International Journal of Ecological Economics & Statistics (IJEES)
Spring 2010, Vol. 17, No. P10; Int. J. Ecol. Econ. Stat.;
ISSN 0973-1385 (Print), ISSN 0973-7537 (Online)

Moment Tests of Fit for Some Discrete Distributions with Ecological Applications

D. J. BEST, [*1] **J. C. W. RAYNER** [2] and **O. THAS** [3]

[*1, 2] School of Mathematical and Physical Sciences, University of Newcastle,
NSW 2308, Australia
[1]Email: John.Best@newcastle.edu.au;
[2]Email John.Rayner@newcastle.edu.au

[3]Department of Applied Mathematics, Biometrics and Process Control,
B-9000 Gent, Belgium
[3]Email: olivier.thas@Ugent.be

ABSTRACT

We present moment orientated tests of fit for some discrete distributions used in ecological applications. The distributions covered are the Poisson, negative binomial, zero-truncated Poisson, logarithmic, zero-inflated Poisson, binomial, beta-binomial, geometric, Neyman Type A and the bivariate Poisson. Each distribution is illustrated by reference to ecological data. The tests of fit we give are often more powerful than traditional chi-squared tests.

Key words: Index of species diversity; Orthonormal polynomials; Parametric bootstrap; Smooth tests.
Mathematics Subject Classification: 62F03
JEL Classification: C12

1. INTRODUCTION

Our aim in this expository note is to present moment-orientated tests of fit for some discrete distributions used in ecological applications. Typically these tests are powerful, and are easy to apply and interpret. Applications include the determination of an appropriate sample size for an ecological experiment, ecological data description and assessment mechanisms to explain ecological data.

A problem with moment-orientated tests of fit based on a single moment is that they may have poor power for alternatives with the same moment. A remedy for this problem is to use a smooth test of fit involving a linear combination of more than one moment. We shall discuss smooth tests of fit in Section 3 below. All of the moment tests we present are in fact special cases of smooth tests. A comprehensive exposition of smooth tests is given in Rayner et al. (2009).

The moment-orientated tests presented here are easy to apply and have, as is typical of smooth tests, competitive power. They are not necessarily the most powerful tests of fit. In general, no one test is

* *Communicating author*

most powerful for all alternatives. The moment-orientated tests presented here usually have substantially better power than a traditional so-called chi-squared test of fit.

In all the applications following method of moments (MOM) estimation is used. While MOM estimators are known to be often less efficient than maximum likelihood (ML) estimators, it is our experience that in testing for fit there is often little, if any, loss of power when using MOM estimators. For some of the distributions we discuss the MOM and ML estimators coincide, and for one distribution, the beta-binomial, they are almost always very close. An advantage of MOM estimators is that they can often be given in closed form, avoiding the convergence problems that often bedevil ML estimation.

All of the moment tests we present have asymptotic χ^2 distributions, and so approximate p-values can easily be obtained. For small samples this approximation may be poor and so it is sensible to check this approximate p-value using the parametric bootstrap discussed in the Appendix. An R package available at
http://www.biomath.ugent.be/~othas/smooth2/
does many of the calculations presented below. The same website also gives FORTRAN code for some of the calculations presented below.

Why verify or test the fit of a model for our ecological data? It does not make sense to summarize the data with an inappropriate model or attempt to explain the data via a mechanism suggested by an inappropriate model. This would be at best misleading and at worst make all the data and the whole study a waste of time and resources. To help emphasize the importance of using an appropriate model for determining the sample size we will now use the example given in Krebs (1998).

Suppose we have the counts of black bean aphids given in Table 1. Suppose further that we wish to do another experiment to estimate the mean infestation ($\bar{x}$) by a 95% confidence interval of length r say, equal to 15% of the mean. If we assume the data are Poisson distributed (so that aphids occur on bean stems in a random fashion) then an approximate sample size, n, based on a standard formula given in Krebs (1998) is

$$n = \frac{4}{r^2}\left(\frac{1}{\bar{x}}\right) = 51 \text{ stems,}$$

since here $r = 0.15$ and $\bar{x} = 3.46$. (Note that we write $\overline{X}$ for the random variable the sample mean and $\bar{x}$ for an observed value of $\overline{X}$ for a given data set.)

However if the data are negative binomial distributed (so that aphids occur on bean stems in a clumped fashion), then using the MOM estimator of the negative binomial parameter $k = 3.19$ and another formula from Krebs (1998)

$$n = \frac{4}{r^2}\left(\frac{1}{\bar{x}} + \frac{1}{k}\right) = 107 \text{ stems.}$$

The sample size to use has more than doubled because of the change in model. The parameter k is defined in section 2 (2) following.

