A NOTE ON A MOMENT TEST OF FIT FOR A MIXTURE OF TWO POISSON DISTRIBUTIONS

D.J. BEST AND J.C.W. RAYNER

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845.
Email: anica@uow.edu.au

A NOTE ON A MOMENT TEST OF FIT FOR A MIXTURE OF TWO POISSON DISTRIBUTIONS

D.J. BEST[1] AND J.C.W. RAYNER[2,1],

University of Newcastle[1], *University of Wollongong*[2,1]

Summary

In this note a new moment test of fit for a mixture of two Poisson distributions is derived. The test is illustrated with (1) a classic data set of deaths per day of women over 80 as recorded in the Times newspaper for the years 1910 to 1912 and (2) a more recent data set concerned with foetal lamb movements. A small indicative size and power study is given.

Keywords: Central moments; deaths of London women data; factorial moments; foetal lamb movements; Pearson X^2 test; zero-inflated Poisson.

1. Introduction

A Poisson process is often used to model count data. Sometimes an underlying mechanism suggests two Poisson processes may be involved. This may be modelled by a two component Poisson mixture model. We will give some examples later. The Poisson probability function, $f(x;\theta)$ say, is given by

$$f(x; \theta) = \exp(-\theta)\theta^{x}/x!,\ x = 0, 1, 2, ..., \text{ in which } \theta > 0,$$

and the two component Poisson mixture model has probability function

$$f^*(x; \theta_1, \theta_2, p) = p\, f(x; \theta_1) + (1 - p)\, f(x; \theta_2),\ x = 0, 1, 2, ...,$$

$$\text{in which } \theta_1 > 0,\ \theta_2 > 0,\ \theta_1 \neq \theta_2 \text{ and } 0 < p < 1.$$

A common test of fit for $f^*(x; \theta_1, \theta_2, p)$ is based on the well-known Pearson's X^2 statistic. If there are l classes X^2 is approximately distributed as χ^2 with $l - 4$ degrees of freedom: $\chi^2_{l\text{-}4}$. A problem with the X^2 test is that different decisions about the suitability of a null distribution can arise with different class pooling.

[1] School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia
e-mail: John.Best@newcastle.edu.au
[2] National Institute for Applied Statistics Research Australia, University of Wollongong, NSW 2522, Australia

A common approach for estimating θ_1, θ_2 and p is based on the method of moments (MOM). If we have n data points x_1, x_2, ... , x_n, $\bar{x} = \sum_{i=1}^{n} x_i / n$ and m_t = $\sum_{i=1}^{n} (x_i - \bar{x})^t / n$, t = 2, 3, ... , the MOM estimators satisfy

$$\bar{p} = \frac{\bar{x} - \tilde{\theta}_2}{\tilde{\theta}_1 - \tilde{\theta}_2},\ \tilde{\theta}_1 = \frac{A - D}{2} \text{ and } \tilde{\theta}_2 = \frac{A + D}{2}$$

in which

$$A = 2\bar{x} + (m_3 - 3m_2 + 2\bar{x})/(m_2 - \bar{x}) \text{ and } D^2 = A^2 - 4A\bar{x} + 4(m_2 + \bar{x}^2 - \bar{x}).$$

This method clearly fails if $D^2 < 0$, if any of θ_1, θ_2 and p are outside their specified bounds, or if m_2 = $\bar{x}$. When the MOM estimates are invalid because one or more of these conditions fail we suggest the mixture of two Poissons model may be inappropriate.

In the following section 2 derives the moment test, section 3 gives two examples and section 4 gives a small size and power study.

2. A New Fourth Moment Test

Consider the statistic $T = m_4 - \tilde{\mu}_4$ where $\tilde{\mu}_4$ is the fourth central moment μ_4, in which all unknown parameters are estimated by their MOM estimators. We need to find var(T) and then T^* = T^2/var(T) is a generalized smooth test as in Rayner et al. (2009, Chapter 11). This means T^* will have some optimum properties and an asymptotic χ_1^2 distribution. Observe that because MOM estimators have been used the first, second and third order generalised smooth test components are all zero. It is straightforward to show that the tth descending factorial moment $\mu'_{[t]}$ of $f^*(x;\ \theta_1,\ \theta_2,\ p)$ is given by $\mu'_{[t]} = p\theta_1^t + (1-p)\theta_2^t$ and that the moments about the origin, μ'_t say, can then be derived using $\mu'_t = \sum_{j=1}^{t} S(t,j)\mu'_{[t]}$ where $S(t, j)$ are the Stirling numbers of the second kind. A table of these numbers is given, for example, by Abramowitz and Stegun (1965, p.835). Then, using the well-known relation $\mu_t = \sum_{j=1}^{t} (-1)^j\, {}^tC_j \mu'_{t-j} \mu'^j$, the central moments, μ_t, can be obtained.

