doi:10.20944/preprints201807.0534.v1

A dispersion test for the modified Borel-Tanner distribution

D.J. Best[a] and J.C.W. Rayner[b,a*]
[a]School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia
[b]National Institute for Applied Statistics Research Australia, University of Wollongong, NSW 2522, Australia

Abstract

Dispersion tests based on the second order component of smooth test statistics are related to Fisher's Index of Dispersion test, used for testing for the Poisson distribution when there are no covariates present. Such tests have been recommended in [1] to test for the Poisson distribution when covariates are present. The modified Borel-Tanner (MBT) distribution seems suited to data with extra zeroes, a monotonic decline in counts and longer tails. Here we recommend a dispersion test for the MBT distribution for both when covariates are absent and when they are present.

AMS Subject Classification: 62G; 62F

Key Words: covariates; crab data; foetal lamb data; orthonormal polynomials; Poisson distribution

1. Introduction

Dispersion tests for count data have long been used in statistical analysis. As the Poisson is the most well-known count data model we begin with discussing a dispersion test for this model. Suppose we wish to see if the counts $y_1, y_2, \ldots , y_n$ are Poisson distributed with mean μ: that is, whether or not the random variable Y has probability function $f(y, \mu) = e^{-\mu}\mu^y/y!$, $y = 0, 1, \ldots$. A dispersion test of whether or not this applies is based on the classic Index of Dispersion statistic $D = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 / \bar{Y}$. See, for example, [2, p.58]. Let the second orthonormal polynomial on the Poisson be

$$h_2(y, \hat{\mu}) = \{(y - \mu)^2 - y\} / \sqrt{2\mu^2}$$

*Corresponding author.
E-mail address: John.Rayner@newcastle.edu.au
Phone: 61 2 49215737

doi:10.20944/preprints201807.0534.v1

in which $\hat{\mu}$ = $\bar{Y}$ is the maximum likelihood estimator of μ. Also let $\hat{V}_2$ = $\sum_{i=1}^{n} h_2(Y_i, \hat{\mu}) / \sqrt{n}$. Then we can link the classic Poisson Index of Dispersion statistic and $\hat{V}_2$ via

$$\hat{V}_2 = (D - n) / \sqrt{2n} .$$

The statistic $\hat{V}_2$ is the second component of the smooth test of fit statistic for the Poisson. See [3].

The discussion of the previous paragraph applies to no covariates Poisson tests of fit. If there are covariates to consider then it is shown in [4] that

$$\hat{V}_2 = \sum_{i=1}^{n} h_2(Y_i, \hat{\mu}_i) / \sqrt{n}$$

is the second order smooth test statistic component in a Poisson regression. Note the remarkable similarity with the no covariate Poisson second component smooth test statistic $\hat{V}_2$ given just above. The only difference is that $\hat{\mu}$ has become $\hat{\mu}_i$. Observe in [1] (Table 1) that her P_C is just $\hat{V}_2$. Again, as in the no covariates case, $\hat{V}_2$ is simply expressed in terms of a well-known Poisson dispersion test. The links between well-established Poisson dispersion tests and the $\hat{V}_2$ dispersion test suggest that the $\hat{V}_2$ second order component smooth tests of fit for both the no covariates and the with covariate cases might be usefully generalized to a new one-parameter count distribution: the modified Borel-Tanner (MBT) distribution recently proposed in [5].

Section 2 defines the MBT distribution and gives examples where it is a good model for real count data. Section 3 considers the no covariates case and section 4 the with covariates case. Section 5 gives some concluding comments.

2. The modified Borel-Tanner distribution

The modified Borel-Tanner (MBT) distribution introduced in [5] has probability function

$$f(y,\mu) = {}^{2y}C_y \mu^y (1+\mu)^{1+y} / \{(y+1)(1+2\mu)^{1+2y}\} \text{ for } y = 0,\ 1,\ 2,\ \dots\ .$$

Observe that if $\alpha = \mu/(1 + \mu)$ then

$$f(0, \mu) = 1/(1 + \alpha) \text{ and}$$
$$f(y,\mu) = 2\alpha(2y-1) f(y-1,\mu) / \{(1+\alpha)^2 (y+1)\} \text{ for } y > 0.$$

The first four cumulants are

$$\kappa_1 = \mu = \alpha/(1\text{-}\alpha),\ \kappa_2 = \alpha\,(1 + \alpha)/(1 - \alpha)^3,\ \kappa_3 = \alpha(1 + \alpha)(1 + 6\alpha - \alpha^2)/(1 - \alpha)^5$$
$$\text{and } \kappa_4 = \alpha(1 + \alpha)(1 + 21\alpha + 36\alpha^2 + 3\alpha^3 - \alpha^4)/(1 - \alpha)^7 - 3\kappa_2^2 .$$

We suppose n observations are available.

