Using Computational and Mathematical Methods to Explore a New Distribution: The ν-Poisson

Galit Shmueli Thomas P. Minka Joseph B. Kadane Sharad Borle

Peter Boatwright *

Abstract

A new distribution (the ν-Poisson) and its conjugate density are introduced and explored using computational and mathematical methods. The ν-Poisson is a two-parameter extension of the Poisson distribution that generalizes some well-known discrete distributions (Poisson, Bernoulli, Geometric). It also leads to the generalization of distributions derived from these discrete distributions (viz. the Binomial and Negative Binomial). We use mathematics as far as we can and then employ computational and graphical methods to explore the distribution and its conjugate density further. Three methods are presented

* Authors are listed in reverse alphabetical order. Galit Shmueli is Visiting Assistant Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. Thomas P. Minka is Visiting Assistant Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. Joseph B. Kadane is Leonard J. Savage University Professor of Statistics and Social Sciences, Carnegie Mellon University, Pittsburgh, PA 15213. His research is supported by National Science Foundation grant DMS-9801401. Sharad Borle is a Ph.D. student, GSIA, Carnegie Mellon University, Pittsburgh, PA 15213. Peter Boatwright is Assistant Professor, GSIA, Carnegie Mellon University, Pittsburgh, PA 15213.

for estimating the ν-Poisson parameters: The first is a fast simple weighted least squares method, which leads to estimates that are sufficiently accurate for practical purposes. The second method of 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 ν-Poisson parameters is easily computed. We derive the necessary and sufficient condition for the conjugate family to be proper. The ν-Poisson is a flexible distribution that can account for over/under dispersion commonly encountered in count data. We also explore an empirical application demonstrating this flexibility of the ν-Poisson to fit count data which does not seem to follow the Poisson distribution.

KEY WORDS: Poisson Distribution; Bernoulli Distribution; Geometric Distribution; Elicitation; Conjugate Family; Exponential Family.

1 Introduction and Motivation

The Poisson is one of the most well utilized discrete densities, since data in multiple research fields often fulfill the Poisson postulates. However, the Poisson's reliance on a single parameter limits its flexibility in many applications. Over-dispersion of data relative to the distribution is a frequent issue in using the density (Breslow, 1990; Dean, 1992). One solution to this limitation is to allow the single parameter of the Poisson itself to be a random variable with some probability distribution (Maceda, 1948; Satterthwaite, 1942), which leads to a hierarchy of distributions. A commonly used hierarchy utilizes a Gamma mixing distribution for the Poisson parameter, leading to a Negative Binomial distribution for the observed data (Manton, Woodbury and Stallard, 1981; Chatfield, Ehrenberg and

Goodhardt, 1966). Consul (1989) introduces a Generalized Poisson distribution which we discuss in Section 5.

A different approach, which we consider in this paper, is to generalize the Poisson to a two parameter distribution (i.e. the ν-Poisson) where the extra parameter ν governs decay. One of the limitations of the Poisson stems from the fact that the ratio $P(X = x-1)/P(X = x)$ is linear in x, being given by x/λ, where λ is the parameter of the Poisson. Many applications may have an empirical distribution that has thicker or thinner tails than the Poisson (the change in probabilities as 'x ' changes may not be linear). Our two parameter extension of the Poisson introduces a second parameter, ν $(\nu \geq 0)$, such that $P(X = x-1)/P(X = x)$ is equal to x^{ν}/λ.Thus, the extra parameter ν can be interpreted as a parameter governing the speed of decay in the distribution. We found such a generalization to be particularly useful in a marketing application (Boatwright, Borle and Kadane, 2001) where the empirical distribution had a long and significant tail, thus necessitating the added flexibility brought about by the parameter ν, a flexibility that could not be achieved without fairly complex alternative extensions of the Poisson.

The received methods of mathematical statistics rely on a small set of distributions with convenient properties. However, the needs of modern statistical modeling often go beyond this set, leading to problems needing to be addressed computationally. This is the approach that we take in exploring the ν-Poisson distribution and its conjugate density; we use mathematics to the extent possible and then use computational and graphical methods to explore these distributions. In some cases, what ultimately became mathematical results were suggested of guided by earlier computations.

The paper is laid out as follows: we introduce the ν-Poisson distribution in Section 2, and the conjugate family in Section 3. The properties of the Bayesian analysis using the conjugate prior are presented in Section 4, and an example is given. The discussion in Section 5 concludes this paper.

2 Introducing the ν-Poisson distribution

2.1 The ν-Poisson probability function

Denote by X a random variable from the Poisson(λ) distribution, with the distribution function given by

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \text{ for } x = 0, 1, 2, \ldots \tag{1}$$

One of the properties of the Poisson distribution is that the ratio of consecutive probabilities is linear in x, or

$$\frac{P(X = x - 1)}{P(X = x)} = \frac{x}{\lambda}. \tag{2}$$

In some applications, this ratio may not decrease linearly in x, i.e. the distribution may have a thicker or thinner tail than the Poisson. Suppose instead of equation (2), we set

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

for a random variable X. The resulting distribution for which equation (3) holds, called here the ν-Poisson distribution, is given by

$$P(X = x \mid \lambda, \nu) = \frac{\lambda^x}{(x!)^\nu} \frac{1}{\sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}}, \text{ for } x = 0, 1, 2, \ldots \tag{4}$$

for $\lambda > 0$ and $\nu \geq 0$ (see exception below). This satisfies the conditions for a probability function.

It can be seen that the series $\frac{\lambda^j}{(j!)^\nu}$ converges for any $\lambda > 0, \nu > 0$, since the ratio of two subsequent terms of the series:

$$\frac{\lambda}{j^\nu} \tag{5}$$

tends to 0 as $j \to \infty$.

From here on, we denote the infinite sum in the denominator by:

$$Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu} \tag{6}$$

The ν-Poisson distribution is a generalization of some well known discrete distributions:

- For $\nu = 1$, $Z(\lambda, \nu) = e^\lambda$, and this distribution is simply the ordinary Poisson(λ).
- As $\nu \to \infty$, $Z(\lambda, \nu) \to 1 + \lambda$, and the distribution approaches a Bernoulli distribution with $P(X = 1) = \frac{\lambda}{1+\lambda}$.
- When $\nu = 0$ and $\lambda < 1$, $Z(\lambda, \nu)$ is a geometric sum:

  $$Z(\lambda, \nu) = \sum_{j=0}^{\infty} \lambda^j = \frac{1}{1-\lambda}, \tag{7}$$

  and the distribution itself is geometric:

  $$P(X = x \mid \lambda, \nu) = \lambda^x (1 - \lambda) \text{ for } x = 0, 1, 2, \ldots \tag{8}$$

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

The ν-Poisson 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 less than one correspond to flatter successive ratios (3) than the Poisson's (2) and hence to longer tails.

The sum of n independent ν-Poisson random variables is also a continuous bridge between three well-known distributions:

- For $\nu = 0$ and $\lambda < 1$, this reduces to the sum of geometric variables, namely, the negative binomial distribution with parameters n and $1 - \lambda$.
- For $\nu = 1$ the sum has a Poisson distribution with parameter $n\lambda$.
- For $\nu = \infty$ the distribution of the sum is Binomial with parameters n and $\lambda/(1+\lambda)$.

See the relation to other distributions in section 2.7.

2.2 Moments of the Distribution

The ν-Poisson belongs to the family of two-parameter power series distributions (Johnson, Kotz, and Kemp, 1992). Moments of this distribution can then be expressed using the recursive formula:

$$E[X^{r+1}] = \begin{cases} \lambda E(X+1)^{1-\nu} & r = 0 \\ \lambda \frac{d}{d\lambda} E[X^r] + E(X)E(X^r) & r > 0 \end{cases} \tag{9}$$

In section 3.3.2 an asymptotic approximation is given for $Z(\lambda, \nu)$. Using the expression in (55), $E[X]$ can be closely approximated by

$$E[X] = \lambda \frac{d \log Z(\lambda, \nu)}{d\lambda} \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \quad . \tag{10}$$

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

Another property that the ν-Poisson inherits as a member of the generalized power series distributions is that its upper tail can be expressed as the lower tail of a continuous distribution (Patil, 1974). The density of this continuous distribution is given by:

$$f(z; x) = \frac{1}{Z(z, \nu) z} \sum_{j=0}^{x} [E(X) - j] \, z^j / (j!)^{\nu} \tag{11}$$

2.3 Sufficient Statistics

The likelihood for a set of n iid observations $x_1, x_2, \ldots, x_n$ is

$$L(x_1, x_2, \ldots, x_n \mid \lambda, \nu) = \frac{\prod_{i=1}^{n} \lambda^{x_i}}{\left(\prod_{i=1}^{n} x_i!\right)^{\nu}} Z^{-n}(\lambda, \nu) = \tag{12}$$

$$= \lambda^{\sum_{i=1}^{n} x_i} e^{-\nu \sum_{i=1}^{n} \log(x_i!)} Z^{-n}(\lambda, \nu) \tag{13}$$

$$= \lambda^{S_1} e^{-\nu S_2} Z^{-n}(\lambda, \nu) \tag{14}$$

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, (14) displays the ν-Poisson as a member of the exponential family.

