NIASRA
NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong

Working Paper

9-17

IMPROVED TESTING FOR THE ZERO INFLATED POISSON USING CHI-SQARED COMPONENTS WITH DATA DEPENDENT CELLS

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

IMPROVED TESTING FOR THE ZERO INFLATED POISSON AND NEGATIVE BINOMIAL USING CHI-SQUARED COMPONENTS WITH DATA DEPENDENT CELLS

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

University of Newcastle and University of Wollongong

Summary

A non-iterative formula for the maximum likelihood estimators of the zero inflated Poisson distribution is given. Examples of real data illustrate how a more sensitive test of goodness of fit uses the components of the Chernoff-Lehmann X^2 test statistic.

Key Words: cold spells; maximum likelihood estimation; partition of X^2; possum abundance; slug counts.

1. Introduction

The zero-inflated Poisson (ZIP) distribution has probability function, $f(x;\, \lambda,\, \omega)$, given by

$$f(x) = \omega + (1 - \omega)e^{-\lambda} \text{ for } x = 0 \text{ and } f(x) = (1 - \omega)\lambda^x e^{-\lambda}/x!$$
$$\text{for } x = 1, 2, \ldots \text{ in which } \lambda > 0 \text{ and } 0 < \omega < 1.$$

The distribution has been used to model animal abundance data as, for example, for the data discussed below. Also, McKendrick (1926) used the ZIP to estimate the proportion of households infected with cholera but not diagnosed in an Indian epidemic. Martin and Katti (1965) fitted the distribution to a variety of ecological data sets. Bohning et al. (1999) used it in the Belo Horizonte study of caries prevention in children.

Rayner et al. (2009) discuss generalised smooth tests of goodness of fit in general and, in section 11.6, for the ZIP distribution in particular. P-values for their omnibus test and its components need to be found by bootstrap, which may be inconvenient. An alternative approach for categorical data is to use an appropriate X^2 test. Rayner et al. (2009, section 2.4) discuss various X^2 tests for composite hypotheses. The Pearson-Fisher and Chernoff-Lehmann tests both use test statistics of the familiar form $\sum\left(observed - expected\right)^2 / expected$. The former uses maximum-likelihood estimators based on the grouped data and under the null hypothesis of the specified distribution has the convenient χ^2_{k-q-1} null distribution, in which k is the number of classes and q is the number of parameters to be estimated. The Chernoff-Lehmann test uses maximum-likelihood estimators based on the ungrouped data and under the null hypothesis has an inconvenient null distribution between χ^2_{k-q-1} and χ^2_{k-1}.

Rayner et al. (2009, Chapter 8) discuss using the Chernoff-Lehmann test and its components to test for the Poisson, binomial and geometric distributions. Here we use this test to assess goodness of fit for the ZIP distribution. As the ungrouped data are often available this leads to convenient estimation of the parameters. Encouraged by the results in Rayner et al. (2009, Chapter 8) we assume that for large samples the Chernoff-Lehmann test statistic has approximate distribution χ^2_{k-3}, and each squared component after the first two has approximate distribution χ^2_1. The first two components will be close to zero, and reflect differences in estimation by the grouped and ungrouped data.

To construct the Chernoff-Lehmann test statistic and its components the ungrouped maximum-likelihood estimators of λ and ω are required. These are given by the equations

$$\hat{\omega}(1-\hat{\omega})e^{-\hat{\lambda}} = f_0 / N \text{ and } \bar{x}(1-e^{-\hat{\lambda}}) = \hat{\lambda}(1-f_0/N)$$

in which f_0/N is the observed proportion of zeroes. The second of these equations can be written as $\hat{\lambda} = a(1-e^{-\hat{\lambda}})$ where $a = \bar{x}/(1-f_0/N)$ and so $\hat{\lambda}$ can be derived from this equation using iteration. It seems universally accepted that there is no closed form to estimate $\hat{\lambda}$. However, using a similar approach to Irwin (1959), we have the new result that

$$\hat{\lambda} = a - \sum_{r=1}^{\infty} r^{r-1}(ae^{-a})^r / r!$$

where the infinite summation converges for $a > 1$ and converges reasonably quickly for $a > 2$. However, it may be more convenient to use a routine from the R software package to find $\hat{\lambda}$ and $\hat{\omega}$. We illustrate use of such a routine in section 3.