The first three orthonormal polynomials on the MBT distribution are

$$h_0(y, \mu) = 1,\ h_1(y, \mu) = t / \sqrt{\kappa_2} \text{ and}$$
$$h_2(y, \mu) = (t^2 - \kappa_3 t / \kappa_2 - \kappa_2) / \sqrt{(\kappa_4 + 2\kappa_2^2 - \kappa_3^2 / \kappa_2)}$$

in which $t = (y - \kappa_1)$. The MLE of the parameter α is $\hat{\alpha} = \overline{Y} / (1+\overline{Y})$.

To demonstrate that the MBT is an important model for real data we observe that data sets well described by the MBT include

- the accident counts in [6, p.115], namely 296*0, 74*1, 26*2, 8*3, 4*4, 4*5, 6 and 8;
- the counts of international terrorism in Turkey in [7, p.412], namely 60*0, 9*1, 4*2, 2*3 and 5;
- the foetal lamb counts 182*0, 41*1, 12*2, 3, 3, 4, 4, 0, 0, 7 discussed in [8] and
- the Carpenter bee larvae data 114*0, 25*1, 15*2, 10*3, 6*4, 3*5, 6, 6, 7, 8, 10 in [9].

3. The no covariates case

As with the Poisson we use the MBT second order orthonormal polynomial to give a dispersion test for the MBT. This uses the test statistic

$$\hat{V}_2 = \sum_{i=1}^{n} h_2(Y_i, \hat{\mu}) / \sqrt{n} \; = \sum_{i=1}^{n} (T_i^2 - \kappa_3 T_i / \kappa_2 - \kappa_2) / \sqrt{(\kappa_4 + 2\kappa_2^2 - \kappa_3^2 / \kappa_2)}$$

where, as before, $T_i = Y_i - \hat{\mu}$ and in the cumulants κ_r the maximum likelihood estimators are used. This dispersion test has an approximate χ_1^2 distribution. For the foetal lamb data we have $n = 240$, $\bar{y} = \hat{\mu} = 0.358$ and $\hat{\alpha} = 0.264$. We find $\hat{V}_2^2$ = 1.08 which is not significant at any of the usual levels of significance (p-value 0.30) and so we might conclude the MBT describes the data well. It is shown in [3, p.237] that the zero inflated Poisson fits the data well if the observation at y = 7 is removed. The one parameter MBT fits the data well without this data removal. The MBT seems suited to data with extra zeroes, a monotonic decline in counts and longer tails.

4. The with covariates case

Agresti [10, p.123] considers Y_i to be the count of the number of male satellite crabs attracted to the ith female crab during the mating season. In the following the width of the shell of the female, X_{1i}, is taken as the covariate for the ith female. There were n = 173 female crabs considered and so 173 (x_{1i}, y_i) observations were available. These are listed in the Appendix. Using the GLM command in R and the `disptest` routine in the `countreg` library, namely

```
fit=glm(y~x;poisson)
disptest(fit,type="scoreNB1")
```

or otherwise, we find $\hat{\mu}_i = \exp(-3.300 + 0.164 x_{1i})$ for a Poisson regression. Further the Poisson dispersion statistic $\hat{V}_2^2$ = 402.6 and using the χ_1^2 approximation gives a p-value of less than 0.0001. The Poisson model is not appropriate.

A one parameter model which allows more dispersion than the Poisson may do better and so we consider the MBT model. In similar fashion to the no covariates case a dispersion test statistic for the MBT regression is

$$\hat{V}_2^2 = \sum_{i=1}^{n} h_2(T_i,\hat{\alpha}_i)/\sqrt{n},\ \hat{T}_i = Y_i - \hat{\mu}\,.$$

To calculate a MBT regression with fitted values $\hat{\mu}_i = \exp(\hat{\beta}_1 + \beta_2 x_{1i})$ we follow [5] and find $\hat{\beta}_1$ and $\hat{\beta}_2$ from the nonlinear equations

$$\sum_{i=1}^{n} z_i = 0 \text{ and } \sum_{i=1}^{n} x_{1i} z_i = 0 \text{ in which}$$
$$z_i = \{1-2p(\hat{\mu}_i)\}\{y_i - (1+2y_i p(\mu_i)\}/\{1-p(\hat{\mu}_i)\} \text{ and } p(\hat{\mu}_i) = \mu_i/\{1+2\hat{\mu}_i\}\,.$$

These $\hat{\beta}_1$ and $\hat{\beta}_2$ are MLEs. We solved these equations for the crab data using IMSL routine NEQNF. Put $\alpha_i = \mu_i/(1+\mu_i)$. The second, third and fourth MBT cumulants are as in section 2 with α replaced by α_i. We find $\hat{\mu}_i = \exp(-5.191 + 0.236x_{1i})$ and can then calculate $\hat{V}_2^2$ = 0.975 with approximate p-value 0.32.