2.4 Generating data from the ν-Poisson distribution

We use property (3) of the ν-Poisson distribution in order to apply the inversion method for simulation purposes. In practice, the ν-Poisson probabilities are summed by multiplying the probability $P(X = 0)$ by $(x+1)^{\nu}/\lambda$ $(x = 0, 1, ...)$ x times, until this sum exceeds the value of a simulated Uniform(0,1) variable. X is then an observation from the ν-Poisson distribution.

For computing the initial probability, $P(X = 0) = Z^{-1}(\lambda, \nu)$, an approximation is used, based on subsection 3.3.2.

2.5 Unimodality of the likelihood

The ν-Poisson log-likelihood function is given by:

$$\mathcal{L}(\lambda, \nu) = \log(\lambda) \sum_{i=1}^{n} X_i - \nu \sum_{i=1}^{n} \log(X_i!) - n \log \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^{\nu}} \tag{15}$$

This function has the same form as the prior distribution in (28), which is described in the next section, with $a - 1 = \sum_{i=1}^{n} X_i$, $b = \sum_{i=1}^{n} \log(X_i!)$, and $c = n$.

In order to find the maximum-likelihood estimates for λ and ν, the method from subsection 3.3 is applied, substituting $(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} \log(X_i!), n)$ for (a, b, c) respectively. In this case, a reasonable starting point for $\underline{\theta}^{(0)}$ is $\lambda^{(0)} = \overline{X}$ and $\nu^{(0)} = 1$ (the ordinary Poisson).

2.6 Fitting the ν-Poisson distribution

We describe a simple and efficient method that enables one to determine whether the ν-Poisson is an adequate model for some dataset, and if so, to estimate the parameters ν, and λ in a simple way.

The method is based on the relation between successive ν-Poisson probabilities, given in (3). Taking a log of both sides of the equation, we get the linear relation

$$\log\left(\frac{p_{x-1}}{p_x}\right) = -\log\lambda + \nu\log x, \tag{16}$$

where p_x denotes $P(X = x)$. The ratio on the left can be estimated by replacing the probabilities with the relative counts of $x - 1$ and x, respectively. The first step is to plot these values versus $\log x$, for all ratios that do not include zero counts. A ν-Poisson would be adequate if the relation appears linear.

If the data do appear to fit a ν-Poisson model, then the parameters can be estimated by fitting a regression, regressing $\log\left(\frac{\hat{p}_{x-1}}{\hat{p}_x}\right)$ on $\log x$. However, two basic assumptions of an ordinary regression model are not true in this case. First, the variance of the dependent variable is not constant. Assuming a Dirichlet prior on the vector of probabilities $\underline{p_x}$ $(x = 0, 1, \ldots)$, with parameter vector $\underline{\alpha_x}$, the posterior variance is approximately

$$\mathrm{Var}\left(\log\frac{p_{x-1}}{p_x} \mid Data\right) \approx \frac{1}{n\hat{p}_x + \alpha_x} + \frac{1}{n\hat{p}_{x-1} + \alpha_{x-1}}, \tag{17}$$

where n is the number non-zero values in the data (see appendix A). With no prior knowledge, $\underline{\alpha_x}$ can be set to $\underline{0}$. In this case the approximation coincides with the non-

Bayesian approximate variance.

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

$$\mathrm{Cov}\left(\log \frac{p_{x-1}}{p_x}, \log \frac{p_x}{p_{x+1}} \mid Data\right) \approx -\frac{1}{n\hat{p}_x + \alpha_x} \tag{18}$$

(see appendix A for details).

To take into account both the varying variance and the one-step dependence, a weighted least square regression can be used. The inverse weight matrix would then have the variances in the diagonal, and the one-step covariances in the first off-diagonal.

The estimates can then be refined by applying the maximum likelihood method that is described in (3.3).

2.7 An Extension: The ν-Binomial distribution

From the ν-Poisson, a variety of derived distributions can be obtained which generalize the classical ones. For example, a ν-Binomial distribution can be defined by dividing two ν-Poisson variables by their sum. The ν-Binomial distribution can represent over-dispersion and under-dispersion relative to the Binomial. Consider the sum of two independent ν-Poisson variables with parameters λ_1 and λ_2:

$$S = X + Y \tag{19}$$

$$P(S=s) = \sum_{x=0}^{s} P(X=x)P(Y=s-x) \tag{20}$$

$$= \sum_{x=0}^{s} \frac{\lambda_x^x}{x!^\nu Z(\lambda_x,\nu)} \frac{\lambda_y^{s-x}}{(s-x)!^\nu Z(\lambda_y,\nu)} \tag{21}$$

$$= \frac{(\lambda_x+\lambda_y)^y}{y!^\nu Z(\lambda_x,\nu)Z(\lambda_y,\nu)} \sum_{x_1=0}^{s} \binom{s}{x}^\nu \left(\frac{\lambda_x}{\lambda_x+\lambda_y}\right)^x \left(\frac{\lambda_y}{\lambda_x+\lambda_y}\right)^{s-x} \tag{22}$$

The distribution of X conditional on the sum is

$$P(X = x|S = s) \quad \propto \quad \binom{s}{x}^{\nu} \left(\frac{\lambda_x}{\lambda_x + \lambda_y}\right)^{x} \left(\frac{\lambda_y}{\lambda_x + \lambda_y}\right)^{s-x} \tag{23}$$

which is ν-Binomial with parameter $p = \frac{\lambda_1}{\lambda_1+\lambda_2}$. A ν-Multinomial distribution over several variables can be similarly defined. The binomial coefficient $\binom{s}{x}$ favors $x = s/2$, so $\nu > 1$ gives the distribution less variance and $\nu < 1$ more variance than a 1-Binomial with the same mean. 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 $(s \pm 1)/2$ for odd s.

The 1-Binomial can be interpreted as a sum of independent Bernoulli variables. The ν-Binomial can be interpreted as a sum of non-independent Bernoulli variables Z_i with joint distribution

$$p(Z_1 = z_1, ..., Z_s = z_s) \quad \propto \quad \binom{s}{x}^{\nu-1} p^x(1-p)^{s-x} \qquad (x = \sum_{i=1}^{s} z_i) \tag{24}$$

For $\nu > 1$ the z's are negatively correlated and for $\nu < 1$ the z's are positively correlated.

3 Conjugate Family

Since the ν-Poisson 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

$$h(\lambda, \nu) = \lambda^{a-1} e^{-\nu b} Z^{-c}(\lambda, \nu)\kappa(a, b, c) \tag{25}$$

for $\lambda > 0$ and $\nu \geq 0$, where $\kappa(a, b, c)$ is the integration constant, and $Z(\lambda, \nu)$ is defined in equation (6). Then the posterior is of the same form, with $a' = a + S_1, b' = b + S_2$, and $c' = c + n$. The density in (25) can be thought of as an extended bivariate Gamma distribution.

3.1 Hyperparameter space for the conjugate density

In order for equation (25) to constitute a density, it must be non-negative and integrate to one. In other words, the values of a, b, and c that lead to a finite $\kappa^{-1}(a, b, c)$, which is given by

$$\kappa^{-1}(a,b,c) = \int_0^\infty \int_0^\infty \lambda^{a-1} e^{-b\nu} Z^{-c}(\lambda,\nu) d\lambda d\nu, \tag{26}$$

will lead to a proper density.

Using Jensen's inequality and the convexity of the log-gamma function, Appendix B shows that a necessary and sufficient condition for a finite $\kappa^{-1}(a, b, c)$ is:

$$b/c > \log(\lfloor a/c \rfloor !) + (a/c - \lfloor a/c \rfloor) \log(\lfloor a/c \rfloor + 1) \tag{27}$$

where $\lfloor k \rfloor$ denotes the floor of k. The set of hyperparameters (a, b, c) satisfying (27) is necessarily closed under sampling, so we restrict attention to this space below.