In Section 2 we consider the Poisson, negative binomial, zero-truncated Poisson, logarithmic, zero-inflated Poisson, binomial, beta-binomial and bivariate Poisson distributions. Section 3 defines smooth tests and points out their relationship with moment tests. The Appendix gives an algorithm for parametric bootstrap p-values.

2. MOMENT TESTS OF FIT FOR SOME DISCRETE DISTRIBUTIONS

Throughout the following sections we assume a random sample $X_1, \ldots, X_n$ is taken from the distribution of interest. In general $m_r = \sum_{i=1}^{n}(X_i - \bar{X})^r$, $r = 2, 3, \ldots$ are the central sample moments. We are typically interested in test statistics $(m_r - \mathrm{E}[m_r])^2/\mathrm{var}((m_r - \mathrm{E}[m_r]))$ where by $\mathrm{E}[m_r]$ and $\mathrm{var}((m_r - \mathrm{E}[m_r]))$ we mean the MOM estimators of $\mathrm{E}[m_r]$ and $\mathrm{var}(m_r - \mathrm{E}[m_r])$ respectively.

(1) *Poisson*

For the Poisson distribution the probability function is

$$e^{-\lambda}\lambda^x / x!, \; x = 0, 1, \ldots, \text{ in which } \lambda > 0.$$

The MOM and ML estimators of λ are the same and will be denoted by $\hat{\lambda}$. Given a random sample $X_1, \ldots, X_n$, $\hat{\lambda} = \bar{X} = (X_1 + \ldots + X_n)/n$. A dispersion test of fit can be based on $D = (m_2 - \hat{\lambda})$. A statistic with an asymptotic χ_1^2 distribution is $D^2/\mathrm{var}(D)$ where $\mathrm{var}(D) = 2\hat{\lambda}^2/n$. We note that $D^2/\mathrm{var}(D)$ is a standardised version of Fisher's classical Poisson Index of Dispersion.

A slightly more arithmetically complicated moment test of fit for the Poisson can be based on

$$R = m_4 - 6m_3 + (11 - 6\hat{\lambda})m_2 + 3\hat{\lambda}(\hat{\lambda} - 2).$$

This statistic is the numerator divided by n of an order four smooth test. As was the case with the dispersion statistic D we have that $R^2/\mathrm{var}(R)$ is asymptotically χ_1^2 distributed. Using the formula in

Section 3, var(R) = 24 $\hat{\lambda}^4/n$. We suggest, first, using D^2/var(D) as the test statistic and second, that R^2/var(R) also be calculated and used in a exploratory data analytic (EDA) fashion as the test based on D has does not have good power when $\bar{x}$ and m_2 are numerically close. We do not consider a third moment test of fit as Best and Rayner (1999) show that such a test does not have good power.

For the black bean aphid data of Table 1 $\bar{x}$ = 3.46, m_2 = 7.21 and D^2/var(D) = 29.34 with χ^2_1 p-value less than 0.001, while R^2/var(R) = 0.50 with χ^2_1 p-value 0.48. The corresponding parametric bootstrap p-values are 0.000 and 0.39. Clearly the Poisson model is poor here. The variance is significantly greater than the mean.

(2) *Negative Binomial (or Gamma-Poisson)*

The negative binomial probability function is

$$\Gamma(k+x)p^k q^x / \{x!\Gamma(k+x)\} \text{ for } x = 0, 1, \ldots, \text{ in which } q = 1 - p \text{ and } k > 0.$$

The MOM estimators are $\hat{k} = \bar{X}^2/(m_2 - \bar{X})$ and $\hat{p} = \hat{k}/(\hat{k} + \bar{X})$. Observe that if $\bar{x} > m_2$ then $\hat{k} < 0$. If this occurs then the negative binomial will not fit the data well and any reasonable test will reject the model. A third moment test statistic given by Anscombe (1950) is

$$T = m_3 - \hat{k}\hat{q}(1+\hat{q})/\hat{p}^3.$$

Using

$$\text{var}(T) = 2\hat{k}(\hat{k}+1)\hat{q}^3(1+\hat{q})(10+3\hat{k}-4\hat{p})/(n\hat{p}^6)$$

we may construct T^2/var(T), which is asymptotically χ^2_1 distributed. In Section 3 we find var(T) using the delta method.

