre $m > 0$ en tant que loi exponentielle de dispersion est donnée par :

$$\frac{\lambda^x}{x!}\exp\{\theta x - \lambda e^\theta\}, \quad x \in \mathbb{N},$$

où $m = \lambda e^\theta$. Les paramètres θ et λ ne sont pas identifiables séparément, nous adopterons donc ici la notation $\mathcal{P}(\lambda e^\theta)$ et non $\mathcal{P}(\theta, \lambda)$ pour désigner la loi exponentielle de dispersion de Poisson.

DÉFINITION 1. – *Soit X une variable aléatoire positive de loi exponentielle de dispersion $ED(\delta, \lambda)$ avec $\delta \in \Delta \subseteq]-\infty, 0]$ et $\lambda \in \Lambda \subseteq]0, +\infty[$; c'est à dire*

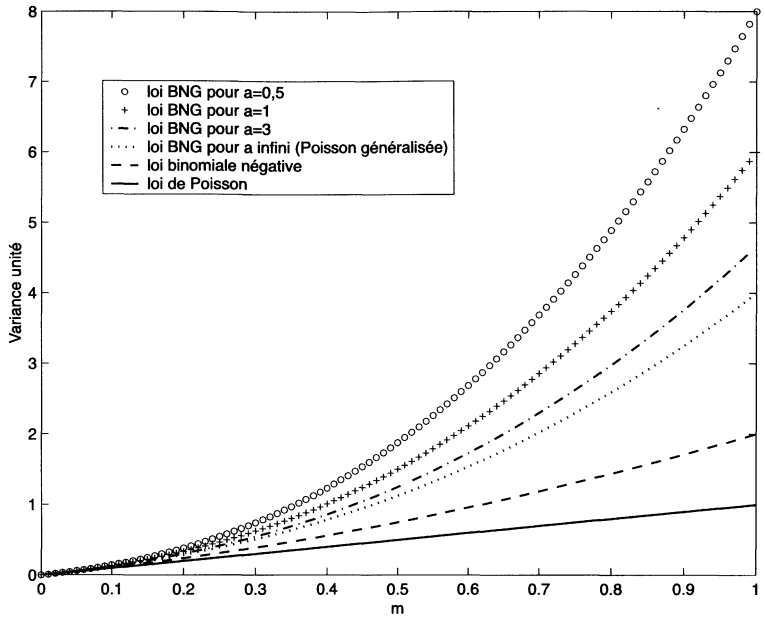


Figure 2. – Graphes de la fonction variance unité pour différentes lois

de densité $f_X(x;\delta,\lambda) = c(x;\lambda)\exp\{\delta x - \lambda\kappa(\delta)\}$. *La variable aléatoire* Y *est la résultante d'un mélange de Poisson par* X *si*

$$Y|(X = x) \textit{ suit une loi } \mathcal{P}(xe^{\theta}),$$

où $\theta \in \mathbb{R}$ *est un paramètre convenable. [Dans le cas où* $x = 0$, $\mathcal{P}(0)$ *est la loi de Dirac en zéro]. La variable aléatoire* X *sera dite mélangeante et* (Y, X) *le couple du mélange de Poisson.*

Montrons maintenant que la résultante d'un mélange de Poisson suit, elle aussi, une loi exponentielle de dispersion.

D'après la définition du couple (X, Y), la densité jointe de X et Y est donnée par

$$f_{X,Y}(x,y;\theta,\delta,\lambda) = \frac{x^y\, c(x;\lambda)}{y!}\exp\{\alpha x + \theta y - \lambda\kappa(\alpha + e^{\theta})\},$$

où $\alpha = \delta - e^{\theta}$. Par conséquent, la densité marginale de Y est, pour $y = 0, 1, ...,$

$$f_Y(y;\theta,\delta,\lambda) = \frac{m(y;\alpha,\lambda)}{y!}\exp\{\theta y - \lambda\kappa(\alpha + e^{\theta})\},$$

où $m(y;\delta,\lambda)e^{-\lambda\kappa(\delta)}$ est le moment d'ordre $y \in \mathbb{N}$ de X.

Ainsi, la proposition suivante devient immédiate.

