A Suitable Discrete Distribution for Modelling Automobile Claim Frequencies

E. Gómez–Déniz[a], A. Hernández–Bastida[b], M.P. Fernández–Sánchez[b]

[a] *Department of Quantitative Methods in Economics, University of Las Palmas de Gran Canaria, Las Palmas, Spain*
[b] *Department of Quantitative Methods in Economics, University of Granada, Granada, Spain*

Abstract

A new discrete distribution, depending on two parameters, is introduced in this paper. A mixing process is utilised, with the discrete Lindley acting as the mixed and the Beta prime as the mixing distribution. The distribution obtained is shown to be unimodal and overdispersed. An equation for the probability density function of the compound version, when claim severities are exponentially distributed, is also derived. After reviewing some of its properties, we investigate the question of parameter estimation. Real frequency data consisting of automobile claim frequencies were fitted successfully using the proposed distribution and the estimated values were used to compute the right tail probabilities of the aggregate claim size distribution when the new distribution acts as the primary distribution. These values are compared with those obtained when the Poisson distribution is assumed to be the primary distribution.

Keywords: Claim, Compound, Poisson, Mixture, Overdispersion, Unimodality

Acknowledgements

This research was partially funded by EGD and AHB: ECO2009-14152 (MICINN, Spain).
The authors are indebted to the referees for helpful comments which, without doubt, helped to improve an earlier version of the paper.

Address for correspondence: Emilio Gómez–Déniz, Department of Quantitative Methods, University of Las Palmas de Gran Canaria, 35017–Las Palmas de Gran Canaria, Spain. E–mail: egomez@dmc.ulpgc.es

A Suitable Discrete Distribution for Modelling Automobile Claim Frequencies

Abstract

A new discrete distribution, depending on two parameters, is introduced in this paper. A mixing process is utilised, with the discrete Lindley acting as the mixed and the Beta prime as the mixing distribution. The distribution obtained is shown to be unimodal and overdispersed. An equation for the probability density function of the compound version, when claim severities are exponentially distributed, is also derived. After reviewing some of its properties, we investigate the question of parameter estimation. Real frequency data consisting of automobile claim frequencies were fitted successfully using the proposed distribution and the estimated values were used to compute the right tail probabilities of the aggregate claim size distribution when the new distribution acts as the primary distribution. These values are compared with those obtained when the Poisson distribution is assumed to be the primary distribution.

Keywords: Claim, Compound, Poisson, Mixture, Overdispersion, Unimodality

1 Introduction

Since Sankaran (1970) presented the discrete Lindley distribution, no study of this distribution has been published until very recently, when some papers have pointed out the usefulness of this simple one-parameter distribution (Ghitany et al., 2008); Ghitany and Al-Mutairi, 2009; Hernández et al., 2011; Gómez–Déniz and Calderín, 2011; Mahmoudi and Zakerzadeh, 2010).

The Poisson Lindley distribution proposed by Sankaran (1970) and studied recently by Ghitany and Al-Mutairi (2009) has the following probability mass function (henceforth, pmf)

$$\Pr(X = x) = \frac{\theta^2(\theta + x + 2)}{(1 + \theta)^{x+3}}, \qquad (1)$$

for $\theta > 0$ and $x = 0, 1, \ldots$. This distribution is built by mixing a Poisson distribution with parameter $\lambda > 0$ and allowing this parameter to follow a continuous Lindley distribution with probability density function (see Lindley, 1958)

$$f(\lambda;\theta) = \frac{\theta^2}{1+\theta}(1+\lambda)\, e^{-\theta\,\lambda}, \quad \lambda > 0,\, \theta > 0. \tag{2}$$

Mahmoudi and Zakerzadeh (2010) suggested an extension of this distribution by including an additional parameter. Henceforth, we shall use $\mathscr{PL}$ to denote the pmf in (1) and $\mathscr{L}$ to denote the probability density function (pdf) in (2).

In his introduction to Rolski et al. (1999), Winkelmann (2000) highlighted the need for richer classes of probability distributions when modelling count data. These monographs place special attention on searching for more flexible distributions, since they can be used as building blocks for improved count data models with immediate application in insurance to describe the accumulated claims. In the present paper, a new discrete distribution is obtained, by mixing the $\mathscr{PL}$ distribution with the Beta prime distribution (also called inverted Beta distribution or beta distribution of the second kind). The proposed distribution can be considered as an alternative to the negative binomial, the strict arcsine (see Kokonendji and Khoudar, 2004), the negative binomial–Pareto (see Meng et al., 1999) and the Poisson–inverse Gaussian (Willmot, 1987), all of which are well known in actuarial literature. The new distribution is unimodal with the possibility of a zero vertex or a mode greater than zero, depending on the values of the parameters of the distribution. Furthermore, because the new distribution can be obtained from a Poisson mixture, it is overdispersed (the variance is greater than the mean). These two features, zero–vertex (high percentage of zero values in the empirical distribution (Boucher et al., 2007 and 2009) and overdispersion are omnipresent in automobile insurance portfolios, which are characterised by overdispersion and zero–inflation. Therefore, the new distribution can be considered as a useful alternative for modelling phenomena of this nature in the context of insurance.

When addressing the aggregate amount of claims for a compound class of policies, and when the new distribution acts as the primary one, a closed expression for the pdf of the aggregate losses is obtained, assuming that the secondary distribution is exponential.

After reviewing some of its properties, we consider the question of parameter estimation, presenting a simple equation that only implements the digamma function when the maximum likelihood method is considered. The expected frequencies are calculated in an example based on automobile claim frequencies, and the estimated values were used to compute the right tail probabilities of the aggregate losses. The results obtained were compared with those when the Poisson distribution is assumed to be the primary distribution.

