Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2008, Article ID 463781, 8 pages
doi:10.1155/2008/463781

Research Article

Tests of Fit for the Logarithmic Distribution

D. J. Best,[1] J. C. W. Rayner,[1] and O. Thas[2]

[1] *School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia*
[2] *Department of Applied Mathematics, Biometrics and Process Control, Ghent University, 9000 Gent, Belgium*

Correspondence should be addressed to D. J. Best, donald.j.best@newcastle.edu.au

Received 21 September 2007; Accepted 25 February 2008

Recommended by Chin Lai

Smooth tests for the logarithmic distribution are compared with three tests: the first is a test due to Epps and is based on a probability generating function, the second is the Anderson-Darling test, and the third is due to Klar and is based on the empirical integrated distribution function. These tests all have substantially better power than the traditional Pearson-Fisher X^2 test of fit for the logarithmic. These traditional chi-squared tests are the only logarithmic tests of fit commonly applied by ecologists and other scientists.

1. Introduction

Species diversity data can sometimes be modeled by a zero-truncated negative binomial distribution with index parameter near zero. Fisher et al. [1] examined the limit as the index parameter of this distribution approached zero and so derived the logarithmic distribution. A random variable X has this distribution if and only if

$$P(X = x) = p_x = \frac{\gamma\beta^x}{x}, \quad x = 1, 2, 3, \ldots \tag{1.1}$$

in which $0 < \beta < 1$ and $\gamma = -1/\ln(1-\beta)$. The logarithmic or log-series distribution is often applied to species diversity data.

As an example ofspecies diversity data which the logarithmic distribution may fit, consider the following data on insect catches from the Sierra Tarahuma, Mexico, reported by Aldrete [2]. Ten species were caught precisely once, three species were caught precisely twice, and so on according to Table 1. The expected line in Table 1 shows the expected counts on fitting a logarithmic distribution. For these data, the alpha index (AI) of diversity is 9.01, where

Table 1: Catch frequencies per species and corresponding expected values assuming a logarithmic model for the Aldrete [2] data.

Times caught	1	2	3	4	5	6	8	10	11	12	13	16	25	69	95	At least 13
No. of species	10	3	4	2	2	2	1	1	1	1	1	1	2	1	1	6
Expected	8.78	4.28	2.78	2.03	1.58	1.28	1.07	0.91	0.62	0.55	—	—	—	—	—	7.63

Table 2: Oscar frequencies per film 1983 to 2000 and corresponding expected values assuming a logarithmic model.

Oscars	1	2	3	4	5	6	7	8	9	10	11	At least 9
No. of films	111	29	14	9	6	1	4	1	2	0	1	3
Expected	103	36	17	9	5	3	2	1	—	—	—	2

$AI = n(1-\widehat{\beta})/\widehat{\beta}$ in which n is the total number of insects and $\widehat{\beta}$ is the maximum likelihood estimator of β. The AI quoted here is defined in Krebs [3, 12.13]; for a discussion of the index of diversity see Krebs [3, Section 12.4.1]. Note that larger AI implies more diversity while smaller AI implies less.

It would seem sensible to test the data for consistency with the logarithmic distribution before quoting an AI value. However, the only statistic that appears to be commonly used by ecologists as a test of fit for the logarithmic distribution is the so-called chi-squared test, which, as Krebs [3, Section 12.4.1] notes, may not always have good power.

“The goodness-of-fit of the logarithmic series . . . can be tested by the usual chi-squared goodness-of-fit test . . . this means low power . . . . Thus in most cases the decision to use the logarithmic series . . . must be made on ecological grounds, rather than statistical goodness-of-fit criteria.” [3, page 429].

In this paper, we will examine a number of statistical tests which are considerably more powerful than the traditional Pearson-Fisher X^2 test. These include tests of fit based on components of Neyman’s smooth test statistic, the Anderson-Darling test discussed by Lockhart et al. [4], an empirical integrated distribution function test given by Klar [5], and a test due to Epps [6] based on a probability generating function (pgf). We suggest that these could be used to help make a decision as to whether or not to use the logarithmic series based on statistical as well as ecological criteria. In particular, the dispersion statistic, $\widehat{V}_2$, defined subsequently, should be useful for identifying the not infrequent case of data for which the abundance species are more abundant than predicted by the logarithmic series.

Our second example is included for its somewhat curious interest and is not involved with conventional species diversity. Collins and Hand [7] have counted the number of times, in the period 1983 to 2000, that a Hollywood film won one Oscar, two Oscars, three Oscars, and so on, giving the data in Table 2. The film with 11 Oscars was “Titanic.”