PROPOSITION 2. – *Avec les notations de la Définition 1, pour tout* $\alpha = \delta - e^{\theta}$ *dans* Δ, Y *suit une loi* $ED(\theta, \lambda)$ *avec* $\theta \in \log(\Delta - \alpha)$, $\lambda \in \Lambda$ *et* $\mathbb{E}(Y) = \lambda e^{\theta} \kappa'(\alpha + e^{\theta})$.

En tant que FED, la résultante d'un mélange de Poisson est donc caractérisée par sa fonction variance unité donnée dans la proposition suivante.

PROPOSITION 3. – *Soit* V *la fonction variance unité d'une mélangeante de Poisson* X *de loi* $ED(\delta, \lambda)$ *avec* $\mathbb{E}(X) = \lambda \kappa'(\delta) > 0$ *où* $\delta \in \Delta$. *Alors, pour* $\alpha \in \Delta$, *la fonction variance unité* V_{α} *de la résultante* Y *suivant une loi* $ED(\theta, \lambda)$ *est donnée par*

$$V_{\alpha}(\upsilon) = \upsilon + V\left(\frac{\upsilon}{\Phi_{\alpha}(\upsilon)}\right) \Phi_{\alpha}^{2}(\upsilon) \, ,$$

pour tout $\upsilon > 0$, *où* Φ_{α} *est la fonction réciproque de* $t \mapsto t\kappa'(\alpha + t)$.

DÉMONSTRATION. – D'après la Proposition 2, on pose $\upsilon = e^{\theta} \kappa'(\alpha + e^{\theta})$ et donc $\alpha + e^{\theta} = \psi(\upsilon / e^{\theta})$ où ψ est la fonction réciproque de $\delta \mapsto \kappa'(\delta)$ telle que $\kappa''(\psi(m)) = V(m)$. En conséquence, la formule de V_{α} découle de (4) :

$$V_{\alpha}\{e^{\theta} \kappa'(\alpha + e^{\theta})\} = \frac{\partial}{\partial \theta}\{e^{\theta} \kappa'(\alpha + e^{\theta})\} \, . \boxdot$$

La résultante Y d'un mélange de Poisson par X, est donc bien un modèle de surdispersion, puisque le deuxième terme de la variance unité est positif.

Parmi les FED du Tableau 1, il est facile de former quelques couples (Y, X) : $(\mathcal{BN}, \mathcal{G})$, $(\mathcal{AS}, \mathcal{GI})$, $(\mathcal{PG}, \mathcal{RK})$, $(\mathcal{BNG}, \mathcal{RK})$ et $(\mathcal{AL}, \mathcal{RK})$, les paramètres de $\mathcal{RK}$ des trois derniers couples étant différents. On remarque que les quatre derniers couples sont associés à des fonctions variance unités strictement cubiques. Ce type de mélange convient mieux que les formes linéaires ou quadratiques pour ajuster certaines données (Efron, 1986).

Remarques. – 1) Dans l'optique d'étudier la $\mathcal{BNG}$ comme extension naturelle de la $\mathcal{BN}$, on obtient alors une interprétation probabiliste du couple $(\mathcal{BNG}, \mathcal{RK})$ par le couple $(\mathcal{BN}, \mathcal{G})$ *via* une loi de premier temps de passage. En effet, d'une part, nous avons déjà vu cette interprétation pour la $\mathcal{BNG}$ par la $\mathcal{BN}$ (§ 2.2.2), et d'autre part, la loi $\mathcal{RK}$ s'interpréte comme la loi de temps du premier passage en un point positif ou nul d'un processus gamma (e.g., Letac et Mora, 1990; Kokonendji, 2001, Théorème 1)

2) Sur l'unimodalité de la résultante d'un mélange de Poisson, Holgate (1970) a démontré que si la loi mélangeante est positive, continue et unimodale, alors la loi résultante du mélange est aussi unimodale. Ce résultat s'applique immédiatement pour le couple $(\mathcal{BN}, \mathcal{G})$, en revanche, dans le cas de la $\mathcal{BNG}$, la loi mélangeante est une $\mathcal{RK}$ dont l'unimodalité n'est pas démontrée. Cependant de nombreuses simulations de la $\mathcal{BNG}$ (Figure 1) laissent supposer que cette loi est effectivement unimodale.

