Appl. Statist. (2005)
54, *Part* 1, *pp.* 127–142

A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution

Galit Shmueli,

University of Maryland, College Park, USA

Thomas P. Minka and Joseph B. Kadane,

Carnegie Mellon University, Pittsburgh, USA

Sharad Borle

Rice University, Houston, USA

and Peter Boatwright

Carnegie Mellon University, Pittsburgh, USA

[Received June 2003. Revised December 2003]

Summary. A useful discrete distribution (the Conway–Maxwell–Poisson distribution) is revived and its statistical and probabilistic properties are introduced and explored. This distribution is a two-parameter extension of the Poisson distribution that generalizes some well-known discrete distributions (Poisson, Bernoulli and geometric). It also leads to the generalization of distributions derived from these discrete distributions (i.e. the binomial and negative binomial distributions). We describe three methods for estimating the parameters of the Conway–Maxwell–Poisson distribution. The first is a fast simple weighted least squares method, which leads to estimates that are sufficiently accurate for practical purposes. The second method, using maximum likelihood, can be used to refine the initial estimates. This method requires iterations and is more computationally intensive. The third estimation method is Bayesian. Using the conjugate prior, the posterior density of the parameters of the Conway–Maxwell–Poisson distribution is easily computed. It is a flexible distribution that can account for overdispersion or underdispersion that is commonly encountered in count data. We also explore two sets of real world data demonstrating the flexibility and elegance of the Conway–Maxwell–Poisson distribution in fitting count data which do not seem to follow the Poisson distribution.

Keywords: Conjugate family; Conway–Maxwell–Poisson distribution; Estimation; Exponential family; Overdispersion; Underdispersion

1. Introduction and motivation

The Poisson distribution is one of the most well-utilized discrete distributions, since data in multiple research fields often fulfil the Poisson postulates. However, its reliance on a single parameter limits its flexibility in many applications. Overdispersion of data relative to the Poisson distribution is a frequent issue (Breslow, 1990; Dean, 1992). One solution to this limitation is to allow the Poisson parameter to be a random variable with some probability distribution (Maceda, 1948; Satterthwaite, 1942), which leads to a hierarchy of distributions. A commonly

Address for correspondence: Galit Shmueli, Department of Decision and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, USA.
E-mail: gshmueli@rhsmith.umd.edu

 0035–9254/05/54127

used hierarchy utilizes a gamma mixing distribution for the Poisson parameter, leading to a negative binomial distribution for the observed data (Manton *et al.*, 1981; Chatfield *et al.*, 1966). Yet in other applications, such as the lengths of words in texts and dictionaries (Wimmer *et al.*, 1994), data are underdispersed relative to the Poisson distribution. Stated differently, many applications have an empirical distribution with thicker or thinner tails than those of the Poisson distribution, where the change in ratios of successive probabilities $P(X=x-1)/P(X=x)$ is non-linear in x.

To illustrate these issues we introduce two sets of real world data that do not fit a Poisson model. The first data set (which is available at `http://www.stat.cmu.edu/COM-Poisson/Sales-data.html`) consists of quarterly sales of a well-known brand of a particular article of clothing at stores of a large national retailer. Even at the quarterly level, sales of a particular clothing brand, style, colour and size can be meager. In many quarters, no units are moved at all. An example of a clothing article at this level of specificity is Mens Gold Toe 'Windsor Wool Midcalf' Black Socks (which come in only one size). Our sales data are for a clothing article other than socks, but it is of the same level of specificity. Retailers use this type of sales data to plan their inventory or order quantities. The average sales are 3.56 per quarter and the sample variance is 11.31, which indicates overdispersion relative to a Poisson distribution.

The second data set contains lengths of words (numbers of syllables) in a Hungarian dictionary (Wimmer *et al.*, 1994). This and other word length examples share the characteristic of underdispersion. In this data set the average number of syllables is 3.30 whereas the sample variance is 1.22, indicating underdispersion relative to a Poisson fit.

These data sets are just two examples in an ocean of non-Poisson data, thus demonstrating the real need for a more flexible alternative.

In this paper we consider a generalization of the Poisson distribution to a two-parameter distribution, which was introduced by Conway and Maxwell (1962) in the context of queuing systems: the Conway–Maxwell–Poisson (CMP) distribution. Although Conway and Maxwell suggested this extension to the Poisson distribution as a solution to handling queuing systems with state-dependent service rates, there appear to have been no probabilistic or statistical characterizations of this distribution, and extremely few applications appear in the literature.

The CMP distribution consists of an extra parameter, which we denote by ν, and which governs the rate of decay of successive ratios of probabilities such that $P(X=x-1)/P(X=x)=x^{\nu}/\lambda$. We found such a generalization to be particularly useful in a marketing application (Boatwright *et al.*, 2003) where the empirical distribution had a long and significant tail, thus necessitating the added flexibility that is brought about by the parameter ν, a flexibility that could not be achieved without fairly complex alternative extensions of the Poisson distribution.

The CMP distribution is appealing from a theoretical point of view since it belongs to the exponential family as well as to the two-parameter power series of distributions family. As such, it allows for sufficient statistics and other properties to be elegantly derived. This is especially useful from a Bayesian perspective. The CMP distribution's parsimonious nature, i.e. its flexibility at the low cost of a single additional parameter, positions its performance favourably among more complicated or more limited distributions that have been suggested in the literature. These include the negative binomial distribution, the generalized Poisson–Pascal distribution (Brockett *et al.*, 1996) and Consul's generalized Poisson distribution (Consul, 1989). Furthermore, the CMP distribution's structure allows for a variety of generalizations such as zero-inflated data and dependence. Its appeal from a practical point of view is even stronger: it is easy to use, flexible for fitting overdispersed and underdispersed data, and performs well in many settings.

The paper is laid out as follows. The CMP distribution with its properties and some generalizations are portrayed in Section 2. Section 3 describes three methods for estimating its

parameters. Section 4 applies the techniques from Section 3 to the two data sets and illustrates how the CMP distribution is fitted, and a discussion in Section 5 concludes the paper. The data that are analysed in the paper can be obtained from

`http://www.blackwellpublishing.com/rss`

2. The Conway–Maxwell–Poisson distribution and its statistical properties

We start by reintroducing the CMP distribution that Conway and Maxwell (1962) suggested and then tackle the important task of characterizing the distribution from a probabilistic and statistical point of view.

2.1. The Conway–Maxwell–Poisson probability function

The CMP distribution generalizes the Poisson distribution, allowing for overdispersion or underdispersion. Its probability function is

$$P(X=x)=\frac{\lambda^x}{(x!)^\nu}\frac{1}{Z(\lambda,\nu)} \qquad x=0,1,2,\ldots, \tag{1}$$

$$Z(\lambda,\nu)=\sum_{j=0}^{\infty}\frac{\lambda^j}{(j!)^\nu} \tag{2}$$

for $\lambda>0$ and $\nu\geqslant 0$ (but see the exception below). This satisfies the conditions for a probability function. This formulation allows for a non-linear decrease in ratios of successive probabilities in the form

$$\frac{P(X=x-1)}{P(X=x)}=\frac{x^\nu}{\lambda}. \tag{3}$$

It can be seen that the series $\lambda^j/(j!)^\nu$ converges for any $\lambda>0$ and $\nu>0$, since the ratio of two subsequent terms of the series λ/j^ν tends to 0 as $j\to\infty$.

