Best , D. J., Rayner, J. C. W., Thas, O., "Goodness of fit for the zero-truncated Poisson distribution". Originally published in Journal of Statistical Computation and Simulation Vol. 77, Issue 7, p. 585-591 (2007).

Available from: http://dx.doi.org/10.1080/10629360600569329

Goodness of Fit for the Zero Truncated Poisson Distribution

D. J. BEST, J. C. W. RAYNER

School of Mathematical and Physical Sciences,
University of Newcastle, NSW 2308, Australia

AND O. THAS

Department of Applied Mathematics, Biometrics and Process Control,
Ghent University, B-9000 Gent, Belgium
olivier.thas@Ugent.be

ABSTRACT

The zero truncated Poisson distribution is an important and appropriate model for many applications. Here we give and assess several tests of fit for this model.

Key Words: Chi-squared components, Ecology applications, Equal catchability, Parametric bootstrap, Smooth tests.

1. INTRODUCTION

A basic discrete probability distribution often studied in introductory statistics courses is the Poisson distribution. Typically sufficient conditions are given for the model to apply. These conditions may occur, except that the recording mechanism is not activated unless at least one event occurs. The corresponding zero truncated distribution has probability function, for $\lambda > 0$,

$$P(X = x) = \lambda^x / \{(e^\lambda - 1)x!\}, \; x = 1, 2, 3, \ldots .$$

There have been many instances of this distribution in the literature, and we now cite four of these.

- Simonoff (2003, p.100) gave counts of the pairs of running shoes owned by 60 members of an athletics club. For 1, 2, 3, 4 and 5 pairs the counts were 18, 18, 12, 7 and 5.
- Finney and Varley (1955) gave counts of flower heads with 1, 2, … , 9 fly eggs. The corresponding counts were 22, 18, 18, 11, 9, 6, 3, 0 and 1.
- Coleman and James (1961) gave counts of sizes of groups of people in public places on a Spring afternoon in Portland, Oregon. For group sizes 1, 2, 3, 4, 5, and 6 the counts were 1486, 694, 195, 37, 10 and 1.
- Matthews and Appleton (1993) gave counts of sites with 1, 2, 3, 4 and 5 particles from immunogold assay data. The counts were 122, 50, 18, 4 and 4.

In section 4 below we will test the fit of the zero truncated Poisson distribution for these four sets of data. Section 2 following gives some definitions while section 3 gives a power comparison of some tests of fit for the zero truncated Poisson distribution. Section 5 considers an ecological application.

2. DEFINITIONS

The first test statistics we consider are components of smooth test statistics. To calculate these statistics we need to find the moments about the mean, μ_r, of the zero truncated Poisson

distribution with parameter λ. We first need the moments about the origin, $\gamma_r^{'}$ say, of the usual untruncated Poisson distribution. These can be derived from

$$\gamma_{r+1}^{'} = \lambda \sum_{i=0}^{r} \binom{r}{i} \gamma_r^{'} \text{ for } r = 0, 1, 2\dots ,$$

with $\gamma_0^{'} = 1$. This result follows by substituting $\gamma_r^{'} = \sum_{x=1}^{\infty} x^r e^{-\lambda} \lambda^x / x!$, exchanging the order of the summations, and simplifying. The moments about the origin of the zero truncated Poisson, $\mu_r^{'}$, are as for the untruncated Poisson distribution, but divided by $\left(1 - e^{-\lambda}\right)$:

$$\mu_r^{'} = \gamma_r^{'}/(1 - e^{-\lambda}) \text{ for } r = 1, 2, 3, \dots .$$

Now the moments about the mean, μ_r, can be derived using the well-known result that if $\mu = \mu_1^{'} = \lambda/(1 - e^{-\lambda})$ and $\mu_0^{'} = 1$, then for $r = 2, 3, \dots ,$

$$\mu_r = \sum_{i=0}^{r} \binom{r}{i} \mu_{r-i}^{'}(-\mu)^i .$$

For a random sample $X_1, X_2, \dots , X_n$ the components of the smooth tests of goodness of fit for the zero truncated Poisson distribution are, for $r = 1, 2, 3, \dots ,$

$$\hat{U}_r = \sum_{j=1}^{n} g_r\left(X_j; \hat{\lambda}\right) / \sqrt{n} .$$

Here $\hat{\lambda}$ is the maximum likelihood estimator (MLE), which is the same as the method of moments estimator (MME) of λ, satisfying