Section 5 deals with the zero inflated negative binomial distribution.

2. Goodness of Fit

The Chernoff-Lehmann X^2 test is an omnibus test and important deviations from the ZIP model may be ‘diluted’ by the test not being focussed on them. To overcome this problem we suggest the test statistic be partitioned into one degree of freedom components. These components may be more powerful than the Chernoff-Lehmann X^2 test for some alternatives to the ZIP model.

Let p_j be the ZIP probabilities for the k grouped classes and let f_j be the corresponding frequencies for such classes. Define the unsquared components of X^2 as

$$\hat{V}_r = \sum_{j=0}^{k-1} f_j g_r(j) / \sqrt{N}$$

for r = 1, 2, ... where the f_j are the frequencies of the k classes of grouped data. If $s^2 = \sum_{j=0}^{k-1} f_j (j - \bar{j})^2 / N$ the $g_r(j)$ are defined as follows. Put

$$g_0(j) = 1, g_1(j) = (j - \bar{j})/s, j = 1, ..., k - 1$$

and for $r > 1$, following Emerson (1968), calculate

$$s_1 = \sum_{j=1}^{k-1} j^2 p_j g_{r-1}^2(j),\ s_2 = \sum_{j=1}^{k-1} j p_j g_{r-1}^2(j),\ s_3 = \sum_{j=1}^{k-1} j p_j g_{r-1}(j) g_{r-2}(j),$$

$$b = \frac{1}{\sqrt{s_1 - s_2^2 - s_3^2}},\ c = -bs_2,\ d = bs_3 \text{ and } g_r(j) = (bj + c)\, g_{r-1}(j) - dg_{r-2}(j).$$

Then, as in Lancaster (1953), $X^2 = \sum_{r=1}^{k-1} \hat{V}_r^2$ where for $r > 2$ the $\hat{V}_r^2$ are approximately χ_1^2 distributed and can be used to further check on the ZIP fit. For $r > 2$ a significant $\hat{V}_r^2$ indicates a possible deviation in the rth moment.

3. Three Examples

3.1. *Rookery slug data*

On the website http://www.bio.ic.ac.uk/research/mjcraw/statcomp/data/slugsurvey.txt Professor Michael Crawley gives the rookery slug data set. Counts of slugs under 40 tiles are shown in Table 1.

To find the maximum likelihood estimators use these commands in the R software package:

```
> library(VGAM)
> y=rep(0:9,c(9,9,8,5,2,4,1,0,1,1))
> vglm(y~1,family=zipoisson)
```

and then two intercept values, –1.572 and 1.011, are obtained. Take $\hat{\omega}$ = exp(–1.572)/(1 + exp(–1.572)) and find $\hat{\lambda}$ = exp(1.011), giving $\hat{\omega}$ = 0.172 and $\hat{\lambda}$ = 2.747. Using classes 5-9 grouped into one class we find X^2 = 5.33 with p-value 0.15. Thus at the 0.05 level of significance the ZIP fit is acceptable. However it is prudent to first check the $\hat{V}_r^2$. See Table 2 where $\hat{V}_3^2$ is significantly large. The grouping of classes here was done based on the rule that class expectations of five or more be used for the X^2 test. We expect that $\hat{V}_1^2$ and $\hat{V}_2^2$ will often be small as they reflect the class grouping. These two components are not approximately χ_1^2 distributed as are the higher order components.

TABLE 1

Counts of slugs under 40 tiles.

# slugs	0	1	2	3	4	5	6	7	8	9
frequency	9	9	8	5	2	4	1	0	1	1

TABLE 2

Partition of X^2 and p-values for slug data.