2. Tests of fit for the logarithmic

A discussion of smooth tests of fit and their components, particularly when testing for the logarithmic, is given in Appendix A. These tests may be derived as a routine application of Rayner and Best [8, Theorem 6.1.1]. The first-order component $\widehat{V}_1$ is identically zero when β is estimated by maximum likelihood, or, equivalently, by method of moments. The test based on the component $\widehat{V}_2$ suggests whether or not the data are consistent with the logarithmic variance while the test based on $\widehat{V}_3$ suggests whether or not the data are consistent with logarithmic moments up to the third. To find p-values for these tests, it is suggested that the parametric bootstrap is to be used as convergence to the asymptotic standard normal distribution is very slow. See Gürtler and Henze [9] and Appendix B for details of the parametric bootstrap in a goodness of fit context.

In Section 3, we give powers for the Anderson-Darling test based on the statistic

$$A^2 = \sum_{j=1}^{\infty} \frac{Z_j^2 p_j}{h_j(1-h_j)} \tag{2.1}$$

in which $Z_j = \sum_{x=1}^{j}(O_x - np_x), h_j = \sum_{x=1}^{j}\widehat{p}_x$, and O_x is the number of observations equal to x. Summation is halted when x is the minimum such that $O_x = 0$ and $\sum_{j=x}^{\infty}\widehat{p}_j < 10^{-3}/n$. We also give powers for a test given by Klar [5], based on the empirical integrated distribution function with test statistic

$$T_n = \sqrt{n} \sup_{1\leq k\leq M} \left| \sum_{j=1}^{k} Z_j \right| \tag{2.2}$$

in which M is the largest observation. Finally, for comparison purposes, we quote powers of the pgf and X^2 tests given by Epps [6].

3. Power comparisons

Random deviates from the logarithmic (L), positive Poisson (P+), and positive geometric (G+) distributions were generated using IMSL [10] routines RNLGR, RNPOI, and RNGEO. Random zeta deviates (Z) and random Yule deviates (Y) were found using algorithms of Devroye [11, pages 551 and 553]. Table 3 gives powers for the same alternatives as used by Epps [6], but with the addition of two Yule alternatives. For convenience, we reproduce the powers given by Epps for his pgf and X^2 tests. The powers we give for A^2, T_n, $\widehat{V}_2^2$, $\widehat{V}_3^2$, and $\widehat{V}_2^2 + \widehat{V}_3^2$ were found using parametric bootstrap with 1000 simulations both for the inner and the outer loops. Note that the calculation of $\widehat{V}_3$ can involve large numbers, and calculation of the pgf and A^2 statistics can involve small numbers. Care with rounding error may be needed. The statistics T_n and $\widehat{V}_2^2$ are less prone to rounding error. Klar [5] notes that the smooth tests, the X^2 test, and the pgf test are not consistent against all alternatives.

From Table 3 our powers for T_n are a little greater than those of Klar [5], and we observe that the power for the Z(1.0) alternative is 0.73, somewhat larger than the 0.40 reported by Klar [5]. Also from Table 3, we see that the X^2 test is not generally competitive with the other tests.

The test based on $\widehat{V}_2^2 + \widehat{V}_3^2$ performs reasonably well. The test based on the T_n statistic has power a little less than that for the pgf- and A^2-based tests. An advantage of the test based on T_n is that Klar [5] showed it is consistent.

The test based on the dispersion statistic $\widehat{V}_2^2$ has good power for the zeta and Yule alternatives, while the A^2 and pgf tests generally have competitive powers for all alternatives. Clearly, the test based on $\widehat{V}_2^2$ will not have good power for alternatives with similar dispersion to the logarithmic distribution. If the test based on $\widehat{V}_2^2$ is not significant but that based on $\widehat{V}_3^2$ is, this *suggests* a skewness departure from the logarithmic distribution. However, if the test

Table 3: Powers of some tests for the logarithmic distribution with $n = 50$ and $\alpha = 0.05$.