Best et al. (2009) find that the test based on the fourth moment statistic R is more powerful than based on T, where

$$R = m_4 + (6 - 12/\hat{p})m_3 - \hat{k}\,\hat{q}\,(3\hat{k}\,\hat{q} - 5\,\hat{p}^2 - 18\,\hat{q})/\hat{p}^4.$$

These authors also give

$$\text{var}(R) = 24\,\hat{k}\,(\hat{k}+1)\,\hat{q}^4\,(3\,\hat{p}^2 - 6\,\hat{p} + \hat{k}^2 + 5\hat{k} + 9)/(n\,\hat{p}^8).$$

For the black bean aphid data of Table 1 $\hat{k}$ = 3.19, $\hat{p}$ = 0.48 and $T^2/\text{var}(T)$ = 1.07 with χ_1^2 p-value 0.30 and $R^2/\text{var}(R)$ = 0.80 with χ_1^2 p-value 0.37. The corresponding parametric bootstrap p-values are 0.16 and 0.26. It appears the negative binomial model is appropriate.

(3) *Zero-truncated-Poisson*

For the zero-truncated-Poisson distribution the probability function is

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

The MOM and ML estimators coincide and satisfy $\overline{X} = \hat{\lambda}/\{1-\exp(-\hat{\lambda})\}$. This equation needs to be solved iteratively. For example, using Newton-Raphson, let

$$f(\lambda) = \lambda/\{1 - \exp(-\lambda)\} - \overline{x}.$$

Then

$$f'(\lambda) = \{1 - \exp(-\lambda)\}^{-1} - \lambda\exp(-\lambda)\{1 - \exp(-\lambda)\}^{-2}$$

and a better estimate than λ_m is $\lambda_{m+1} = \lambda_m - f(\lambda_m)/f'(\lambda_m)$ for m = 0, 1, 2, ... A reasonable starting point for the iteration is $\lambda_0 = \overline{x}$.

A dispersion test statistic is

$$D = m_2 - \overline{X}\left(1 + \hat{\lambda} - \overline{X}\right)$$

for which

$$\text{var}(D) = \overline{X}\hat{\lambda}\{1 + (\hat{\lambda}+1)^2 - \overline{X}(\hat{\lambda}+2)\}/\{n(1+\hat{\lambda}-\overline{X})\}.$$

As before $D^2/\text{var}(D)$ is asymptotically χ_1^2 distributed. We note that $D^2/\text{var}(D)$ is a standardized version of the dispersion statistic given by Rao and Chakravarti (1956).

Finney and Varley (1955) discussed mechanisms for explaining counts of fly eggs on flower heads. They gave the data in Table 2 and used the conventional X^2 test to confirm the data were modelled by the zero truncated Poisson distribution. However if they had used $D^2/\text{var}(D)$ they would not have reached this conclusion. For the Table 2 data we have $\bar{x}$ = 3.03, $\hat{\lambda}$ = 2.86 and $D^2/\text{var}(D)$ = 4.61. Both the χ_1^2 and parametric bootstrap p-values are 0.03.

(4) *Logarithmic Distribution*

The logarithmic probability function is

$$\gamma\beta^x/x \text{ for } x = 1, 2, \ldots, \text{ in which } 0 < \beta < 1 \text{ and } \gamma = -1/\log(1-\beta).$$

The MOM and ML estimators coincide and satisfy $\overline{X} = \hat{\beta}\hat{\gamma}/\left(1-\hat{\beta}\right)$. As in (3) above, Newton-Raphson iteration can be used to find $\hat{\beta}$, by first estimating $g(\beta) = (1-\beta)^{-1}$ and hence $\beta = 1 - 1/g$. For $m = 0, 1, 2, \ldots$ a better estimate than g_m is g_{m+1} given by

$$g_{m+1} = \{1 + \overline{X}\log(g_m - 1)\}/(1 - \overline{X}/g_m)$$

where, following Birch (1963), we take $g_0 = 1 + \{1.5(\overline{X} - 1) + 2\}\log\overline{X}$. A dispersion test statistic is

$$D = m_2 - \overline{X}(1 - \hat{\beta}\hat{\gamma})/(1 - \hat{\beta}).$$

Noting that $n\text{var}(D) = \hat{\gamma}\hat{\beta}^2(2 - 2\hat{\gamma}\hat{\beta} - \hat{\gamma}\hat{\beta}^2)/\{(1-\hat{\beta})^4(1-\hat{\gamma}\hat{\beta})\}$, an appropriate test statistic is $D^2/\text{var}(D)$, which is asymptotically χ_1^2 distributed.

Table 3 gives some species diversity data on insect catches from the Sierra Tarahuma, Mexico, reported by Aldrete (2002). We find $\hat{\beta}$ = 0.9743. The species diversity index given, for example in Krebs (1998, section 12.4.1), is $n(1-\hat{\beta})/\hat{\beta}$ where n is the total number of species. Here this index takes the value 9.01. It is suggested that there is no point in quoting this index unless the data are well modelled by the logarithmic distribution. We find we have $D^2/\text{var}(D)$ = 0.24 with a χ_1^2 p-value of 0.52 and parametric bootstrap p-values 0.55. The logarithmic model seems acceptable.