Statistic	Value	P-value
$\hat{V}_1^2$	0.218	-
$\hat{V}_2^2$	0.569	-
$\hat{V}_3^2$	4.289	0.038
$\hat{V}_4^2$	0.001	0.975
$\hat{V}_5^2$	0.256	0.611
X^2	5.333	0.149

As $\hat{V}_3^2$ is significantly large we look for an alternative distribution. A simple possibility with a different third moment is the negative binomial distribution with

$$f(x) = \binom{k+x-1}{x}\left(\frac{\mu}{\mu+k}\right)^x\left(1+\frac{\mu}{k}\right)^{-k}, x = 0, 1, 2, ..., \mu > 0, k > 0.$$

To find the maximum likelihood estimators of μ and k we can use the MASS library in the R software package:

```
> library(MASS)
> x=rep(0:9,c(9,9,8,5,2,4,1,0,1,1))
> fitdistr(x,"negative binomial")
```

which gives $\hat{\mu}$ = 2.275 and $\hat{k}$ = 1.92895. Using classes 5-9 grouped as above we find X^2 = 0.977 and so the negative binomial is a better fit than the zero-inflated Poisson. As $X^2 < 3.841$ none of the V_r^2 can be significant.

3.2. *Leadbeater's possum data*

Welsh et al. (1996) in their Figure 1 give counts of the possum species, Leadbeater's possum, in 151 sites each of area three hectares. This rare Australian animal had its population diminished by severe bushfires. Table 3 gives the counts.

TABLE 3

Counts of Leadbeater's possum in 151 sites.

# possums	0	1	2	3	4	5	6	7	8	9	10
frequency	95	10	10	12	8	10	0	5	0	0	1

TABLE 4

Partition of X^2 and p-values for possums data.

Statistic	Value	P-value
$\hat{V}_1^2$	0.003	-
$\hat{V}_2^2$	0.540	-
$\hat{V}_3^2$	1.286	0.257
$\hat{V}_4^2$	0.076	0.793
$\hat{V}_5^2$	4.537	0.033
$\hat{V}_6^2$	1.245	0.265
$\hat{V}_7^2$	2.536	0.110
X^2	10.233	0.069

We find $\hat{\omega}$ = 0.642 and $\hat{\lambda}$ = 3.565. Using classes 7-10 grouped into one class we find X^2 = 10.23 with p-value 0.07. Table 4 shows the $\hat{V}_r^2$. Note again that according to the test based on X^2 the ZIP fit is acceptable at the 0.05 level of significance, but consider $\hat{V}_5^2$ in Table 4. We see $\hat{V}_5^2$ is significantly large and might conclude the ZIP fit is not good. Welsh et al. (1996) used covariates to improve the fit. Here we grouped classes so that class expectations were at least one. Our experience suggests the five or more rule used for the Rookery slug data can often lose too much information.

3.3. *Uppsala cold spells data*

Eggers (2015) in her Figure 1.1 gives counts of cold spells in Uppsala, Sweden for the years 1840-2112. Table 5 gives these counts. We find maximum likelihood estimators $\hat{\omega}$ = 0.399 and $\hat{\lambda}$ = 1.377. Observing that $\omega = 1 - p$, where p is the notation used by Eggers (2015), our estimators are close to the moment estimators given by Eggers (2015). With the number of spells for 5 and 6 grouped we find X^2 = 1.271 with an approximate p-value of 0.736 based on the χ_3^2 approximation. Clearly none of the components can be significant with such a small X^2 value. The values of the components V_3^2, V_4^2 and V_5^2 are 0.693, 1.106 and 0.528 respectively. Their non-significant p-values can be found using the χ_1^2 distribution.

TABLE 5

Counts of cold spells in Uppsala, 1840-2012.

# spells	0	1	2	3	4	5	6
frequency	93	35	25	12	2	1	1

Perhaps the extra zeroes in Table 5 are due to global warming. A time series plot of the number of cold spells versus year would be interesting, but Eggers (2015) does not give this.

In the first two examples of this section although X^2 was not significant at least one component was, stressing the need to consider components. The more focused component tests were able to detect alternatives masked by the X^2 test, which is diluted by seeking alternatives in a much larger parameter space. In the final example none of X^2 and its components were significantly large. In effect here the X^2 test is not masking a lower dimensional effect.