The Fisher information matrix for the MBT regression is

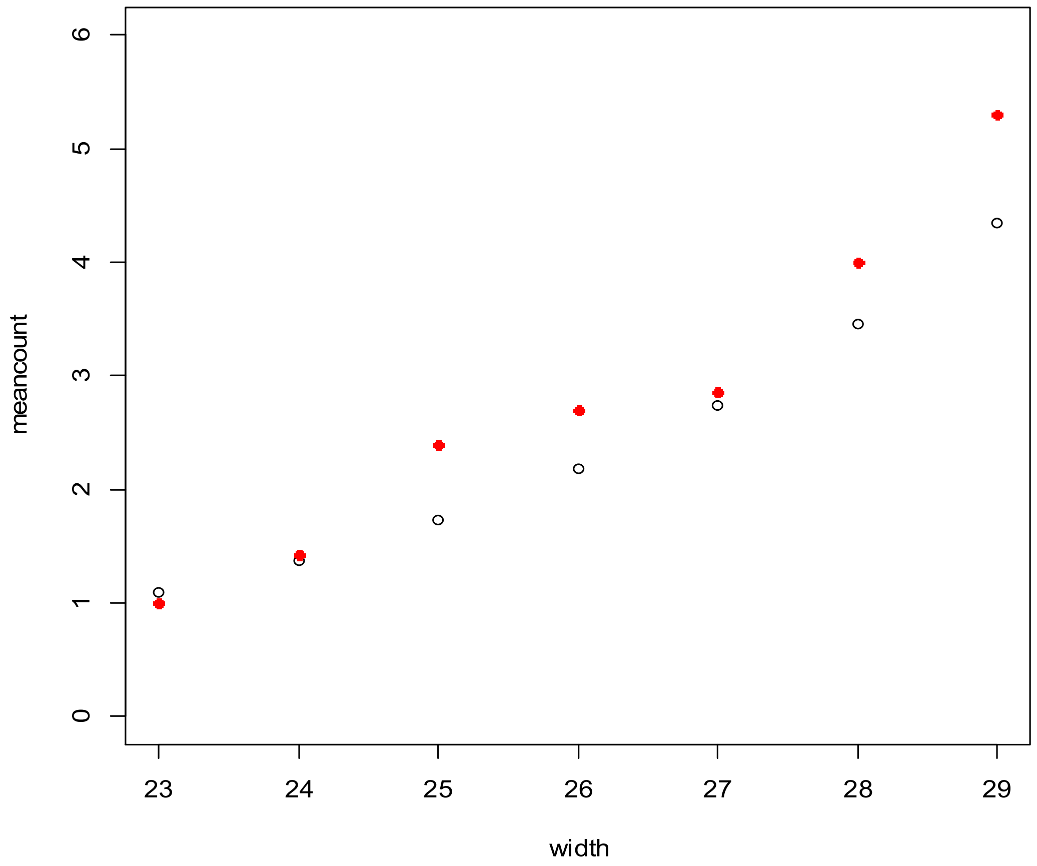
$$10^3 \begin{pmatrix} 0.1432 & 3.789 \\ 3.789 & 100.9 \end{pmatrix}$$

and so the asymptotic standard error of $\hat{\beta}_2$ is SE($\hat{\beta}_2$) = 0.040. Agresti [10, p.156] finds SE($\hat{\beta}_2$) = 0.020 for a Poisson regression and SE($\hat{\beta}_2$) = 0.048 for a negative binomial regression. He explains that the Poisson regression SE($\hat{\beta}_2$) is smaller because it does not take into account the overdispersion in the data. The MBT regression, like the negative binomial regression, does take into account the overdispersion. In Figure 1 we have plotted fitted values for the regression equation at the shell widths shown (closed circles) and means of satellite counts in the range shell width +/- 0.5 (open circles). Figure 1 suggests the linear model for the covariate is reasonable. Agresti [10, p.126] gives a similar figure to our Figure 1 which compares link functions whereas our figure compares the log link MBT model fit at mid ranges of shell widths with the mean of satellite counts for the same ranges of shell widths. Our figure attempts to relate the MBT regression model to the data.

5. Comment

Although we have referred to $\hat{V}_2^2$ as a dispersion test statistic, we observe that this statistic can be large because higher order effects than the dispersion effect can differ from what is expected. However, it could be said that $\hat{V}_2^2$ may *suggest* over or under dispersion and that large values indicate deviations from the MBT model. While not always diagnosing dispersion differences $\hat{V}_2^2$ is a dispersion test statistic in that it compares sample and population dispersions or variances.

