JOURNAL OF APPLIED MATHEMATICS AND DECISION SCIENCES, 5(3), 181–191

Power of the Neyman Smooth Tests for the Uniform Distribution

GLEN D. RAYNER† gdrayner@deakin.edu.au
School of Computing and Mathematics, Deakin University, Waurn Ponds, VIC3217, Australia

JOHN C. W. RAYNER john_rayner@uow.edu.au
School of Mathematics and Applied Statistics, University of Wollongong, NSW2522, Australia

Abstract. This paper compares and investigates the generalised Neyman smooth test, its components, and the classical chi-squared test with a variety of equiprobable classes. Each test is evaluated in terms of its power to reject a wavelike alternative to the uniform distribution, chosen to quantify the *complexity* of the alternative. Results indicate that if broadly focused tests (rather than strongly directional or weakly omnibus) are sought, then smooth tests of order about four, or the chi-squared test with between five and ten classes, will perform well.

Keywords: Neyman smooth tests, Goodness-of-fit, Uniform Distribution.

1. Introduction

Neyman (1937) constructed his smooth tests specifically to test for the continuous uniform distribution. Uniformity testing is important in a range of applications, including the assessment of random number generators. Moreover any goodness of fit test for a completely specified alternative reduces, via the probability integral transformation, to testing for uniformity. Neyman's construction has been generalised and, of interest here, evaluated several times. See, for example, Quesenberry and Miller (1977) and Miller and Quesenberry (1979), and Best and Rayner (1985). The articles involving Quesenberry and Miller do not consider the tests based on the components of Neyman's smooth tests. Moreover we are now in a position to take advantage of modern fast computers to make assessments that were not previously so easily carried out.

Neyman's smooth tests have been generalised to testing for arbitrary distributions. See for example, Rayner and Best (1989). Recent advice on

† Requests for reprints should be sent to Glen D. Rayner, School of Computing and Mathematics, Deakin University, Waurn Ponds, VIC3217, Australia

such testing recommends using the sum of the squares of the first two, three or four components of the appropriate Neyman smooth test, augmented by using the components themselves in a data analytic fashion. See Rayner and Best (2000). The manner in which the components give information about the alternatives to uniformity can perhaps be best interpreted in terms of the parameter space spanned by the alternatives. Use of orthonormal functions means this space is decomposed into orthogonal one dimensional spaces. The r-th component assess differences between the data and the hypothesised distribution in, by definition, the r-th order, and this may be thought of as in the r-th moment. Although this correspondence isn't exact, it leads to a useful and insightful interpretation of the data. See, for example, Rayner, Best and Mathews (1995).

Carolan and Rayner (2000) looked at the smooth tests for normality, and showed that even when differences from the hypothesised distribution are generated from an *order r alternative* (see Carolan and Rayner, 2000 for the precise meaning), earlier components may be significant. What seems to be happening is similar to a polynomial of degree six say, being reasonably well approximated, over a specified domain, by a combination of polynomials of degree say, one, two and five. Rayner, Best and Dodds (1985) looked at what Pearson's chi-squared test of equiprobability and its components best detect, and it seems timely to look at the Neyman smooth tests for the uniform distribution and its components.

The study by Kallenberg et al. (1985) into how to construct the classes for the Pearson chi-squared test characterised the alternatives in part by tail weight (heavy or light). We do not look for an answer in terms of tail weight, but in terms of how complicated the alternative may be.

2. Notation

The Neyman smooth tests were constructed to be asymptotically locally uniformly most powerful symmetric, unbiased and of specified size against specified *smooth* alternatives (Neyman, 1937). To define these alternatives, note that the Legendre polynomials $\{\pi_r(y)\}$ are the polynomials that are orthonormal on the uniform $U(0,1)$ distribution. Explicitly the first few polynomials are given by

$$\begin{aligned}
\pi_0(y) &= 1 \\
\pi_1(y) &= \sqrt{3}(2y-1) \\
\pi_2(y) &= \sqrt{5}(6y^2-6y+1) \\
\pi_3(y) &= \sqrt{7}(20y^3-30y^2+12y-1) \text{ and} \\
\pi_4(y) &= 3(70y^4-140y^3+90y^2-20y+1).
\end{aligned} \tag{1}$$

The *order k smooth alternative* to $U(0,1)$, the uniform continuous distribution on $(0,1)$, is

$$C(\theta)\exp\left\{\sum_{r=1}^{k}\theta_r\pi_r(y)\right\},\ \text{for } 0<y<1,\ \text{zero otherwise,} \tag{2}$$

in which $C(\theta)$ is a normalising constant that ensures the probability density function integrates to one. The *smooth test of order k for uniformity* is based on the statistic

$$S_k = V_1^2 + \ldots + V_k^2 \tag{3}$$

which has *components*

$$V_r = \sum_{j=1}^{n}\pi_r(Y_j)/\sqrt{n},\ \text{for } r=1,\ldots,k. \tag{4}$$

To calculate Pearson's chi-squared test statistic X_P^2, we assume the n data points are categorised into m classes, with class probabilities $p_1,\ldots,p_m$ and class counts $N_1,\ldots,N_m$. Then

