JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods Inquires

APPROXIMATING THE POISSON PROBABILITY DISTRIBUTION BY THE CONWAY-MAXWELL POISSON DISTRIBUTION

N. E. ARUA
Graduate Student
Department of Statistics, University of Botswana

E-mail: eke2892@yahoo.com

R. M. SAKIA
Associate Professor of Statistics
Department of Statistics, University of Botswana

E-mail: sakiar@mopipi.ub.bw

Abstract

The aim of this research was to approximate the Poisson distribution by the COM-Poisson as a way to induce equi-dispersion in the model and hence, make some inferences by taking advantage of the close-form moments of the Poisson distribution. This was achieved by relating the approximate moments of the COM-Poisson distribution to that of the Poisson distribution to determine the relationship between their respective parameters. The estimates of the Poisson parameters were found to induce equi-distribution to the observed data. The advantage of the estimation is that closed-form moments of the Poisson can then be used to make inferences on the data. It is recommended that the COM-Poisson distribution should be applied to induce equi-distribution when the data does not conform to the Poisson distribution.

Keywords: *Poisson distribution, COM-Poisson distribution, equi-dispersion*

Introduction

When counting items that arise independently of one another at random in space (or time), the Poisson distribution may be appropriate. If the items (for example, insects, lesions on a leaf, or weeds) occur at a constant average rate of θ per unit area, and if a large number of unit areas are counted, the actual number of items in each unit being X, then the distribution of X follows the Poisson distribution. If the rate per unit area θ does not remain constant over a complete population of units being studied then the Poisson distribution will not be a suitable model. Another case where it does not work is when the items being counted are not fully independent of one another but tend to arise in groups. Theoretically, the mean and variance of the Poisson are equal, so we can say that the Poisson distribution has equal dispersion or equi-dispersion. According to Sellers, Borle and Shmueli (2012), this is hardly the case in real life count data. This has led to the popularization of the negative

binomial distribution which can capture over-dispersion. Feng-Chang and Bo-Cheng (2010) declare that for over dispersed-data, the negative binomial can be used, and, further, that the generalized Poisson regression model is one of the few distributions that can be used for both under- and over-dispersed count data. Berk and MacDonald (2008) conclude that if apparent over-dispersion results from specification errors in the systematic part of the Poisson model, resorting to the negative binomial distribution does not help. It can make things worse by giving a false sense of security when the fundamental errors in the model remain.

It is ideal to avoid over-dispersion or under-dispersion in count data. Hinde and Demetrio (1998) discuss the consequences of over-dispersion. They state that, firstly, the standard errors obtained from the model will not be correct and may be seriously underestimated and, consequently, that we may incorrectly assess the significance of individual regression parameters. Also, changes in deviance associated with model terms will be too large and this will lead to the selection of overly complex models. Finally, our interpretation of the model will be incorrect and any predictions will be too precise. Cameron and Trivedi (2001) state that over-dispersed and under-dispersed data will lead to the standard errors of model parameters being inconsistent.

Of particular focus in this study is the Conway-Maxwell-Poisson (COM-Poisson). The Conway-Maxwell-Poisson (COM-Poisson) model is another such technique for such count data. The distribution was first introduced in 1962 by Richard W. Conway and William L. Maxwell, but only recently have the statistical and probabilistic properties of the distribution been published by Shmueli, Minka, Kadane, Borle and Boatwright (2005). So it can be said to be a relatively new distribution. The COM-Poisson distribution adds a new parameter ν which governs the rate of decay of successive probability ratios (Shmueli et al, 2005). Since then, further advancements on the distribution have been produced. Sellers and Shmueli (2010) used COM-Poisson regression to predict censored count data. Lord, Guikema and Geedipally (2008) applied the generalized COM-Poisson linear model to the analysis of motor vehicle crashes using a flexible GLM that could model both under-dispersed and over-dispersed data sets. Rodrigues et al (2009) developed a flexible cure rate survival model by assuming that the number of competing causes of the event of interest follows a COM-Poisson distribution.