The structure of the paper is as follows. Section 2 presents the new discrete distribution and describes its properties. Section 3 discusses aggregate losses, and Section 4 addresses the estimation of the two parameters of the new model. In Section 5 the expected frequencies are calculated for a set of data concerning automobile claim frequencies, and the distribution was found to provide a very satisfactory fit for computing the right tail probabilities of the aggregate losses. In the final section, some conclusions are drawn.

2 Main results

In this section we define the new pmf and study some of its more important properties.

We begin by defining the beta prime distribution, also known as inverted beta distribution or beta distribution of the second kind. This is an absolutely continuous probability distribution defined for $\theta > 0$ with two parameters $\alpha > 0$ and $\beta > 0$, having the pdf

$$f(\theta;\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\frac{\theta^{\alpha-1}}{(1+\theta)^{\alpha+\beta}}, \quad \alpha > 0,\ \beta > 0,\ \theta > 0, \tag{3}$$

where $B(x,y)$ denotes the usual beta function. Henceforth, a random variable with pdf (3) is denoted by $\theta \sim \mathscr{BP}(\alpha,\beta)$. Some interesting properties of this distribution are considered in Goovaerts and De Pril (1980) and in Gómez–Déniz et al. (2008).

Definition 1 *A random variable X is said to have a Poisson Lindley– Beta prime distribution if it accepts the stochastic representation:*

$$X|\theta \sim \mathscr{PL}(\theta), \tag{4}$$
$$\theta \sim \mathscr{BP}(\alpha,\beta), \tag{5}$$

with $\alpha > 0$, $\beta > 0$. We denote this distribution by $X \sim \mathscr{PLBP}(\alpha,\beta)$.

The next theorem provides closed formulas for the pmf and for the factorial moments.

Theorem 1 *Let $X \sim \mathscr{PLBP}(\alpha,\beta)$ be a Poisson Lindley–Beta prime distribution defined in (4)-(5). Some basic properties are:*

(a) The pmf is given by

$$\Pr(X=x)=\frac{\alpha\,(1+\alpha)\,\Gamma(\alpha+\beta)\,\Gamma(\beta+x)}{\Gamma(\beta)\,\Gamma(\alpha+\beta+x+3)}\left[(\beta+x)(2+x)+\alpha+2\right], \tag{6}$$

with $x=0,1,2,\dots$ and $\alpha>0$, $\beta>0$.

(b) The factorial moment of order k is given by

$$\mu_{[k]}(X)=k!\frac{\Gamma(\alpha-k)\,\Gamma(\beta+k)}{\Gamma(\alpha)\,\Gamma(\beta)}\,\frac{\alpha+\beta(k+1)+k^2}{\alpha+\beta},\ \alpha>k, \tag{7}$$

with $k=1,2,\dots$

Proof: If $X \sim \mathscr{PL}(\theta)$ and $\theta \sim \mathscr{BP}(\alpha,\beta)$, the pmf of X can be obtained by using the well–known compound formula,

$$\Pr(X=x)=\int_0^\infty \Pr(X=x|\theta)f(\theta;\alpha,\beta)\,d\theta,$$

where $f(\theta;\alpha,\beta)$ is the pdf of the Beta prime distribution defined in (3). The factorial moment of order k of the $\mathscr{PL}$ distribution is given by (see Gómez–Déniz et al., 2012)

$$\mu_{[k]}(X;\theta)=E[X(X-1)\dots(X-k+1)]=k!\,\frac{\theta+k+1}{\theta^k(\theta+1)},\quad k=1,2,\dots,$$

from which we can obtain the factorial moment of the $\mathscr{PLBP}$ distribution, again by the compound formula

$$\mu_{[k]}(X)=\int_0^\infty k!\frac{\theta+k+1}{\theta^k(\theta+1)}f(\theta;\alpha,\beta)\,d\theta$$

and after some algebra. ■

Some examples of the pmf given in (6) for different values of the parameters α and β are shown in Figure 1.

The moments about the origin can be obtained from (7). In particular, the mean and the second moment about the origin are given by

$$E(X) = \mu(\alpha,\beta) = \frac{\beta(2\beta+\alpha+1)}{(\alpha+\beta)(\alpha-1)}, \quad \alpha > 1, \tag{8}$$

$$E(X^2) = \frac{\beta[6+2\beta(5+3\beta)+\alpha(\alpha+4\beta+1)]}{(\alpha+\beta)(\alpha-1)(\alpha-2)}, \quad \alpha > 2. \tag{9}$$

In addition, the probabilities in (6) can be computed recursively by using

$$\Pr(X=x) = \frac{(\beta+x)(x+2)+\alpha+2}{(\beta+x-1)(x+1)+\alpha+2}\,\frac{\beta+x-1}{\alpha+\beta+x+2}\Pr(X=x-1)$$

for $x = 1, 2, \ldots$, with

$$\Pr(X=0) = \frac{\alpha(\alpha+1)(2\beta+\alpha+2)}{(\alpha+\beta+2)(\alpha+\beta+1)}(\alpha+\beta). \tag{10}$$

From (8) it is straightforward to obtain that

$$\frac{\partial\mu(\alpha,\beta)}{\partial\alpha} = \beta\left[\frac{1}{(\alpha+\beta)^2} - \frac{2}{(\alpha-1)^2}\right]. \tag{11}$$

Since $\alpha+\beta > \alpha-1$, it can be seen that $\frac{\partial\mu(\alpha,\beta)}{\partial\alpha} < 0$ and therefore the mean decreases always with α. Furthermore,

$$\frac{\partial\mu(\alpha,\beta)}{\partial\beta} = \frac{2}{\alpha-1} - \frac{\alpha}{(\alpha+\beta)^2} = \frac{\alpha^2+2\beta^2+4\alpha\beta+\alpha}{(\alpha-1)(\alpha+\beta)^2} > 0.$$

Therefore, the mean always increases with β.