$$\overline{X} = \hat{\lambda} / \left(1 - e^{-\hat{\lambda}}\right).$$

Moreover $\{g_r(x; \lambda): r \geq 0\}$ is the set of polynomials orthonormal on the zero truncated Poisson, and hence satisfying

$$\sum_{x=1}^{\infty} g_i(x) g_j(x) \lambda^x / \left\{\left(e^{\lambda} - 1\right) x!\right\} = 0 \text{ for } i \neq j \text{ and } = 1 \text{ for } i = j.$$

The first four orthonormal polynomials are defined as follows:

$$g_0(x) = 1 \text{ for all } x, g_1(x) = (x - \mu)/\sqrt{\mu_2} ,$$
$$g_2(x) = \{ (x - \mu)^2 - \mu_3(x - \mu)/\mu_2 - \mu_2 \}/\sqrt{\mu_4 - \mu_3^2/\mu_2 - \mu_2^2} \text{ and}$$
$$g_3(x) = \frac{(x - \mu)^3 - a(x - \mu)^2 - b(x - \mu) - c}{\sqrt{\mu_6 - 2a\mu_5 + \left(a^2 - 2b\right)\mu_4 + 2(ab - c)\mu_3 + \left(b^2 + 2ac\right)\mu_2 + c^2}}$$

2

in which

$$a = \left(\mu_5 - \mu_3\mu_4 / \mu_2 - \mu_2\mu_3\right)/d \,, \; b = \left(\mu_4^2 / \mu_2 - \mu_2\mu_4 - \mu_3\mu_5 / \mu_2 + \mu_3^2\right)/d \,,$$
$$c = \left(2\mu_3\mu_4 - \mu_3^3 / \mu_2 - \mu_2\mu_5\right)/d \text{ and } d = \mu_4 - \mu_3^2 / \mu_2 - \mu_2^2.$$

Further polynomials may be given using the recurrence relations of Emerson (1968).

The first non-zero component of the omnibus smooth test for the untruncated Poisson distribution - Fisher's Index of Dispersion - generally had good power in the study of Best and Rayner (1999), except when the alternatives had approximately equal mean and variance. In such cases good power was obtained by the Anderson-Darling test. For the zero truncated Poisson this can be based on the test statistic A^2, given by

$$A^2 = \sum_{j=1}^{\infty} Z_j^2 \hat{\lambda}^j / \left\{ h_j\left(1 - h_j\right)\left(e^{\hat{\lambda}} - 1\right)j!\right\}$$

where $Z_j = \sum_{x=1}^{j}\left(O_x - n\hat{\lambda}^x / \left\{\left(e^{\hat{\lambda}} - 1\right)x!\right\}\right)$ and $h_j = \sum_{x=1}^{j}\hat{\lambda}^x / \left\{\left(e^{\hat{\lambda}} - 1\right)x!\right\}$ in which O_x is the

number of observations equal to x. Summation is halted when $\hat{\lambda}^x / \left\{\left(e^{\hat{\lambda}} - 1\right)x!\right\} < 10^{-3}/n$ and $O_x = 0$.

Unlike the $\hat{U}_r^2$, the Anderson-Darling test does not produce biased tests for some alternatives. We suggest A^2 (with summation beginning at one rather than zero) be also used for formal goodness of fit testing of the zero truncated Poisson distribution while $\hat{U}_2^2$ be used in the spirit of exploratory data analysis to examine whether or not the data are under, equally or over-dispersed. Rao and Chakravarti (1956) gave a dispersion statistic

$$D = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 / \left\{\overline{X}\left(1 + \hat{\lambda} - \overline{X}\right)\right\}$$

for the zero truncated distribution, and, in fact, $\hat{U}_2 = (D - n)/\sqrt{(2n)}$.

In the next section we give powers of the tests based on $\hat{U}_2^2$, $\hat{U}_3^2$ and A^2. These powers are compared with powers of a Pearson's X^2 test and with a probability generating function based test, denoted subsequently by PGF. Both tests are described in Epps (1995), and the powers there are reproduced here.