The CMP distribution is a generalization of some well-known discrete distributions. When $\nu=1$ (and thus $Z(\lambda,\nu)=\exp(\lambda)$) an ordinary Poisson($\lambda$) distribution results. As $\nu\to\infty$, $Z(\lambda,\nu)\to 1+\lambda$, and the CMP distribution approaches a Bernoulli distribution with $P(X=1)=\lambda/(1+\lambda)$. When $\nu=0$ and $\lambda<1$, $Z(\lambda,\nu)$ is a geometric sum,

$$Z(\lambda,\nu)=\sum_{j=0}^{\infty}\lambda^j=\frac{1}{1-\lambda}, \tag{4}$$

and the distribution itself is geometric:

$$P(X=x)=\lambda^x(1-\lambda) \qquad \text{for } x=0,1,2,\ldots. \tag{5}$$

When $\nu=0$ and $\lambda\geqslant 1$, $Z(\lambda,\nu)$ does not converge, and hence the distribution is undefined.

The CMP distribution can thus be thought of as a continuous bridge between the geometric (that counts failures until the first success) ($\nu=0$), the Poisson ($\nu=1$) and the Bernoulli ($\nu=\infty$) distributions. Values of ν that are less than 1 correspond to flatter successive ratios than the Poisson distribution's and hence to longer tails or overdispersion.

2.2. Moments of the distribution

The CMP distribution belongs to the family of two-parameter power series distributions (Johnson *et al.*, 1992). Moments of this distribution can then be expressed by using the recursive

formula

$$E(X^{r+1}) = \begin{cases} \lambda\, E(X+1)^{1-\nu} & r=0, \\ \lambda \frac{\mathrm{d}}{\mathrm{d}\lambda} E(X^r) + E(X)\, E(X^r) & r>0. \end{cases} \tag{6}$$

Using an asymptotic approximation for $Z(\lambda,\nu)$ (see equation (41) in Appendix B) $E(X)$ can be closely approximated by

$$E(X) = \lambda \frac{\mathrm{d}[\log\{Z(\lambda,\nu)\}]}{\mathrm{d}\lambda} \approx \lambda^{1/\nu} - \frac{\nu-1}{2\nu}. \tag{7}$$

The approximation is especially good for $\nu \leqslant 1$ or $\lambda > 10^{\nu}$.

2.3. Sufficient statistics

The likelihood for a set of n independent and identically distributed observations $x_1, x_2, \ldots, x_n$ is

$$\begin{aligned} L(x_1, x_2, \ldots, x_n \mid \lambda, \nu) &= \frac{\prod_{i=1}^{n} \lambda^{x_i}}{\left(\prod_{i=1}^{n} x_i!\right)^{\nu}} Z^{-n}(\lambda,\nu) \\ &= \lambda^{\sum_{i=1}^{n} x_i} \exp\left\{-\nu \sum_{i=1}^{n} \log(x_i!)\right\} Z^{-n}(\lambda,\nu) \\ &= \lambda^{S_1} \exp(-\nu S_2) Z^{-n}(\lambda,\nu) \end{aligned} \tag{8}$$

where $S_1 = \sum_{i=1}^{n} x_i$ and $S_2 = \sum_{i=1}^{n} \log(x_i!)$. By the factorization theorem, (S_1, S_2) are sufficient statistics for $x_1, x_2, \ldots, x_n$. Furthermore, equation (8) displays the CMP distribution as a member of the exponential family.

2.4. Extensions

From the CMP distribution a variety of derived distributions can be obtained which generalize classical ones and allow even more flexibility. We describe just a few such extensions in this section to show the richness and versatility of the CMP distribution.

2.4.1. *Sums of Conway–Maxwell–Poisson variables*

The sum of n independent CMP random variables forms a continuous bridge between three well-known distributions:

(a) for $\nu = 0$ and $\lambda < 1$, the sum of CMP variables reduces to the sum of geometric variables, which follows a negative binomial distribution with parameters n and $1-\lambda$;
(b) for $\nu = 1$ the sum has a Poisson distribution with parameter $n\lambda$;
(c) for $\nu = \infty$ the distribution of the sum is binomial with parameters n and $\lambda/(1+\lambda)$.

2.4.2. *The Conway–Maxwell-Poisson–binomial distribution*

A CMP–binomial distribution that can represent overdispersion and underdispersion relative to the ordinary binomial distribution can be defined by a CMP distribution conditional on the sum of that CMP distribution with another independent CMP distribution. Consider the sum of two independent CMP variables with parameters λ_1 and λ_2:

$$S = X + Y, \tag{9}$$

$$P(S=s)=\sum_{x=0}^{s} P(X=x)\ P(Y=s-x) \tag{10}$$

$$=\sum_{x=0}^{s} \frac{\lambda_x^x}{(x!)^\nu\ Z(\lambda_x,\nu)} \frac{\lambda_y^{s-x}}{\{(s-x)!\}^\nu\ Z(\lambda_y,\nu)} \tag{11}$$

$$=\frac{(\lambda_x+\lambda_y)^y}{(y!)^\nu\ Z(\lambda_x,\nu)\ Z(\lambda_y,\nu)} \sum_{x_1=0}^{s} \binom{s}{x}^\nu \left(\frac{\lambda_x}{\lambda_x+\lambda_y}\right)^x \left(\frac{\lambda_y}{\lambda_x+\lambda_y}\right)^{s-x}. \tag{12}$$

The distribution of X conditional on the sum is

$$P(X=x|S=s)\propto \binom{s}{x}^\nu \left(\frac{\lambda_x}{\lambda_x+\lambda_y}\right)^x \left(\frac{\lambda_y}{\lambda_x+\lambda_y}\right)^{s-x} \tag{13}$$

which is a CMP–binomial distribution with parameter $p=\lambda_1/(\lambda_1+\lambda_2)$. A CMP–multinomial distribution over several variables can be similarly defined. The binomial coefficient $\binom{s}{x}$ favours $x=s/2$, so $\nu>1$ gives the distribution a smaller variance and $\nu<1$ a larger variance than an ordinary binomial distribution with the same mean has. For $\nu=0$, the most likely count is extreme: 0 or s. In the other direction, as $\nu\to\infty$, the count is always $s/2$ for even s and $(s\pm1)/2$ for odd s.