Alternative	$\widehat{V}_2^2$	$\widehat{V}_3^2$	$\widehat{V}_2^2 + \widehat{V}_3^2$	PGF	X^2	T_n	A^2
P + (1.05)	0.39	0.54	0.51	0.56	0.43	0.50	0.57
P + (1.2)	0.49	0.68	0.63	0.72	0.58	0.61	0.73
P + (1.3)	0.57	0.78	0.69	0.79	0.63	0.68	0.79
G + (0.25)	0.45	0.07	0.05	0.78	0.51	0.52	0.77
G + (0.33)	0.32	0.37	0.30	0.61	0.39	0.38	0.58
G + (0.4)	0.21	0.39	0.33	0.46	0.30	0.30	0.48
Z(1.0)	0.85	0.37	0.54	0.84	0.08	0.73	0.72
Z(1.3)	0.74	0.63	0.72	0.72	0.16	0.68	0.69
Z(2.0)	0.50	0.38	0.43	0.44	0.17	0.48	0.42
Y(2.75)	0.55	0.38	0.50	—	—	0.49	0.43
Y(3.0)	0.50	0.35	0.45	—	—	0.46	0.35

based on $\widehat{V}_2^2$ is significant, then this *suggests* that the test based on $\widehat{V}_3^2$ may be significant due to either a dispersion or a skewness departure of the data from the logarithmic distribution. Notice that we say the test based on $\widehat{V}_3^2$ *suggests* how the data deviate from the logarithmic. We do not claim that the data actually do deviate in this manner. See the comments of Henze and Klar [12].

On the basis of Table 3 powers, we suggest that the tests based on A^2 and Klar's T_n are considered as tests of fit for the logarithmic distribution. These tests have good power and are consistent. We recommend that the tests based on T_n and A^2 are augmented by the use of $\widehat{V}_2^2$ and $\widehat{V}_3^2$ in a data analytic fashion.

4. Examples

In the following parametric bootstrap, p-values for the tests based on $\widehat{V}_2^2$, $\widehat{V}_3^2$, A^2, and T_n are given. These use 1000 random samples of the logarithmic distribution with parameter $\widehat{\beta}$ as given below. We give $\widehat{V}_2$ and $\widehat{V}_3$ values because they may *suggest* how the data deviate from the logarithmic. We give the A^2 and T_n values because the tests based on these statistics are consistent and have good power.

4.1. Insect data

From the data in Table 1, we find $\widehat{\beta} = 0.9743$, $AI = 9.0104$, $\widehat{V}_2 = 0.4879$ with p-value 0.52, $\widehat{V}_3 = -0.8791$ with p-value 0.19, $A^2 = 0.2613$ with p-value 0.82, and $T_n = 204.8192$ with p-value 0.33. It appears that the logarithmic distribution is a good fit. In agreement with this, the Pearson-Fisher statistic takes the value 4.56 on 11 degrees of freedom when data greater than 12 have been combined.

4.2. Oscars data

We find that $\widehat{V}_2 = 0.49$, $\widehat{V}_3 = -1.27$, and $X^2 = 7.40$ on 7 degrees of freedom if the classes greater than or equal to 9 are combined. The corresponding p-values are 0.58, 0.08, and 0.62. It appears that a logarithmic distribution with $\widehat{\beta} = 0.7044$ fits the data reasonably well.However, the p-value for $\widehat{V}_3$ suggests that the data may not be quite as skewed as would be expected for the logarithmic distribution. Collins and Hand [7] suggest a Yule distribution fits the data well. In addition, we note that $A^2 = 0.9727$ with p-value 0.12 and $T_n = 120.0203$ with p-value 0.26.

5. Conclusion

In this paper, we have examined a number of statistical tests which are considerably more powerful than the traditional Pearson-Fisher X^2 test. We suggest that these could be used to help make a decision as to whether or not to use the logarithmic series based on statistical as well as ecological criteria. A test of fit could be done before quoting the index of diversity. In particular, the dispersion statistic, $\widehat{V}_2$, should be useful for identifying the not infrequent case of data for which the abundance species are more abundant than predicted by the logarithmic series.

Appendices

A. The smooth tests and their components

For distributions from exponential families the smooth tests can be derived as score statistics for testing H_0: $\theta = 0$ against $K : \theta \neq 0$ for observations $X_1, \ldots, X_n$ from the model

$$C(\theta, \beta) \exp\left\{ \sum_{i=q+1}^{q+k} \theta_i g_i(x; \beta) \right\} f(x; \beta) \tag{A.1}$$

in which

(i) $f(x;\beta)$ is a probability density function that depends on a $q \times 1$ vector of nuisance parameters β and for which we test;

(ii) $\{g_i(x;\beta)\}$ is a complete orthonormal set on $f(x;\beta)$;

(iii) $C(\theta;\beta)$ is a normalizing constant.

For details see Rayner and Best [8].

The score test statistic has a particularly appealing form

$$\widehat{S}_k = \widehat{V}_{q+1}^2 + \cdots + \widehat{V}_{q+k}^2, \tag{A.2}$$

where

$$\widehat{V}_r = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} g_r(x_j; \widehat{\beta}), \quad r = q+1, \ldots, q+k. \tag{A.3}$$

Here, $\widehat{\beta}$ is the maximum likelihood estimator of β assuming that H_0 is true.