3.2. *Surdispersion et problèmes d'estimation*

Dans ce paragraphe, nous abordons le problème de l'estimation des paramètres de la $\mathcal{BNG}$, lié au phénomène de surdispersion.

Famoye (1997) a étudié plusieurs méthodes pour estimer simultanément les trois paramètres de la loi $\mathcal{BNG}(a, \lambda, p)$. Il s'agit des méthodes du maximum de vraisemblance, des moments, des proportions de zéro, du minimum de khi-deux. Il a notamment comparé leur efficacité asymptotique à partir d'un échantillon simulé. Cependant, il semble que la surdispersion (liée à cette loi) pose en soi le problème de la représentativité de cette estimation, ce dont Famoye (1997) n'a pas tenu compte.

Si on se reporte à la Figure 2, on constate que la surdispersion est inversement proportionnelle à la valeur de a. Donc, afin de prendre en compte le phénomène de surdispersion dans l'estimation, nous proposons de fixer *a priori* le paramètre a. On se ramène ainsi à une FED (cf. § 2.3.2). On a alors le résultat suivant sur l'estimation des paramètres d'une loi exponentielle de dispersion $ED(\theta, \lambda)$.

THÉORÈME 4. – *Soit une loi exponentielle de dispersion $ED(\theta, \lambda)$, de densité donnée par (2) et de fonction variance unité V. Soit maintenant un échantillon aléatoire $\mathbf{y} = (y_1, \cdots, y_n)$ de valeurs de Y suivant la loi $ED(\theta, \lambda)$, où les paramètres $\theta \in \text{int}\Theta$ et $\lambda > 0$ sont tous les deux inconnus. Nous désignons la moyenne et la variance de l'échantillon respectivement par*

$$\overline{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ et } s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_n)^2.$$

On a alors : (i) L'estimation λ_n^* de λ par la méthode des moments est la solution strictement positive de

$$s_n^2 = \lambda\, V(\overline{y}_n/\lambda).$$

(ii) L'estimation $\widehat{\lambda}_n$ de λ par la méthode du maximum de vraisemblance est la solution strictement positive de

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\lambda}\log c(y_i;\lambda) = \int_0^{\overline{y}_n/\lambda}\frac{t\,dt}{V(t)}\,.$$

(iii) Si la fonction

$$\theta_n(\lambda) = \theta_0 + \int_{\theta_0}^{\overline{y}_n/\lambda}\frac{dt}{V(t)}$$

est définie de $]0, +\infty[$ sur $\text{int}\Theta$ avec $\theta_0 = \sup\{\theta; \theta \in \text{int}\Theta\}$, l'estimation θ_n^* [resp. $\widehat{\theta}_n$] de θ par la méthode des moments [resp. du maximum de vraisemblance] est donnée par $\theta_n(\lambda_n^*)$ [resp. $\theta_n(\widehat{\lambda}_n)$].

La démonstration du théorème est immédiate par une manipulation adroite des propriétés des FED.

CONJECTURE 5. – *Sous les hypothèses du Théorème 4, les estimations des paramètres θ et λ par les méthodes des moments et du maximum de vraisemblance existent si et seulement si* $s_n^2 > \overline{y}_n$.

Dans le cas de l'estimation des paramètres par la méthode des moments, la conjecture est vérifiée très simplement au cas par cas pour les modèles du Tableau 1, en calculant explicitement les estimateurs. En revanche, dans le cas de l'estimation des paramètres par la méthode du maximum de vraisemblance, la Conjecture 5 n'est pas aussi simple à vérifier. En effet, le cas de la binomiale négative a été conjecturé dans Anscombe (1950) et n'a été démontré que quelques années plus tard (e.g., Levin et Reeds, 1977). Pour la loi $\mathcal{BNG}$ qui est une généralisation lagrangienne de la binomiale négative, le résultat est présenté dans Kokonendji (1999) ainsi que celui de Poisson généralisée.