The 1–binomial distribution can be interpreted as a sum of independent Bernoulli variables. The CMP–binomial distribution can be interpreted as a sum of non-independent Bernoulli variables Z_i with joint distribution

$$p(Z_1=z_1,\ldots,Z_s=z_s)\propto \binom{s}{x}^{\nu-1} p^x(1-p)^{s-x} \qquad \left(x=\sum_{i=1}^{s} z_i\right). \tag{14}$$

For $\nu>1$ the zs are negatively correlated and for $\nu<1$ the zs are positively correlated.

2.4.3. Mixtures (zero-inflated and zero-deflated data)

The term 'zero inflated' is used in the literature to describe data with a large frequency of 0s (e.g. Lambert (1992) and Brockett *et al.* (1996)). This usually occurs because of a mixture or contamination of the modelled process with a different process. Although the term is widely used, it is misleading in the sense that it is model specific. By definition, an 'inflation' or 'deflation' of 0s is relative to what would have been expected under a given model with given parameters.

In some cases three-parameter distributions such as the generalized Poisson–Pascal distribution were used to model data that arise from mixtures. Although such models can sometimes fit the data, they do not model the mixture directly. To model them directly a fourth parameter is required (Brockett *et al.*, 1996).

To fit a CMP distribution to data that come from a mixture of a CMP process with another process (δ_0), a third parameter is required. This parameter p is used to construct an explicit mixture model of the form

$$p\delta_0+(1-p)Y.$$

Two options for Y are as follows.

(a) For cases where *all* zero counts come from a different process, a shifted CMP distribution of the form $Y_{\text{shift}}\sim \text{CMP}(\lambda,\nu)+1$ can be used. In practice this means that the CMP distribution is fitted to counts of 1 and upwards.
(b) Alternatively, Y is a conditional CMP variable, conditional on its being positive.

In both cases the maximum likelihood estimate for p is the proportion of 0s in the data. It is interesting that the conditional model reduces, under the assumption that the data are zero inflated, to a mixture where some of the 0s in the data are inherent to the process modelled and others are assumed to come from a different process.

3. Estimating the Conway–Maxwell–Poisson parameters

We introduce three methods for estimating the CMP parameters. The first combines a simple graphical technique with a computationally simple and cheap least squares method. The second method offers more refined estimates at a higher computational cost. The third method, which is Bayesian, is very accurate and computationally simple (the estimates can be calculated by hand!) provided that the user has some prior knowledge on the distribution of λ and ν. The use of each of the three methods applied to real world data is illustrated in Section 4.

3.1. Quick and crude method: Conway–Maxwell–Poissonness plot and weighted least squares

The following method is simple and computationally efficient and enables us to determine whether the CMP distribution is an adequate model for some data set and, if so, to estimate the parameters ν and λ in a simple way.

The method is based on the relationship between successive CMP probabilities, given in equation (3). Taking logarithms of both sides of the equation, we obtain a linear relationship between the log-ratios and $\log(x)$:

$$\log(p_{x-1}/p_x) = -\log(\lambda) + \nu \log(x), \tag{15}$$

where p_x denotes $P(X = x)$. The ratio on the left-hand side can be estimated by replacing the probabilities p_{x-1} and p_x with the relative frequencies of $x-1$ and x respectively. The first step is to plot these values against $\log(x)$, for all the ratios that do not include zero counts. A CMP distribution would be adequate if the points on the *Conway–Maxwell–Poissonness plot* fall on a straight line. In the simple Poisson case, the intercept of this line would be 0. This means that the Conway–Maxwell–Poissonness plot does not reduce to the well-known Poissonness plot (Hoaglin, 1980). Although the ideas behind the two plots are similar, in the Conway–Maxwell–Poissonness plot λ determines only the intercept, whereas in Hoaglin's Poissonness plot it determines both the intercept and the slope. The Conway–Maxwell–Poissonness plot is more similar to the Ord plot (Ord, 1967), which also looks at ratios of successive probabilities.

If the data do appear to fit a CMP model, then the parameters can be estimated by fitting a regression of $\log(\hat{p}_{x-1}/\hat{p}_x)$ on $\log(x)$. However, two basic assumptions of ordinary regression models are violated here. First, the variance of the dependent variable is not constant. This variance is approximately

$$\operatorname{var}\left\{\log\left(\frac{\hat{p}_{x-1}}{\hat{p}_x}\right)\right\} \approx \frac{1}{n p_x} + \frac{1}{n p_{x-1}}, \tag{16}$$

where n is the number of non-zero values in the data (see appendix A).

The second deviation from ordinary regression assumptions is the fact that the 'observations' are not independent. In fact, every two successive observations are negatively correlated:

$$\operatorname{cov}\left\{\log\left(\frac{\hat{p}_{x-1}}{\hat{p}_x}\right), \log\left(\frac{\hat{p}_x}{\hat{p}_{x+1}}\right)\right\} \approx -\frac{1}{n p_x} \tag{17}$$

(see appendix A for details).

To take into account both the heteroscedasticity and the first-order dependence, a weighted least square regression can be used. The inverse weight matrix would then have the variances on the diagonal, and the one-step covariances on the first off-diagonal. This method performs quite well, in the absence of too many low values with zero counts. Unlike the weighted least squares approach that was suggested for the Ord plot (Friendly, 1995), the weights here are not chosen heuristically. These weighted least squares estimates can then be refined by applying the maximum likelihood method that is described next.

3.2. Accurate and intensive method: maximum likelihood

The CMP likelihood function is given by equation (8). Since it is a member of the exponential family its likelihood function can be expressed in the form

$$L(X \mid \boldsymbol{\theta}) = \gamma(\boldsymbol{\theta})\, \phi(x) \exp\left\{ \sum_{i=1}^{k} \pi_i(\theta)\, t_i(x) \right\} \tag{18}$$

where $\Pi(\boldsymbol{\theta}) = (\pi_1(\theta), \ldots, \pi_k(\theta))$ is the natural parameter and $T(X) = (t_1(x), \ldots, t_k(x))$ is the natural sufficient statistic. For the CMP distribution these are

$$\Pi(\lambda, \nu) = [\log(\lambda), -\nu], \tag{19}$$

$$T(X) = \left[\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} \log(x_i!) \right]. \tag{20}$$

The values of λ and ν at the point of maximum satisfy

$$E(X) = \lambda \bar{X}, \tag{21}$$

$$E\{\log(X!)\} = \overline{\log(X!)}. \tag{22}$$

Since these equations cannot be solved analytically, an iterative method such as the Newton–Raphson method can be used (see Gelman *et al.* (1995), pages 272–273). In each iteration the expectations, variances and covariance of X and $\log(X!)$ are computed (or approximated) by plugging the estimates for λ and ν from the previous iteration into the expression

$$E\{f(X)\} = \sum_{j=0}^{\infty} f(j) \frac{\lambda^j}{(j!)^\nu\, Z(\lambda, \nu)}. \tag{23}$$

In most cases the above sums must be truncated to reach numerical values (except for $\nu = 1$, where $E(X) = \lambda$ and $E(X^2) = \lambda(1 + \lambda)$, and for $\nu = 0$ and $\nu = \infty$). Note that each of these expectations is a ratio of infinite sums, since $Z(\lambda, \nu)$ is an infinite sum itself. Although it is not always possible to calculate these ratios of infinite sums exactly, they can be calculated to any prespecified precision (see Minka *et al.* (2003)).

