NIASRA
NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

07-14

Smooth tests of fit for more flexible alternatives to the exponential and Poisson: the Lindley and Poisson-Lindley distributions

D.J. Best and J.C.W Rayner

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email:
anica@uow.edu.au

Smooth tests of fit for more flexible alternatives to the exponential and Poisson: the Lindley and Poisson-Lindley distributions

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

We consider the little-known one parameter Lindley and Lindley-Poisson distributions. These distributions may be of interest as they appear to be more flexible than the exponential and Poisson distributions, the Lindley fitting more data than the exponential and the Lindley-Poisson fitting more data than the Poisson. We give smooth tests of fit for each of these distributions. The smooth test for the Lindley has power comparable with the Anderson-Darling test. Advantages of the smooth test are discussed. Examples that illustrate the flexibility of the two distributions are given.

Key Words: Ecological data; Lifetime data; Orthonormal polynomials; Powers; Shelf life data.

1. Introduction

Ghitany and co-workers [3] give many properties of the Lindley distribution. They suggest it is often a better model than the traditional exponential distribution that is commonly used to model lifetime or waiting time data. The Lindley distribution is not well known, so there may be other applications. We hope this article will help bring the Lindley distribution to statisticians notice. Ghitany et al. (2008) examine the fit of the Lindley distribution to some waiting time data by looking at plots and by showing the Lindley likelihood is better than the exponential likelihood. However, this does not prove the Lindley distribution fits the data well, only that it fits better than the exponential. Assessment of the plots is subjective and here we derive a smooth test of fit to give a more objective assessment of goodness of fit of the Lindley model. We also examine the use of the Anderson-Darling test.

Just as the Lindley is an alternative model to the continuous exponential distribution, the Poisson-Lindley is an alternative to the discrete Poisson distribution. We also derive a smooth test for the Poisson-Lindley distribution.

The Lindley distribution has probability density function

$$f(x;\,\theta) = \frac{\theta^2}{\theta+1}\,(1+x)\,e^{-\theta x} \text{ for } x > 0 \text{ and zero otherwise, in which } \theta > 0$$

and cumulative distribution function

$$F(x;\,\theta) = 1 - \frac{\theta+1+\theta x}{\theta+1}\,(1+x)\,e^{-\theta x} \text{ for } x > 0 \text{ and zero otherwise.}$$

Smooth tests of fit can be found using the second and third order smooth test components. See [6] for a discussion of smooth tests.

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

Section 2 gives the smooth test statistics. Section 3 looks at the approach to the asymptotic chi-squared distributions of the smooth test statistics and finds them to be quite slow. It is suggested that p-values be found using the parametric bootstrap. A slightly expanded version of an algorithm in [3] generating random Lindley variates is given in section 3 so that these p-values can be calculated. The powers of the smooth test and the Anderson-Darling test are compared in section 4 that also gives an example. In section 5 we discuss the Poisson-Lindley distribution. The required orthonormal polynomials are given in an appendix.

2. The smooth test statistics

Smooth tests of goodness of fit are extensively discussed in [6]. Components of smooth tests, V_r, $r = 1, 2, ...,$ are defined as

$$V_r = \sum_{i=1}^{n} g_r(x_i) / \sqrt{n}$$

where there are n data elements $x_1, \ldots, x_n$, $g_r(x)$ is the rth orthonormal polynomial on the distribution under investigation and any nuisance parameters (suppressed in this notation) are appropriately estimated. The components give focused tests. For example here the second order component detects dispersion differences between the data and the hypothesised Lindley distribution. Sums of squares of components give more omnibus tests.

For the Lindley distribution note that the mean and variance are given by $\mu = (\theta + 2)/\{\theta(\theta + 1)\}$ and $\mu_2 = (\theta^2 + 4\theta + 2)/\{\theta^2(\theta^2 + 1)\}$ respectively.

The Lindley distributions form a one-parameter exponential family of distributions. For such distributions the method of moments estimator is the same as the maximum likelihood estimator. For the Lindley this estimator is

$$\hat{\theta} = \frac{-(\bar{X}-1) + \sqrt{(\bar{X}-1)^2 + 8\bar{X}}}{2\bar{X}} \text{ provided } \bar{X} > 0. \qquad (2.1)$$

It is shown in [6] that when non-degenerate, the V_r^2 have an asymptotic χ_1^2 distribution. Here then V_1 is degenerate because method of moments estimation is being used, so that

$$V_1 = \sum_{i=1}^{n} (x_i - \hat{\mu}) / \sqrt{n\mu_2} = (\bar{x} - \hat{\mu})\sqrt{n/\mu_2} \equiv 0 .$$

In the next section we briefly consider the approach of V_2^2, V_3^2 and $S = V_2^2 + V_3^2$ to the asymptotic distribution.