(5) *Zero Inflated Poisson Distribution*

The zero inflated Poisson has probability function $f(x)$ given by

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

A number of rare animal species were affected by the recent severe bushfires in the montane ash forests of the Central Highlands of Victoria, south eastern Australia. One such species was Leadbeater's Possum, *Gymnobelideus leadbeateri*. Welsh et al. (1996) had previously reported on abundance models for this species giving counts shown in Table 4 from 151 sites of three hectares.

Suppose we wish to fit the zero inflated Poisson distribution to these data. We find

$$\hat{w} = \left(m_2 - \overline{X}\right)/\left(m_2 - \overline{X} + \overline{X}^2\right) \text{ and } \hat{\lambda} = \left(m_2 - \overline{X} + \overline{X}^2\right)/\overline{X}.$$

As before, we use $T^2/\text{var}(T)$ where $T = m_3 - \hat{\mu}_3$ and $\mu_3 = \varphi\{1 + 3\lambda + \lambda^2 - 3\varphi(1 + \lambda) + 2\varphi^2\}$ in which $\varphi = \lambda(1 - w)$. Then

$$n\,\text{var}(T) = 2\hat{\varphi}\left(\hat{\lambda}^3 + 3\hat{\lambda}^2\right).$$

For the Table 4 data $\hat{w} = 0.6415$, $\hat{\lambda} = 3.5648$, $T = 1.20$, $\text{var}(T) = 1.41$, $T^2/\text{var}(T) = 1.02$ and using the χ^2_1 approximation the p-value is 0.32 while the parametric bootstrap p-value is 0.26. It appears the model is satisfactory.

(6) *Binomial Distribution*

The binomial probability function is

$$\binom{n^*}{x} p^x q^{n^*-x} \text{ for } x = 0, 1, 2, \ldots, n^* \text{ in which } 0 < p < 1 \text{ and } q = 1 - p.$$

Given a random sample $x_1, \ldots, x_n$ the MOM and ML estimators coincide with $\hat{p} = \overline{x}/n$. A dispersion test statistic is $D = m_2 - n^*\hat{p}\hat{q}$ and $\text{var}(D) = 2n^*\hat{p}^2\hat{q}^2(n^* - 1)/n$. Note that $D^2/\text{var}(D)$ is asymptotically χ^2_1 distributed and is a standardised version of Fisher's Binomial Index of Dispersion.

Cochran (1936) discussed a classical data set concerning spotted wilt disease on tomato plants. There were 160 plots each containing nine tomato plants all grown at the Waite Agricultural Institute, South Australia. Counts of diseased plants are shown in Table 5.

We find $\hat{p}$ = 0.18 and D^2/var(D) = 15.11 with both χ_1^2 and parametric bootstrap p-values less than 0.01, and so we do not accept the binomial model. As D is positive the data are over-dispersed, possibly due to clumps of diseased plants.

(7) *Beta-binomial Distribution*

The *beta-binomial* has probability function

$$f(x;\,\alpha,\beta) = \binom{n^*}{x} \mathrm{B}(\alpha + x, \beta + n^* - x)/\mathrm{B}(\alpha,\beta)$$

for x = 0, 1, ..., n^* in which $\alpha > 0$ and $\beta > 0$.

Using MOM we have

$$\hat{\alpha} = \left(n^* - \overline{X} - m_2/\overline{X}\right)/\left(n^* m_2/\overline{X} + \overline{X} - n^*\right) \text{ and } \hat{\beta} = \left(n^* - \overline{X}\right)\hat{\alpha}/\overline{X}.$$

As before use T^2/var(T) where $T = m_3 - \hat{\mu}_3$ and, in terms of the parameters $\theta = 1/(\alpha + \beta + 1)$ and $\pi = \alpha/(\alpha + \beta)$,