Define

$$\frac{\partial f}{\partial x} = \frac{4\tilde{\mu}^3 - 12\tilde{\mu}^2\tilde{\mu}_2 + 12\tilde{\mu}\tilde{\mu}_2^2 - 2\tilde{\mu}^2 + 4\tilde{\mu}\tilde{\mu}_2 - 4\tilde{\mu}_3^2 + 2\tilde{\mu}_2\tilde{\mu}_3 - 3\tilde{\mu}_2^2 - \tilde{\mu}_3^2}{(\tilde{\mu} - \tilde{\mu}_2)^2},$$

$$\frac{\partial f}{\partial y} = \frac{6\tilde{\mu}^2\tilde{\mu}_2 - 4\tilde{\mu}^3 - 4\tilde{\mu}\tilde{\mu}_2 + 3\tilde{\mu}^2 - 2\tilde{\mu}_2^3 + 2\tilde{\mu}_2^2 - 2\tilde{\mu}\tilde{\mu}_3 + \tilde{\mu}_3^2}{(\tilde{\mu} - \tilde{\mu}_2^2)^2} \text{ and}$$

$$\frac{\partial f}{\partial z} = \frac{(\tilde{\mu}_3 - \tilde{\mu})}{(\tilde{\mu} - \tilde{\mu}_2)}.$$

Then using the delta method

$$n\ \mathrm{var}(T) = \delta^{\mathrm{T}}\sum\delta$$

in which $\delta^{\mathrm{T}} = (\partial f/\partial x, \partial f/\partial y, \partial f/\partial z, 1)$ and $\sum$ is the variance-covariance matrix of $x = \bar{x}$, $y = m_2$, $z = m_3$ and m_4 evaluated using the MOM estimates. Note $\partial f/\partial x$ equals the partial derivative of T with respect to x, etc., evaluated at the expected values of x, y and z. Stuart and Ord (2005, section 10.5), for example, give details of the delta method.

3. Examples

(1) *Deaths of London Women During 1910 to 1912*

A classic data set, possibly first considered in connection with a mixture of two Poisson distributions by Schilling (1947), considers deaths per day of women over 80 in London during the years 1910, 1911 and 1912 as recorded in the Times newspaper. Table 1 shows the data and expected counts for $(\tilde{\theta}_1, \tilde{\theta}_2, \tilde{p})$ = (1.10, 2.58, 0.29). Possibly due to different death rates in summer and winter, T^* = 0.29 indicates a good fit by a mixture of two Poisson distributions. If a single Poisson is used to describe the data then X^2 = 27.01 with a χ^2_4 p-value of less than 0.01.

Table 1. Deaths per day of London women over 80 during 1910 to 1912

Number of deaths	0	1	2	3	4	5	6	7	8	9
Count	162	267	271	185	111	61	27	8	3	1
Mixture expected	161	271	262	191	114	58	25	9	3	1
Poisson expected	127	273	295	212	114	49	18	5	1	0

(2) *Foetal Lamb Movements*

Douglas et al. (1994) fitted the zero-inflated Poisson (ZIP) distribution to the data on foetal lamb movements shown in Table 2. The ZIP model is defined for $x = 0$ as $g(0; \lambda, \omega) = \omega + (1 - \omega)\exp(-\lambda)$ and for $x = 1, 2, \ldots$ as $g(x; \lambda, \omega) = (1 - \omega)\exp(-\lambda)\lambda^x/x!$. Douglas et al. (1994) used an X^2 test where rejection of the ZIP model is not clear. This is because the one observation of seven movements has been pooled with the latter classes and information has been lost.