3. The approach to the asymptotic distribution

In Table 1 we look, for $\theta = 0.5$ and 1.5, at the approach to the asymptotic χ_1^2 distribution of V_2^2 and V_3^2 and the approach to the asymptotic χ_2^2 distribution of $S = V_2^2 + V_3^2$. The results in Table 1 are 5% critical values found using 100,000 simulations of Monte Carlo samples of size *n*. A random variate generator, given below, is needed for these results, the powers of the next section and bootstrap p-values.

TABLE 1

5% critical values based on 100,000 simulations of samples of size n for V_2^2, V_3^2 and S when θ = 0.5 and 1.5.

n	$\theta = 0.5$			$\theta = 1.5$		
	V_2^2	V_3^2	S	V_2^2	V_3^2	S
20	2.59	1.93	4.38	2.65	1.93	4.33
100	3.47	2.71	5.06	3.41	2.52	4.98
200	3.69	2.95	5.42	3.66	2.90	5.73
1,000	3.83	3.59	5.72	3.83	3.28	5.88
10,000	3.89	3.90	6.02	3.87	3.89	6.06
∞	3.84	3.84	5.99	3.84	3.84	5.99

The convergence to the asymptotic values is quite slow and so we suggest in applications the parametric bootstrap will be needed to find p-values. The Table 1 results are similar for $\theta = 0.5$ and $\theta = 1.5$.

To generate random Lindley values we follow [2] and observe that the Lindley distribution is a mixture of an exponential (θ) distribution and a gamma (2, θ) distribution:

$$f(x;\, \theta) = \frac{\theta}{\theta+1}\theta e^{-\theta x} + \frac{1}{\theta+1}\theta^2 x e^{-\theta x}.$$

To obtain a random *x* value we need four uniform (0, 1) values, u_1, u_2, u_3 and u_4 say, and take $x = -\{\log(u_1\, u_2)\}/\theta$ unless $u_4 \le \theta/(\theta + 1)$, in which case $x = -(\log u_3)/\theta$.

To find parametric bootstrap p-values generate Lindley samples of size *n* many times (say 10,000 times) and take the p-values as the proportion of the samples with test statistics greater than or equal to the values of the test statistics for the original data set.

4. Power comparisons and an example

In Table 2, for a significance level of $\alpha = 0.05$ and a sample size of $n = 20$, we find some powers for V_2^2, V_3^2, *S* and *AD* where *AD* is the Anderson-Darling test statistic

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\{\log z_{(i)} + \log(1 - z_{(n+1-i)})\}$$

in which $\{z_{(i)}\}$ are ordered values of $\{z_i\}$, where $z_i = F(x_i;\ \theta)$.

TABLE 2

Powers of tests based on V_2^2, V_3^2, S and AD for α = 0.05 and n = 20.

Alternative	V_2^2	V_3^2	S	AD
Lindley (0.5)	0.05	0.05	0.05	0.05
$\chi^2_{0.75}$	0.83	0.76	0.88	0.92
χ^2_1	0.63	0.62	0.72	0.80
χ^2_2	0.18	0.16	0.18	0.15
χ^2_3	0.06	0.04	0.06	0.05
χ^2_4	0.08	0.10	0.09	0.09
χ^2_8	0.53	0.55	0.59	0.60
Weibull (0.8)	0.43	0.34	0.44	0.43
Weibull (1.5)	0.21	0.26	0.27	0.28
Weibull (2.0)	0.75	0.84	0.84	0.84
Beta (1, 2)	0.14	0.18	0.18	0.15
Beta (1, 3)	0.07	0.11	0.11	0.11
Beta (2, 3)	0.87	0.93	0.94	0.90
Uniform (0, 1)	0.51	0.57	0.57	0.54

The tests based on *S* and *AD* have similar powers but *S* has the advantage that its approximate null distribution is the convenient χ^2_2. As all four of the tests we examined had χ^2_3 power approximately the same as the test size it appears that for these values of θ the χ^2_3 distribution is, in this sense, close to the Lindley. The test based on V_2^2 has slightly less power than the tests based on *S* and *AD*. It has little power if the alternative has similar variance to the Lindley variance, which is quite reasonable as it is testing for distributions with the Lindley variance. The test based on V_3^2 is, roughly, testing for distributions with the Lindley skewness. This is why it is useful to apply V_2^2 and V_3^2 together, either separately as in exploratory data analysis, or more formally together, via *S*.

Product shelf life data.

Data consisting of days to be judged unacceptable by an expert panel for a food product are given in [1]. This data set relates to a food product shelf life study and could lead to determination of 'use-by' dates for food products. The original 14 data points but with 19 and 20 added, are

19, 20, 21, 23, 25, 38, 43, 43, 52, 56, 61, 63, 67, 69, 70, 107.