$$\begin{aligned}\text{var}(T) = {} & 2(n^* - 1)(n^* - 2)(1 - \hat{\pi})(\hat{\theta} - 1)(\hat{\pi}\,\hat{\theta} - \hat{\pi} + 1)(\hat{\pi}\,\hat{\theta} - \hat{\pi} - \hat{\theta})(1 - \hat{\theta} + n^*\hat{\theta})(230n^*\hat{\pi}^2\hat{\theta}^4 \\ & - 6n^*\hat{\pi}^2\hat{\theta} + 37n^*\hat{\pi}^2\hat{\theta}^2 - 230n^*\hat{\pi}\,\hat{\theta}^4 + 28n^{*2}\hat{\pi}^2\hat{\theta}^3 - 242n^*\hat{\pi}^2\hat{\theta}^5 + 124n^*\hat{\pi}\,\hat{\theta}^3 - \\ & 3n^{*2}\hat{\pi}^2\hat{\theta}^2 + 80n^{*2}\hat{\pi}^2\hat{\theta}^5 + 3n^{*2}\hat{\pi}\,\hat{\theta}^2 - 124n^*\hat{\pi}^2\hat{\theta}^3 + 242n^*\hat{\pi}\,\hat{\theta}^5 - 80n^{*2}\hat{\pi}\,\hat{\theta}^5 - \\ & 70n^{*2}\hat{\pi}^2\hat{\theta}^4 + 6n^*\hat{\pi}\,\hat{\theta} + 70n^{*2}\hat{\pi}\,\hat{\theta}^4 - 28n^{*2}\hat{\pi}\hat{\theta}^3 - 37n^*\hat{\pi}\,\hat{\theta}^2 + 6\hat{\theta} + 141n^*\hat{\pi}^2\hat{\theta}^6 - \\ & 47n^{*2}\hat{\pi}^2\hat{\theta}^6 - 9\hat{\pi}\,\hat{\theta} - 183\hat{\pi}^2\hat{\theta}^4 + 9\hat{\pi}^2\hat{\theta} - 56\hat{\pi}^2\hat{\theta}^2 + 56\hat{\pi}\,\hat{\theta}^2 + 16n^*\hat{\theta}^2 + 62n^*\hat{\theta}^4 - \\ & 14n^{*2}\hat{\theta}^4 - 22n^*\hat{\theta}^3 + 10n^{*2}\hat{\theta}^3 + 183\hat{\pi}\,\hat{\theta}^4 - 142\hat{\pi}\,\hat{\theta}^3 + 142\hat{\pi}^2\hat{\theta}^3 + 3\hat{\pi} - 8\hat{\theta}^2 - 3\hat{\pi}^2 - \\ & 44\hat{\theta}^4 + 40\hat{\theta}^3 + 161\hat{\pi}^2\hat{\theta}^5 - 161\hat{\pi}\,\hat{\theta}^5 + 24n^{*2}\hat{\theta}^5 + 46\hat{\theta}^5 - 68n^*\hat{\theta}^5 - 94\hat{\pi}^2\hat{\theta}^6 + 94\hat{\pi}\,\hat{\theta}^6 \\ & - 12n^{*2}\hat{\theta}^6 + 36n^*\hat{\theta}^6 - 141n^*\hat{\pi}\,\hat{\theta}^6 + 47n^{*2}\hat{\pi}\hat{\theta}^6 - 24\hat{\theta}^6 + 24\hat{\pi}^2\hat{\theta}^7 - 24\hat{\pi}\,\hat{\theta}^7 - \\ & 36n^*\hat{\pi}^2\hat{\theta}^7 + 12n^{*2}\hat{\pi}^2\hat{\theta}^7 + 36n^*\hat{\pi}\,\hat{\theta}^7 - 12n^{*2}\hat{\pi}\,\hat{\theta}^7)n^*\hat{\pi}/\{(1 + 4\hat{\theta})(1 + 3\hat{\theta})(1 + 2\hat{\theta})(1 + \\ & \hat{\theta})^5 n\}.\end{aligned}$$

Observe that for the beta-binomial $\mu_3 = \sum_{x=0}^{n^*}(x-\mu)^3 f(x;\alpha,\beta)$ in which $\mu = n^*\alpha/(\alpha + \beta + 1)$. Then, for the data of Table 5, T^2/var(T) = 0.296 with χ_1^2 p-value 0.42 and parametric bootstrap p-value 0.56. The beta-binomial model describes the data better than the binomial model.

(8) *Geometric Distribution*

The geometric probability function is

$$qp^{x-1} \text{ for } x = 1, 2, \ldots, \text{ in which } p = 1 - q \text{ and } 0 < q < 1.$$

We have $\hat{q} = 1/\overline{X}$ and as before use $D^2/\text{var}(D)$ where $D = m_2 - \hat{p}/\hat{q}^2$ and $\text{var}(D) = 4\hat{p}^2/(n\hat{q}^4)$.

Pielou (1962) examined runs of one tree species with respect to another. For runs of the tree species *Pseudotsuga menziesii* with respect to *Pinus ponderosa* the data in Table 6 were obtained.

We find $\overline{X} = 2.3$, $\hat{q} = 0.435$, $D^2/\text{var}(D) = 99.1$ with χ^2_1 p-value and parametric bootstrap p-values less than 0.001. As D is positive the data are overdispersed compared to the geometric.

