NOVA

University of Newcastle Research Online

nova.newcastle.edu.au

Best, D. J.; Rayner, J. C. W. "Improved testing for the binomial distribution using chi-squared components with data-dependent cells" Journal of Statistical Computation and Simulation Vol. 76, Issue 1, p. 75-81 (2006)

Available from: <http://dx.doi.org/10.1080/00949650412331320891>

This is an Accepted Manuscript of an article published in Journal of Statistical Computation and Simulation on 19/08/2006, available online: <http://www.tandfonline.com/10.1080/00949650412331320891>

Accessed from: <http://hdl.handle.net/1959.13/34548>

IMPROVED TESTING FOR THE BINOMIAL DISTRIBUTION USING CHISQUARED COMPONENTS WITH DATA DEPENDENT CELLS

D.J. BEST¹ and J.C.W. RAYNER²

The University of Wollongong

SUMMARY

A power study suggests that a good test of fit analysis for the Binomial distribution is provided by a data dependent Chernoff-Lehmann X^2 test with class expectations greater than unity, and its components. These data dependent statistics involve arithmetically simple parameter estimation, convenient approximate distributions, and provide a comprehensive assessment of how well the data agree with a Binomial distribution. We suggest that a well-performed single test of fit statistic is the Anderson-Darling statistic.

Key Words: Goodness of fit; dispersion test; non-central chisquared distribution; partition of X^2 .

1. Introduction

The Binomial distribution is one of the most important discrete distributions. Tests of fit for the Binomial distribution have been compared recently by Spinelli (1994) and Best and Rayner (1997). Spinelli (1994) shows that if nothing is known about the alternative distribution then the Anderson-Darling test provides a powerful test of fit; we agree with this assessment. Here we add to the previous comparisons by examining the performance of tests based on the components of a particular Chernoff-Lehmann X^2 statistic. These statistics have not been included in previous comparisons but, as we show, they have good power properties. There are also convenient approximations to the null distributions of the test statistics. The Chernoff-Lehmann X^2 statistic is an important statistic to consider as textbook writers commonly use it to test for the Binomial distribution. See, for example, Snedecor and Cochran (1989, p.205) or Mead, Curnow and Hasted (2003, p.341). Examination of the components of this statistic improves its use by suggesting how a given data set differs from a Binomial data set. Section 3 illustrates this point and section 5 illustrates how the components can be used to provide an alternative model if the Binomial assumption is rejected.

For testing the fit of data to a Binomial distribution we recommend a comprehensive analysis based on the Chernoff-Lehmann X^2 statistic and its components. This can be done using critical values based on standard chisquared distributions. We also recommend that formal inference to complement this X^2 analysis be based on the Anderson-Darling test.

¹ *School of Mathematics and Applied Statistics, The University of Wollongong, Wollongong, NSW 2522, Australia*

Email: jbest@uow.edu.au

² *School of Mathematics and Applied Statistics, The University of Wollongong, Wollongong, NSW 2522, Australia*

Email: john_rayner@uow.edu.au

2. Definitions

The Binomial probability function is

$$p_j = {}^k C_j p^j (1-p)^{k-j}, j = 0, 1, \dots, k.$$

Suppose we wish to test the hypothesis that n observations come from a Binomial distribution $\mathcal{F}$, against the general alternative, not $\mathcal{F}$. A common approach is to form $k+1$ classes from the data with associated counts $N_j, j = 0, 1, \dots, k$, and to use the familiar Pearson goodness of fit test based on the statistic

$$X^2 = \sum_{j=0}^k \frac{(N_j - np_j)^2}{np_j},$$

where $n = N_0 + \dots + N_k$.

We focus on the important case where the unknown parameter p is estimated by

$$\hat{p} = \bar{Z},$$

where $\bar{Z}$ is the mean of an ungrouped random sample of size n from the Binomial distribution, namely $Z_1, \dots, Z_n$. The estimator $\hat{p}$ is thus the maximum likelihood estimator based on the *ungrouped* data. It is known that X^2 based on $\hat{p}$, X_{CL}^2 say, is a Chernoff-Lehmann (1954) statistic under $\mathcal{F}$ and does not have an asymptotic χ^2 distribution. The null distribution of X_{CL}^2 is asymptotically a linear combination of χ^2 variables and is bounded between χ_{k-1}^2 and χ_k^2 . See Kimber (1987) for an easy to read discussion that supplements the original Chernoff and Lehmann (1954) article. Note that Snedecor and Cochran (1989, p.205) and Mead, Curnow and Hasted (2003, p.341), as do other authors, use just the χ_{k-1}^2 distribution.