3.2 Point of maximum of the conjugate family

Following (25), the log-prior/posterior can be expressed as

$$\mathcal{H} = \log[h(\lambda,\nu)] = (a-1)\log(\lambda) - \nu b - c \log Z(\lambda,\nu) + \log(\kappa(a,b,c)) \tag{28}$$

The first partial derivatives of (28) with respect to λ and ν can be expressed as:

$$\frac{\partial \mathcal{H}}{\partial \lambda} = \frac{1}{\lambda}[a - 1 - cE(X)] \tag{29}$$

$$\frac{\partial \mathcal{H}}{\partial \nu} = -b + cE(\log(X!)) \tag{30}$$

The second partial derivatives and the cross-derivative, calculated at the maximum, can be expressed as:

$$\frac{\partial^2 \mathcal{H}}{\partial \lambda^2} = -\frac{c}{\lambda^2} V(X) \tag{31}$$

$$\frac{\partial^2 \mathcal{H}}{\partial \nu^2} = -cV(\log(X!)) \tag{32}$$

$$\frac{\partial^2 \mathcal{H}}{\partial \lambda \partial \nu} = \frac{c}{\lambda} Cov(X, \log(X!)) \tag{33}$$

Note that the ν-Poisson is a member of the exponential family, and its likelihood function can be expressed in the form:

$$L(X \mid \underline{\theta}) = \gamma(\underline{\theta}) \phi(x) \exp(\sum_{i=1}^{k} \pi_i(\theta) t_i(x)) \tag{34}$$

where $\Pi(\underline{\theta}) = [\pi_1(\theta), ..., \pi_k(\theta)]$ is the natural parameter and $T(X) = [t_1(x), ..., t_k(x)]$ is the natural sufficient statistic. For the ν-Poisson these are

$$\Pi(\lambda, \nu) = [\log \lambda, -\nu] \tag{35}$$

$$T(X) = [\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} \log(x_i!)] \tag{36}$$

In general, moments of the sufficient statistics can then be calculated by taking derivatives of $\log \gamma(\underline{\theta})$, and in particular

$$E(T_i) = \frac{-\partial}{\partial \theta_i} \log \gamma(\underline{\theta}) \tag{37}$$

$$Cov(T_i, T_j) = \frac{-\partial^2}{\partial \theta_i \partial \theta_j} \log \gamma(\underline{\theta}) \tag{38}$$

(Schervish, 1995, Theorem 2.64). This leads to a general form for the first derivatives of the log-likelihood (or log-posterior):

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \theta_i} &= \frac{\partial \pi_i(\theta)}{\partial \theta_i} \frac{\partial}{\partial \pi_i(\theta)} \log \gamma(\underline{\theta}) + t_i(x) \frac{\partial \pi_i(\theta)}{\partial \theta_i} = \\ &= \frac{\partial \pi_i(\theta)}{\partial \theta_i} (-E(T_i) + t_i(x)) \end{aligned} \tag{39}$$

Thus the maximum likelihood estimators are characterized by matching two sample averages, (in the ν-Poisson $\overline{X}$ and $\overline{log(X!)}$) to their expectations. While computing these sample averages is easy, finding the expectations is not always straightforward. Under the ν-Poisson the calculation of these expectations requires more work, as is explored in subsection 3.3.

The second derivatives, at the maximum point (where the first derivatives are zero) are:

$$\begin{aligned}\frac{\partial^2 \mathcal{L}}{\partial\theta_i\partial\theta_j} &= \frac{\partial\pi_i(\theta)}{\partial\theta_i}\frac{\partial\pi_j(\theta)}{\partial\theta_j}\frac{\partial^2}{\partial\pi_i(\theta)\pi_j(\theta)}\log\gamma(\underline{\theta}) = \\ &= \frac{\partial\pi_i(\theta)}{\partial\theta_i}\frac{\partial\pi_j(\theta)}{\partial\theta_j}(-\mathrm{Cov}(T_i,T_j)) \qquad (40)\end{aligned}$$

Since the natural sufficient statistic also belongs to the exponential family (Schervish, 1995, Lemma 2.58), and is nondegenerate, the coordinates of X and $\log(X!)$ are not linearly dependent. Therefore, the second derivatives (31) and (32) are negative and their product exceeds the squared cross derivative:

$$\frac{\partial^2\mathcal{H}}{\partial\lambda^2}\times\frac{\partial^2\mathcal{H}}{\partial\nu^2}-\left(\frac{\partial^2\mathcal{H}}{\partial\lambda\partial\nu}\right)^2=\left(\frac{c}{\lambda}\right)^2 \mathrm{V}ar(X)\mathrm{Var}(\log(X!))\left[1-\rho^2(X,\log(X!))\right]>0 \qquad (41)$$

These conditions imply that the posterior distribution is unimodal.

3.3 Computing the point of maximum of the conjugate family

In order to find the values of λ and ν at the point of maximum, it is required to equate (29)-(30) to zero. Since these equations can not be solved analytically, the Newton-Raphson method is used instead (see Gelman et al.,1995 pp. 272-273).

Using the score vector $U(\lambda,\nu)$, and the information matrix $I(\lambda,\nu)$, we compute approximations for the values of λ and ν at the point of maximum iteratively. Denote the vector of values for λ and ν at the kth iteration by $\underline{\theta}^{(k)}$. Then,

$$\underline{\theta}^{(k+1)}=\underline{\theta}^{(k)}-I^{-1}(\underline{\theta}^{(k)})U(\underline{\theta}^{(k)}) \qquad (42)$$

In each iteration the expectations and variances of X and $\log(X!)$ are computed (or approximated), as well as the covariance between them. These computations are performed by plugging the estimates for λ and ν from the previous iteration into the following expression:

$$E[f(X)] = \sum_{j=0}^{\infty} f(j) \frac{\lambda^j}{(j!)^\nu Z(\lambda,\nu)} \tag{43}$$

3.3.1 Approximating the expectations

In most cases the above sums must be truncated in order 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 pre-specified precision. This can be done by recognizing that the tails of both the numerator and denominator of each of the expectations decrease more rapidly than a geometric series.

3.3.2 The denominator $Z(\lambda,\nu)$

An upper bound on $Z(\lambda,\nu)$: As shown in subsection 2.1, the series $\frac{\lambda^j}{(j!)^\nu}$ converges. In addition, $\lim \frac{\lambda^j}{(j!)^\nu} = 0$ as $j \to \infty$. Therefore there exists a value K such that, for $k > K$,

$$\frac{\lambda}{k^\nu} < 1. \tag{44}$$

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 (44). Thus, $Z(\lambda,\nu)$ can be approximated by truncating the series at some kth term such that (44) holds, i.e.

$$Z(\lambda,\nu) = \sum_{j=0}^{k} \frac{\lambda^j}{(j!)^\nu} + R_k, \tag{45}$$

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

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

$$\frac{\lambda}{(k+1)^\nu} < \epsilon_k. \tag{46}$$

R_k is then bounded by

$$\frac{\lambda^{k+1}}{((k+1)!)^\nu(1-\epsilon_k)}. \tag{47}$$

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

$$R_k / \sum_{j=0}^{k} \frac{\lambda^j}{(j!)^\nu}. \tag{48}$$

The relative truncation error can be bounded by

$$\frac{\lambda^{k+1}}{((k+1)!)^\nu(1-\epsilon_k)} \times \frac{1}{\sum_{j=0}^{k} \frac{\lambda^j}{(j!)^\nu}} \tag{49}$$

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

$$\hat{Z}^{-1}(\lambda, \nu) = \left(\sum_{i=0}^{k} \frac{\lambda^j}{(j!)^\nu} \right)^{-1} \tag{50}$$

leads to a relative error that is

$$\frac{\hat{Z}^{-1} - Z^{-1}}{\hat{Z}^{-1}} = \frac{\sum_{j=k}^{\infty} \frac{\lambda^j}{(j!)^\nu}}{\sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}}, \tag{51}$$

which is smaller than (49). Thus, the relative error for computing $Z^{-1}(\lambda, \nu)$ by truncation is bounded by the same bound as that for $Z(\lambda, \nu)$.

An asymptotic approximation of $Z(\lambda, \nu)$: Defining $i = \sqrt{-1}$, we have the identity

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{e^{ia}} e^{-iaj} da \quad = \quad \frac{1}{j!} \tag{52}$$

which means

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{e^{ia}} Z(\lambda e^{-ia},\nu)da \;=\; \sum_{j=0}^{\infty}\frac{\lambda^j}{j!^{\nu}}\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{e^{ia}} e^{-iaj}da \;=\; Z(\lambda,\nu+1) \tag{53}$$

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

$$Z(\lambda,\nu) \;=\; \frac{1}{(2\pi)^{\nu-1}}\int\cdots\int \exp\left(\sum_{k=1}^{\nu-1}\exp(ia_k)+\lambda\exp(-\sum_{k=1}^{\nu-1} ia_k)\right) da_1\cdots da_{\nu-1} \tag{54}$$

The behavior of this integral for large λ can be determined by making the change of variable $ia_j = ib_j + \frac{1}{\nu}\log(\lambda)$ and applying Laplace's method (Bleistein and Handelsman, 1986, Ch 5.1). The result is:

$$Z(\lambda,\nu) \;=\; \frac{\exp(\nu\lambda^{1/\nu})}{\lambda^{\frac{\nu-1}{2\nu}}(2\pi)^{(\nu-1)/2}\sqrt{\nu}}(1+O(\lambda^{-1/\nu})) \tag{55}$$

This asymptotic formula has been derived for integer ν, but numerical studies suggest that it holds for all real $\nu > 0$. It is especially accurate for $\nu \leq 1$ or $\lambda > 10^{\nu}$.