Proposition 1 *The discrete distribution with pmf given in (6) is a Poisson (with parameter $\lambda > 0$) mixture distribution with mixing distribution with pdf*

$$\pi(\lambda) = \frac{\alpha\,(1+\alpha)\,\Gamma(\alpha+\beta)}{\Gamma(\beta)}\,(1+\lambda)\,\mathscr{U}(2+\alpha, 2-\beta, \lambda), \quad \lambda \geq 0, \tag{12}$$

where $\mathscr{U}(a,b,z)$ represents the confluent hypergeometric function given by $(a, z > 0)$:

$$\mathscr{U}(a,b,z) = \frac{1}{\Gamma(a)}\int_0^\infty e^{-zs}s^{a-1}(1+s)^{b-a-1}ds \tag{13}$$

(see Gradshteyn and Ryzhik (1994), page 1085, formula 9211–4).

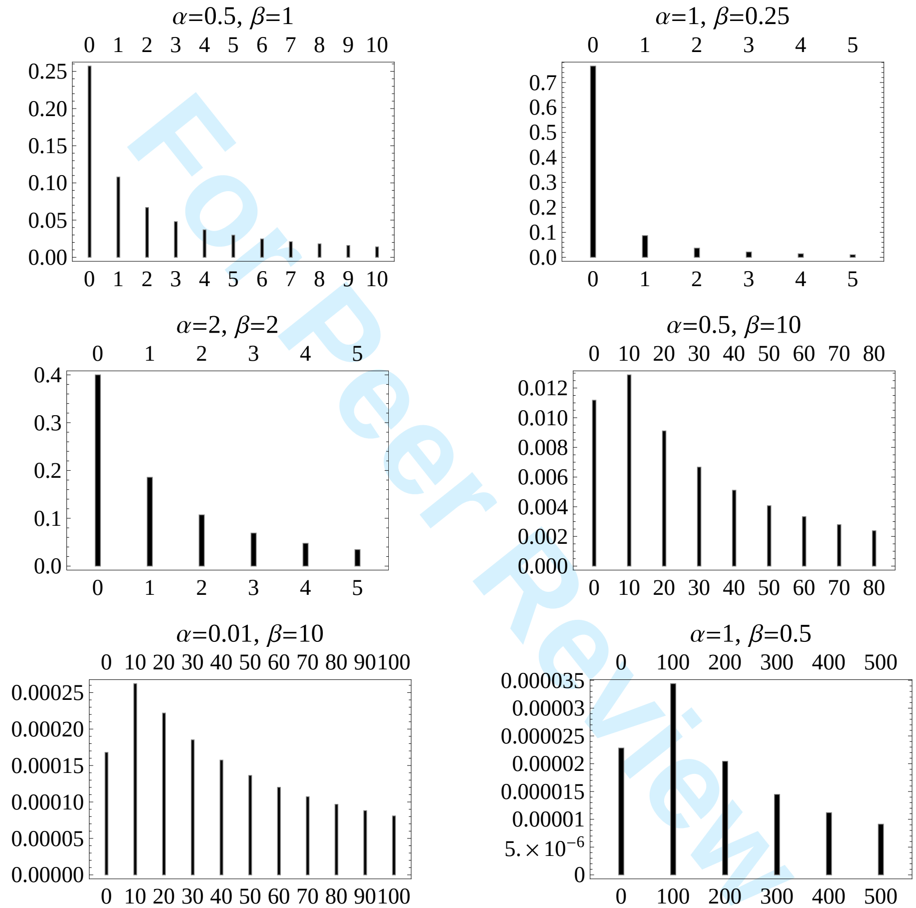

Figure 1: Some examples of probability mass functions of the new distribution for different values of the parameters α and β

Proof: It is obvious that

$$\begin{aligned} Pr(X=x) &= \mathscr{P}\mathscr{L}(\theta)\bigwedge_{\theta}\mathscr{B}\mathscr{P}(\alpha,\beta) = \left(\mathscr{P}(\lambda)\bigwedge_{\lambda}\mathscr{L}(\theta)\right)\bigwedge_{\theta}\mathscr{B}\mathscr{P}(\alpha,\beta) \\ &= \mathscr{P}(\lambda)\bigwedge_{\lambda}\left(\mathscr{L}(\theta)\bigwedge_{\theta}\mathscr{B}\mathscr{P}(\alpha,\beta)\right). \end{aligned}$$

Now, the mixture $\mathscr{L}(\theta)\bigwedge_{\theta}\mathscr{B}\mathscr{P}(\alpha,\beta)$ is given by

$$\pi(\lambda) = \frac{1+\lambda}{B(\alpha,\beta)}\int_0^\infty \frac{\theta^{\alpha+1}e^{-\theta\lambda}}{(1+\theta)^{\alpha+\beta+1}}\,d\theta,$$

from which the result follows after simple algebra by using (13). ■

Using Proposition 10 in Karlis and Xekalaki (2005) the posterior expectation of λ^r given x can be computed using the expression

$$E(\lambda^r|x) = \frac{\Gamma(x+r)\ \Pr(X=x+r)}{\Gamma(x+1)\ \Pr(X=x)},$$

for r taking positive or negative values.

Furthermore, because the new distribution arises from a mixture of a Poisson distribution, the variance–to–mean ratio is greater than one (see Karlis and Xekalaki (2005) and Sundt and Vernic, (2009), p.66) which means that the new distribution is overdispersed.