In Table 3 we give both the bootstrap p-values and approximate p-values based on the asymptotic chi-squared distribution. The latter are useful as a first approximation; the former have greater validity. We see that the exponential is not a good fit. However the smallest p-value for the Lindley fit is 0.12. This indicates a more reasonable fit for the Lindley distribution, in line with the suggestion in [3] that the Lindley distribution is more flexible than the exponential. That is, compared to the exponential, the Lindley distribution may model more data well.

TABLE 3

P-values for shelf life data.

Statistic	Asymptotic exponential p-value	Bootstrap exponential p-value	Asymptotic Lindley p-value	Bootstrap Lindley p-value
V_2^2	0.13	0.03	0.21	0.12
V_3^2	0.08	0.01	0.30	0.14
S	0.06	0.02	0.27	0.12
AD	–	0.01	–	0.16

Hough and co-workers [4] have suggested use of lifetime distributions to find an optimum amount of food ingredient, such as sugar, to use in food product development. The Lindley distribution may be useful for this application also.

5. Smooth tests for the Poisson-Lindley distribution

In section 1 above it was suggested that the continuous Lindley distribution might be a better one-parameter model than the classical exponential distribution and an example was given supporting this suggestion. In the discrete case the Poisson-Lindley model may similarly be a better model than the classical one-parameter Poisson distribution. The Poisson-Lindley model was introduced in [7] and has probability density function

$$f(x;\theta) = \frac{\theta^2(x+\theta+2)}{(\theta+1)^{x+3}} \text{ for } x = 0, 1, \dots, \text{ in which } \theta > 0.$$

As above, smooth tests can be found using the second and third order smooth test components. Again see [6] for a discussion of smooth tests. We now describe their use for the Poisson-Lindley distribution.

To generate a random Poisson-Lindley value first generate a random Lindley value, λ say, and then generate a random Poisson (λ) value. We note that many properties of the Poisson-Lindley distribution are given in [2] where it is also shown that the method of moments (MOM) and maximum likelihood (ML) estimators are almost equally efficient. In the following we will use MOM estimators so that the second component has a dispersion detecting interpretation and because MOM and

ML are numerically very similar. Note that $\mu = (\theta + 2)/\{\theta(\theta + 1)$ as for the Lindley distribution and so $\hat{\theta}$ will be given by (2.1).

Ecological Example.

An ecological example concerning quadrats in a Scottish pasture with frequencies (f) of x earthworms is given in [5]. If we add one quadrat with 7 earthworms to the Krebs data then $\{x\} = \{0, 1, 2, 3, 4, 5, 6, 7\}$ and $\{f\} = \{4, 8, 2, 5, 2, 3, 1, 1\}$. This parallels the shelf life data where the extra data are informative. Here with the adjusted data the Poisson is not a good one-parameter model, but the Poisson-Lindley is. Table 4 gives the p-values.

TABLE 4

P-values for earthworm data.

Statistic	Asymptotic Poisson p-value	Bootstrap Poisson p-value	Asymptotic Poisson-Lindley p-value	Bootstrap Poisson-Lindley p-value
V_2^2	0.03	0.03	0.23	0.17
V_3^2	0.26	0.16	0.76	0.91
S	0.05	0.04	0.47	0.59

7. Conclusion

We have given a smooth test of fit statistic S for the Lindley and Poisson-Lindley distributions. Two examples illustrate the flexibility of the distributions. We suggest that p-values be given using the parametric bootstrap. Executable code may be obtained from the first author.

Appendix: Orthonormal polynomials

Let $z = x - \mu$. No matter what the distribution, the orthonormal polynomials of orders 0, 1, 2 and 3 are

$$g_0(x) = 1 \text{ for all } x,$$
$$g_1(x) = z/\sqrt{\mu_2}$$
$$g_2(x) = (z^2 - a_2 z - \mu_2)/\sqrt{d_2} \text{ and}$$
$$g_3(x) = (z^3 - a_3 z^2 - b_3 z - c_3)/\sqrt{d_3}.$$

The values taken by the constants vary with the distribution.