3.3.3 The numerators

Using the same logic as for the denominator, the tails of the infinite sums of the form (43) converge faster than a geometric series. For instance, the ratio of two subsequent terms of the series in the numerator of $E(X)$ is

$$\frac{\lambda(k+1)^{1-\nu}}{k} \tag{56}$$

This ratio tends to 0 as $k \to \infty$, and in addition $\lim \frac{j\lambda^j}{(j!)^{\nu}} = 0$ as $j \to \infty$. Thus there exists a value K such that, for $k > K$, the ratio (56) is smaller than 1. Also, note that this ratio is monotonically decreasing, meaning that for $k > K$, this series converges faster than a geometric series with a multiplier given by (56).

Similarly, the ratio of subsequent terms of the series in the numerator of $E(X^2)$ is

$$\frac{\lambda(k+1)^{2-\nu}}{k^2} \tag{57}$$

the limit of which is 0 for $\nu \geq 0$. Similar results hold for the remaining expectations.

3.3.4 Bounding the ratios

Geometric series can be used to specify upper and lower bounds for expectations of the form(43). Write the expectation in the form:

$$E[f(X)] = \frac{\sum_{i=1}^{\infty} a_i}{\sum_{i=1}^{\infty} b_i} = \frac{\sum_{i=1}^{k} a_i + \sum_{i=k+1}^{\infty} a_i}{\sum_{i=1}^{k} b_i + \sum_{i=k+1}^{\infty} b_i}. \tag{58}$$

Then an upper bound is given by

$$U_k = \frac{\sum_{i=1}^{k} a_i + \frac{a_{k+1}}{1-a_{k+2}/a_{k+1}}}{\sum_{i=1}^{k} b_i} \tag{59}$$

and a lower bound is given by

$$L_k = \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} b_i + \frac{b_{k+1}}{1-b_{k+2}/b_{k+1}}}. \tag{60}$$

One may then choose k so that $U_k - L_k < \epsilon$, for any desired value of ϵ, and then approximate $E[f(X)]$ by

$$E[f(X)] \approx \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} b_i} \tag{61}$$

Alternatively, the relative error can be bounded, setting

$$(U_k - L_k)\frac{\sum_{i=1}^{k} b_i}{\sum_{i=1}^{k} a_i} < \epsilon \tag{62}$$

To avoid reaching machine limits for computing (59) and (60), the ratio can be normalized each time the sum reaches machine limits. This enables a numerical approximation with high precision even when each of the sums exceeds machine limits.

3.4 The distribution $h(\lambda, \nu)$

3.4.1 Marginal and conditional distributions

We investigate the marginal and conditional distributions of λ and ν that arise from a bivariate distribution of the form:

$$h(\lambda, \nu) = \lambda^{a-1} e^{-b\nu} [Z(\lambda, \nu)]^{-c} \kappa(a, b, c) \tag{63}$$

where $a > 0$, $b > 0$, and $c > 0$ are hyperparameters. The marginal distribution of λ can then be expressed as:

$$h_1(\lambda) = \int_{\nu=0}^{\infty} h(\lambda, \nu) d\nu \quad = \quad \lambda^{a-1} \kappa(a, b, c) \int_{\nu=0}^{\infty} e^{-b\nu} [Z(\lambda, \nu)]^{-c} d\nu \tag{64}$$

The conditional density of ν given λ is then:

$$h(\nu \mid \lambda) = \frac{h(\lambda, \nu)}{h_1(\lambda)} \propto e^{-b\nu} [Z(\lambda, \nu)]^{-c} \tag{65}$$

Using the same logic, the marginal distribution of ν can be expressed as

$$h_2(\nu) = \int_{\lambda=0}^{\infty} h(\lambda, \nu) d\lambda = e^{-b\nu} \kappa(a, b, c) \int_{\lambda=0}^{\infty} \lambda^{a-1} [Z(\lambda, \nu)]^{-c} d\lambda \tag{66}$$

The conditional density of λ given ν is then:

$$h(\lambda \mid \nu) = \frac{h(\lambda, \nu)}{h_2(\nu)} \propto \lambda^{a-1} [Z(\lambda, \nu)]^{-c} \tag{67}$$

For the three well known cases ($\nu = 0, 1, \infty$) this reduces to:

$$\lambda \mid_{\nu=0} \quad \sim \quad Beta(a, c+1) \tag{68}$$

$$\lambda \mid_{\nu=1} \quad \sim \quad \Gamma(a, c) \tag{69}$$

$$\lambda \mid_{\nu=\infty} \quad \sim \quad F(2a, 2(c-a)) \tag{70}$$

The result in (70) is obtained by noticing that

$$\frac{\lambda}{1+\lambda} \mid_{\nu=\infty} \sim Beta(a, c-a), \tag{71}$$

and using the relation between the Beta and F distributions (Mood, Greybill, & Boes, 1974, p. 249).

3.4.2 Generating data from $h(\lambda, \nu)$

Posterior computations are greatly facilitated by employing the reparameterization (m, ν) where $m = \lambda^{1/\nu}$. The posterior distribution transforms to

$$h(m, \nu) = \nu m^{a\nu-1} e^{-b\nu} Z(m^\nu, \nu)^{-c} \kappa(a, b, c) \tag{72}$$

By exploiting properties of Z, a rejection method can be derived for sampling from the exact conditionals. Numerical studies suggest the following conjectures:

1. If $\nu \geq 1$, $\log Z(m^\nu, \nu)$ is convex in m.
2. If $\nu < 1$, $\log Z(\lambda, \nu)$ is convex in λ.
3. $\log Z(m^{e^\gamma}, e^\gamma)$ is convex in γ.

Convexity means that all tangent lines are lower bounds. Therefore these assumptions lead to three different lower bounds for Z which immediately give rejection algorithms as described below. These rejection algorithms can be embedded in a Gibbs sampling scheme to obtain bivariate samples from the posterior $h(m, \nu)$.

Sampling from $h(m|\nu)$, $\nu \geq 1$ From the first convexity property, we know that if $\nu \geq 1$ then for any m_0 we can draw a tangent line to $\log Z(m^\nu, \nu)$ and get a lower bound:

$$\log Z(m^\nu, \nu) \geq q(m_0)(m - m_0) + \log Z(m_0^\nu, \nu) \tag{73}$$

$$Z(m^\nu, \nu) \geq \exp(q(m_0)(m - m_0)) Z(m_0^\nu, \nu) \tag{74}$$

$$\text{where } q(m_0) = \left. \frac{d \log Z(m^\nu, \nu)}{dm} \right|_{m=m_0} \tag{75}$$

$$= \frac{\nu}{m_0} E[x; \lambda = m_0^\nu, \nu] \tag{76}$$

Substituting this bound into (72) gives a Gamma approximation to $p(m|\nu)$ as well as the minimal scale factor to make it dominate the true conditional. This gives the following rejection algorithm:

1. Choose m_0 and compute $q(m_0)$.
2. Draw m from the Gamma distribution $\Gamma(a\nu, cq(m_0))$.
3. Draw a uniform variate $u \sim U(0,1)$ and accept m if

$$u \leq \frac{Z(m^\nu,\nu)^{-c}}{\exp(-cq(m_0)(m-m_0))Z(m_0^\nu,\nu)^{-c}} \tag{77}$$

4. If m is rejected, repeat from step 2.

The dominating curve makes contact with $h(m|\nu)$ at m_0. The best m_0 is the mode of $h(m|\nu)$, because this minimizes the supremum of the dominating curve. The mode can be found using the same methods for finding the mode of $h(\lambda,\nu)$ presented earlier.

Sampling from $h(m|\nu)$, $\nu < 1$ If $\nu < 1$, then the second convexity property implies for any m_0:

$$Z(m^\nu,\nu) \geq \exp(q(m_0)(m^\nu - m_0^\nu))Z(m_0^\nu,\nu) \tag{78}$$

$$\text{where } q(m_0) = \left.\frac{d\log Z(\lambda,\nu)}{d\lambda}\right|_{\lambda=m_0^\nu} \tag{79}$$

The rejection algorithm is

1. Choose m_0 and compute $q(m_0)$.
2. Draw m from the Weibull distribution $p(m|\nu) \propto m^{a\nu-1}\exp(-cq(m_0)m^\nu)$.
3. Draw a uniform variate $u \sim U(0,1)$ and accept m if

$$u \leq \frac{Z(m^\nu,\nu)^{-c}}{\exp(-cq(m_0)(m^\nu-m_0^\nu))Z(m_0^\nu,\nu)^{-c}} \tag{80}$$

4. If m is rejected, repeat from step 2.

Figure 1(a) shows the dominating curve vs. the exact conditional $p(m|\nu)$ for the example data given in section 4.3. The value of ν was 0.126 (near the mode of $h(m,\nu)$) and m_0 was placed at the mode of the conditional.