The X^2 Binomial test may lose information if too much pooling is done. On the other hand, the χ^2 approximation to the null distribution of X^2 may be poor unless at least some pooling is done. Clearly, choice of classes is a problem for the X^2 Binomial test. Further, the X^2 test may also have low power because it is an omnibus test: it has some power against many alternatives, but may not have substantial power against any particular alternative.

There are various rules for choosing the number of classes for the X_{CL}^2 test. Here we take this number to be as large as possible such that each class has expectation at least unity and call the test statistic the X_{CLE1+}^2 statistic. This may involve grouping from above as well as grouping from below and agrees with the suggestion of Douglas (1994). For the reason given at the end of this section, if the number of classes that results from following the expectation greater than unity rule is less than four, then we reduce the expectation of each class to force the number of classes to be exactly four. However we will still call the resulting statistic X_{CLE1+}^2 .

The X^2 statistic can be partitioned into useful components. For many alternatives some of these components have much greater power than X^2 itself. Put

$$\mu = \sum_{j=0}^k j p_j \text{ and } \mu_r = \sum_{j=0}^k (j - \mu)^r p_j \text{ for } r = 2, 3, \dots$$

and define the initial orthonormal polynomials by

$$g_0(j) = 1, g_1(j) = (j - \mu)/\sqrt{\mu_2}, g_2(j) = \{(j - \mu)^2 - (j - \mu)\mu_3/\mu_2 - \mu_2\}/\sqrt{\mu_4 - \mu_3^2/\mu_2 - \mu_2^2}, \text{ and}$$

$$g_3(x_j) = \frac{(j - \mu)^3 - a(j - \mu)^2 - b(j - \mu) - c}{\sqrt{\mu_6 - 2a\mu_5 + (a^2 - 2b)\mu_4 + 2(ab - c)\mu_3 + (b^2 + 2ac)\mu_2 + c^2}}$$

in which

$$a = (\mu_5 - \mu_3\mu_4/\mu_2 - \mu_2\mu_3)/d, b = (\mu_4^2/\mu_2 - \mu_2\mu_4 - \mu_3\mu_5/\mu_2 + \mu_3^2)/d,$$

$$c = (2\mu_3\mu_4 - \mu_3^2/\mu_2 - \mu_2\mu_5)/d \text{ and } d = \mu_4 - \mu_3^2/\mu_2 - \mu_2^2.$$

It is possible to derive $g_4(j), \dots, g_k(j)$ so that $\{g_r(j)\}$ is a set of orthonormal functions by using the recurrence relations in Emerson (1968). Take

$$U_r = \sum_{j=0}^k N_j g_r(j) / \sqrt{n}, r = 1, \dots, k.$$

Then, as in Lancaster (1953),

$$X^2 = U_1^2 + \dots + U_k^2.$$

A consequence of estimation of p by $\hat{p}$ is that U_1 will be close to zero if not much pooling is done. Unlike $U_2, \dots, U_k$, U_1 will *not* have an asymptotic standard normal distribution.

If $\pi_j, j = 0, 1, \dots, k$ are the cell or class probabilities specified by the alternative hypothesis then approximate powers for X^2 , U_2^2 and U_3^2 can be obtained by using non-central χ^2 distributions with degrees of freedom $k - 1$, 1 and 1 respectively and non-centrality parameters

$$n \sum_{j=0}^k (p_j - \pi_j)^2 / \pi_j, n \left\{ \sum_{j=0}^k g_2(j) \pi_j \right\}^2 \text{ and } n \left\{ \sum_{j=0}^k g_3(j) \pi_j \right\}^2$$

respectively. If we are to look at powers of X^2 , U_2^2 and U_3^2 , we require at least four classes: $k \geq 3$. With fewer classes U_3 cannot be defined. Moreover a discrete distribution with four or fewer classes may be consistent with the Binomial, but it will also be consistent with many other distributions, so testing for the Binomial in this situation is somewhat artificial.

3. Example

Lindsay (1995, p.131) gives frequencies of sons in 6115 families of 12 children - the classical data of Geissler. For 0, 1, 2, $\dots$, 12 sons per family the counts are

3, 24, 104, 286, 670, 1033, 1343, 1112, 829, 478, 181, 45, and 7.