3.3. Immediate and simple method: Bayesian estimation

Since the CMP distribution is in the exponential family, there is a conjugate family of priors such that, whatever the data, the posterior is of the same form. For this distribution, the conjugate prior is of the form

$$h(\lambda, \nu) = \lambda^{a-1} \exp(-\nu b)\, Z^{-c}(\lambda, \nu)\, \kappa(a, b, c) \tag{24}$$

for $\lambda > 0$ and $\nu \geqslant 0$, where $\kappa(a, b, c)$ is the integration constant. Thus, given a prior on λ and ν and given the data the posterior is of the same form as equation (24), substituting $a' = a + \Sigma_{i=1}^{n} X_i$,

$b' = b + \Sigma_{i=1}^{n} \log(X_i!)$ and $c' = c + n$ for a, b and c respectively. See Kadane *et al.* (2003) for further details.

4. Fitting the Conway–Maxwell–Poisson distribution: examples

To illustrate the usefulness and flexibility of the CMP distribution, we fit it to the two sets of real world data that we described in Section 1. The first describes quarterly sales of a well-known brand of a particular article of clothing, and the second contains lengths of words (numbers of syllables) in a Hungarian dictionary. In both cases the data do not fit a Poisson model, and we compare the fit of several other discrete models.

At first, a Conway–Maxwell–Poissonness plot is constructed (Figs 1 and 2), plotting the logarithms of the ratios of successive frequencies against $\log(x)$, as described in Section 3.1. Ratios involving zero counts are omitted. It can be seen in Figs 1 and 2 that a linear relationship is reasonable for both data sets. However, the data do not seem to follow a Poisson model in either case, as the angle of the Poisson line is much smaller than 45° for the sales data (indicating $\nu < 1$, or overdispersion) and larger than 45° for the lengths of words (indicating $\nu > 1$, or underdispersion). Next, a straight line is fitted to the data by using weighted least squares. The estimates for the sales data are $\hat{\lambda} = 0.97$ and $\hat{\nu} = 0.124$, and for the word length data are $\hat{\lambda} = 7.68$ and $\hat{\nu} = 2.14$. (Following Wimmer *et al.* (1994), the count for 1 was treated as a count for 0, and in general the count for x was treated as $x - 1$, as though the data were generated by adding 1 to a CMP distribution.) From the error bars (± 1 standard error) it can be seen that points on the right, which represent counts at the tail, have a larger variance, and thus less influence on the weighted least squares estimates. In the next step, the weighted least squares estimates are refined by feeding them as initial values into the maximum likelihood procedure (described in Section 3.2). The refined estimates for the sales data are $\hat{\lambda} = 0.97$ and $\hat{\nu} = 0.126$, and for the word length data are $\hat{\lambda} = 7.74$ and $\hat{\nu} = 2.15$.

Finally, to illustrate the third estimation method, we use a prior of $a = 1$, $b = 1$ and $c = 1$, which conveys a relative lack of prior knowledge about λ and ν. The posterior is given by equa-

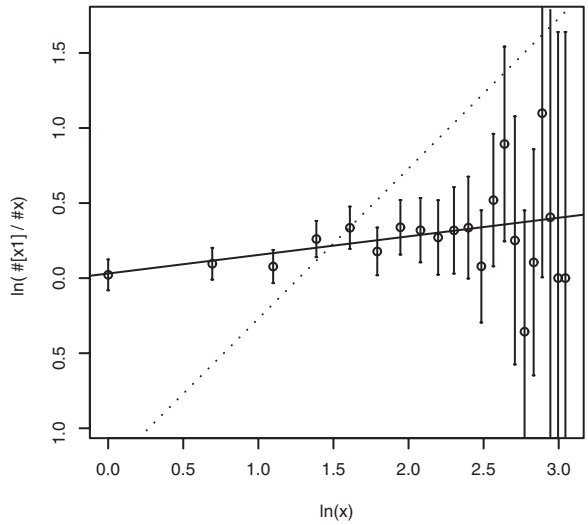

Fig. 1. Conway–Maxwell–Poissonness plot for the quarterly sales data: ○, data; ·····, Poisson(3.56); ———, CMP(0.97, 0.124)

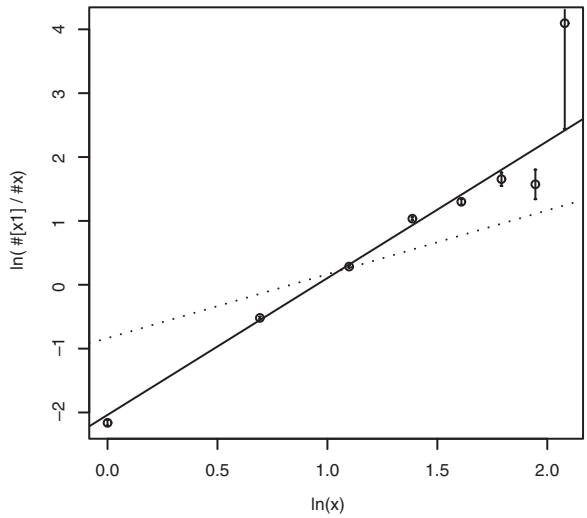

Fig. 2. Conway–Maxwell–Poissonness plot for the word length data: ○, data; ·····, Poisson(2.30); ———, CMP(7.68, 2.14)

tion (24) with $a' = 1 + \Sigma_{i=1}^{n} x_i$, $b' = 1 + \Sigma_{i=1}^{n} \log(x_i!)$ and $c' = 1 + n$, which yield for the sales data $a' = 11\,278$, $b' = 12\,000.82$ and $c' = 3169$, and for the word length data $a' = 189\,873$, $b' = 134\,792.95$ and $c' = 57\,460$.

Fig. 3 contains a contour plot and a three-dimensional plot of the posterior distribution for the sales data. The values accompanying the contours and the third axis of the three-dimensional plot are the values of the integrals of the region inside the contour, i.e. the cumulative posterior probability of the regions plotted.