Sampling from $h(\nu|m)$ From the third convexity property we know that for any ν_0:

$$\begin{aligned}
\log Z(m^{e^\gamma}, e^\gamma) &\geq q(\nu_0)(\gamma - \log(\nu_0)) + \log Z(m^{\nu_0}, \nu_0) && (81)\\
Z(m^\nu, \nu) &\geq \left(\frac{\nu}{\nu_0}\right)^{q(\nu_0)} Z(m^{\nu_0}, \nu_0) && (82)\\
\text{where } q(\nu_0) &= \left.\frac{d \log Z(m^{e^\gamma}, e^\gamma)}{d\gamma}\right|_{\gamma=\log(\nu_0)} && (83)\\
&= \nu_0 E[x; \lambda = m^{\nu_0}, \nu_0] \log(m) - \nu_0 E[\log(x!); \lambda = m^{\nu_0}, \nu_0] && (84)
\end{aligned}$$

The rejection sampler is

1. Choose ν_0 and compute $q(\nu_0)$.
2. Draw ν from the Gamma distribution $\Gamma(2 - cq(\nu_0), b)$.
3. Draw a uniform variate $u \sim U(0,1)$ and accept ν if

$$u \leq \frac{Z(m^\nu, \nu)^{-c}}{\left(\frac{\nu}{\nu_0}\right)^{-cq(\nu_0)} Z(m^{\nu_0}, \nu_0)^{-c}} \tag{85}$$

4. If ν is rejected, repeat from step 2.

The best ν_0 is the mode of $p(\nu|m)$. Figure 1(b) shows the dominating curve vs. the exact conditional $p(\nu|m)$ for the example data given in section 4.3. The value of m was 0.7921 (near the mode of $h(m,\nu)$) and ν_0 was placed at the mode of the conditional.

Insert Figure 1 here

4 The ν-Poisson with a Conjugate Prior

4.1 Predictive distribution

Given values for the hyperparameters a, b and c, the predictive probability function is given by

$$\begin{aligned} P(X = x \mid a,b,c) &= \int_0^\infty \int_0^\infty P(X = x \mid \lambda, \nu) h(\lambda, \nu \mid a,b,c) d\lambda d\nu \\ &= \int_0^\infty \int_0^\infty \left[\frac{\lambda^x}{(x!)^\nu Z(\lambda,\nu)} \right] \left[\lambda^{a-1} e^{-b\nu} Z^{-c}(\lambda,\nu) \kappa(a,b,c) \right] d\lambda d\nu \\ &= \kappa(a,b,c) \int_0^\infty \int_0^\infty \lambda^{a+x-1} e^{-\nu(b+\log(x!))} [Z(\lambda,\nu)]^{-(c+1)} d\lambda d\nu = \\ &= \frac{\kappa(a,b,c)}{\kappa(a+x, b+\log(x!), c+1)} \end{aligned} \tag{86}$$

where $\kappa(a, b, c)$ is defined in (26).

In practice, there are several issues that must be handled in order to compute this distribution. The first issue involves the calculation of the two double integrals in (86). This can be done by using a non-equally spaced grid over the λ, ν space. The numerical routines that aid in the specification of this region turn out to be more robust in the transformed space

$$\lambda^* = \log(\lambda) \tag{87}$$

$$\nu^* = \log(\nu) \tag{88}$$

4.2 An Elicitation Program

To make the theory more accessible to practitioners, we have created an application to help the practitioner select the hyperparameters a, b, and c or to aid the statistician in eliciting these values from the expert who understands the data better than the theory. See Kadane and Wolfson (1997) for a review. Even when a prior is of the form of the well used Gaussian distribution, practitioners often have difficulty setting covariance parameters of the prior (Barnard, McCulloch, and Meng, 2000). Here, where the distribution is new to the practitioner, it is even more difficult to give meaningful values to a, b, and c. Because the practitioner is very likely to have some degree of knowledge about $P(X = x)$, our application calculates and plots the predictive distribution for selected values of a, b, and c. Our application aid can be found on the web at `http://www.stat.cmu.edu/~jacklin/`. The application has slider bars which allow the user to select and change values of a, b, and c. The application ensures that the constraints on a, b, and c are met (equation (27)). If the constraints are not met, the user is warned in a text window that they should alter their choice of the triplet (a, b, c). If the user ignores the warning, the program will give an error message. The application calculates and plots a histogram of the predictive density for each triplet (a, b, c). The practitioner can then use knowledge of $P(X = x)$ to select reasonable values for a, b, and c.

Calculation of the predictive density involves integrals, which are calculated numerically in the elicitation application. The numerical integration requires edges of the region over which the integral is to be taken. Since this approximates an open space with a closed region, it is necessary to ensure that the integrated region contains almost all of the probability mass. One way to choose such boundaries is to integrate over the rectangular region defined by

$$\hat{\lambda^*} \pm l\sqrt{Var(\hat{\lambda^*})} \tag{89}$$

$$\hat{\nu^*} \pm l\sqrt{Var(\hat{\nu^*})} \tag{90}$$

where $(\hat{\lambda^*}, \hat{\nu^*})$ are the maximum likelihood estimators, while $Var(\hat{\lambda^*})$ and $Var(\hat{\nu^*})$ are the diagonal elements of the inverse of the negative Hessian.

To find the point of maximum (λ^*, ν^*) we use a globally convergent Newton Raphson routine (a double precision version of the function newt of Numerical Recipes), as suggested in section 3.3. For the purpose of approximating the expectation of the form in (43), the truncation point k is selected so that $U_k - L_k < \epsilon$, where we set ϵ to 1.0e-8.

Since λ^* and ν^* are highly correlated, integration over a rectangular space is inefficient since many calculations are made over regions with trivial probability mass. We therefore set integration boundaries using the marginal distributions $f(\nu^*|\lambda^*)$ of the bivariate normal distribution with mean $(\hat{\lambda^*}, \hat{\nu^*})$ and covariance matrix equal to the inverse of the negative Hessian. We retain the bounds on λ^* as given in equation (89) and use the mean and variance of $\nu^*|\lambda^*$ to replace $\hat{\nu^*}$ and $Var(\nu^*)$ in equation (90). In our elicitation program, we set $l = 8$.

A snapshot of the predictive distribution plot, created by our web application, is given in Figure (2). The prior hyperparameters entered in this case were $a = b = c = 1$, which correspond to a relatively non-informative prior.

Insert Figure 2 here

4.3 Fitting the ν-Poisson distribution: An Example

To illustrate the usefulness and flexibility of the ν-Poisson distribution, we compare the fit of the ν-Poisson with that of some other discrete models on an actual data set. The data in this illustration are the quarterly sales of a particular brand of a particular article of clothing at stores of a large national retailer. For items of such specificity, quarterly sales at individual stores are small in quantity and often zero.

At first, the logs of the ratios of successive counts are plotted vs. $\log x$, as described in (2.6). Ratios involving zero counts are omitted. Figure (3) illustrates this for the quarterly sales data. It can be seen that a linear relation is plausible. However, the data do not seem to follow a Poisson model (with the MLE $\hat{\lambda} = 3.56$), as the angle of the Poisson line is too large. Next, a linear line is fit to the data using weighted least squares. The estimates in this case turn out to be $\lambda = 1$ and $\nu = 0.15$. 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.

Insert Figure 3 here

In the next step, these estimates are refined by applying the ML method described in section 3.3. The refined estimates are $\hat{\lambda} = 0.97$ and $\hat{\nu} = 0.126$. The WLS estimates can be used as initial values for the maximum likelihood procedure. Figure (4) gives a plot of the posterior of the data (with a prior of a=1, b=1, c=1). In the contour plot, the figures accompanying the contours are the integrals of the region inside of the contour. The third axis of the three dimensional plot has analogous labels. Thus both plots indicate the cumulative posterior probability of the plotted regions.

Insert Figure 4 here

The maximum posterior estimates are $\hat{\lambda} = 0.97$ and $\hat{\nu} = 0.126$. Since $\nu = 1$ is not near the mass of the distribution this implies once again that the ordinary Poisson distribution should not be considered as a reasonable distribution to fit these data. Figure (5) yields the same conclusion, showing that the Poisson model provides terrible sales estimates. The geometric distribution also is a poor choice, for it greatly overestimates the number of 0s and 1s in the data. The Negative Binomial distribution fits the model better than the Poisson and geometric models, but its shape differs from that of the data. In addition, the parameter estimation process for the Negative Binomial model is not very simple or efficient. The simpler methods (e.g., the method of Moments, the Mean-and-Zero-Frequency method) are not suitable for many areas of the parameter space. Maximum likelihood estimation is done iteratively (Bliss & Fisher, 1953) and is rather tedious and inefficient if the initial estimates are poor (Johnson, Kotz, and Kemp, 1992, p. 216). The ν-Poisson, a distribution that is between the geometric and the Poisson, provides an extremely close fit to the data. The parameter estimation using the simple initial method leads to relatively accurate estimates. These can then be refined using ML.