4. Improving the Chi-Squared Approximation

The approach we have adopted for finding p-values of X^2 in the previous section is a common one. See, for example, Mead et al. (2003). However, it can be improved upon by noting that for $r > 2$ that the $\hat{V}_r$ are very nearly uncorrelated and standard normal distributed. See, for example, Rayner et al. (2009, Chapter 8) for some supporting evidence of this assertion. Given this knowledge then $Q = X^2 - \hat{V}_1^2 - \hat{V}_2^2$ is more likely to have the χ^2_{k-3} distribution than X^2. Use of Q rather than X^2 in the examples of the previous section does not alter the discussion there but could be important with other data. These comments also apply to testing fit for other distributions.

5. The Zero Inflated Negative Binomial Distribution

The probability function for the zero inflated negative binomial (ZINB) distribution, $g(x)$ say, is

$$g(0) = \pi + (1 - \pi) f(0),\; g(x) = (1 - \pi) f(x),$$

in which $f(x)$ is the negative binomial probability function given previously in section 3.1. Millar (2011, p.57) considers the number of roots produced on apples cultivars for a number of treatments. For one particular treatment the results for 40 cultivars were

TABLE 6

Number of roots produced on an apple cultivar.

# roots	0	1	2	3	4	5	6	7	8	9
frequency	19	2	2	4	3	1	4	3	0	2

To find the MLEs we use the VGAM package in R with the following commands:

```
> library(VGAM)
> y=rep(0:9,c(19,2,2,4,3,1,4,3,0,2))
> vglm(y~1,family=zinegbinomial)
```

which gives three intercepts with values –0.139, 1.522 and 2.732. From these $\hat{\pi}$ = $e^{-0.139}/(1+e^{-0.139})$, $\hat{\mu}=e^{1.522}$ and $\hat{k}=e^{2.732}$, giving $\hat{\pi}$ = 0.465, $\hat{\mu}$ = 4.581 and $\hat{k}$ = 15.364. We calculate expected frequencies of 19, 1.4, 2.6, 3.4, 3.6, 3.2, 2.5, 1.8, 1.1, 1.4 and $X^2 = 5.269$ with six degrees of freedom. The following table checks that none of the components is significantly large (for example, exceeds 3.841, say).

TABLE 7
Partition of X^2

Statistic	Value
$\hat{V}_1^2$	0.007
$\hat{V}_2^2$	0.148
$\hat{V}_3^2$	0.001
$\hat{V}_4^2$	0.241
$\hat{V}_5^2$	0.489
$\hat{V}_6^2$	1.562
$\hat{V}_7^2$	2.339
$\hat{V}_8^2+\hat{V}_9^2$	0.380
X^2	5.269

We conclude the ZINB fits the data well.

6. References

Bohning, D., Dietz, E., Schlattmann,P., Mendonca, L. and Kirchner, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *JRSS* A, 162, 195-209.

Eggers, J. (2015). On statistical methods for zero-inflated models. UUDM Project Report 2015:9, Uppsala University, Sweden.

Emerson, P.L. (1968) Numerical construction of orthogonal polynomials from a general recurrence formula. *Biometrics*, 24, 695-701.

Irwin, J.O. (1959). On the estimation of the mean of a Poisson distribution from a sample with the zero class missing. *Biometrics*, 15, 324-326.

Lancaster, H.O. (1953). A reconciliation of chi-squared from metrical and enumerative aspects. *Sankhya*, 13, 1-10.

Martin, D.C. and Katti, S.K. (1965). Fitting of some contagious distributions to some available data by the maximum likelihood method. *Biometrics*, 18, 354-364.

McKendrik, A.G. (1926). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44, 98-130.

Mead, R., Curnow, R. and Hasted, A. (2003). *Statistical Methods in Agriculture and Experimental Biology.* Chapman & Hall, Boca Raton, FL.

Millar, R. (2011). *Maximum Likelihood Estimation and Inference.* Wiley, New York.

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

Welsh, A.H., Cunningham, R.B., Donnelly, C.F. and Lindenmayer, D.B. (1996**).** Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, 88, 297-308.