The maximum posterior estimates for the sales data are $\hat{\lambda} = 0.97$ and $\hat{\nu} = 0.126$. The marginal posterior over ν has essentially no mass near 1.0, once again implying that the ordinary Poisson distribution should not be considered as a reasonable distribution to fit these data. Figs 4 and 5 compare the fit of several distributions with the quarterly sales data. For each distribution we display the fitted counts *versus* the actual counts (Figs 4(a), 4(c), 4(e), 5(a), 5(c) and 5(e)) as well as the deviations between the two (Figs 4(b), 4(d), 4(f), 5(b), 5(d) and 5(f)). The first three fitted distributions show that the Poisson model provides poor sales estimates. The geometric distribution, which might have been considered owing to the proximity of ν to 0, also turns out to be a poor choice, for it greatly overestimates the number of 0s and 1s in the data. The negative binomial distribution does not fit the data as well as the CMP distribution for the first three counts. The generalized Poisson distribution of Consul (1989) is included as another two-parameter distribution, but its fit is even worse than that of the negative binomial distribution. The generalized Poisson–Pascal distribution (Brockett *et al.*, 1996) is a three-parameter distribution which fits as well as the CMP distribution but involves additional complexity. All these distributions were fitted by using maximum likelihood. Although the models differ for the first 10 counts, they seem to perform similarly for larger counts.

For the word length example we do not provide plots, since most of the above distributions are not suitable for fitting underdispersed data, and thus they do not fit the word length data well. The CMP distribution provides an extremely close fit to the data. Even the simple weighted least squares method leads to relatively accurate estimates. These can then be refined by using maximum likelihood.

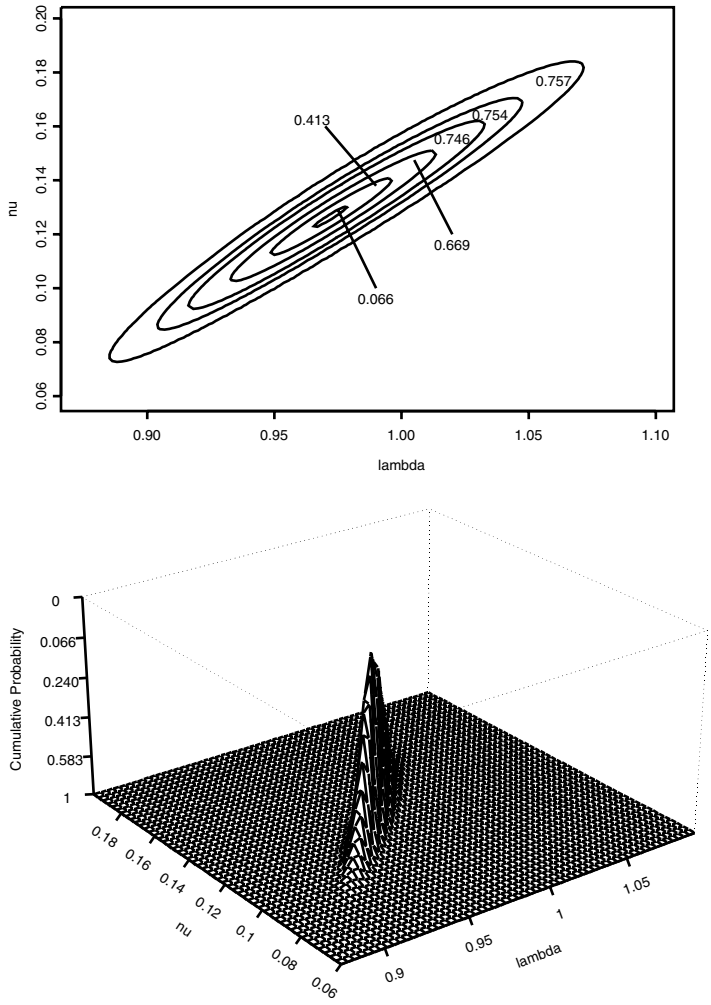

Fig. 3. Posterior of the quarterly sales data

Although in this illustration ν is clearly not 0 or 1, the CMP distribution can also be fitted to data where 1 and/or 0 is a plausible value for ν. Posterior plots such as that in Fig. 3 can serve as a diagnostic to indicate the range of possible distributions that can be fitted to the data.

5. Discussion

The CMP distribution originated in 1962 from a need to model queuing systems with dependent service times. Although it was introduced more than 40 years ago, it has rarely been explored and its statistical and probabilistic probabilities have not been published. From the scarce literature it also seems that the distribution has hardly been applied to real world data. However, we believe that this distribution is too useful to be forgotten.

With its second parameter (ν), the CMP distribution allows for flexibility in the magnitude of decay in the distribution, compared with the Poisson distribution. Furthermore, it enables fitting distributions that do not resemble the well-known discrete distributions yet are on a