As an alternative to tests based on $\hat{U}_2^2$, $\hat{U}_3^2$ and A^2, we consider tests based on Pearson's X_{CLE1+}^2, and its second and third order components $\hat{V}_2^2$ and $\hat{V}_3^2$. For a complete background and discussion see Best and Rayner (2003, 2005a and 2005b). It is sufficient to note that there are various rules for choosing the number of classes for the Chernoff-Lehmann X_{CL}^2 test. Here we take this number, k say, to be as large as possible such that each class has expectation at least unity and call the test statistic X_{CLE1+}^2. This may involve grouping from above as well as grouping from below and agrees with the suggestion of Douglas (1994). As in Best and Rayner (2005a and 2005b) the test statistic X_{CLE1+}^2 has null distribution well approximated by the χ_{k-2}^2 distribution. The X_{CLE1+}^2 test statistic can be partitioned into useful components. For some alternatives some of these components have much greater power than X^2 itself. Put

$$\mu = \sum_{j=1}^{k} j p_j \text{ and } \mu_r = \sum_{j=1}^{k} (j - \mu)^r p_j \text{ for } r = 2, 3, \dots,$$

where now $p_j = \hat{\lambda}^j / \left\{ \left(e^{\hat{\lambda}} - 1 \right) j! \right\}$ for $j = 1, \dots, k-1$ and $p_k = 1 - p_1 - \dots - p_{k-1}$. Using these multinomial central moments we may calculate $\{g_r(j; \lambda)\}$ defined as previously, and hence define

$$\hat{V}_r = \sum_{j=1}^{k} O_j g_r \left(j; \hat{\lambda} \right) \sqrt{n}, \, r = 1, \dots, k.$$

Then, as in Lancaster (1953),

$$X_{CL}^2 = \hat{V}_1^2 + \dots + \hat{V}_k^2.$$

This is a Chernoff-Lehmann test statistic because the estimation of λ is achieved by using the uncategorized MLE. Notice that estimation of λ by $\hat{\lambda}$ implies that $\hat{V}_1$ will be close to zero if not much pooling is done. Unlike $\hat{V}_2, \dots, \hat{V}_k$, $\hat{V}_1$ will *not* have an asymptotic standard normal distribution. Also notice that if the bound on the class expectation is successively reduced, then k becomes larger and larger, and the $\hat{V}_r$ will approach the corresponding $\hat{U}_r$.

3. POWERS

In the power study reported here random deviates for the geometric (G+) alternatives were found using the IMSL(1995) routine RNGEO and similarly random deviates for the logarithmic series (L) distribution were found using the routine RNLGR from the same source. Random shifted binomial deviates (B+) were found by adding unity to values returned by the routine RNBIN and zero truncated Poissons (P+) were found by using the routine RNPOI and discarding zeroes. Random zeta deviates (Z) were found using an algorithm in Devroye (1986, p.551).

Table 1. Powers of some tests for the zero truncated Poisson distribution with $n = 50$ and $\alpha = 0.05$ (10,000 simulations of the outer loop, 1,000 for the inner loop).

Alternative	$\hat{U}_2^2$	$\hat{U}_3^2$	A^2	PGF	X^2
P+(3)	0.054	0.050	0.054	0.059	0.057
G+(0.4)	0.885	0.566	0.807	0.84	0.64
G+(0.33)	0.969	0.714	0.940	0.93	0.84
B+(10, 0.2)	0.492	0.075	0.444	0.53	-
L(0.3)	0.237	0.176	0.193	0.22	0.14
L(0.5)	0.552	0.364	0.447	0.52	0.28
L(0.7)	0.913	0.680	0.859	0.91	0.73
Z(2)	0.786	0.674	0.747	0.76	-

The parametric bootstrap powers in Table 1 are based on 10,000 simulations of a parametric bootstrap p-value which used 1,000 simulations. Gürtler and Henze (2000) give

4

details of the parametric bootstrap in the goodness of fit context. The Table 1 powers show the smooth dispersion test based on $\hat{U}_2^2$ is usually best for the alternatives considered. The PGF test is almost as good, with the test based on A^2 almost as good as the PGF test. The tests based on $\hat{U}_3^2$ and X^2 fared worst for these alternatives. Epps (1995) notes that the PGF test, like those based on the $\hat{U}_r^2$, $\hat{V}_r^2$ and X^2, can have poor power for some alternatives.