It is well known that the unimodality property is very important in many statistics models. The following result shows that the discrete distribution proposed in this paper is unimodal.

Proposition 2 *The discrete distribution with pmf given in (6) is unimodal.*

Proof: By using the following equalities

$$\begin{aligned} \mathscr{U}(a,b,z) &= z^{1-b}\,\mathscr{U}(a-b+1,2-b,z), \\ \frac{\partial}{\partial z}\mathscr{U}(a,b,z) &= -a\,\mathscr{U}(a+1,b+1,z), \end{aligned}$$

it is simple to obtain that

$$\begin{aligned} \frac{d\,\pi(\lambda)}{d\lambda} &= \frac{\alpha(1+\alpha)\Gamma(\alpha+\beta)}{\Gamma(\beta)}\mathscr{U}(2+\alpha,2-\beta,\lambda)\left[1\right. \\ &- \left.\left(1+\frac{1}{\lambda}\right)(\alpha+2)\frac{\mathscr{U}(\alpha+\beta+1,\beta-1,\lambda)}{\mathscr{U}(\alpha+\beta+1,\beta,\lambda)}\right]. \end{aligned}$$

Now, it can be seen that $\mathscr{U}(\alpha+\beta+1,\beta-1,\lambda) > \mathscr{U}(\alpha+\beta+1,\beta,\lambda)$ and therefore $\frac{d}{d\lambda}\pi(\lambda) < 0$, i.e. it is unimodal with a modal value at $\lambda = 0$.

Now, the proposition is the direct consequence of applying Proposition 1 and Theorem in Holgate (1970). ■

Using

$$\frac{\Pr(X=x)}{\Pr(X=x-1)} = \frac{(\beta+x)(x+2)+\alpha+2}{(\beta+x-1)(x+1)+\alpha+2}\,\frac{\beta+x-1}{\alpha+\beta+x+2}, \quad x = 1,2,\ldots,$$

it can be shown that, provided that

$$m(\alpha,\beta) = (1+\alpha\beta)^2 - 4(1+\alpha)\left[(2+\alpha)^2 + (3+\alpha)\beta - \beta^2\right] \geq 0$$

the mode is at the origin if $\mathscr{M}(\alpha,\beta) < 0$, where

$$\mathscr{M}(\alpha,\beta) = \frac{-1-\alpha\beta+\sqrt{m(\alpha,\beta)}}{2(1+\alpha)}.$$

When $\mathscr{M}(\alpha,\beta) > 0$ the mode is at $[\mathscr{M}(\alpha,\beta)]$, where $[\cdot]$ denotes the integer part. If $[\mathscr{M}(\alpha,\beta)]$ is an integer, then there are joint modes at $\mathscr{M}(\alpha,\beta)-1$ and $\mathscr{M}(\alpha,\beta)$. Finally, when $\mathscr{M}(\alpha,\beta) < 0$ or when $m(\alpha,\beta) < 0$ the mode is at zero.

3 $\mathscr{PLBP}$ as the primary distribution

In collective risk theory, the random variable of interest is the total or aggregate amount of claims of the insurance portfolio, $Y = \sum_{i=1}^{X} Y_i$, where X is the random variable denoting the number of claims and Y_i, for $i = 1,2,\ldots$ is the random variable denoting the size or amount of the i–th claim. Assuming that $Y_1, Y_2, \ldots$, are independent and identically distributed random variables which are also independent of the random number of claims X, it is well known (Klugman et al., 2008; Rolski et al., 1999; Antzoulakos and Chadjiconstantinidis, 2004; among others) that the pdf of the sum Y is given by $g(y) = \sum_{x=0}^{\infty} p_x f^{*x}(y)$, where p_x denotes the probability of x claims (primary distribution) and $f^{*x}(y)$ is the x–th fold convolution of $f(y)$, the pdf of the claim amount (secondary distribution).

There exists a considerable body of literature dealing with compound mixture Poisson distributions, including Willmot (1986, 1993) and Antzoulakos and Chadjiconstantinidis (2004). An extensive review of the topic can be found in Sundt and Vernic (2009). Based on the recursion provided by Panjer (1981) for the Poisson distribution, Sundt and Vernic (2009, chapter 3, p.68) developed a simple algorithm to determine the probabilities of the random variable Y when the amount of the single claim follows a discrete distribution with a given pmf.

Numerical evaluation of compound distributions is one of the most important numerical tasks in insurance mathematics. Two widely–used techniques are Panjer's recursion algorithm and transform methods. With respect to the latter, many authors have pointed out that aliasing errors imply the need to consider the whole distribution, which is a potential drawback, especially for heavy–tailed distributions, for which the question of large claims is very important, particularly in relation to reinsurance.

It is well known in actuarial statistics that the use of the extrapolated tail to estimate an extremely large claim probability, after fitting a distribution to claim size data, could mislead us into believing that the value is larger than it actually is. This is particularly important in the framework of reinsurance premiums. Obviously, a reinsurer must not use a distribution whose tail declines to zero too quickly. For this reason, Pareto and log–normal distributions are commonly used in reinsurance premium computation. Nevertheless, the compound Poisson model has traditionally been considered when the size of a single claim is modelled by an exponential distribution, chiefly because of the complexity of the collective risk model under these kinds of distributions. In this section, an alternative expression is developed by considering the $\mathscr{PLBP}$ distribution proposed in the previous sections.

The collective theory of risk is based on the assumption that the counting process representing the number of claims is a Poisson process and the associated cumulative, which represents the compound process, is thus a compound Poisson. The following result shows an alternative to this classical model, providing a closed form for the pdf of the total costs of claims when the $\mathscr{PLBP}$ and exponential distributions are assumed as primary and secondary distributions, respectively.