(9) *Neyman-Type A Distribution*

The Neyman-Type A probability function is

$$\frac{e^{-\lambda}\varphi^x}{x!}\sum_{j=0}^{\infty}\frac{\left(\lambda e^{-\varphi}\right)^j j^x}{j!} \text{ for } x = 0, 1, 2, \ldots, \text{ in which } \lambda > 0 \text{ and } \varphi > 0.$$

We have $\hat{\varphi} = m_2/\overline{X} - 1$, $\hat{\lambda} = \overline{X}/\hat{\varphi}$ and if we define $T = m_3 - m_2 + \overline{X} - m_2^2/\overline{X}$, we can use $T^2/\text{var}(T)$ as a test statistic to assess goodness of fit.

In section 2 (5) above we found that the Leadbeater's Possum data was reasonably described by a zero inflated Poisson distribution. Neyman's Type A distribution can model zero-inflated data as well as data with more than one mode. To see how well the Type A model applies to the Table 4 data we need $\text{var}(T)$, which is given by

$$\text{var}(T) = \hat{\lambda}\hat{\varphi}^3\left(6\hat{\lambda}^2 + 5\hat{\lambda}\hat{\varphi} + 18\hat{\lambda} + 2\hat{\varphi} + 6 + 25\hat{\lambda}\hat{\varphi}^2 + 2\hat{\lambda}\hat{\varphi}^3 + 18\hat{\lambda}^2\hat{\varphi} + 18\hat{\lambda}^2\hat{\varphi}^2 + 6\hat{\lambda}^2\hat{\varphi}^3\right)/n.$$

This formula was given incorrectly by Evans (1953).

We find $T^2/\text{var}(T) = 0.81$ and using the χ^2_1 approximation, the p-value is 0.37 while the parametric bootstrap p-value is 0.33. This indicates that, like the zero-inflated Poisson, the Neyman Type A model describes the possum data fairly well.

(10) *Bivariate Poisson Distribution*

The bivariate Poisson distribution has probability function

$$f(x, y; \lambda_1, \lambda_2, \lambda_3) = \exp(-\lambda_1 - \lambda_2 + \lambda_3) \sum_{i=0}^{\min(x,y)} \frac{(\lambda_1 - \lambda_3)^{x-i}(\lambda_2 - \lambda_3)^{y-i}\lambda_3^i}{(x-i)!(y-i)!i!}$$

in which $x, y = 0, 1, 2, \ldots$ and $\lambda_1, \lambda_2, \lambda_3 > 0$.

Also we take $\hat{\lambda}_1 = \overline{X}$, $\hat{\lambda}_2 = \overline{Y}$ and $\hat{\lambda}_3 = r\sqrt{\overline{XY}}$ in which r is the Pearson product moment correlation. It is known that X is Poisson (λ_1) and Y is Poisson (λ_2). This suggests, following section 2 (1) above, that in constructing tests of fit we start with $D_X = m_{20} - \overline{X}$ and $D_Y = m_{02} - \overline{Y}$, where $m_{20} = \sum_i (X_i - \overline{X})^2 / n$ and $m_{02} = \sum_i (Y_i - \overline{Y})^2 / n$. Take $Z = (D_X, D_Y)^T$ and let V be the estimated covariance matrix of Z. It can be shown that

$$V = \frac{2}{n}\begin{pmatrix} \hat{\lambda}_1^2 & \hat{\lambda}_3^2 \\ \hat{\lambda}_3^2 & \hat{\lambda}_2^2 \end{pmatrix}$$

and that $Z^T V^{-1} Z$ has an asymptotic χ_2^2 distribution. This statistic is well defined whereas in the usual Pearson-Fisher X^2 test it is not clear which is the best way to form classes.

Holgate (1966) gives counts of the plant species *Lacisema aggregatum* and *Protium guianense* in each of 100 quadrants. The data are given in Table 7. We calculate $Z^T V^{-1} Z$ = 12.04 with p-value 0.002 using the χ_2^2 approximation while the parametric bootstrap p-value is 0.001. The bivariate Poisson distribution is not a good model for these data. The marginal X data appear too dispersed.

3. DETERMINATION OF THE VARIANCE OF THE TEST STATISTICS

In the previous section we quoted the variances of various statistics. These variances were found using two approaches.