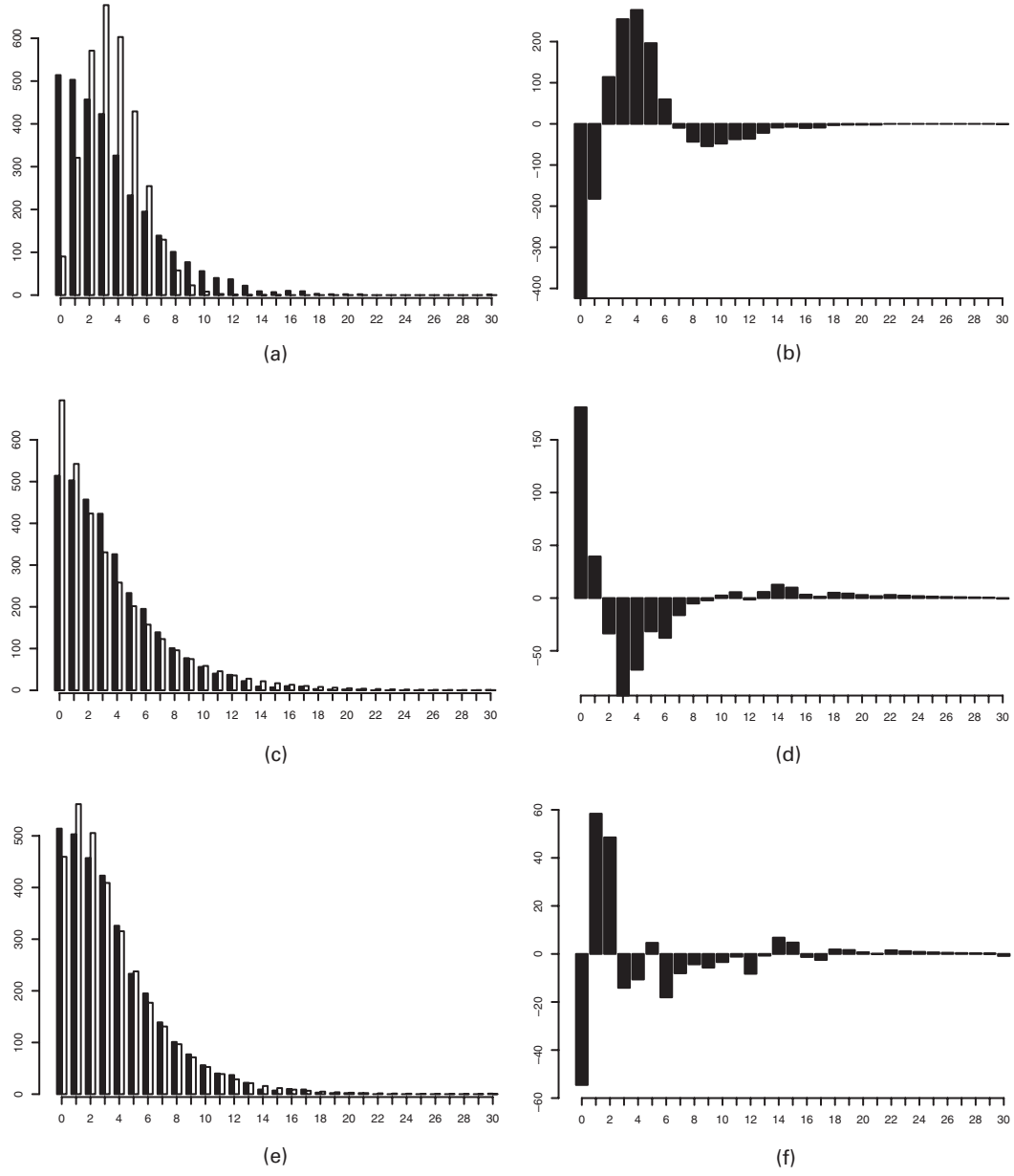

Fig. 4. Comparing observed (▮) *versus* expected (▯) counts for quarterly sales and residuals plots for various distributions: (a) Poisson counts; (b) Poisson residuals; (c) geometric counts; (d) geometric residuals; (e) generalized Poisson counts; (f) generalized Poisson residuals

continuum between a geometric distribution and a Bernoulli distribution. The form of this distribution required us to blend analytical and numerical calculations, a combination that is feasible because of recent advances in computational power.

In practice, not many discrete distributions are used to fit discrete data. The most widely used is the Poisson distribution, whereas for overdispersed data the negative binomial distribution is often used. When none of the known distributions seem appropriate (such as for underdis-

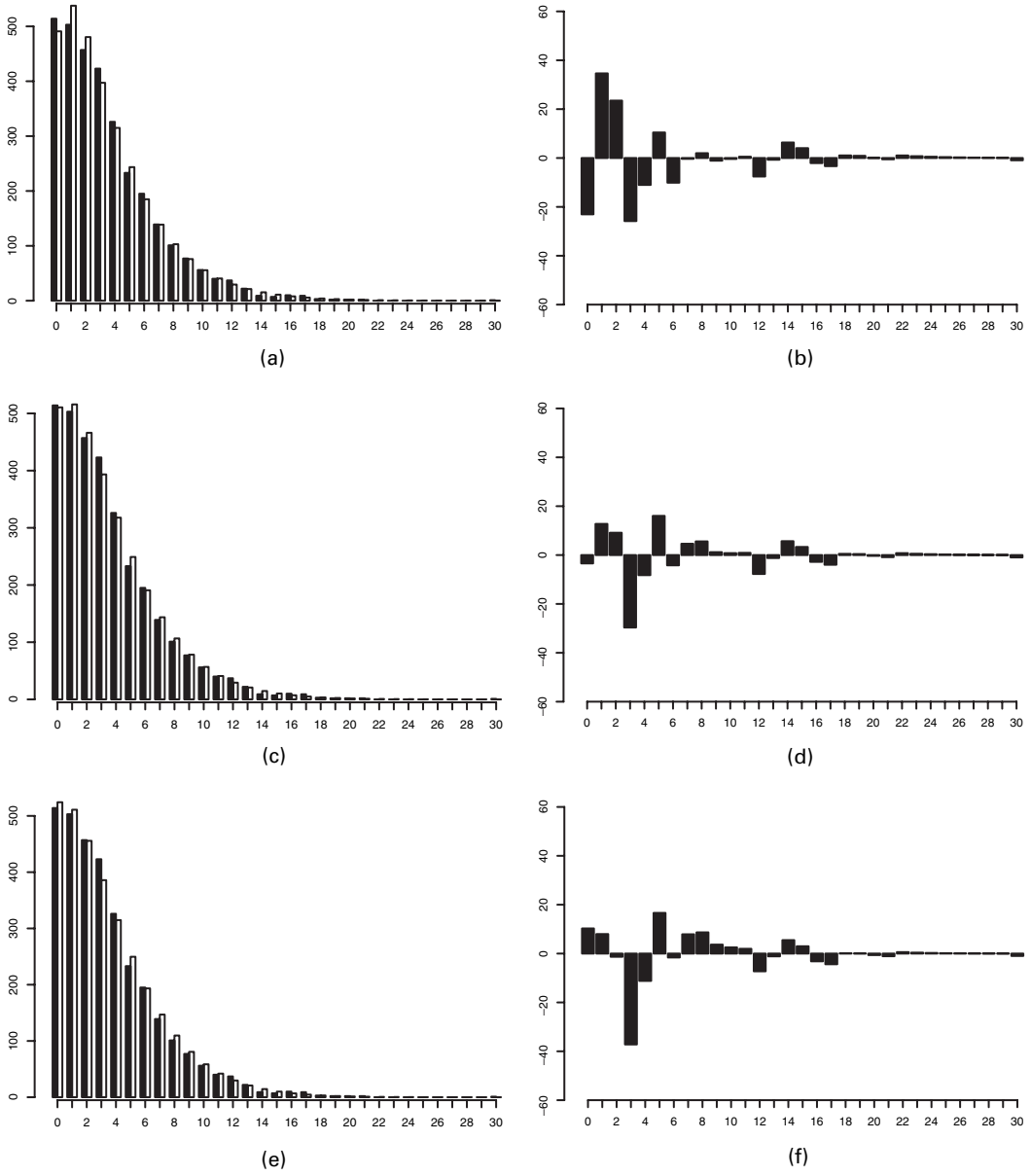

Fig. 5. Comparing observed (▮) *versus* expected (▯) counts for quarterly sales and residuals plots for various distributions: (a) negative binomial counts; (b) negative binomial residuals; (c) generalized Poisson–Pascal counts; (d) generalized Poisson–Pascal residuals; (e) CMP counts; (f) CMP residuals

persion), variations are invented. The CMP distribution, in this sense, increases the variety of discrete distributions that are available for modelling data.

Our data fitting examples in Section 4 illustrate the ability to interpret a fitted CMP model, by comparing the value of ν with the well-known 0, 1 and ∞ values (corresponding to a geometric, Poisson and Bernoulli distribution). Moreover, since these three distributions are nested

within the CMP distribution, the hierarchical structure can be used to make inferences and comparisons. The generalized Poisson distribution that was defined by Consul (1989) also has flexibility in terms of modelling overdispersion relative to the Poisson distribution. It has simple expressions for the normalizing constant and moments. However, it cannot handle underdispersion and is not in the exponential family, which makes analysis more difficult. Numerical studies show that for every CMP distribution with $0.75 < \nu < 1$ (or so) there is a generalized Poisson distribution with very similar form. But for $\nu < 0.75$ the two families differ markedly, especially for the quarterly sales example in Section 4 where $\nu = 0.126$ matches the data well.