Insert Figure 5 here

Although in this illustration ν is clearly not 0 or 1, the ν-Poisson can also be fit to data where 1 and/or 0 is a plausible value for ν. Posterior plots such as that in Figure (4) can serve as a diagnostic to indicate the range of possible distributions that can be fit to the data.

5 Discussion

The ν-Poisson distribution originated from a need to model discrete data that did not fit an ordinary Poisson distribution. The ν-Poisson generalization, with its second parameter (ν), allows for flexibility in the magnitude of decay in the distribution, as compared to the Poisson. Furthermore, it enables fitting distributions that do not resemble the well known discrete distributions yet are on a 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 due to the recent advances in computational power.

In practice, there are not many discrete distributions that are used to fit discrete data. The most widely used is the Poisson, whereas for over or under dispersed data the Negative Binomial distribution is often used. When none of the known distributions seem appropriate, different variations are invented. The ν-Poisson, in this sense, increases the variety of discrete distributions available for data modelling.

Our data fitting example in subsection 4.3 illustrates the ability to interpret a fitted ν-Poisson model, by comparing the value of ν to the well known $0, 1$, and ∞ values (corresponding to a geometric, Poisson, and Bernoulli distribution). Moreover, since these three distributions are nested within the ν-Poisson, the hierarchical structure can be used to make inferences and comparisons. The Generalized Poisson distribution defined by Consul (1989) also has flexibility in terms of modeling underdispersion or overdispersion with respect to the Poisson. It has simple expressions for the normalizing constant and moments. However, it is not in the exponential family, which makes analysis more difficult. Numerical studies show that for every ν-Poisson with $\nu > 0.75$ (or so) there exists a Generalized Poisson with very similar distribution. But for $\nu < 0.75$ the two families differ markedly, especially for the example in section (4.3) where $\nu = 0.126$ matches the data well.

Besides its practical usefulness, the ν-Poisson has several appealing theoretical properties. As a member of several families of distributions, it inherits some favorable theoretical properties: First, as an exponential family member, the existence of 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 ν-Poisson's moments and tail were derived.

The infinite sum, $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{j!^\nu}$ is a generalization of several well known infinite sums (such as e^λ and $\frac{1}{1-\lambda}$). This function plays a major role in different computations that are needed in the ν-Poisson context. The ν-Poisson 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 ν-type distributions arise from the ν-Poisson. Although we described briefly only the ν-Binomial distribution, we have been studying some additional distributions (e.g., sums of ν-Poissons) and processes (the ν-Poisson process), and their applications. 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.

We have explored various numerical methods in order to learn about the ν-Poisson and characterize it. In some cases, the methods themselves shed light on the nature of the distribution. For example, when computing the predictive distribution (numerically), the more efficient methods for integration (i.e. transforming into log space) hinted at properties of the posterior function.

Although discovering "new" distributions is no longer in fashion, we have explored this new distribution using a modern approach that combines theory and numerical methods. This was possible only due to the advanced computational power available today. We have thus used modern tools to arrive at a new discrete distribution that generalizes some well-known distributions and is flexible enough for practical use.

A Variances and Covariances computations

Formulas (17) and (18) are derived by considering the sequence of probabilities ($p_x, x = 0, 1, ..., \infty$) to be independent random variables (except for the constraint that they sum to unity) and finding their posterior distribution given the data. Let n be the total number of samples and $n\hat{p}_x$ be the number of times we have seen value x. The latter is a Binomial variable with parameter p_x. Let

$$p_x = \frac{g_x}{\sum_{x=0}^{\infty} g_x} \qquad g_x \sim \Gamma(\alpha_x, 1) \qquad (g_x \text{ independent}) \tag{91}$$

where α_x is some positive sequence whose sum is finite: $\sum_x \alpha_x = \alpha$. Then the p_x's sum to 1 and their prior density is proportional to $\prod_x p_x^{\alpha_x - 1}$. The posterior density is proportional to $\prod_x p_x^{n\hat{p}_x + \alpha_x - 1}$, which we recognize as the distribution of

$$p_x = \frac{h_x}{\sum_x h_x} \qquad h_x \sim \Gamma(n\hat{p}_x + \alpha_x, 1) \qquad (h_x \text{ independent}) \tag{92}$$

Using the Delta method, $\operatorname{Var}\log(h_x)$ can be approximated by

$$\operatorname{Var}(\log(h_x)) \approx \frac{1}{n\hat{p}_x + \alpha_x} \tag{93}$$

which means

$$\begin{aligned}
\operatorname{Var}\left(\log\frac{p_{x-1}}{p_x}|Data\right) &= \operatorname{Var}(\log h_{x-1}|Data) + \operatorname{Var}(\log h_x|Data) && (94)\\
&\approx \frac{1}{n\hat{p}_{x-1} + \alpha_{x-1}} + \frac{1}{n\hat{p}_x + \alpha_x} && (95)\\
\operatorname{Cov}\left(\log\frac{p_{x-1}}{p_x}, \log\frac{p_x}{p_{x+1}}|Data\right) &= -\operatorname{Var}(\log h_x|Data) && (96)\\
&\approx -\frac{1}{n\hat{p}_x + \alpha_x} && (97)
\end{aligned}$$

The parameters α_x smooth the empirical counts. Taking the limit $\alpha_x \to 0$ removes the smoothing. If an empirical count is zero ($\hat{p}_x = 0$), then with no smoothing $\operatorname{Var}(\log(h_x)|Data) = \infty$ which implies omitting the bin from the weighted least squares procedure.

B Propriety of the conjugate density

In this section, we derive upper and lower bounds on the normalizing constant $\kappa^{-1}(a, b, c)$ for the ν-Poisson conjugate density. For the conjugate density to be proper, it is necessary that the lower bound be finite and sufficient that the upper bound be finite. The lower bound turns out to be finite only under condition (27), which is also sufficient for the upper bound to be finite. Therefore it is a necessary and sufficient condition for propriety.

B.1 Proof that (27) is a Necessary Condition

A lower bound on $\kappa^{-1}(a, b, c)$ is obtained via an upper bound on $Z(\lambda, \nu)$, which in turn comes from a lower bound on the factorial function. Because the log-gamma function is convex, i.e. has positive second derivative, we can lower bound the log-factorial with a linear function. In particular, the following bound is valid for all integers j and x:

$$\log j! \geq \log x! + (j - x)\log(x + 1) \tag{98}$$

The bound is tight at $j = x$ and $j = x + 1$. This gives the following bound $\overline{Z}(\lambda, \nu)$ on $Z(\lambda, \nu)$:

$$\tag{99}$$

$$Z(\lambda,\nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{j!^\nu} \leq \sum_{j=0}^{\infty} \frac{\lambda^j}{x!^\nu (x+1)^{(j-x)\nu}} = \overline{Z}(\lambda,\nu) \tag{100}$$

$$\overline{Z}(\lambda,\nu) = \frac{(x+1)^{x\nu}}{x!^\nu} \sum_{j=0}^{\infty} \left(\frac{\lambda}{(x+1)^\nu} \right)^j \tag{101}$$

$$= \begin{cases} \dfrac{(x+1)^{x\nu}}{x!^\nu} \dfrac{1}{1 - \lambda (x+1)^{-\nu}} & \text{if } \lambda < (x+1)^\nu \\ \infty & \text{otherwise} \end{cases} \tag{102}$$

From this we obtain a lower bound on the double integral:

$$\kappa^{-1}(a,b,c) \geq \int_0^\infty e^{-b\nu}\int_0^\infty \frac{\lambda^{a-1}}{\overline{Z}(\lambda,\nu)^c} d\lambda d\nu \tag{103}$$

$$= \int_0^\infty \frac{e^{-b\nu} x!^{c\nu}}{(x+1)^{cx\nu}} \int_0^{(x+1)^\nu} \lambda^{a-1}\left(1-\lambda(x+1)^{-\nu}\right)^c d\lambda d\nu \tag{104}$$

Change from λ to $\omega = \lambda(x+1)^{-\nu}$:

$$\kappa^{-1}(a,b,c) \geq \int_0^\infty \frac{e^{-b\nu} x!^{c\nu}(x+1)^{a\nu}}{(x+1)^{cx\nu}} d\nu \int_0^1 \omega^{a-1}(1-\omega)^c d\omega \tag{105}$$

The integral over ω is always finite. The integral over ν is finite only if

$$b/c > \log x! + (a/c - x)\log(x+1) \tag{106}$$

This condition is necessary for every x, including $x = \lfloor a/c \rfloor$, which gives condition (27).