(1) *MOM and ML Estimation Coincide*

Suppose X is a Poisson (λ) random variable, and if the Poisson orthogonal polynomial of fourth order is

$$R = s^4 + a_4 s^3 + b_4 s^2 + c_4 s + d_4$$

where $s = x - \lambda$, $a_4 = -6$, $b_4 = 11 - 6\lambda$, $c_4 = 14\lambda - 6$ and $d_4 = 3\lambda(\lambda - 2)$

then

$$E[R^2] = \text{var}(R) = \mu_8 + 2a_4\mu_7 + (a_4^2 + 2b_4)\mu_6 + 2(a_4b_4 + c_4)\mu_5 + (b_4^2 + 2a_4c_4 + 2d_4)\mu_4 +$$
$$+ 2(2a_4d_4 + b_4c_4)\mu_3 + 2(c_4^2 + 2b_4d_4)\mu_2 + d_4^2 = 24\lambda^4$$

on using

$$\mu_2 = \lambda,\ \mu_3 = \lambda,\ \mu_4 = 3\lambda^2 + \lambda,\ \mu_5 = 10\lambda^2 + \lambda,\ \mu_6 = 15\lambda^3 + 30\lambda^2 + \lambda,$$
$$\mu_7 = 105\lambda^3 + 62\lambda^2 + \lambda,\ \mu_8 = 105\lambda^4 + 1025\lambda^3 + 126\lambda^2 + \lambda.$$

These are known results for a Poisson (λ) distribution. Note that R in the previous section is divided by n in the present section. Other variances when MOM and ML estimators coincide can be found in a similar fashion.

(2) *MOM and ML Estimation Don't Coincide*

In this case we can resort to the so-called 'delta' method. Suppose we consider, as above, a negative binomial (k, p) distribution and, if we write $X = \overline{X}$, $Y = m_2$ and $Z = m_3$,

$$T = m_3 - \hat{\mu}_3 = m_3 - \hat{k}\hat{q}(1 + \hat{q})/\hat{p}^3 = m_3 - m_2(2m_2/\overline{X} - 1) = f(X, Y, Z) = Z - 2Y^2/X + Y.$$

From the delta method

$$\text{var}(T) = \left(\frac{\partial f}{\partial X}\right)^2 \text{var}(X) + \left(\frac{\partial f}{\partial Y}\right)^2 \text{var}(Y) + \left(\frac{\partial f}{\partial Z}\right)^2 \text{var}(Z) +$$
$$2\frac{\partial f}{\partial X}\frac{\partial f}{\partial Y}\text{cov}(X,Y) + 2\frac{\partial f}{\partial Y}\frac{\partial f}{\partial Z}\text{cov}(Y,Z) + 2\frac{\partial f}{\partial Z}\frac{\partial f}{\partial X}\text{cov}(Z,X)$$

where the partial differentials are evaluated at E[X], E[Y] and E[Z]. See, for example, Stuart and Ord (2005, p.350, equation 10.12).

It is also known, for example Stuart and Ord (2005, pp. 349-350), that, to order n^{-1},

$$\text{var}(X) = \mu_2/n,\ \text{var}(Y) = (\mu_4 - \mu_2^2)/n,\ \text{var}(Z) = (\mu_6 - \mu_3^2 - 6\mu_2\mu_4 + 9\mu_3^2)/n,$$
$$\text{cov}(X, Y) = \mu_3/n,\ \text{cov}(Y, Z) = (\mu_5 - 4\mu_2\mu_3)/n,\ \text{cov}(Z, X) = (\mu_4 - 3\mu_2^2)/n.$$

For a negative binomial (k, p) distribution it is known that

$$\mu = \mu_1 = kq/p,\ \mu_2 = kq/p^2,\ \mu_3 = kq(1+q)/p^3,\ \mu_4 = (kq/p^2)(6/p^2 - 6/p + 1) + 3k^2q^2/p^4,$$

$$\mu_5 = (kq/p^2)(24/p^3 - 36/p^2 + 14/p - 1) + 10k^2q^2(1+q)/p^5 \text{ and}$$

$$\mu_6 = (kq/p^2)(120/p^4 - 240/p^3 + 150/p^2 - 30/p + 1) + 15(k^2q^2/p^4)(6/p^2 - 6/p + 1) + 10k^2q^2(1+q)^2/p^6 + 15k^3q^3/p^6.$$

It follows that

$$\text{var}(T) = 2\hat{k}(\hat{k}+1)\hat{q}^3(1+\hat{q})(10+3\hat{k}-4\hat{p})/(n\hat{p}^6)$$

as given in the previous section.

REFERENCES

Aldrete, A.N.G. 2002, *Psocoptera (Insecta) from the Sierra Tarahumara, Chihuahua, Mexico*, Anales del Instituto de Biologica, Universidad Nacional Autonoma de Mexico, Serie Zoologia, 73(2), 145-156.

Anscombe, F.J. 1950, *Sampling theory of the negative binomial and logarithmic distributions*, Biometrika, 37, 358-382.

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., Rayner, J.C.W. and Thas, O. 2009, *Anscombe's tests of fit for the negative binomial distribution*, Journal of Statistical Theory and Practice, Special Issue on Modern Goodness of Fit Methods. Rayner, J.C.W., Thas, O. and Best, D.J. (Eds.), Greensboro, North Carolina: Grace Scientific Publishing, 3, 555-565.