For the Lindley distribution

$$\mu_2 = \frac{(\theta^2 + 4\theta + 2)}{\theta^2(\theta + 1)^2},$$

$$a_2 = \frac{2(\theta^3 + 6\theta^2 + 6\theta + 2)}{\theta(\theta+1)(\theta^2 + 4\theta + 2)},$$

$$d_2 = \frac{4(\theta^3 + 9\theta^2 + 18\theta + 6)}{\theta^4(\theta+1)(\theta^2 + 4\theta + 2)},$$

$$a_3 = \frac{6(\theta^4 + 11\theta^3 + 3\theta^2 + 24\theta + 6)}{\theta(\theta^4 + 10\theta^3 + 27\theta^2 + 24\theta + 6)},$$

$$b_3 = -\frac{3(\theta^3 + 13\theta^2 + 40\theta + 24)}{(\theta+1)^2(\theta^3 + 9\theta^2 + 18\theta + 6)},$$

$$c_3 = -\frac{4(\theta^6 + 15\theta^5 + 75\theta^4 + 164\theta^3 + 162\theta^2 + 72\theta + 12)}{(\theta+1)^3\theta^3(\theta^3 + 9\theta^2 + 18\theta + 6)},$$

$$d_3 = \frac{36(\theta^4 + 16\theta^3 + 72\theta^2 + 96\theta + 24)}{(\theta+1)\theta^6(\theta^3 + 9\theta^2 + 18\theta + 6)}.$$

For the exponential distribution the constants required for orthonormal polynomials are

$$\mu_2 = 1/\theta^2,\ a_2 = 2/\theta,\ d_2 = 2/\theta^2,\ a_3 = 6/\theta,\ b_3 = 3/\theta^2,\ c_3 = 4/\theta^3 \text{ and } d_3 = 9/\theta^3.$$

These polynomials, needed in section 4, are much simpler than the Lindley orthonormal polynomials.

For the Poisson-Lindley distribution the required constants are

$$\mu_2 = \frac{(\theta^3 + 4\theta^2 + 6\theta + 2)}{\theta^2(\theta+1)^2}$$

$$a_2 = \frac{(\theta+2)(\theta^4 + 5\theta^3 + 12\theta^2 + 8\theta + 2)}{\theta(\theta+1)(\theta^3 + 4\theta^2 + 6\theta + 2)},$$

$$d_2 = \frac{4(\theta^5 + 8\theta^4 + 27\theta^3 + 42\theta^2 + 30\theta + 6)}{\theta^4(\theta^3 + 4\theta^2 + 6\theta + 2)},$$

$$a_3 = \frac{3(\theta+2)(\theta^6 + 9\theta^5 + 37\theta^4 + 75\theta^3 + 78\theta^2 + 36\theta + 6)}{\theta(\theta+1)(\theta^5 + 8\theta^4 + 27\theta^3 + 42\theta^2 + 30\theta + 6)},$$

$$b_3 = -\frac{(2\theta^7 + 23\theta^6 + 124\theta^5 + 385\theta^4 + 744\theta^3 + 888\theta^2 + 588\theta + 156)}{(\theta^7 + 10\theta^6 + 44\theta^5 + 104\theta^4 + 141\theta^3 + 108\theta^2 + 42\theta + 6)},$$

$$c_3 = -\frac{2(\theta+2)(\theta^9 + 13\theta^8 + 79\theta^7 + 278\theta^6 + 612\theta^5 + 862\theta^4 + 762\theta^3 + 396\theta^2 + 108\theta + 12)}{(\theta+1)^3\theta^3(\theta^5 + 8\theta^4 + 27\theta^3 + 42\theta^2 + 30\theta + 6)},$$

$$d_3 = \frac{36(\theta+1)(\theta^7 + 13\theta^6 + 74\theta^5 + 224\theta^4 + 384\theta^3 + 360\theta^2 + 168\theta + 24)}{\theta^6(\theta^5 + 8\theta^4 + 27\theta^3 + 42\theta^2 + 30\theta + 6)}.$$

For the Poisson distribution the constants required for orthonormal polynomials are

$$\mu_2 = \lambda,\ a_2 = 1,\ d_2 = 2\lambda^2,\ a_3 = 3,\ b_3 = 2 - 3\lambda,\ c_3 = 2\lambda \text{ and } d_3 = 6\lambda^3.$$

References

[1] M.C. Gacula, J. Singh, J. Bi, S. Altan, Statistical Methods in Food and Consumer Research, second ed., Academic Press, New York, 2009.

[2] M.E. Ghitany, D.K. Al-Mutairi, Estimation methods of the discrete Poisson-Lindley distribution, J. Statist. Comput. & Simul., 79 (2009) 1-9.

[3] M.E. Ghitany, B.Atieh, S. Nadarajah, Lindley distribution and its applications, Math & Comput in Simul., 78 (2008) 493-506.

[4] G. Hough, K. Langohr, G. Gomez, A.M. Curia, Survival analysis applied to sensory shelf life of foods, J. Food Sci., 68 (2003) 359-362.

[5] Krebs, C.J. (1999). Ecological Methodology, 2nd ed., Addison Wesley Longman, New York, 1999.

[6] J.C.W. Rayner, O. Thas, D.J. Best, Smooth Tests of Goodness of Fit: Using R, 2nd ed., Wiley, Singapore, 2009.

[7] M. Sankaran, The discrete Poisson-Lindley distribution, Biometrics, 26 (1970) 145-149.