The powers in Table 2 are intended to be directly comparable with those in Table 1. They are based on 10,000 simulations of samples of size n and use χ^2 critical values. As in Best and Rayner (2005a) the minimum number of classes for X_{CLE1+}^2 and its components was taken to be 5 even if this meant some classes had expectation less than 1.0. The powers in Table 2 are generally not quite as good as those in Table 1 but nevertheless those for $\hat{V}_2^2$ and X_{CLE1+}^2 are close enough to those of $\hat{U}_2^2$ and A^2 to suggest they offer an acceptable approach when ease of calculation of the p-values is important.

***Table* 2**. Powers of some categorised tests for the zero truncated Poisson distribution: X_{CLE1+}^2 and its components $\hat{V}_2^2$ and $\hat{V}_3^2$, with $n = 50$ and $\alpha = 0.05$

Alternative	$\hat{V}_2^2$	$\hat{V}_3^2$	X_{CLE1+}^2
P+(3)	0.052	0.048	0.046
G+(0.4)	0.753	0.063	0.613
G+(0.33)	0.915	0.065	0.821
B+(10, 0.2)	0.550	0.029	0.224
L(0.3)	0.199	0.114	0.224
L(0.5)	0.472	0.220	0.484
L(0.7)	0.823	0.089	0.742
Z(2)	0.759	0.436	0.772

4. EXAMPLES

We now apply the tests based on $\hat{U}_2^2$, $\hat{V}_2^2$, A^2 and X_{CLE1+}^2 to the four data sets given in the Introduction. The p-values for $\hat{U}_2^2$ and A^2 are based on 1,000 simulations using the parametric bootstrap, while the p-values for $\hat{V}_2^2$ and X_{CLE1+}^2 use the appropriate approximating χ^2 distribution.

Running Shoes Example. We find $\hat{\lambda} = 2.088$, $\hat{U}_2^2 = 0.128$ with p-value 0.73 and $A^2 = 0.124$ with p-value 0.89. Both tests suggest these data are well described by the zero truncated Poisson distribution. This is supported by the tests based on $\hat{V}_2^2$ and X_{CLE1+}^2. The former takes the value 0.029 with p-value 0.87, while the latter takes the value 3.169 with p-value 0.53.

Flower Heads Example. Finney and Varley (1955), on the basis of a Pearson X^2 test, suggest these data are consistent with the zero truncated Poisson distribution. We find $\hat{\lambda} = 2.860$, $\hat{U}_2^2 = 4.62$ with p-value 0.02 and $A^2 = 1.162$ with p-value 0.04. Both tests suggest the zero truncated Poisson is not a good model. Our approach has identified the data are too dispersed for this model to be adequate. We also find $\hat{V}_2^2 = 5.712$ with p-value 0.02 and $X_{CLE1+}^2 = 6.865$ with p-value 0.23. This is consistent with the conclusion of Finney and Varley (1955), and with the test based on $\hat{U}_2^2$.

Public Places Example. The zero truncated Poisson distribution seems to fit these data. We find $\hat{\lambda} = 0.8925$, $\hat{U}_2^2 = 0.773$ with p-value 0.40, $\hat{V}_2^2 = 0.731$ with p-value 0.39, $A^2 = 0.405$ with p-value 0.30 and $X_{CLE1+}^2 = 2.995$ with p-value 0.56.

Immunogold Assay Example. Use of A^2 alone suggests the zero truncated Poisson distribution provides an appropriate model, but $\hat{U}_2^2$ suggests overdispersion. We find $\hat{\lambda} = 0.9906$, $\hat{U}_2^2 = 3.896$ with p-value 0.05 and $A^2 = 0.742$ with p-value 0.13. We also find $\hat{V}_2^2 = 5.301$ with p-value 0.02 and $X_{CLE1+}^2 = 8.875$ with p-value 0.03.

5. EQUAL CATCHABILITY

To estimate animal abundance ecologists perform mark-recapture studies. In these it is of interest to ascertain whether or not animals are caught in a random fashion. The ecological literature refers to caught at random or not as testing for equal catchability.

In a study carried out by Keith and Meslow (1968), snowshoe hares were captured over seven days. After a hare was captured it was marked and released. Subsequently the same hare may or may not have been recaptured. Those that have been captured on a previous day were identified by the marking done on their first day of capture. There were 261 hares caught over the seven days. Of these 184 were caught once, 55 were caught twice, 14 were caught three times, 4 were caught four times, and 4 were caught five times.