The assumption of equi-distribution of a Poisson data does not hold in most experimental situations. To circumvent this problem, the COM-Poisson probability model has been proposed because of its assumed ability to remedy the violation of equi-distribution. A major hurdle in its use is the lack of a closed form moment generating function for which the exact moments could be obtained (Shmueli et al, 2005). This problem also arises in deriving some closed form estimates of the model parameters (for example, MLEs) as well as deriving some inferential results from the model, such as, best critical region of a test and tests of hypotheses. An attempt is therefore made to relate the COM-Poisson parameters to the Poisson parameters by using the approximate moments of the COM-Poisson in order to induce equi-distribution to some count data.

Methodology

Poisson Model

The Poisson distribution is a discrete probability distribution used to describe the number of occurrences in a given small interval of time and/or space if these events occur

with a known average rate and the occurrence of one event is independent of the occurrence of others.

The probability mass function of the Poisson distribution is:

$$P(X = x) = \frac{\theta^x e^{-\theta}}{x!} \qquad x = 0,1,2,\ldots, \quad \theta > 0 \qquad (1)$$

The mean and variance of the Poisson distribution is given by:

$$E(X) = Var(X) = \theta \qquad (2)$$

The limitation of the Poisson model is that it requires the variance to be equal to the mean which, as was stated earlier, is hardly satisfied in real life count data.

COM-Poisson Model

The COM-Poisson probability function according to Shmueli et al (2005) is given as:

$$P(X = x) = \frac{\lambda^x}{(x!)^\nu} \frac{1}{Z(\lambda, \nu)} \qquad \lambda > 0, \quad \nu \geq 0 \quad x = 0,1,2,\ldots, \qquad (3)$$

where $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu} \qquad \lambda > 0, \ \nu \geq 0 \qquad (4)$

This satisfies the conditions for a probability function. The formulation allows for a non-linear decrease in the ratios of successive probabilities in the form:

$$\frac{P(X = x-1)}{P(X = x)} = \frac{x^\nu}{\lambda} \qquad (5)$$

υ is the shape parameter of the COM-Poisson distribution. The condition $\upsilon > 1$ corresponds to under-dispersed data, $\upsilon < 1$ to over-dispersed data, and $\upsilon = 1$ to equi-dispersed (Poisson) data. The series $\frac{\lambda^j}{(j!)^\nu}$ converges for any $\lambda > 0$ and $\nu > 0$ as the ratio of the subsequent terms of the series $\frac{\lambda}{j^\upsilon}$ tends to 0 as $j \to \infty$.

The first central moment of the COM-Poisson distribution is given by:

$$E(X) = \frac{\partial log Z}{\partial log \lambda} \qquad (6)$$

The second central moment is given by:

$$Var(X) = \frac{\partial^2 log Z}{\partial log^2 \lambda} \qquad (7)$$

The COM-Poisson distribution does not have a closed-form expression for its moments in terms of the parameters λ and ν. By using an asymptotic expression for Z in (4) the mean and variance can be approximated (Shmueli et al, 2005) in the form:

$$E(X) \approx \lambda^{1/\upsilon} + \frac{1}{2\upsilon} - \frac{1}{2} \qquad (8)$$

$$Var(X) \approx \frac{1}{\nu} \lambda^{1/\nu} \qquad (9)$$

The approximations are especially good for $\nu \leq 1$ and $\lambda > 10^\nu$ (Shmueli et al, 2005).

Estimation of the poisson parameter

The moments of the COM-Poisson will be related to that of the Poisson distribution to determine the relationship between them by equating equation (2) to (8) and (9) as:

$$\theta = \lambda^{1/\nu} + \frac{1}{2}\left(\frac{1}{\nu} - 1\right) \qquad (10)$$

$$\theta = \frac{1}{\nu}\lambda^{1/\nu} \quad (11)$$

Solving for θ in terms of λ as well as ν:

$$\hat{\theta}_1 = \frac{1}{2\upsilon} \quad (12)$$

$$\hat{\theta}_2 = \frac{0.5\ln(0.5)}{ln\lambda} \quad (13)$$

As noted by Shmueli et al (2005), a simple and computationally efficient method of finding estimates of λ and ν is the linearizing of equation (5) as:

$$ln\left[\frac{P(X=x-1)}{P(X=x)}\right] = -ln\lambda + \nu\ln(x) \quad (14)$$

Ignoring all counts with zero frequencies in the data, the ratio $r = \frac{P(X=x-1)}{P(X=x)}$ is to be computed and displayed in Table 1.