Theorem 2 *If we assume a $\mathscr{PLBP}$ distribution as the primary distribution and an exponential distribution with parameter $\gamma > 0$ as the secondary distribution, then the pdf of the random variable $Y = \sum_{i=1}^{X} Y_i$ is given by*

$$g(y) = \frac{\gamma\beta\alpha(\alpha+1)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+4)} \left[(\alpha+2)\, {}_1F_1(3+\alpha; 4+\alpha+\beta; -\gamma\, y)\right.$$

$$+ \quad 3(\beta+1)e^{-\gamma y}\,{}_2F_2(\{\beta+2,4\},\{\alpha+\beta+4,3\},\gamma y)\Big], \quad y>0, \quad (14)$$

and $g(0)$ is given in (10). Here, ${}_1F_1(a;b;z)$ represents the Kummer confluent hypergeometric function given by ${}_1F_1(a;b;z) = \sum_{k=0}^{\infty}(a)_k/(b)_k z^k/k!$ and ${}_2F_2(\{a_1,a_2\},\{b_1,b_2\},z) = \frac{(a_1)_k(a_2)_k}{(b_1)_k(b_2)_k}z^k/k!$ is the generalized hypergeometric function, where $(a)_k = \Gamma(a+k)/\Gamma(a)$ is the Pochhammer symbol.

Proof: By assuming that the claim amount follows an exponential distribution with parameter $\gamma > 0$, the x–th fold convolution of exponential distribution has a closed form and is given by (see Klugman et al. (2008), Rolski et al. (1999) and Gómez–Déniz and Calderín (2011))

$$f^{*x}(y) = \frac{\gamma^x}{(x-1)!}y^{x-1}e^{-\gamma y}, \; x = 1,2,\dots$$

i.e. it is a gamma distribution with parameters x and γ. Then, for $y > 0$ we have that

$$\begin{aligned}
g(y) &= \frac{e^{-\gamma y}}{y}\frac{\alpha(\alpha+1)\Gamma(\alpha+\beta)}{\Gamma(\beta)}\sum_{x=1}^{\infty}\frac{(\gamma y)^x}{\Gamma(x)}\Pr(X=x) \\
&= \frac{e^{-\gamma y}}{y}\frac{\gamma y\alpha(\alpha+1)\Gamma(\alpha+\beta)}{\Gamma(\beta)}\left[(\alpha+2)\sum_{j=0}^{\infty}\frac{\Gamma(\beta+j+1)}{\Gamma(\alpha+\beta+j+4)}\frac{(\gamma y)^j}{j!}\right. \\
&+ \left.\sum_{j=0}^{\infty}\frac{\Gamma(\beta+j+2)}{\Gamma(\alpha+\beta+j+4)}\frac{\Gamma(j+4)}{\Gamma(j+3)}\frac{(\gamma y)^j}{j!}\right] \\
&= \frac{\gamma e^{-\gamma y}\alpha(\alpha+1)\Gamma(\alpha+\beta)}{\Gamma(\beta)}\left[\frac{(\alpha+2)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+4)}\,{}_1F_1(\beta+1,\alpha+\beta+4,\gamma y)\right. \\
&+ \left.\frac{\Gamma(\beta+2)\Gamma(4)}{\Gamma(\alpha+\beta+4)\Gamma(3)}\,{}_2F_2(\{\beta+2,4\},\{\alpha+\beta+4,3\},\gamma y)\right].
\end{aligned}$$

Now, using Kummer's first theorem, which states

$${}_1F_1(a;b;z) = e^z\,{}_1F_1(b-a,b,-z),$$

and after simple computations the theorem holds. ∎

Observe that the pdf of the aggregate claim amount has a size jump given by $\Pr(X=0)$ at the origin.

The best–known aggregate claim size distribution is that obtained when the primary and the secondary distribution are the Poisson and the exponential distribution, respectively. In this case (see Rolski et al. (1999) and

Hernández et al. (2009); among others) the distribution of the random variable aggregate claim size is given by

$$g(y) = \sqrt{\frac{\gamma\nu}{y}} e^{-(\nu+\gamma y)} I_1\left(2\sqrt{\nu\gamma y}\right), \qquad y > 0, \tag{15}$$

while $g(0) = e^{-\nu}$. Here, $\nu > 0$ and $\gamma > 0$ are the parameters of the Poisson and exponential distributions, respectively and

$$I_c(z) = \sum_{k=0}^{\infty} \frac{(z/2)^{2k+c}}{\Gamma(k+1)\Gamma(c+k+1)}, \quad z \in \mathbb{R},\ c \in \mathbb{R},$$

represents the modified Bessel function of the first kind.

It is also well known that in this model, the mean of the random variable aggregate claim amount is $E(Y) = E(X)E(Y_1)$, which gives

$$\begin{aligned} E(Y) &= \frac{\beta(2\beta+\alpha+1)}{\gamma(\alpha+\beta)(\alpha-1)}, \\ E(Y) &= \frac{\nu}{\gamma}, \end{aligned}$$

for the compound Poisson Lindley Beta prime model, henceforth denoted as $\mathscr{CPLBP}$ and for the compound Poisson, denoted as $\mathscr{CP}$ model, respectively.

The pdf's given in (14) and (15) with the same mean and different values of α and β assuming, without loss of generality, $\gamma = 1$ are shown in Figure 2. As can be seen, the pdf (14) declines to zero more slowly than (15), and therefore is suitable for modelling extreme data. The new model seems to capture potentially severe tail loss events, having the longest tail and thus involving greatest uncertainty.