Besides its practical usefulness, the CMP distribution has several appealing theoretical properties. As a member of several families of distributions, it inherits some favourable theoretical properties. First, as an exponential family member, the existence of sufficient statistics and a conjugate family of priors is guaranteed. This membership was also used to prove the unimodality of the likelihood. Second, as a generalized power series distribution, expressions for the CMP distribution's moments and tail were derived. A detailed analysis of the CMP conjugate family is given in Kadane *et al.* (2003).

The infinite sum, $Z(\lambda, \nu) = \Sigma_{j=0}^{\infty} \lambda^j/(j!)^\nu$, is a generalization of several well-known infinite sums (such as $\exp(\lambda)$ and $1/(1-\lambda)$). This function plays a major role in different computations that are needed in the CMP context. The CMP distribution relies both on closed form expressions and on numerical approximations for the $Z(\lambda, \nu)$ function. By truncating this sum and bounding the error, the difficulty of computing an infinite sum is overcome for all practical purposes. The numerical approximation does not require specialized software and can be programmed easily in any language. (We used C++ and S-PLUS programs for our application and illustrations.)

Other Conway–Maxwell-type distributions arise from the CMP distribution. We described briefly the CMP–binomial distribution and the distribution of the sum of CMP variables, but a variety of other distributions, processes and generalizations can be derived. In some cases these distributions turn out to be generalizations of other well-known distributions which are useful in various applications where ordinary distributions are inadequate.

Although the revival of an old and almost forgotten distribution is unusual, we believe that our uncovering of its statistical properties sheds light on the usefulness and elegance of the CMP distribution. Our exploration of the CMP distribution uses a modern approach that combines theory and numerical methods. This was possible only because of the advanced computational power that is available today but was not back in the 1960s.

Appendix A: Variances and covariances of successive probability ratios

Formulae (16) and (17) are derived by using the delta method and properties of the multinomial distribution. The vector of counts in the difference bins has a multinomial distribution, and therefore the count in a single bin x is a binomial variable with parameters n and p_x, where n is the number of bins. Therefore

$$\operatorname{var}(\hat{p}_x) = p_x(1-p_x)/n, \tag{25}$$

$$\operatorname{cov}(\hat{p}_x, \hat{p}_y) = -p_x p_y/n, \qquad x \neq y. \tag{26}$$

Using the delta method, the following approximation is derived:

$$\operatorname{var}\left\{\log\left(\frac{\hat{p}_{x-1}}{\hat{p}_x}\right)\right\} = \operatorname{var}\{\log(\hat{p}_{x-1})\} + \operatorname{var}\{\log(\hat{p}_x)\} - 2\operatorname{cov}\{\log(\hat{p}_{x-1})\log(\hat{p}_x)\}$$

$$\approx \frac{\text{var}(\hat{p}_{x-1})}{p_{x-1}^2} + \frac{\text{var}(\hat{p}_x)}{p_x^2} - \frac{2\,\text{cov}(\hat{p}_{x-1}, \hat{p}_x)}{p_{x-1}p_x}$$
$$= \frac{1}{np_{x-1}} + \frac{1}{np_x}. \tag{27}$$

Similarly,

$$\text{cov}\left\{\log\left(\frac{\hat{p}_{x-1}}{\hat{p}_x}\right), \log\left(\frac{\hat{p}_{x+y-1}}{\hat{p}_{x+y}}\right)\right\} = \text{cov}\{\log(\hat{p}_{x-1}), \log(\hat{p}_{x+y-1})\} + \text{cov}\{\log(\hat{p}_x), \log(\hat{p}_{x+y})\}$$
$$- \text{cov}\{\log(\hat{p}_{x-1}), \log(\hat{p}_{x+y})\} - \text{cov}\{\log(\hat{p}_x), \log(\hat{p}_{x+y-1})\}. \tag{28}$$

It can be seen that except for $y=0$ (the one-step covariance) for all $y>1$ the covariance is 0. The one-step covariance can be approximated by

$$\text{cov}\left\{\log\left(\frac{\hat{p}_{x-1}}{\hat{p}_x}\right), \log\left(\frac{\hat{p}_{x+y-1}}{\hat{p}_{x+y}}\right)\right\} = -\frac{1}{n}[1 - \text{var}\{\log(\hat{p}_x)\}] \approx -\frac{1}{n\hat{p}_x}. \tag{29}$$

Appendix B: Bounding and approximating $Z(\lambda, \nu)$

B.1. An upper bound on $(Z\lambda, \nu)$

As shown in Section 2.1, the series $\lambda^j/(j!)^\nu$ converges. In addition, $\lim\{\lambda^j/(j!)^\nu\} = 0$ as $j \to \infty$. Therefore there is a value K such that, for $k > K$,

$$\lambda/k^\nu < 1. \tag{30}$$

Also, note that this ratio is monotonically decreasing, meaning that, for $k > K$, this series converges faster than a geometric series with multiplier given by inequality (30). Thus, $Z(\lambda, \nu)$ can be approximated by truncating the series at some kth term such that inequality (30) holds, i.e.

$$Z(\lambda, \nu) = \sum_{j=0}^{k} \frac{\lambda^j}{(j!)^\nu} + R_k, \tag{31}$$

where $R_k = \Sigma_{j=k+1}^{\infty} \lambda^j/(j!)^\nu$ is the absolute truncation error.

An upper bound can be found, based on the fact that the series $\lambda^j/(j!)^\nu$ $(j = 0, 1, 2, \ldots)$ decreases at a faster rate than a geometric series. Thus, there exists $0 < \varepsilon_k < 1$ for all $k > K$ so that

$$\lambda/(k+1)^\nu < \varepsilon_k. \tag{32}$$

R_k is then bounded by

$$\frac{\lambda^{k+1}}{\{(k+1)!\}^\nu (1-\varepsilon_k)}. \tag{33}$$

Another computational improvement, which increases efficiency, is to bound the *relative* truncation error given by

$$R_k \Big/ \sum_{j=0}^{k} \frac{\lambda^j}{(j!)^\nu}. \tag{34}$$

The relative truncation error can be bounded by

$$\frac{\lambda^{k+1}}{\{(k+1)!\}^\nu (1-\varepsilon_k)} \frac{1}{\sum_{j=0}^{k} \lambda^j/(j!)^\nu}. \tag{35}$$

B.2. Bounding $Z^{-1}(\lambda, \nu)$

Computing the inverse function $Z^{-1}(\lambda, \nu)$ by truncating the infinite sum

$$\hat{Z}^{-1}(\lambda,\nu)=\left\{\sum_{j=0}^{k}\frac{\lambda^j}{(j!)^\nu}\right\}^{-1} \tag{36}$$

leads to a relative error that is

$$\frac{\hat{Z}^{-1}-Z^{-1}}{\hat{Z}^{-1}}=\frac{\sum_{j=k}^{\infty}\lambda^j/(j!)^\nu}{\sum_{j=0}^{\infty}\lambda^j/(j!)^\nu}, \tag{37}$$

which is smaller than expression (35). Thus, the relative error for computing $Z^{-1}(\lambda,\nu)$ by truncation is bounded by the same bound as that for $Z(\lambda,\nu)$.