Tableau 2. – Estimation de λ à partir de 1000 échantillons de taille 100

a	p	λ	taux d'échantillons surdispersés en %	λ_n^*	$var(\lambda_n^*)$
1	0,33	**10**	100	**10,39**	2,68
1	0,10	**0,50**	81,5	**0,68**	0,37
0,50	0,50	**1**	100	**1,24**	0,14
5	0,10	**1**	100	**1,16**	0,10
15	0,05	**0,10**	98,3	**0,19**	0,007

Tableau 3. – Estimation de λ à partir de 1000 échantillons de taille 1000

a	p	λ	taux d'échantillons surdispersés en %	λ_n^*	$var(\lambda_n^*)$
1	0,33	**10**	100	**10,04**	0,26
1	0,10	**0,50**	99,8	**0,62**	0,13
0,50	0,50	**1**	100	**1,02**	0,015
5	0,10	**1**	100	**1,02**	0,01
15	0,05	**0,10**	100	**0,12**	0,0005

Les Tableaux 2 et 3 donnent l'estimation du paramètre λ de la $\mathcal{BNG}(a, \lambda, p)$ par la méthode des moments (i) du Théorème 4, calculé à partir des échantillons simulés qui sont surdispersés (e.g., Kokonendji, 1999; Théorème 1 pour la formule explicite de λ_n^* à a connu). Par la suite et à partir de (8), on pourrait estimer p par $p_n^* = \overline{y}_n/(\overline{y}_n(1+a)+a\lambda_n^*)$. L'estimation du paramètre θ de la FED associée s'obtient enfin à l'aide de la formule (iii) du Théorème 4 ou bien simplement à l'aide de la formule (6).

Dans ces Tableaux 2 et 3, on remarque que pour certains paramètres a, p et λ de $\mathcal{BNG}(a, \lambda, p)$, tous les échantillons ne sont pas surdispersés (au sens $s_n^2 > \overline{y}_n$). Ce phénomène s'explique par le fait que dans ces cas le moment d'ordre 1 (8) est très faible (ici $m < 0,4$ avec $\lambda \leqslant 0,5$). Comme on peut le constater sur la Figure 2, les courbes de variances unités ($\lambda = 1$) sont presque confondues pour m très

petit, d'où un risque non nul (dû essentiellement au calculateur) de sousdispersion sur l'échantillon simulé.

3.3. Discussion

D'un point de vue pratique, dès qu'un échantillon est surdispersé ($s_n^2 > \overline{y}_n$), comment peut-on juger de son degré de surdispersion? Autrement dit, quand peut-on parler de petite, moyenne ou grande surdispersion? L'exemple du "nombre d'accidents enregistrés auprès de conducteurs Belges" décrit dans Gelfand et Dalal (1990) pose un problème du choix de modèle surdispersé adéquat; car on a bien $s_n^2 = 0,029$ supérieur à $\overline{y}_n = 0,021$. L'avantage du modèle $\mathcal{BNG}$ dans ce cas, est que le degré de surdispersion est automatiquement pris en compte par les deux paramètres a et λ (cf. début de § 3).

Enfin, la terminologie "surdispersée" prend parfois dans la littérature un tout autre sens que celui défini dans cet article. Elle peut être utilisée pour dire que la variance d'un échantillon observé est supérieure à la variance fournie par le modèle considéré. Nous pouvons l'illustrer à l'aide de l'exemple classique de la fréquence des mâles dans 6115 "sibships" de taille 12 en Saxony (e.g., Gelfand et Dalal, 1990). On observe sur l'échantillon que $\overline{y}_n = 6,23$ et $s_n^2 = 3,49$; on est donc dans un cas sousdispersé et le choix d'un modèle binomial envisagé par les auteurs semble justifié. Cependant, ils parlent alors de "surdispersion" car, dans ce cas, la variance du modèle binomial est sous estimée par la variance de l'échantillon : $n\widehat{p}(1-\widehat{p}) = 2,4 < s_n^2$.

Remerciements. – Les auteurs remercient le comité de rédaction de la revue pour leurs commentaires sur cet article.

Références bibliographiques

ANSCOMBE, F.J. (1950), Sampling theory of the negative binomial and logarithmic series distributions, *Biometrika* ***37***, 358-382.

BRESLOW, N. (1984), Extra-Poisson variation in log-linear models, *Applied Statistics* ***33***, 38-44.

CASTILLO, J. ET PÉREZ-CASANY, M. (1998), Weighted poisson distribution for overdispersion and underdispersion situations, *Ann. Inst. Statist. Math.* ***50***, 567-585.

CONSUL, P.C. (1989), *Generalized Poisson Distributions,* Marcel Dekker, New York.