The probability of a son is 0.51912. For a binomial model $X^2 = 113.2$ on 11 degrees of freedom which is very highly significant. Following our suggested rule we should combine counts for 0

and 1 sons but as the expectation for zero sons is close to unity we have not done so. Most of the X^2 value is accounted for by $U_2^2 = 90.60$. As U_2 is negative the distribution of counts is more peaked than expected for a Binomial distribution. This straightforward description of how the data differ from the Binomial model is not available from most tests of fit for the Binomial model.

In section 5 we will use the components of X^2 to develop adjusted Binomial models.

4. Size And Power Comparisons

Table 1 gives a size comparison of the smooth tests V_2^2 , V_3^2 , V_4^2 defined in Best and Rayner (1997) with the components of the X_{CLE1+}^2 statistic U_2^2 , U_3^2 and U_4^2 . In all cases the critical values used were based on the approximating χ^2 distribution: that is, the χ_1^2 distribution was used for all U_i^2 and all V_i^2 , $i = 2, 3, 4$, and the χ_{k-1}^2 distribution was used for X_{CLE1+}^2 . Note that for the $(n, \alpha, 0.1)$ values shown the actual sizes of U_4^2 and X_{CLE1+}^2 are closer to the nominal size than the actual sizes of V_4^2 and X^2 (no pooling). For $\alpha = 0.05$ and $p = 0.5$ the sizes of the pooled and unpooled components are similar. For $\alpha = 0.10$ the superiority of X_{CLE1+}^2 and its components is more apparent. The powers of A^2 shown in Table 2 are based on accurate Monte Carlo critical values that give actual sizes very close to the nominal α value. In Best and Rayner (1997) it was suggested that P-values for X^2 , V_2^2 , V_3^2 and V_4^2 be found via Monte Carlo methods. It is more convenient to be able to use standard chi-squared probabilities. Table 1 indicates the chisquared approximation is more reasonable for X_{CLE1+}^2 , U_2^2 , U_3^2 and U_4^2 .

In Table 2 three types of alternative distribution are used:

- (i) $BB(a, b, m^*)$: a beta-binomial distribution with parameters a , b , and m^* . Observations from this distribution come from a composite binomial distribution with parameters m^* and p^* , where p^* has a standard beta distribution with parameters a and b ;
- (ii) $DU(i, j)$: a discrete uniform distribution defined on integers $i, i + 1, \dots, j$. The probability of a value x occurring is $(j - i + 1)^{-1}$, where $i \leq x \leq j$;
- (iii) $M(p_0, p_1, \dots, p_k)$: a standard multinomial with specified $p_0, p_1, \dots, p_k$.

TABLE 1: Sizes, based on χ^2 critical values, of the pooled (U_i) and unpooled (V_i) components used as tests of the Binomial assumption when $n = 20$.

α	p	k	X_{CLE1+}^2	X^2	U_2^2	V_2^2	U_3^2	V_3^2	U_4^2	V_4^2
0.05	0.1	5	0.051	0.028	0.040	0.035	0.045	0.018	-	0.019
	0.5	5	0.047	0.042	0.059	0.046	0.044	0.048	0.048	0.042
	0.1	20	0.042	0.014	0.050	0.036	0.045	0.047	0.054	0.007
	0.5	20	0.042	0.049	0.048	0.038	0.045	0.047	0.047	0.030
0.10	0.1	5	0.083	0.034	0.069	0.062	0.072	0.031	-	0.035
	0.5	5	0.100	0.086	0.101	0.098	0.097	0.095	0.093	0.085
	0.1	20	0.094	0.016	0.104	0.081	0.097	0.074	0.105	0.015
	0.5	20	0.090	0.061	0.095	0.083	0.095	0.078	0.093	0.049

Missing entries for U_4^2 are because the maximum number of classes is 4 and U_4^2 requires 5 at least classes.

To enable comparisons with powers presented in Best and Rayner (1997), as well as the powers given in Spinelli (1994), powers were calculated for $n = 20$ and $\alpha = 0.05$. These are

shown in Table 2. Details of how to generate the random deviates used in Tables 1 and 2 were given in Best and Rayner (1997). Notice that the critical values were again based on the approximating χ^2 distributions, except for the Anderson-Darling A^2 statistic powers. These were not calculated here, but taken directly from Spinelli (1994). The previous power studies of Spinelli (1994) and Best and Rayner (1997) indicated that the Anderson-Darling statistic A^2 had good power for testing Binomial fit and so we include it here as a benchmark. It is, of course, not appropriate to compare two or more statistics collectively with A^2 . However, we suggest A^2 be used for more formal inference and X^2_{CLE1+} and its components U_2^2 and U_3^2 be calculated to suggest how an empirical data set differs from the Binomial assumption.