To define $\widehat{V}_r$, central moments of $f(x;\beta)$ up to order $2r$ are required. For example, to directly define components up to $\widehat{V}_3$ to test for the logarithmic, we note that the equation to estimate β is $\widehat{V}_1 \equiv 0$, as discussed below. To define $\widehat{V}_2$ and $\widehat{V}_3$ requires $g_2(x;\beta)$ and $g_3(x;\beta)$, which in turn require central logarithmic moments up to order six. These are given by

$$\begin{aligned}
\mu &= \frac{\gamma\beta}{(1-\beta)},\\
\mu_2 &= \frac{\gamma\beta(1-\gamma\beta)}{(1-\beta)^2},\\
\mu_3 &= \frac{\gamma\beta(1+\beta-3\gamma\beta+2\gamma^2\beta^2)}{(1-\beta)^3},\\
\mu_4 &= \frac{\gamma\beta(1+4\beta-4\gamma\beta+\beta^2-4\gamma\beta^2+6\gamma^2\beta^2-3\gamma^3\beta^3)}{(1-\beta)^4},\\
\mu_5 &= \frac{\gamma\beta(1+11\beta+11\beta^2-5\gamma\beta-20\gamma\beta^2+10\gamma^2\beta^2+\beta^3-5\gamma\beta^3+10\gamma^2\beta^3-10\gamma^3\beta^3+4\gamma^4\beta^4)}{(1-\beta)^5},\\
\mu_6 &= \gamma\beta(1+26\beta-6\gamma\beta+66\beta^2-66\gamma\beta^2+15\gamma^2\beta^2+26\beta^3-66\gamma\beta^3+60\gamma^2\beta^3-20\gamma^3\beta^3+\beta^4\\
&\qquad -6\gamma\beta^4+15\gamma^2\beta^4-20\gamma^3\beta^4+15\gamma^4\beta^4-5\gamma^5\beta^5)/(1-\beta)^6.
\end{aligned} \tag{A.4}$$

To calculate further orthonormal polynomials directly, we could use the result that for the logarithmic, $k_{r+1} = \beta\partial k_r/\partial\beta$ generates cumulants and hence central moments, but it is more efficient to use recurrence as described in Rayner et al. [13]. Proceeding directly, the first six central moments can be used to calculate

$$\begin{aligned}
g_1(x;\beta) &= \frac{(x-\mu)}{\sqrt{\mu_2}},\\
g_2(x;\beta) &= \frac{\{(x-\mu)^2-\mu_3(x-\mu)/\mu_2-\mu_2\}}{\sqrt{\mu_4-\mu_3^2/\mu_2-\mu_2^2}},\\
g_3(x;\beta) &= \frac{(x-\mu)^3-a(x-\mu)^2-b(x-\mu)-c}{\sqrt{\mu_6-2a\mu_5+(a^2-2b)\mu_4+2(ab-c)\mu_3+(b^2+2ac)\mu_2+c^2}}
\end{aligned} \tag{A.5}$$

in which

$$\begin{aligned}
a &= \frac{(\mu_5-\mu_3\mu_4/\mu_2-\mu_2\mu_3)}{d},\\
b &= \frac{(\mu_4^2/\mu_2-\mu_2\mu_4-\mu_3\mu_5/\mu_2+\mu_3^2)}{d},\\
c &= \frac{(2\mu_3\mu_4-\mu_3^3/\mu_2-\mu_2\mu_5)}{d},\\
d &= \mu_4-\frac{\mu_3^2}{\mu_2}-\mu_2^2.
\end{aligned} \tag{A.6}$$

These formulas give the first three orthonormal polynomials for any univariate distribution.

The components $\widehat{V}_r$ can be called smooth components as they are analogous to the components of the smooth test for uniformity introduced by Neyman [14]. His smooth components also used orthonormal polynomials. When testing for distributions from exponential families these components are asymptotically independent and asymptotically have the standard normal distribution.