Table 2

Frequencies of foetal lamb movements

Outcome	0	1	2	3	4	5	6	7
Frequency	182	41	12	2	2	0	0	1

For the Table 2 data Rayner et al. (2009, p.237) calculate a generalized smooth statistic V_3^{*2} for assessing the ZIP model. As an aside we note that V_3^{*2} is $(m_3 - \tilde{\mu}_3)^2 / \{2(1-\tilde{\omega})\tilde{\lambda}^3(\tilde{\lambda}+3)\}$ where μ_3, ω and λ are evaluated using MOM estimators for the ZIP. Use of V_3^{*2} avoids the pooling for the X^2 test noted above. The ZIP only fits well if the count of 7 is removed: with the count of 7 included the p-value is less than 0.01 and with this count removed the p-value is 0.34.

Are the Table 2 data fitted well by a mixture of two Poisson distributions? For the two Poisson mixture $\tilde{\theta}_1$ = 0.247, $\tilde{\theta}_2$ = 3.032 and $\tilde{p}$ = 0.960 with T* = 0.076 and p-value 0.74 based on 10,000 simulations. The mixture of two Poisson distributions is an excellent model even with the count of seven included. This suggests two biological mechanisms are needed to explain the Table 2 data.

For the deaths of women data the chi-squared p-value is 0.59 and the parametric bootstrap p-value is 0.53 when 10,000 samples of size *n* are used. For the foetal lamb data chi-squared p-value is 0.78 and parametric bootstrap p-value is 0.72. In both examples there is reasonable agreement. For the deaths of women example Suesse et al. (2015) give a parametric bootstrap p-value of 0.47 for their fourth order component compared with our parametric bootstrap p-value of 0.53.

4. Indicative Sizes and Powers

For nominal α = 0.05, θ_1 = 2.0, θ_2 = 5.0 and p = 0.5 Table 3 shows estimates of actual sizes for the chi-squared approximation with 1 degree of freedom. These actual estimates were found using 100,000 Monte Carlo samples of size *n*. It appears the chi-squared approximation for T^* is reasonable for *n* > 500 and quite good for *n* > 5000. Other similar calculations, not shown, are in agreement with this suggestion.

Table 3

Estimated actual test sizes of T^* for α = 0.05, θ_1 = 2.0, θ_2 = 5.0 and p = 0.5

n	100	200	500	1000	5000	10000
Size	0.005	0.017	0.026	0.035	0.046	0.049

Table 4 gives a small indicative power comparison of the fourth order MOM based T^* statistic with the fourth order MLE based $\hat{V}_4^2$ component of Suesse et al (2015). We use a negative binomial alternative, NB(*k*, p), and a Neyman Type A alternative, NTA(λ_1, λ_2). Critical values used were 0.56 for $\hat{V}_4^2$ and 1.40 for T*.

Table 4
100 * powers based on 10,000 Monte Carlo samples for n = 100, α = 0.05, θ_1 = 2.0, θ_2 = 5.0 and p = 0.5

Alternative	$\hat{V}_4^2$	T^*
NB(2, 0.4)	40	35
NB(3, 0.5)	20	18
NB(4, 0.5)	24	20
NTA(1, 2)	45	36
NTA(2, 2)	55	55
NTA(2, 1)	22	14
NTA(1, 3)	70	66

Generally the $\hat{V}_4^2$ powers are marginally better but T^* has two advantages: its sampling distribution may be approximated by a chi-squared distribution, and a large T* implies an alternative probability model that could differ in the fourth moment. In the two examples above, p-values for $\hat{V}_4^2$ and T* were similar.

References

Abramowitz, M. and Stegun, I. (1965). *Handbook of Mathematical Functions*. Dover, New York.

Douglas, J., Leroux, B. and Puterman, M. (1994). Empirical fitting of discrete distributions. *Biometrics,* **50**, 576-9.

Rayner, J.C.W., Thas, O. and Best D.J. (2009). *Smooth Tests of Goodness of Fit: Using R* (2nd ed.).Singapore: Wiley.

Schilling, W. (1947). A frequency distribution represented as the sum of two Poisson distributions. *Journal of the American Statistical Association*, **42**, 407-24.

Stuart, A. and Ord, K. (2005). *Kendall's Advanced Theory of Statistics, Volume 1.* (6th ed.) Hodder-Arnold, London.

Suesse, T., Rayner, J.C.W. and Thas, O. (2015). Smooth tests of fit for finite mixture distributions. Working Paper 04/15, NIASRA, University of Wollongong, Wollongong NSW.