B.2 Proof that (27) is a Sufficient Condition

An upper bound on $\kappa^{-1}(a,b,c)$ is obtained by breaking the integral into two parts and bounding each part:

$$\kappa^{-1}(a,b,c) = \int_0^\infty e^{-b\nu}\int_0^1 \frac{\lambda^{a-1}}{Z(\lambda,\nu)^c} d\lambda d\nu + \int_0^\infty e^{-b\nu}\int_1^\infty \frac{\lambda^{a-1}}{Z(\lambda,\nu)^c} d\lambda d\nu \tag{107}$$

$$= I_1 + I_2 \tag{108}$$

Since $Z(\lambda,\nu) \geq 1$, $I_1 \leq \int_0^\infty e^{-b\nu}\int_0^1 \lambda^{a-1} d\lambda d\nu$, which is finite for all $a > 0$ and $b > 0$. We use a different lower bound on $Z(\lambda,\nu)$ to handle I_2. Because $\log(x)$ is concave, i.e. the second derivative is always negative, we know from Jensen's inequality that

$$\log\left(\sum_{j=0}^\infty q_j a_j\right) \geq \sum_{j=0}^\infty q_j \log a_j \qquad \text{if } \sum_{j=0}^\infty q_j = 1 \tag{109}$$

Therefore by introducing variables q_j we have

$$\log Z(\lambda,\nu) = \log \sum_{j=0}^{\infty} q_j \frac{\lambda^j}{q_j (j!)^\nu} \geq \sum_{j=0}^{\infty} q_j \log \left(\frac{\lambda^j}{q_j (j!)^\nu} \right) \tag{110}$$

$$= \left(\sum_{j=0}^{\infty} j q_j \right) \log \lambda - \nu \left(\sum_{j=0}^{\infty} q_j \log(j!) \right) - \sum_{j=0}^{\infty} q_j \log q_j \tag{111}$$

Let Q be a random variable on the non-negative integers with probability mass function $Pr(Q = j) = q_j$. Then the bound can be written succinctly as

$$\underline{Z}(\lambda,\nu) = \lambda^{E[Q]} e^{-E[\log Q!]\nu} \prod_{j=0}^{\infty} q_j^{q_j} \tag{112}$$

Now we have an upper bound on the double integral:

$$I_2 \leq \int_0^{\infty} e^{-b\nu} \int_1^{\infty} \frac{\lambda^{a-1}}{\underline{Z}(\lambda,\nu)^c} d\lambda d\nu \tag{113}$$

$$= \int_0^{\infty} e^{-b\nu} e^{cE[\log(Q!)]\nu} d\nu \int_1^{\infty} \lambda^{a-1} \lambda^{-cE(Q)} d\lambda \prod_{j=0}^{\infty} q_j^{-cq_j} \tag{114}$$

This integral, and therefore $\kappa^{-1}(a,b,c)$, is finite if

$$E(Q) > a/c \quad \text{and} \quad E(\log Q!) < b/c \tag{115}$$

Given (27), we just have to show that there exists a distribution satisfying (115). Let

$$q_j = \begin{cases} 1 - (a/c - \lfloor a/c \rfloor + \epsilon) & \text{if } j = \lfloor a/c \rfloor \\ a/c - \lfloor a/c \rfloor + \epsilon & \text{if } j = \lfloor a/c \rfloor + 1 \\ 0 & \text{otherwise.} \end{cases} \tag{116}$$

Then

$$E[Q] = a/c + \epsilon \tag{117}$$

and

$$E[\log Q!] = \log\lfloor a/c \rfloor! + (a/c - \lfloor a/c \rfloor + \epsilon) \log(\lfloor a/c \rfloor + 1) \tag{118}$$

This q_j will satisfy (115) if we choose $\epsilon > 0$ small enough. Therefore (27) is a sufficient condition.

Bibliography

Barnard, J., McCulloch, R., and Xiao-Li, M. (2000) “Modelling covariance Matrices in Terms of Standard Deviations and Correlations with Application to Shrinkage”, *Statistica Sinica*, vol.10, no. 4, pp. 1281-1311 .

Bliss, C. I. and Fisher, R. A. (1953), “Fitting the Negative Binomial Distribution to Biological Data”, *Biometrics*, vol. 9, pp. 176-200.

Bleistein, N. and Handelsman, R. (1986), *Asymptotic Expansions of Integrals*, Dover, NY.

Boatwright, P., Borle, S., and Kadane, J.B. (2001) “ A Model of the Joint Distribution of Purchase Quantity and Timing”, Submitted for publication.

Breslow, N. (1990) “Tests of Hypotheses in Overdispersed Poisson Regression and Other Quasi-Likelihood Models”, *Journal of the American Statistical Association*, Vol. 85, pp. 565-571.

Chatfield, C., Ehrenberg, A. S. C., and Goodhardt, G. J. (1966), “Progress on a Simplified Model of Stationary Purchasing Behaviour”, *Journal of the Royal Statistical Society, Series A*, Vol. 129, No. 3, pp. 317-367.

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

Dean, C. B. (1992), “Testing for Overdispersion in Poisson and Binomial Regression Models”, *Journal of the American Statistical Association*, Vol. 87, pp. 451-457.

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

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

Kadane J. B. and Wolfson, L. J. (1997), “Experiences in Elicitation”, *The Statistician*, Vol. 47, pp. 3-19.

Maceda, E. C. (1948) “On the Compound and Generalized Poisson Distributions”, *Annals of Mathematical Statistics*, Vol. 19, No. 3. (Sept.), pp. 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*, vol. 37, pp. 259-269.

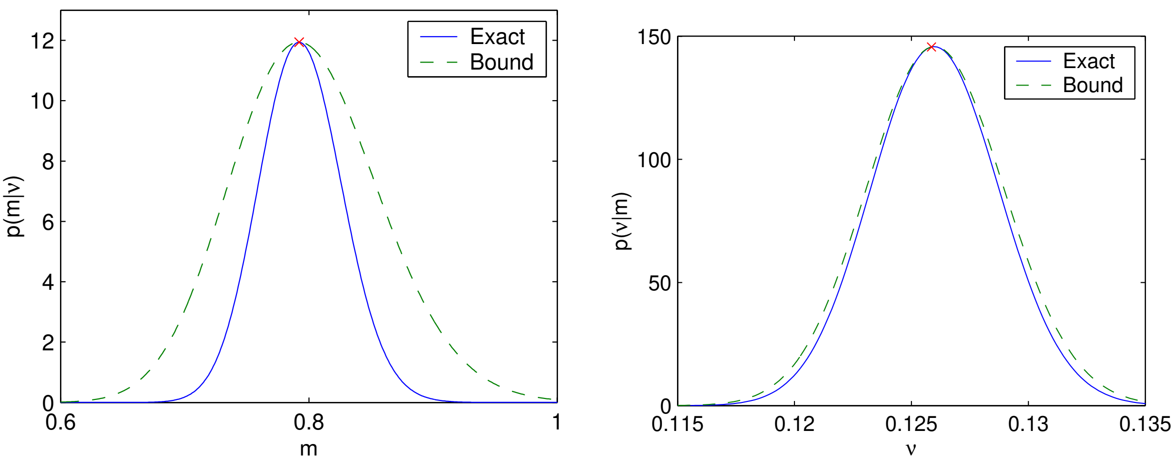
Mood, A. M., Graybill, F. A., Boes, D. C. (1974), *Introduction to the Theory of Statistics*, 3rd Edition, McGraw Hill Kogakusha.

Riordan J. (1968), *Combinatorial Identities*, Wiley NY.

Satterthwaite, F.E. (1942) ”Generalized Poisson Distribution”, *Annals of Mathematical Statistics*, Vol. 13, No. 4. (Dec.), pp. 410-417.

Schervish M. J. (1995), *Theory of Statistics* Springer-Verlag.

Figure 1: Dominating curves derived in this section for the conditional distributions $p(m|\nu)$ and $p(\nu|m)$, using the example data of section 4.3. The contact points m_0 and ν_0 are marked with an 'x'.



(a) (b)

Figure 2: The predictive distribution $P(X = x|a = 1, b = 1, c = 1)$, obtained using the web application.