For these data Krebs (1998, p.52) finds $\hat{\lambda} = 0.7563$ and $X^2 = 7.77$ on two degrees of freedom with p-value 0.02. From section 3 above we know that X^2 is a reasonable omnibus test statistic for the zero truncated Poisson but that the dispersion test based on $\hat{U}_2^2$ is also quite powerful. To calculate $\hat{U}_2^2$ note that the mean and variance of the data are 1.4253 and 0.6300 respectively. We calculate $D = 347.2$ and hence $\hat{U}_2^2 = (D - 262)^2/522 = 14.65$. Using the χ_1^2 approximation to the distribution of $\hat{U}_2^2$ the p-value is 0.0001, and using the parametric bootstrap based on 1,000 simulations a p-value of 0.002 is obtained. The two p-values for $\hat{U}_2^2$ are reasonably consistent. Notice that the test based on $\hat{U}_2^2$ is more critical of the data than that based on X^2. However, as in the power study above, a moment test such as that based on $\hat{U}_2$ may sometimes lack power, and it is advisable to calculate A^2 or X^2 as well as $\hat{U}_2^2$.

While testing for the zero truncated Poisson is appropriate for these snowshoe hare data, we suggest that a doubly truncated Poisson distribution may be a more appropriate model if at most one capture a day is possible, as at most seven captures are possible over the study. Johnson, Kotz and Kemp (1992, p.186) discuss how to estimate λ for a doubly truncated Poisson distribution.

REFERENCES

Best, D.J. and Rayner, J.C.W. (1999). Goodness of fit for the Poisson distribution. *Statistics and Probability Letters*, 44, 259-265.

Best, D.J. and Rayner, J.C.W. (2003). Tests of fit for the geometric distribution. *Communications in Statistics, Simulation and Computation*, 32 (4), 1065-1078.

December 5, 2005

Best, D.J. and Rayner, J.C.W. (2005a). Improved testing for the Poisson distribution using chisquared components with data dependent cells. *Communications in Statistics, Simulation and Computation*, 34(1), 85-96.

Best, D.J. and Rayner, J.C.W. (2005b). Improved testing for the binomial distribution using chisquared components with data dependent cells. To appear in the *Journal of Statistical Computation and Simulation*.

Coleman, J.S. and James, J. (1961). The equilibrium size distribution of freely-forming groups. *Sociometry*, 24, 36-45.

Devroye, L. (1986). *Non-Uniform Random Variate Generation.* New York: Springer-Verlag.

Douglas, J.B. (1994). Empirical fitting of discrete distributions. *Biometrics*, *50*, 576-579.

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

Epps, T.W. (1995). A test of fit for lattice distributions. *Communications in Statistics, Theory and Methods*, 24 (6), 1455-1479.

Finney, D.J. and Varley, G.C. (1955). An example of the truncated Poisson distribution. *Biometrics*, 11, 387-394.

Gürtler, N. and Henze, N. (2000). Recent and classical goodness-of-fit tests for the Poisson distribution. *Journal of Statistical Planning and Inference*, 90, 207-225.

IMSL (1995). *Fortran Numerical Libraries.* Houston: Visual Numerics.

Johnson, N.L., Kotz, S. and Kemp, A.W. (1992). *Univariate Discrete Distributions* (2nd ed.). New York: Wiley.

Keith, L.B. and Meslow, E.C. (1968). Trap response by snowshoe hares. *Journal of Wildlife Management*, 32, 795-801.

Krebs, C.J. (1998). *Ecological Methodology* (2nd ed.). New York: Addison Wesley Longman.

Lancaster, H.O. (1953). A reconciliation of χ^2 from metrical and enumerative aspects. *Sankhya*, 13, 1-10.

Matthews, J.N.S. and Appleton, D.R. (1993). An application of the truncated Poisson distribution to immunogold assay. *Biometrics*, 49, 617-621.

Rao, C.R. and Chakravarti, I.M. (1956). Some small sample tests of significance for a Poisson distribution. *Biometrics*, 12, 264-282.

Simonoff, J.S. (2003). *Analyzing Categorical Data*. New York: Springer.

December 5, 2005