4 Estimation of the parameters

In practical situations, we need to estimate/elicit the two parameters α and β. This can be done by using the first two moments about the origin given in (8) and (9). Since no closed expressions exist for these estimates, a numerical method can be used to achieve them. When the proportion of zero frequencies is very large, it is also possible to use the mean, given in (8), and the proportion of zero frequencies given in (10) to estimate the parameters.

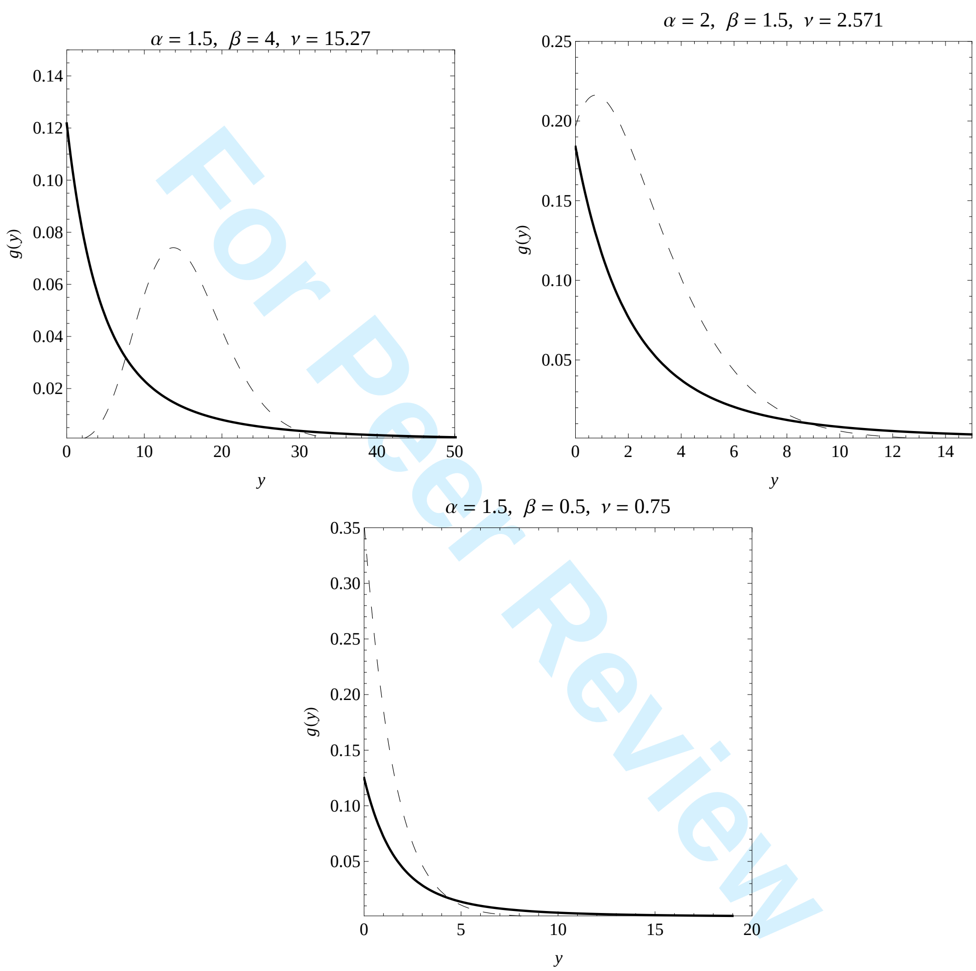

Figure 2: Probability density function in the aggregate claim model: $\mathscr{CPLBP}$ model and $\mathscr{CP}$ model (dashed) for selected values of parameters

Nevertheless, we propose to directly use the maximum likelihood method, because the normal equations are very simple and depend only on the digamma function, which is available within standard Mathematica packages. To do so, assume that $\underline{x} = (x_1, ..., x_n)$ is a random sample of size n from the random variable with pmf (6). The log–likelihood function is:

$$\begin{aligned} \ell &= n\left[\log\alpha + \log(1+\alpha) + \log\Gamma(\alpha+\beta) - \log\Gamma(\beta)\right] \\ &+ \sum_{i=1}^{n}\{\log\Gamma(\beta+x_i) + \log\left[(\beta+x_i)(x_i+2)+\alpha+2\right] \\ &- \log\Gamma(\alpha+\beta+x_i+3)\}\,. \end{aligned} \tag{16}$$

Hence, from (16), we obtain the likelihood equations, which are given by

$$\begin{aligned} \frac{\partial\ell}{\partial\alpha} &= n\left[\frac{1}{\alpha} + \frac{1}{1+\alpha} + \psi(\alpha+\beta)\right] + \sum_{i=1}^{n}\left[\frac{1}{(\beta+x_i)(x_i+2)+\alpha+2}\right. \\ &- \left.\psi(\alpha+\beta+x_i+3)\right] = 0, \end{aligned} \tag{17}$$

$$\begin{aligned} \frac{\partial\ell}{\partial\beta} &= n\left[\psi(\alpha+\beta) - \psi(\beta)\right] + \sum_{i=1}^{n}\left[\psi(\beta+x_i)\right. \\ &+ \left.\frac{x_i+2}{(\beta+x_i)(x_i+2)+\alpha+2} - \psi(\alpha+\beta+x_i+3)\right] = 0, \end{aligned} \tag{18}$$

where $\psi(\omega) = d/d\omega\,\log\Gamma(\omega)$ is the digamma function. It is a straightforward matter to obtain the parameters, since most software packages such as Mathematica provide very fast routines for evaluating the digamma function.

Maximum likelihood estimators (MLEs) can be obtained by solving the system of equations (17) and (18) in α and β.