COX, D.R. (1983), Some remarks on overdispersion, *Biometrika* ***70***, 269-274.

DIEUDONNÉ, J. (1971). *Infinitesimal Calculus*, Houghton Mifflin, Boston.

EFRON, B. (1986), Double exponential families and their use in generalized linear regression, *J. Amer. Statist. Assoc.* ***81***, 709-721.

FAMOYE, F. (1997), Parameter estimation for generalized negative binomial distribtution, *Comm. Statist.-simula* ***26***, 269-279.

GELFAND, A.E ET DALAL, S.R.A. (1990), A note on overdispersed exponential families, *Biometrika* ***77***, 55-64.

GIANO, L.M. ET SCHAFER, D.W. (1992), Diagnostics for overdispersion, *J. Amer. Statist. Assoc.* ***87***, 795-804.

GREENWOOD, M. ET YULE, G.U. (1920), An inquiry into the nature of frequency distributions representative of multiple happenings with particular referee to the occurrence of multiple attacks of disease or of repeated accidents, *J. R. Statist. Soc. Ser. A* ***83***, 255-279.

GUPTA, R.C. (1977). Minimum variance unbiased estimation in a modified power series distributions, *Comm. Statist.* ***10***, 977-991.

HARRIS, T.E. (1963). *The Theory of Branching Processes*, Springer-Verlag, Berlin.

HASSAIRI, A. (1992), La classification des familles exponentielles naturelles sur $\mathbb{R}^n$ par l'action du groupe linéaire de $\mathbb{R}^{n+1}$. *C. R. Acad. Sci. Paris Sér. I* ***315***, 207-210.

HINDE, J.P. ET DEMÉTRIO, C.G.B. (1996), Modelling with overdispersion, In *Statistical Modelling Proceedings of the 11th International Workshop on Statistical Modelling*, (ed. Fratra, A.), 200-207.

HOLGATE, P. (1970), The modality of some compound Poisson distributions, *Biometrika* ***57***, 666-667.

JAIN, G.C. ET CONSUL, P.C. (1971), A generalized negative binomial distribution, *SIAM J. Appl. Math.* ***21***, 501-513.

JOHNSON, N.L., KOTZ, S. ET KEMP, A.W. (1992), *Univariate Discrete Distributions,* Second Edition, John Wiley & Sons, New York.

JøRGENSEN, B. (1997), *The theory of dispersion models,* Chapman & Hall, London.

KOKONENDJI, C.C. (1994), Exponential families with variance functions $\sqrt{\Delta}P(\sqrt{\Delta})$: Seshadri's class, *Test* ***3***, 123-172.

KOKONENDJI, C.C. (1999), Le problème d'Anscombe pour les lois binomiales négatives généralisées, *Canadian J. Statist.* ***27***, 199-205.

KOKONENDJI, C.C. (2001), First passage times on zero and one and natural exponential families, *Statist. Probab. Letters* ***51***, 293-298.

KOTZ, S., BALAKRISHNAN, N. ET JOHNSON, N.L. (2000), *Continuous multivariate distributions*, Volume 1 : Models and Applications, Wiley, NewYork, 2nd édition.

LETAC, G. ET MORA, M. (1990), Natural real exponential families with cubic variance functions, *Ann. Statist.* ***18***, 1-37.

LEVIN, B. ET REEDS, J. (1977), Compound multinomial likelihood functions are unimodal : proof of a conjecture of I. J. Good, *Ann. Statist.* ***5***, 79-87.

MCCULLAGH, P. ET NELDER, J.A. (1989), *Generalized Linear Models*, 2nd ed. Chapman & Hall, London.

MORRIS, C.N. (1982), Natural exponential families with quadratic variance functions, *Ann. Statist.* ***10***, 65-80.

TAKÁCS, L. (1962), A generalization of the ballot problem and its applications in the theory of queues, *J. Amer. Statist. Assoc.* ***57,*** 327-337.

WEI, B-G. (1998), *Exponential Family Nonlinear Models*, Lecture Notes in Statistics No. 130, Springer-Verlag, Singapore.

YANAGIMOTO, T. (1989), The inverse binomial distribution as statistical model, *Comm. Statist.- Theory Methods* ***18***, 3625-3633.