TABLE 2: Powers of Pooled Component Tests for $n = 20$, $\alpha = 0.05$ a) $p = 0.1$, $k = 5$

Alternative	σ^2	A^2	X^2_{CLE1+}	U_2^2	U_3^2
BB(5, .7, 6.3)	0.675	0.33	0.31	0.32	0.14
BB(5, .3, 2.7)	0.900	0.64	0.58	0.61	0.24
DU(0, 1)	0.25	0.42	0.25	0.25	0.02
M(.7, .1, .2, 3*0)	0.655	0.76	0.66	0.32	0.64
M(.65, .2, .15, 3*0)	0.551	0.28	0.20	0.06	0.24
B(5, .1)	0.45	0.05	0.05	0.04	0.05

b) $p = 0.5$, $k = 5$

Alternative	σ^2	A^2	X^2_{CLE1+}	U_2^2	U_3^2
BB(5, 2*3.5)	1.875	0.31	0.18	0.27	0.05
BB(5, 2*1.5)	2.5	0.73	0.53	0.66	0.05
DU(0, 5)	2.92	0.91	0.74	0.84	0.04
DU(1, 4)	1.25	0.14	0.16	0.14	0.15
M(.083, 0, 2*.417, 0, .083)	1.25	0.79	0.67	0.44	0.32
B(5, .5)	1.25	0.05	0.05	0.06	0.04

c) $p = 0.1$, $k = 20$

Alternative	σ^2	A^2	X^2_{CLE1+}	U_2^2	U_3^2
BB(20, 1.667, 10.5)	4.5	0.77	0.53	0.65	0.06
BB(20, 3.7222, 33.5)	2.7	0.25	0.13	0.21	0.05
DU(0, 4)	2.0	0.27	0.30	0.22	0.25
DU(1, 3)	0.6667	0.72	0.61	0.96	0.00
M(.2333, .1, .2, .3667, .1, 16*0)	1.8	0.60	0.65	0.15	0.74
B(20, .1)	1.8	0.05	0.04	0.05	0.05

d) $p = 0.5$, $k = 20$

Alternative	σ^2	A^2	X^2_{CLE1+}	U_2^2	U_3^2
BB(20, 18.5, 18.5)	7.5	0.27	0.12	0.22	0.06
BB(20, 5.8333, 5.8333)	12.5	0.83	0.49	0.72	0.06
DU(6, 14)	6.667	0.35	0.26	0.38	0.12
DU(8, 12)	2.0	0.45	0.35	0.80	0.00
M(7*0, .2, .15, 3*0, 3*.1, .15, .2, 7*0)	5.0	0.31	0.57	0.05	0.35
B(20, .5)	5.0	0.05	0.04	0.05	0.05

The power of the A^2 statistic is competitive for all alternatives. The power of U_2^2 is close to the power of A^2 except when the alternative has similar dispersion to the binomial. In such cases the power of X_{CLE1+}^2 is close to the A^2 power. The power of U_3^2 equals the size for symmetric alternatives, and generally is less than the power of at least one of X_{CLE1+}^2 and U_2^2 .

Given the reasonable sizes for X_{CLE1+}^2 and for U_2^2 and U_3^2 in Table 1, it is of interest to see how well the non-central χ^2 approximation for the power of the alternatives performs. For $n = 20$ we have taken at least four classes so that U_3^2 can be calculated. As indicated by the (l, u) column of Table 3 these powers based on non-central chisquared use a fixed assignment of classes. This means they cannot be expected to match exactly the powers of Table 2. In fact the differences in power between Tables 2 and 3 are sufficiently large that we caution against the use of the non-central χ^2 approximation to the powers.