Table 1. A layout of the data

X	x_1	x_2	…	x_n
P(x)	p_1	p_2	…	p_n
r	-	r_1	…	r_{n-1}

A simple linear regression of (14) will enable one to obtain estimates of $\hat{\lambda}$ and $\hat{\nu}$ where $P(X=x)$ and $P(X=x-1)$ are replaced by the respective relative frequencies. Note that $p_i = \frac{f_i}{\sum f_i}$ where f_i is the corresponding non-zero frequency.

Considering the two estimates of θ, then the corresponding Poisson distributions are:

$$P_j(X=x) = \frac{e^{-\hat{\theta}_j}\hat{\theta}_j{}^x}{x!} \text{ for j=1,2} \quad (15)$$

We compute the estimates of the probabilities for the two cases above and deduce the corresponding frequencies $\hat{f}_i$. Then we compute the mean and variance of the Poisson using the estimates $\hat{f}_i$ for the two cases and note the extent of the closeness (or lack of it) of the mean and variance.

Description of the datasets

The following data sets were used in the study because of the varying inherent different levels of dispersion.

Dataset 1

The data set consists of quarterly sales of a well known brand of a particular article of clothing at stores of a large national retailer. This data set was published by Shmueli et al (2005) and is available at http://www.stat.cmu.edu/COM-Poisson/sales-data.html

Dataset 2

Gilchrist (1984) refers to an experiment in which a total of 33 insect traps were set out across sand dunes and the number of insects caught in a fixed time was counted. The data consists of the number of traps containing various numbers of the taxa ***staphylinoidea***.

Dataset 3

The data gives the fertility of eggs of the CP strain of ***Drosophila melanogaster*** raised in 100 vials of 10 eggs in a study by Sokal (1966) and reproduced in Sokal and Rohlf (2003; pp. 96)

Dataset 4

It is well known that there is a tendency for unisexual sibships to result in a clumped distribution of observed frequencies. In an extensive study by Geissler (1889), the sex ratio of 6115 sibships of 12 children were recorded from actual hospital records in Saxony, Germany. The data consists of the number of females per sibship X. The data is reproduced in Sokal and Rohlf (2003; pp 80)

Results and discussion

Dispersion of the original data

Table 2 gives the extent of dispersion of the raw datasets on the basis of their mean and variance. Note that over-dispersion occurs when the variance exceeds the mean.

Table 2. Means and Variances for the data sets

Dataset	Mean	Variance	**Dispersion**
Shmueli	3.56	11.31	Over-dispersed
Gilchrist	1.64	2.74	Moderately Over-dispersed
Geissler	5.77	3.49	Under-dispersed
Sokal	5.91	5.56	Moderately under-dispersed

Estimates for λ, ν and the corresponding θ

The estimates were obtained using the regression run in equation (14) and substituted in equations (15) and (16). The results are shown in Table 3 below.

Table 3. Estimates of λ, ν and the corresponding $\theta's$

Data Set	Estimate of ν	Estimate of λ	Estimate of θ_1	Estimate of θ_2
Shmueli	0.135	0.887	3.704	2.890
Gilchrist	0.109	0.768	4.587	1.310
Geissler	1.476	10.890	0.339	-0.145
Sokal	0.557	2.889	0.897	-0.330

It should be noted that the $\hat{\theta}_2$ estimates for the Geissler and Sokal dataset do not provide valid estimated of a Poisson parameter since they are negative and hence are ignored.

Assessment of the means and variances for the estimated Poisson model

The means and variances for all the data sets were recomputed using the estimated probability for a Poisson parameter $\hat{\theta}$. The results are presented in Table 4 for $\hat{\theta}_1$ and Table 5 for $\hat{\theta}_2$.