Figure 1. Mean Count vs Shell Width



References

1. Dean, C. Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association* **1992**, *87*, 451-457.
2. Fisher, R.A. *Statistical Methods for Research Workers*, 9th ed.; Oliver and Boyd: Edinburgh, **1944**.
3. Rayner, J.C.W.; Thas, O.; Best, D.J. *Smooth Tests of Goodness of Fit: Using R*, 2nd ed.; Wiley: Singapore, **2009**.
4. Rippon, P. Application of Smooth Tests of Goodness of Fit to Generalized Linear Models. Ph.D. Thesis, University of Newcastle, Callaghan, Australia, **2013**.
5. Gomez-Deniz, E., Vazquez - Polo, F. and Garcia, V. The modified Borel-Tanner regression model. *Revstat* **2017**, *15*, 425-442.
6. Simonoff, J. *Analyzing Categorical Data*; Springer: New York, **2003**.

doi:10.20944/preprints201807.0534.v1

7. Hoaglin, D., Mostellar, F. and Tukey, J. *Exploring Data Tables, Trends, and Shapes*; Wiley: New York, **1985**.
8. Douglas, J., Leroux, B. and Puterman, M. Empirical fitting of discrete distributions. *Biometrics* **1994**, *50*, 576-579.
9. Jayarama, K. *A Statistical Manual for Forestry Research.* FAO United Nations: Bangkok **1999**.
10. Agresti, A. *Categorical Data Analysis*, 3rd ed.; Wiley: New York, **2013**.

Appendix

The 173 (x_i , y_i) , crab observations were

(28.3, 8), (22.5, 0), (26.0, 9),(24.8, 0), (26.0, 4), (23.8, 0),(26.5, 0), (24.7, 0), (23.7,0,), (25.6, 0), (24.3, 0), (25.8, 0), (28.2, 11), (21.0, 0), (26.0, 14), (27.1, 8), (25.2, 1), (29.0, 1), (24.7, 0), (27.4, 5), (23.2, 4), (25, 3), (22.5, 1), (26.7, 2), (25.8, 3), (26.2, 0), (28.7, 3), (26.8, 5), (27.5, 0), (24.9, 0), (29.3,4),(25.8, 0), (25.7, 0), (25.7, 8), (26.7, 5), (23.7, 0),(26.8, 0), (27.5, 6), (23.4, 0), (27.9, 6), (27.5, 3), (26.1, 5), (27.7, 6), (30.0, 5), (30.0, 5), (28.5, 9), (28.9, 4), (28.2, 6), (25.0, 4), (28.5, 3), (30.3, 3), (24.7, 5), (27.7, 5), (27.4, 6), (22.9, 4), (25.7, 5), (28.3, 15), (27.2, 3), (26.2, 3),(27.8, 0), (25.5,0),(27.1, 0), (24.5, 5), (27.0, 3), (26.0, 5), (28.0, 1), (30.0, 8), (29.0, 10), (26.2, 0), (26.5, 0), (26.2, 3), (25.6, 7),(23.0, 1), (23.0, 0), (25.4, 6), (24.2, 0), (22.9, 0), (26.0, 3,), (25.4, 4), (25.7, 0), (25.1, 5), (24.5, 0), (27.5, 0), (23.1, 0), (25.9, 4), (25.8, 0), (27.0, 3), (28.5, 0), (25.5, 0), (23.5, 0), (24.0, 0), (29.7, 5), (26.8, 0), (26.7, 0), (28.7, 0), (23.1, 0), (29.0, 1), (25.5, 0), (26.5, 1), (24.5, 1), (28.5, 1), (28.2, 1), (24.5, 1), (27.5, 1), (24.7, 4), (25.2, 1), (27.3, 1), (26.3, 1), (29.0, 1), (25.3, 2), (26.5, 4), (27.8, 3), (27.0, 6), (25.7, 0), (25.0, 2), (31.9, 2), (23.7, 0), (29.3, 12), (22.0, 0), (25.0, 5), (27.0, 6), (23.8, 6), (30.2, 2), (26.2, 0), (24.2, 2), (27.4, 3), (25.4, 0), (28.4, 3), (22.5, 4), (26.2, 2), (24.9, 6), (24.5, 6), (25.1, 0), (28.0, 4), (25.8,10), (27.9, 7), (24.9, 0), (28.4, 5), (27.2, 5), (25.0, 6), (27.5, 6), (33.5, 7), (30.5, 3), (29.0, 3), (24.3, 0), (25.8, 0), (25.0, 8), (31.7, 4), (29.5, 4), (24.0, 10), (30.0, 9), (27.6, 4), (26.2, 0), (23.1, 0), (22.9, 0), (24.5, 0),(24.7,4), (28.3, 0), (23.9, 2), (23.8, 0), (29.8, 4), (26.5, 4), (26.0, 3), (28.2, 8), (25.7, 0), (26.5, 7), (25.8, 0), (24.1, 0), (26.2, 2), (26.1, 3), (29.0, 4), (28.0, 0), (27.0, 0), (24.5, 0).