Birch, M.W. 1963, *Note 194: An algorithm for the logarithmic series distribution*, Biometrics, 19, 651-652.

Cochran, W.G. 1936, *The statistical analysis of field counts of diseased plants*, Supplement of the Journal of the Royal Statistical Society, 3, 49-67.

Evans, D.A. 1953, *Experimental evidence concerning contagious distributions in ecology*, Biometrika, 40, 186-196.

Devroye, L. 1986, *Non-Uniform Random Variate Generation*, Springer, New York.

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.

Holgate, P. 1966, *Bivariate generalizations of Neyman's Type A distribution*, Biometrika, 53, 241-245.

IMSL 1995, *Users' Manual*, IMSL, Houston.

Krebs, C.J. 1998, *Ecological Methodology*, Addison Wesley Longman, New York.

Pielou, E.C. 1962, *Runs of one species with respect to another in transects through plant populations*, Biometrics, 18(4), 579-593.

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

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

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

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

Appendix

P-Values via the Parametric Bootstrap

Gürtler and Henze (2000, p. 223) suggest p-values can be obtained using an analogue of the parametric bootstrap. If W_n denotes a test statistic calculate $w_n := W_n(x_1, x_2, \ldots, x_n)$ where $x_1, x_2, \ldots, x_n$ denotes, as usual, the data. Find estimates from the data and conditional on this estimate, generate B = 10,000 say pseudo-random samples of size n, each having the appropriate distribution. For j = 1, ..., B compute the value $W^*_{n,j}$ on each random sample. The parametric bootstrap p-value is then the proportion of the $W^*_{n,j}$ that are at least the observed w_n, namely $\sum_{j=1}^{B} I\left(W^*_{n,j} \geq w_n\right) / B$.

The above requires random values from various distributions. Devroye (1986) outlines algorithms for generating random deviates from many distributions. Alternatively, the routines from IMSL (1995) can be used. To obtain p-values for two-tailed tests proceed as above and find the p-value, P say. Then if $P \leq 0.5$ the two-tailed p-value is $2P$, while if $P > 0.5$ the two-tailed p-value is $2(1 - P)$.

Algorithms for random values from the four distributions below are not given in IMSL (1995) and so we now outline possible approaches.

(a) ***Zero Inflated Poisson***

i. Generate a random x value from a binomial (n, $\hat{w}$) distribution and take $y = n - x$.

ii. Generate y random values from a Poisson ($\hat{\lambda}$) distribution.

iii. These y random values and the x zeros constitute the random sample of size n.

(b) ***Neyman Type A***

i. Generate a random value m from a Poisson ($\hat{\lambda}$) distribution and then generate m values x_1, x_2, ..., x_m from a Poisson ($\hat{\varphi}$) distribution. A random Neyman Type A value is then $y = x_1 + x_2 + \ldots + x_m$.

ii. Repeat n times.

(c) ***Bivariate Poisson***

i. Generate a random values x, y and z from Poisson ($\hat{\lambda}_1 - \hat{\lambda}_3$), Poisson ($\hat{\lambda}_2 - \hat{\lambda}_3$) and Poisson ($\hat{\lambda}_3$) distributions respectively. A bivariate Poisson random value is (u, v) where $u = x + z$ and $v = y + z$.

ii. Repeat n times.

(d) ***Beta-Binomial***

i. Generate a random value p from beta $(\hat{\alpha}, \hat{\beta})$ distribution. Generate a random beta-binomial value as a random value from a binomial (n^*, p) distribution.

ii. Repeat n times.

Table 1. Counts of black bean aphids

# aphids per stem	0	1	2	3	4	5	6	7	8	9
frequency	6	8	9	6	6	2	5	3	1	4

Table 2. Counts of eggs on flower heads

# eggs	1	2	3	4	5	6	7	8	9
frequency	22	18	18	11	9	6	3	0	1

Table 3. Catch frequencies per species

Times caught	1	2	3	4	5	6	8	10	11	12	13	16	25	69	95	≥ 13
# species	10	3	4	2	2	2	1	1	1	1	1	1	2	1	1	6

Table 4. Counts of the abundance of Leadbeater's Possum

# 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 5. Counts of plants per plot with spotted wilt

# plants	0	1	2	3	4	5	6	7	8	9
frequency	36	48	38	23	10	3	1	1	0	0

Table 6. Distribution of run lengths

Run length	1	2	3	4	5	7	10	12	15	23
frequency	66	22	8	3	8	2	1	1	1	1

Table 7. Counts of two plant species

$Y \setminus X$	0	1	2	3	4	5	6
0	34	12	4	5	2	0	1
1	8	13	3	3	0	0	0
2	3	6	1	2	0	0	0
3	1	1	0	1	0	0	0