TABLE 3. Comparison of Powers for X_{CLE1+}^2 and U_2^2 based on Monte Carlo Simulations and the non-central χ^2 approximation with pooling specified by the range (l, u)

p	k	(l, u)	Alternative	X_{CLE1+}^2	noncent X_{CLE1+}^2	U_2^2	noncen U_2^2
0.1	5	(0, 3)	BB(5, .7, 6.3)	0.31	0.18	0.32	0.32
0.1	5	(0, 3)	BB(5, .3, 2.7)	0.58	0.49	0.61	0.72
0.5	5	(1, 4)	BB(5, 2*3.5)	0.18	0.17	0.27	0.23
0.5	5	(1, 4)	BB(5, 2*1.5)	0.53	0.47	0.66	0.58
0.5	5	(1, 4)	DU(0, 5)	0.74	0.70	0.84	0.77
0.1	20	(0, 4)	BB(20, 1.667, 10.5)	0.53	0.41	0.65	0.63
0.1	20	(0, 4)	BB(20, 3.722, 33.5)	0.13	0.12	0.21	0.19
0.1	20	(0, 4)	DU(0, 4)	0.30	0.22	0.22	0.33
0.5	20	(7, 13)	BB(20, 2*18.5)	0.12	0.10	0.22	0.18
0.5	20	(7, 13)	BB(20, 2*5.8333)	0.49	0.34	0.72	0.62

TABLE 4. Comparison of three models for Geissler data

Sons	Observed	Round ($n \hat{p}_j$)	Round ($n \hat{\pi}_j$)	Double Binomial
0	3	1	3	2
1	24	12	25	19
2	104	72	111	95
3	286	259	319	299
4	670	628	659	658
5	1033	1085	1028	1060
6	1343	1367	1241	1285
7	1112	1266	1169	1187
8	829	854	852	836
9	478	410	469	445
10	181	133	185	176
11	45	26	47	53
12	7	2	6	1

5. Alternative Models

In section 3 above we saw that the Geissler data had a significant U_2 value and that this was the only significant component of X_{CLE1+}^2 for these data. Table 4 compares the fit of these data for the model with the alternative model $\hat{\pi}_j = C\hat{p}_j \exp(n^{-1/2}U_2g_2(j))$ where C is chosen so that the sum of the $\hat{\pi}_j$ (over all j) is unity.

The $\hat{\pi}_j$ model improves on the p_j Binomial model as can be seen by direct observation in Table 4 and by the decrease in X^2 from 113.2 to 16.4. These X^2 values are based on the rounded expected values shown in Table 4. The p_j Binomial model X^2 has approximate χ^2 P-value of less than 0.001, while the $\hat{\pi}_j$ model has approximate χ^2 P-value of 0.09. Lindsay (1995, p.132) fits a double Binomial to this data, with $X^2 = 51.4$ and approximate χ^2 P-value again less than 0.001. Clearly our $\hat{\pi}_j$ model does somewhat better than both the Binomial and double Binomial models.

6. Conclusion

A data dependent Chernoff-Lehmann X^2 statistic, X_{CLE1+}^2 , with class expectations greater than unity was defined. This statistic, and its dispersion component U_2^2 , were shown to have good power. For under or over-dispersed alternatives U_2^2 had power similar to the Anderson-Darling A^2 test, while for equally dispersed alternatives X_{CLE1+}^2 had similar power to A^2 . The χ^2 approximation to the pooled statistic distributions was seen to be an improvement over a χ^2 approximation to the distributions of the unpooled tests. An example used the components of X_{CLE1+}^2 . Adjusted Binomial models based on the components were illustrated using this example. The non-central χ^2 approximation to the powers of X_{CLE1+}^2 and U_2^2 does not appear to be good enough for practical use.

Bibliography

- BEST, D.J. and RAYNER, J.C.W. (1997). Goodness of fit for the Binomial distribution. *Australian Journal of Statistics*, **39**, 355-364.
- CHERNOFF, H. and LEHMANN, E.L. (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Ann. Math. Statist.*, **25**, 579-586.
- 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.
- IMSL (1995). *Users' Manual*. Houston: IMSL.
- KIMBER, A.C. (1987). When is a χ^2 not a χ^2 ? *Teaching Statistics*, **9**, 74-77.
- LANCASTER, H.O. (1953). A reconciliation of χ^2 from metrical and enumerative aspects. *Sankhya*, **13**, 1-10.
- LINDSAY, J.K. (1995). *Modelling Frequency and Count Data*. Oxford: Oxford University Press.
- MEAD, R., CURNOW, R.N. and HASTED, A.M. (2003). *Statistical Methods in Agriculture and Experimental Biology*, 3rd edition. London: Chapman & Hall.
- SNEDECOR, G.W. and COCHRAN, W.G. (1989). *Statistical Methods*, 8th edition. Ames, IA: Iowa State University Press.
- SPINELLI, J.J. (1994). *Cramer-von Mises Statistics for Discrete Distributions*. Unpublished PhD Thesis, Simon Fraser University, Canada.