B.3. An asymptotic approximation of $Z(\lambda,\nu)$

Defining $\mathrm{i}=\sqrt{-1}$, we have the identity

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\exp\{\exp(\mathrm{i}a)\}\exp(-\mathrm{i}aj)\,\mathrm{d}a=\frac{1}{j!} \tag{38}$$

which means that

$$\begin{aligned}\frac{1}{2\pi}\int_{-\pi}^{\pi}&\exp\{\exp(\mathrm{i}a)\}\,Z\{\lambda\exp(-\mathrm{i}a),\nu\}\,\mathrm{d}a\\&=\sum_{j=0}^{\infty}\frac{\lambda^j}{(j!)^\nu}\frac{1}{2\pi}\int_{-\pi}^{\pi}\exp\{\exp(\mathrm{i}a)\}\exp(-\mathrm{i}aj)\,\mathrm{d}a=Z(\lambda,\nu+1).\end{aligned} \tag{39}$$

Thus we can represent $Z(\lambda,\nu)$ for integer $\nu>0$ as a multiple integral:

$$Z(\lambda,\nu)=\frac{1}{(2\pi)^{\nu-1}}\int\cdots\int\exp\left\{\sum_{k=1}^{\nu-1}\exp(\mathrm{i}a_k)+\lambda\exp\left(-\sum_{k=1}^{\nu-1}\mathrm{i}a_k\right)\right\}\mathrm{d}a_1\ldots\mathrm{d}a_{\nu-1}. \tag{40}$$

The behaviour of this integral for large λ can be determined by making the change of variable $\mathrm{i}a_j=\mathrm{i}b_j+(1/\nu)\log(\lambda)$ and applying Laplace's method (Bleistein and Handelsman (1986), section 5.1). The result is

$$Z(\lambda,\nu)=\frac{\exp(\nu\lambda^{1/\nu})}{\lambda^{(\nu-1)/2\nu}(2\pi)^{(\nu-1)/2}\sqrt{\nu}}\{1+O(\lambda^{-1/\nu})\}. \tag{41}$$

Table 1. Percentage errors for the $Z(\lambda,\nu)$ approximation†

λ	*Percentage errors for the following values of ν:*									
	0.1	*0.3*	*0.5*	*0.7*	*0.9*	*1.1*	*1.3*	*1.5*	*1.7*	*1.9*
0.1	−100	−79	−36	−11	−1	0	−2	−5	−9	−13
0.3	−98	−38	−7	1	1	−1	−4	−7	−10	−13
0.5	−83	−12	3	4	1	−1	−4	−7	−9	−11
0.7	−49	2	6	4	1	−1	−4	−6	−8	−10
0.9	−9	8	7	4	1	−1	−3	−5	−7	−8
1.1	10	9	6	3	1	−1	−3	−5	−6	−7
1.3	5	7	5	3	1	−1	−3	−4	−6	−7
1.5	1	5	4	3	1	−1	−2	−4	−5	−6
1.7	0	3	3	2	1	−1	−2	−3	−5	−6
1.9	0	2	3	2	1	−1	−2	−3	−4	−5

†A negative number means that the approximations is less than the exact value.

This asymptotic formula has been derived for integer ν, but numerical studies suggest that it holds for all real $\nu > 0$. Table 1 gives the percentage errors from approximating the $Z(\lambda, \nu)$ function for a variety of λ- and ν-values. It can be seen that the approximation improves as λ increases. It is especially accurate for $\lambda > 10^{\nu}$ and least accurate when ν and λ are both small.

References

Bleistein, N. and Handelsman, R. A. (1986) *Asymptotic Expansions of Integrals*. New York: Dover Publications.

Boatwright, P., Borle, S. and Kadane, J. B. (2003) A model of the joint distribution of purchase quantity and timing. *J. Am. Statist. Ass.*, **98**, 564–572.

Breslow, N. (1990) Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *J. Am. Statist. Ass.*, **85**, 565–571.

Brockett, P. L., Golden, L. L. and Panjer, H. L. (1996) Flexible purchase frequency modeling. *J. Marktng Res.*, **33**, 94–107.

Chatfield, C., Ehrenberg, A. S. C. and Goodhardt, G. J. (1966) Progress on a simplified model of stationary purchasing behaviour (with discussion). *J. R. Statist. Soc.* A, **129**, 317–367.

Consul, P. C. (1989) *Generalized Poisson Distributions: Properties and Applications*. New York: Dekker.

Conway, R. W. and Maxwell, W. L. (1962) A queuing model with state dependent service rates. *J. Indstrl Engng*, **12**, 132–136.

Dean, C. B. (1992) Testing for overdispersion in Poisson and binomial regression models. *J. Am. Statist. Ass.*, **87**, 451–457.

Friendly, M. (1995) Plots for discrete distributions. York University, Downsview. (Available from `http://www.math.yorku.ca/SCS/Courses/grcat/grc1.html`.)

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) *Bayesian Data Analysis*. New York: Chapman and Hall.

Hoaglin, D. C. (1980) A Poissonness plot. *Am. Statistn*, **34**, 146–149.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992) *Univariate Discrete Distributions*. New York: Wiley.

Kadane, J. B., Shmueli, G., Minka, T. P., Borle, S. and Boatwright, P. (2003) Conjugate analysis of the Conway-Maxwell-Poisson distribution. To be published.

Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Maceda, E. C. (1948) On the compound and generalized Poisson distributions. *Ann. Math. Statist.*, **19**, 414–416.

Manton, K. G., Woodbury, M. A. and Stallard, E. (1981) A variance components approach to categorical data models with heterogenous cell populations: analysis of spatial gradients in lung cancer mortality rates in North Carolina counties. *Biometrics*, **37**, 259–269.

Minka, T. P., Shmueli, G., Kadane, J. B., Borle, S. and Boatwright, P. (2003) Computing with the COM-Poisson distribution. *Technical Report 775*. Department of Statistics, Carnegie Mellon University, Pittsburgh. (Available from `http://www.stat.cmu.edu/tr/`.)

Ord, J. K. (1967) Graphical methods for a class of discrete distributions. *J. R. Statist. Soc.* A, **130**, 232–238.

Satterthwaite, F. E. (1942) Generalized Poisson distribution. *Ann. Math. Statist.*, **13**, 410–417.

Wimmer, G., Kohler, R., Grotjahn, R. and Altmann, G. (1994) Toward a theory of word length distributions. *J. Quant. Ling.*, **1**, 98–106.