The second partial derivatives are given by

$$\begin{aligned} \frac{\partial^2\ell}{\partial\alpha^2} &= n\left[-\frac{1}{\alpha^2} + \psi'_\alpha(1+\alpha) + \psi'_\alpha(\alpha+\beta)\right] \\ &- \sum_{i=1}^{n}\left[\frac{1}{((\beta+x_i)(x_i+2)+\alpha+2)^2} + \psi'_\alpha(\alpha+\beta+x_i+3)\right], \\ \frac{\partial^2\ell}{\partial\alpha\,\partial\beta} &= n\psi'_\beta(\alpha+\beta) \\ &- \sum_{i=1}^{n}\left[\frac{x_i+2}{((\beta+x_i)(x_i+2)+\alpha+2)^2} + \psi'_\beta(\alpha+\beta+x_i+3)\right], \end{aligned}$$

$$\frac{\partial^2 \ell}{\partial \beta^2} = n\left[\psi'_\beta(\alpha+\beta) - \psi'_\beta(\beta)\right] + \sum_{i=1}^{n} \left\{ \psi'_\beta(\beta + x_i) \right.$$
$$\left. - \frac{x_i+2}{((\beta+x_i)(x_i+2)+\alpha+2)^2} - \psi'_\beta(\alpha+\beta+x_i+3) \right\},$$

where $\psi'_\alpha(\cdot)$ and $\psi'_\beta(\cdot)$ represent the derivative of the digamma function with respect to α and β respectively, and this, too, can be computed using Mathematica.

5 Applications

To see how the proposed model works for fitting automobile claim frequency data, we chose a data set which is well known in studies of automobile insurance claim calculations. These data concern the number of automobile liability policies in Belgium during the years 1975–76 and are shown in Table 1. The empirical mean and variance are $\bar{x} = 0.10$ and $s^2 = 0.15$ respectively, showing the overdispersion phenomena. Therefore, the $\mathscr{PLBP}$ distribution seems suitable for fitting them. Note also that there is a high proportion of zero values in the observed data.

Using equations (17) and (18), we estimated the parameters of the model by the maximum likelihood method; the expected frequencies are shown in the third column of Table 1. This Table also shows the parameters estimated, the maximum value of the log–likelihood function, and the Pearson goodness–of–fit chi-squared value, among others. In parentheses we show the standard error computed from

$$E\left(-\frac{\partial^2 \ell}{\partial \alpha^2}\right) \approx -\left.\frac{\partial^2 \ell}{\partial \alpha^2}\right|_{\widehat{\alpha}},$$
$$E\left(-\frac{\partial^2 \ell}{\partial \beta^2}\right) \approx -\left.\frac{\partial^2 \ell}{\partial \beta^2}\right|_{\widehat{\beta}},$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ are the maximum likelihood estimators of α and β, respectively.

As it can be observed the maximum of the log–likelihood function is similar to the one obtained with the negative binomial and the Poisson–inverse Gaussian distribution (see Willmot, 1987). Note that, although mathematically straightforward, it is mathematically difficult to work with the Poisson–inverse Gaussian distribution since its pmf depends on the modified Bessel

Table 1: Fit of automobile claim data. Willmot (1987)

Number of claims	Observed	Fitted
0	3719	3718.54
1	232	234.26
2	38	35.50
3	7	8.05
4	3	2.32
5	1	0.80
Total	4000	4000
$\widehat{\alpha} = 10.103\,(2.02)$; $\widehat{\beta} = 0.682\,(0.15)$		
Maximum log–likelihood	–1183.56	
Pearson goodness–of–fit		
Chi–squared	0.203	
Degrees of freedom	1	
p–value	0.651	

function of the third kind. Furthermore, by fitting the Poisson–Lindely distribution we have obtained that the maximum of the log–likelihood function is –1207.65 and the Chi–squared value is 63.29 with two degree of freedom. Therefore, a considerable fit improvement is achieved with the new discrete distribution proposed here as compared with the Poisson–Lindley distribution.

We then used the parameters estimate to obtain the cumulative probability of the right tail of the aggregate claim size distribution, using expressions (14) and (15) for different values of the γ parameter. The value of the ν parameter of the Poisson distribution using the maximum likelihood method was found to be 0.0865 (Table 2).

The fact that the tails of the pdf given in (14) are heavier than those of the exponential distribution, given in expression (15), is confirmed by the numerical values shown in this Table, at least for $y > 1$.

6 Final comments

This paper proposes a new two–parameter discrete distribution with an infinite and non–negative integer range as a possible alternative to the negative binomial, the strict arcsine and the Poisson–inverse Gaussian distributions, which have been previously considered in actuarial literature. The new distribution is unimodal (an explicit solution for this is provided), or has two consecutive integer modal values, and it is overdispersed. For the actuarial data considered in this paper, the distribution provides a similar fit to the one obtained when the negative binomial, the strict arcsine and the Poisson–inverse Gaussian distributions are used. Finally, a closed expression for the pdf of the aggregate claim size distribution is obtained when the new distribution is considered as the primary distribution and the exponential as the secondary distribution, as part of collective risk theory.

References

Antzoulakos, D., Chadjiconstantinidis, S. (2004). On Mixed and Compound Mixed Poisson Distributions. *Scandinavian Actuarial Journal*, 3, 161–188.

Boucher, J.P., Denuit, M. and Guillén, M. (2007). Risk classification for claim counts: a comparative analysis of various zero-inflated mixed Poisson and hurdle models. *North American Actuarial Journal*, 11, 4, 110–131.

Boucher, J.P., Denuit, M. and Guillén, M. (2009) Number of accidents or number of claims? An approach with zero–inflated Poisson models data. *Journal of Risk and Insurance*, 76, 4, 821–846.