Table 4. Means and Variances for the datasets after calculating frequency estimates for $\hat{\theta}_1$

Data Set	Mean	Variance
Shmueli	3.704	3.705
Gilchrist	1.310	1.340
Geissler	0.339	0.339
Sokal	0.910	0.907

Table 5. Means and Variance for the data sets after calculating frequency estimates for $\hat{\theta}_2$

Data set	Mean	Variance
Shmueli	2.89	2.89
Gilchrist	3.156	3.173
Geissler	-	-
Sokal	-	-

In comparison to results in Table 2, the empirical results clearly shows that the Poisson parameters estimated by the COM –Poisson to have effectively induced the equi-distribution property of the Poisson probability distribution which is a prerequisite for analysing count data which is assumed to follow that distribution.

Conclusions

Despite its usefulness when it comes to handling count data, the Poisson distribution is impractical to use because its assumptions of equi-distribution are rarely met in real-life count data. The COM-Poisson has been found to be flexible when handling count data as it caters for both over- and under-dispersion. However, a major deficiency for the distribution is the lack of closed form moments which in turn renders it impossible for use in testing hypotheses about the parameters λ and ν. For example, the test statistic for the Neyman-Pearson lemma is impossible to derive. In this case, the test may be approximated by the estimated Poisson model. In view of the results in this study it is recommended that care be taken when analyzing data that is deemed to follow a Poisson process. Exploratory analysis should be undertaken to check whether the data indeed conform to the Poisson distribution. If not, then the COM-Poisson distribution should be applied to induce equi-distribution which is a key requirement for any Poisson process.

References

1. Berk, R. and MacDonald, J. **Overdispersion and Poisson Regression**. Journal of Quantitative Criminology, Vol. 24, No. 3, 2008, pp. 269-284
2. Cameron, A. C. and Trivedi, P. K. **Essentials of Count Data Regression**, in Baltagi, H. B. (Ed.) "A Companion to Theoretical Econometrics", Blackwell Publishing, 2001
3. Feng-Chang, X. and Bo-Cheng, W. **Influence analysis for count data based on generalized Poisson regression models,** Statistics, Vol. 44, 2010, pp. 341–360
4. Geissler, A. **Beitraege zur Frage des Geschlechtshaeltisses der Geborenen**. Z. K. Saechs. Stat. Bur., Vol. 35, 1889, pp. 1-24
5. Gilchrist, W. **Statistical modelling,** Statistical Modelling, London: Wiley, 1984
6. Hinde, J. and Demetrio, C. G. **Overdispersion: Models and Estimation,** Computational Statistics & Data Analysis, Vol. 27, No. 2, 1998, pp. 151-170
7. Lord, D., Guikema, S. D. and Geedipally, S. R. **Application of the Conway-Maxwell-Poisson Generalized Linear Model for Analyzing Motor Vehicle Crashes,** Accident Analysis & Prevention, Vol. 40, No. 3, 2008, pp. 1123–1134
8. Rodrigues, J., de Castro, M., Cancho, V. G. and Balakrishnan, N. **COM-Poisson cure rate survival models and an application to cutaneous melanoma data,**

Journal of Statistical Planning and Inference, Vol. 139, No. 10, 2009, pp. 3605-3611

9. Sellers, K. F. and Shmueli, G. **Predicting Censored Count Data with COM-Poisson Regression,** Robert H. Smith School of Business, 2010
10. Sellers, K. F., Borle, S. and Shmueli, G. **The COM-Poisson Model for Count Data: A Survey of Methods and Applications,** Applied Statistic Models in Business and Industry, Vol. 28, No. 2, 2012, pp. 104-116
11. Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. and Boatwright, P. **A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution,** Applied Statistics, 2005, pp. 127-142
12. Sokal, R. R. **Pupation site difference in Drosophila melanogaster,** Univ. Kansas Sci. Bull., Vol. 46, 1966, pp. 697-715
13. Sokal, R. R. and Rohlf, F. J. **Biometry: The Principles and Practice of Statistics in Biological Research,** New York: W.H. Freeman and Company, 2003
14. Winkelmann, R. and Zimmerman, K. F. **Recent Developments in Count Data Modelling: Theories and Applicatons,** Journal of Economic Surveys, Vol. 9, No. 1, 1995, pp. 1-24
15. Zou, Y. **Over- And Under-Dispersed Crash Data: Comparing the Conway-Maxwell-Poisson and Double-Poisson Distributions,** Msc Thesis, Texas A&M University, 2012