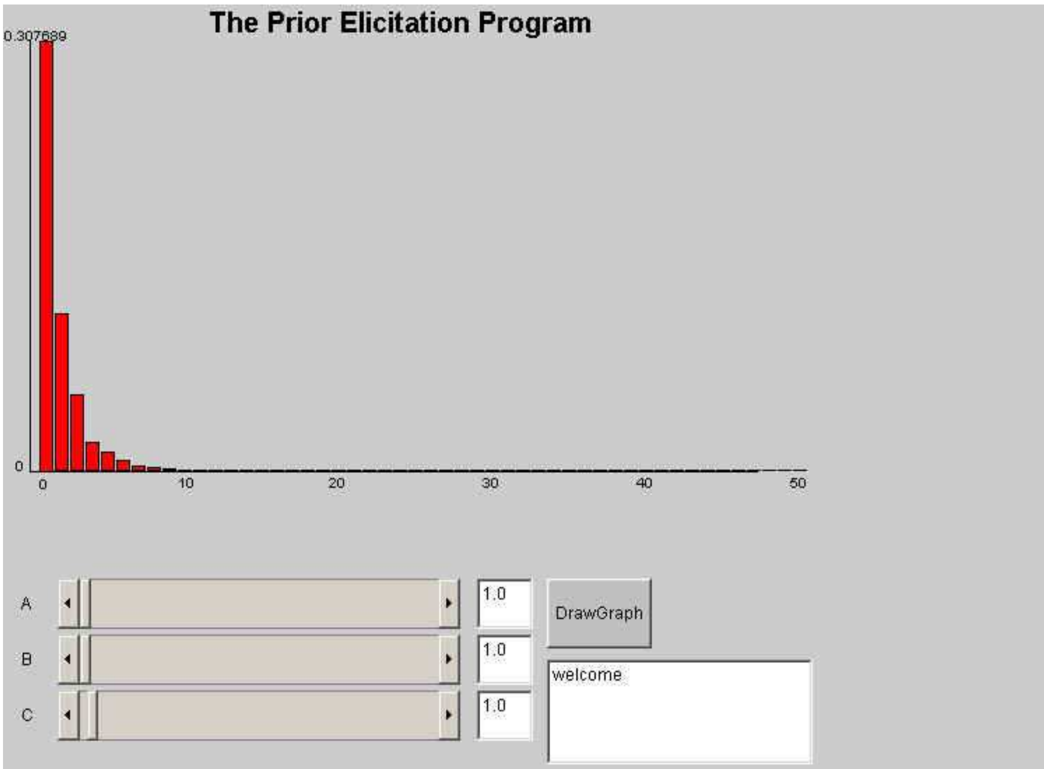


Figure 3: Fit of a ν-Poisson distribution to the quarterly sales data using the graphical method from section 2.6. The log-ratios of successive counts $(\log \hat{p}_{x-1}/\hat{p}_x)$ are plotted vs. $\log x$. The ν-Poisson fit at the WLS estimates is compared to an ordinary Poisson distribution.

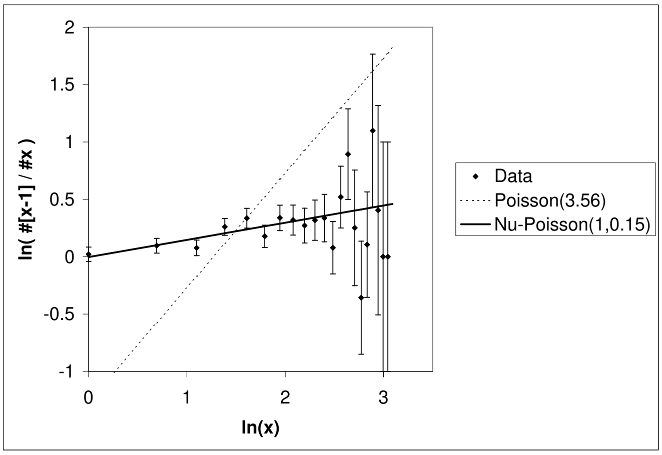
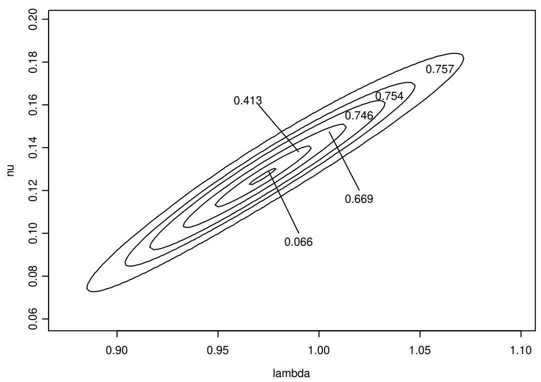


Figure 4: Contour and 3D plots of the posterior for the quarterly sales data, using a conjugate prior with $a = b = c = 1$. The values on the contours and the third axis are the cumulative probability from the maximum of the distribution downwards. $\nu = 1$ is clearly not near the mass of the distribution.



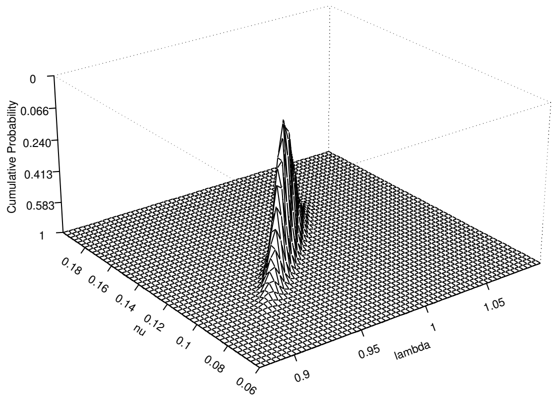
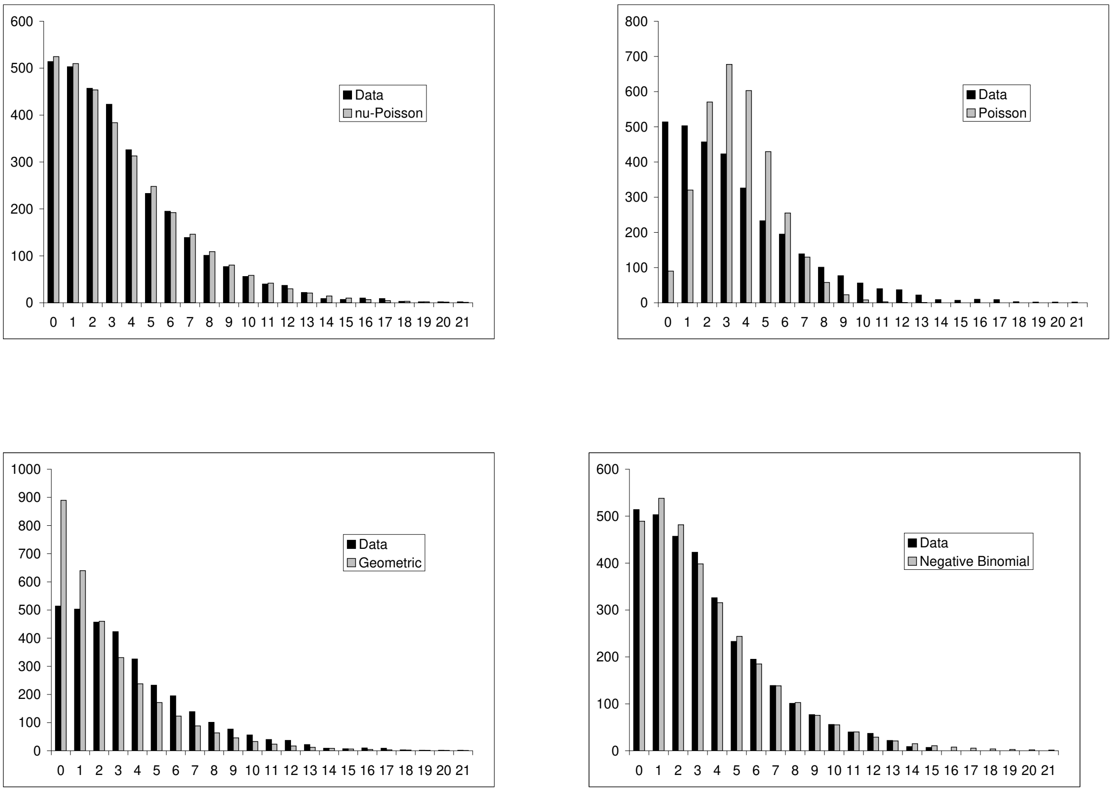


Figure 5: Comparison of the quarterly sales counts with the expected counts for four distributions: The Poisson, geometric, Negative binomial, and ν-Poisson. In all cases the ML estimates were used. The ν-Poisson seems to follow the shape and the actual counts closer than any of the other distributions



Captions for Figures

Figure 1: Dominating curves derived in this section for the conditional distributions $p(m|\nu)$ and $p(\nu|m)$, using the example data of section 4.3. The contact points m_0 and ν_0 are marked with an 'x'.

Figure 2: The predictive distribution $P(X = x|a = 1, b = 1, c = 1)$, obtained using the web application.

Figure 3: Fit of a ν-Poisson distribution to the quarterly sales data using the graphical method from section 2.6. The log-ratios of successive counts ($\log \hat{p}_{x-1}/\hat{p}_x$) are plotted vs. $\log x$. The ν-Poisson fit at the WLS estimates is compared to an ordinary Poisson distribution.

Figure 4: Contour and 3D plots of the posterior for the quarterly sales data, using a conjugate prior with $a = b = c = 1$. The values on the contours and the third axis are the cumulative probability from the maximum of the distribution downwards. $\nu = 1$ is clearly not near the mass of the distribution.

Figure 5: Comparison of the quarterly sales counts with the expected counts for four distributions: The Poisson, geometric, Negative binomial, and ν-Poisson. In all cases the ML estimates were used. The ν-Poisson seems to follow the shape and the actual counts closer than any of the other distributions