Ghitany, M.E., Al–Mutairi, D.K. and Nadarajah, S. (2008). Zero–truncated Poisson-Lindley distribution and its application. *Mathematics and Computers in Simulation*, 79, 3, 279–287.

Ghitany, M.E. and Al–Mutairi, D.K. (2009). Estimation methods for the discrete Poisson-Lindley distribution. *Journal of Statistical Computation and Simulation*, 79, 1, 1–9.

Gómez–Déniz, E., Sarabia, J.M., Vázquez, F. and Pérez, J.M. (2008). Using a Bayesian Hierarchical Model for Fitting Automobile Claim Frequency Data. *Communications in Statistics–Theory and Methods*, 37, 1425–1435.

Gómez–Déniz, E., Calderín, E. (2011). The discrete Lindley distribution: properties and applications. *Journal of Statistical Computation and Simulation*, 81, 11, 1405–1416.

Gómez–Déniz, E., Sarabia, J.M. and Balakrishnan, N. (2012). A multivariate discrete Poisson–Lindley distribution: extensions and actuarial applications. *Astin Bulletin*, 42, 2, 655–678.

Goovaerts, M.J., De Pril, N. (1980). Survival probabilities based on Pareto claim distributions. *Astin Bulletin*, 11, 154–157.

Gradshteyn, I.S., Ryzhik, I.M. (1994). *Table of Integrals, Series, and Products.* Alan Jeffrey, Editor. Fifth Edition. Academic Press, Boston.

Hernández, A., Gómez–Déniz, E., Pérez, J.M. (2009). Bayesian robustness of the compound Poisson distribution under bidimensional prior: an application to the collective risk model. *Journal of Applied Statistics*, 36, 8, 853–869.

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

Karlis, D., Xekalaki, E. (2005). Mixed Poisson distributions. *International Statistical Review*, 73, 35–58.

Klugman, S.A., Panjer, H.H. and Willmot, G.E. (2008). Loss Models: From Data to Decisions, 3rd Ed., Wiley, New York.

Kokonendji, C. and Khoudar, M. (2004). On strict arcsine distribution. *Communications in Statistics–Theory and Methods*, **33**, 5, 993–1006.

Lindley, D.V. (1958). Fiducial Distributions and Bayes's Theorem. *Journal of the Royal Statistical Society. Series B*, **20**, 1, 102–107.

Mahmoudi, E. and Zakerzadeh, H. (2010). Generalized Poisson–Lindley distribution. *Communications in Statistics–Theory and Methods*, 39, 1785–1798.

Meng, S., Yuan, W., Whitmore, G.A. (1999). Accounting for individual over–dispersion in a bonus–malus automobile insurance system. *Astin Bulletin*, **29**, 2, 327–337.

Panjer, H. (1981). Recursive evaluation of a family of compound distributions. *Astin Bulletin*, 12, 1 22–26.

Rolski, T., Schmidli, H. Schmidt, V. and Teugel, J. (1999). *Stochastic processes for insurance and finance*. John Wiley & Sons.

Sankaran, M. (1970). The Discrete Poisson–Lindley Distribution. *Biometrics*, 26, 1, 145–149.

Sundt, B., Vernic, R. (2009). *Recursions for Convolutions and Compound Distributions with Insurance Applications*. Springer–Verlag, New York.

Willmot, G.E. (1986). Mixed compound Poisson distributions. *Astin Bulletin*, 16, 59–79

Willmot, G.E. (1987). The Poisson–inverse Gaussian distribution as an alternative to the negative binomial. *Scandinavian Actuarial Journal*, 113–127.

Willmot, G.E. (1993). On recursive evaluation of mixed Poisson probabilities and related quantities. *Scandinavian Actuarial Journal*, 114–133.

Winkelmann, I. (2000). *Econometric Analysis of Count Data*. 3rd Edition. Springer–Verlag: Berlin.

Table 2: Cumulative probability of the right tail of the aggregate claim size distribution, i.e. $\Pr(Y \geq y)$, for different values of γ

y	$\gamma = 0.10$		$\gamma = 0.50$		$\gamma = 0.75$	
	$\mathscr{CP}$	$\mathscr{CPLBP}$	$\mathscr{CP}$	$\mathscr{CPLBP}$	$\mathscr{CP}$	$\mathscr{CPLBP}$
1	0.075298	0.064752	0.051338	0.046493	0.040407	0.037842
2	0.068423	0.059593	0.031802	0.030828	0.019697	0.020519
3	0.062175	0.054853	0.019697	0.020519	0.009598	0.011225
4	0.056498	0.050497	0.012198	0.013711	0.004676	0.006201
5	0.051338	0.046493	0.007553	0.009200	0.002277	0.003461
6	0.046650	0.042813	0.004676	0.006201	0.001108	0.001953
7	0.042389	0.039429	0.002894	0.004198	0.000539	0.001115
8	0.038517	0.036319	0.001791	0.002856	0.000262	0.000644
9	0.034999	0.033459	0.001108	0.001953	0.000127	0.000377
10	0.031802	0.030828	0.000686	0.001342	0.000062	0.000224

y	$\gamma = 1.0$		$\gamma = 1.5$		$\gamma = 2.0$	
	$\mathscr{CP}$	$\mathscr{CPLBP}$	$\mathscr{CP}$	$\mathscr{CPLBP}$	$\mathscr{CP}$	$\mathscr{CPLBP}$
1	0.031802	0.030828	0.019697	0.020519	0.012198	0.013711
2	0.012198	0.013711	0.004676	0.006201	0.001791	0.002856
3	0.004676	0.006201	0.001108	0.001953	0.000262	0.000644
4	0.001791	0.002856	0.000262	0.000644	0.000038	0.000159
5	0.000686	0.001342	0.000062	0.000224	0.000000	0.000043