For the logarithmic distribution, the maximum likelihood and method of moments estimators $\widehat{\beta}$ of β coincide, given by $\widehat{V}_1 \equiv 0$ or

$$\overline{X} = -\frac{\widehat{\beta}}{(1-\widehat{\beta})\ln(1-\widehat{\beta})} = \frac{\widehat{\beta}\widehat{\gamma}}{(1-\widehat{\beta})} = \widehat{\mu}. \qquad \text{(A.7)}$$

To solve this equation, the Newton-Raphson algorithm can be used. An initial estimate of $\widehat{\beta}$ and other details helpful in the solution are given in Birch [15]. Note also that for the logarithmic, $\widehat{V}_2$ is proportional to $(m_2 - \widehat{\mu}_2)$ where $m_2 = \sum_{j=1}^{n}(x_j - \overline{x})^2/n$, so the test based on $\widehat{V}_2$ tests for the dispersion of the logarithmic distribution. Similarly, if $m_3 = \sum_{j=1}^{n}(x_j - \overline{x})^3/n$, then the numerator of $\widehat{V}_3$ is of the form $n(m_3 - am_2 - c)$, so the test based on $\widehat{V}_3$ assesses whether the data are consistent with moments of the logarithmic up to the third.

B. *P*-values via the parametric bootstrap

Gürtler and Henze [9, page 223] suggest that 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 $x_1, x_2, \ldots, x_n$ denote, as usual, the data. Find an estimate $\widehat{\beta}$ from the data and conditional on this estimate, generate $B = 10\,000$ say pseudorandom samples of size n, each having the logarithmic $(\widehat{\beta})$ distribution. For $j = 1, \ldots, B$ 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, $\sum_{j=1}^{B} I(W_{n,j}^* \geq w_n)/B$.

The above requires random logarithmic (β) values. Devroye [11, page 547] outlines an algorithm for generating random logarithmic deviates. Alternatively, the routine RNLGR from IMSL [10] can be used. To obtain p-values for two-tailed tests proceed as above and find the p-value, say P. Then if $P \leq 0.5$, the two-tailed p-value is $2P$, while if $P > 0.5$, the two-tailed p-value is $2(1-P)$.

Acknowledgment

Research of Olivier Thas was supported by IAP research network Grant no. P6/03 of the Belgian Government (Belgian Science Policy).

References

[1] R. A. Fisher, A. S. Corbet, and C. B. Williams, "The relation between the number of species and the number of individuals in a random sample of an animal population," *Journal of Animal Ecology*, vol. 12, no. 1, pp. 42–58, 1943.

[2] A. N. G. Aldrete, "Psocoptera (insecta) from the Sierra tarahumara, Chihuahua, Mexico," *Anales del Instituto de Biología, Universidad Nacional Autónoma de México, Serie Zoología*, vol. 73, no. 2, pp. 145–156, 2002.

[3] C. J. Krebs, *Ecological Methodology*, Addison Wesley Longman, New York, NY, USA, 1998.

[4] R. A. Lockhart, J. J. Spinelli, and M. A. Stephens, "Cramér-von Mises statistics for discrete distributions with unknown parameters," *Canadian Journal of Statistics*, vol. 35, no. 1, pp. 125–133, 2007.

[5] B. Klar, "Goodness-of-fit tests for discrete models based on the integrated distribution function," *Metrika*, vol. 49, no. 1, pp. 53–69, 1999.

[6] T. W. Epps, "A test of fit for lattice distributions," *Communications in Statistics: Theory and Methods*, vol. 24, no. 6, pp. 1455–1479, 1995.

[7] A. Collins and C. Hand, "Vote clustering in tournaments: what can Oscar tell us?" *Creativity Research Journal*, vol. 18, no. 4, pp. 427–434, 2006.

[8] J. C. W. Rayner and D. J. Best, *Smooth Tests of Goodness of Fit*, The Clarendon Press, Oxford University Press, New York, NY, USA, 1989.

[9] N. Gürtler and N. Henze, "Recent and classical goodness-of-fit tests for the Poisson distribution," *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 207–225, 2000.

[10] IMSL, *Users' Manual*, IMSL, Houston, Tex, USA, 1995.

[11] L. Devroye, *Non-Uniform Random Variate Generation*, Springer, New York, NY, USA, 1986.

[12] N. Henze and B. Klar, "Properly rescaled components of smooth tests of fit are diagnostic," *Australian & New Zealand Journal of Statistics*, vol. 38, no. 1, pp. 61–74, 1996.

[13] J. C. W. Rayner, O. Thas, and B. De Boeck, "A generalised Emerson recurrence relation," to appear in *Australian & New Zealand Journal of Statistics* .

[14] J. Neyman, ""Smooth test" for goodness of fit," *Skandinavisk Aktuarietidskrift*, vol. 20, pp. 149–199, 1937.

[15] M. W. Birch, "194 Note: an algorithm for the logarithmic series distribution," *Biometrics*, vol. 19, no. 4, pp. 651–652, 1963.