$$X_P^2 = \sum_{j=1}^{m}\left(N_j - np_j\right)^2/\left(np_j\right). \tag{5}$$

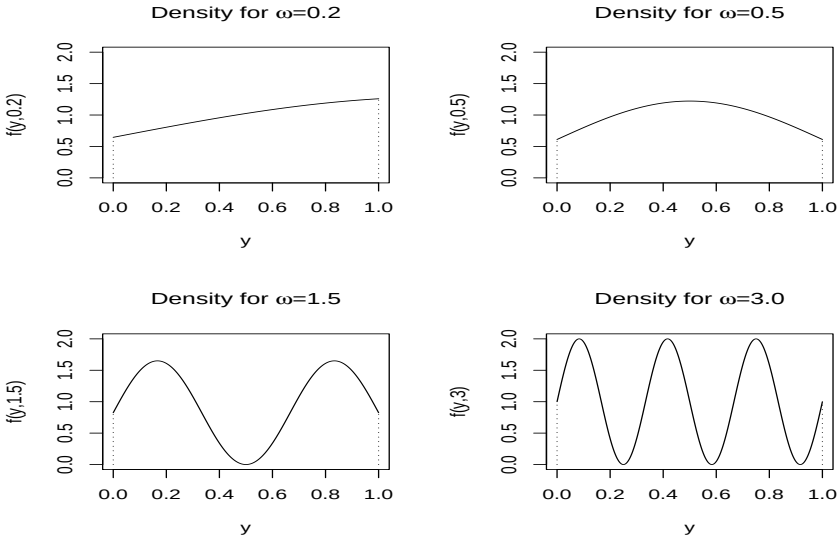
The Pearson chi-squared test statistic X_P^2 based on m equiprobable classes ($p_j = 1/m$ for all j) will be denoted by X_{Pm}^2.

The alternatives to uniformity that we will consider are not the ones for which the test was constructed to be optimal. We take

$$f_Y(y;\omega) = \frac{2\pi\omega\left[1+\sin(2\pi\omega y)\right]}{1+2\pi\omega-\cos(2\pi\omega)},\ \text{for } 0<y<1,\ \text{zero otherwise} \tag{6}$$

where the 'complexity' parameter $\omega > 0$. Note that this distribution is $U(0,1)$ in the limit as ω approaches zero. See Figure 1 for a plot of some of these alternatives. It seems to us that probability density functions of this form cannot be characterised in terms of tail weight as envisaged by Kallenberg et al. (1985). We expect that small values of ω will reflect low order alternatives to uniformity, and will be better detected by the earlier components, while larger values of ω reflect higher order alternatives, and will be better detected by the later components. For a discussion of order and effective order, see Rayner, Best and Dodds (1985) or Rayner and Best (1989, section 4.3).

Figure 1. The probability density function $f_Y(y;\omega)$ for $\omega = 0.2, 0.5, 1.5$ and 3.0.



3. Size and Power Study

Here we present powers, estimated using 100,000 simulations, of the tests based on V_1^2, V_2^2, V_3^2, V_4^2, S_4, X_{P2}^2, X_{P5}^2, X_{P10}^2, X_{P20}^2 and the Anderson-Darling test (D'Agostino and Stephens, 1986, p101, 104-105), based on the statistic AD. The alternatives used assume the distribution with probability density function $f_Y(y;\omega)$ with $0 < \omega < 3$. Observations are taken from a random sample of $n = 25$, and a significance level of 5% is used. We also looked at a significance level of 1% and random samples of $n = 50$. The results presented here are typical. The critical values used, except for the Anderson-Darling test, were from the asymptotic χ^2 distribution as these

are what will be used in practice and (judging by the $\omega = 0$ results) don't seem to unduly advantage any statistic. For the Anderson-Darling statistic, critical values of 2.492 (5%) and 3.857 (1%) were used (see D'Agostino and Stephens, 1986, Table 4.2, p105). Especially for $n = 50$ it was expected that the difference between actual and nominal sizes would be minimal. This was largely the case. When there was a discrepancy, this reflects what would happen in practice.

To allow efficient simulation, rather than transforming $U(0,1)$ variates using the inverse of the alternative cumulative distribution function (which requires computationally expensive numerical root finding), random variates from the probability density function $f_Y(y;\omega)$ are generated using the acceptance-rejection method. We follow the treatment given in Lange (1999; pp.272-276) with $c = 2$ and $g(y)$ the probability density function of the uniform $U(0,1)$ distribution, since $f_Y(y;\omega) \leq 2g(y)$. Here, a $U(0,1)$ random variate X is taken to come from the probability density function $f_Y(y;\omega)$ if another independent associated $U(0,1)$ random variate Z is such that $f_Y(X;\omega) \geq 2Z$; otherwise X and Z are discarded.

For $0 < \omega < 0.25$, $f_Y(y;\omega)$ is strictly increasing, and may be thought of as crudely linear. For $0.25 < \omega < 0.75$, $f_Y(y;\omega)$ increases and then decreases, and may be thought of as crudely quadratic. For $0.75 < \omega < 1.25$, $f_Y(y;\omega)$ increases, then decreases, and then increases again, and may be thought of as crudely cubic. And so on. If more complicated alternatives arise than may be crudely modeled by $\omega < 3$, we would hope that this could have been anticipated from the context, and smooth tests based on trigonometric or some other functions used. Except for the tests based on X^2_{P10} and X^2_{P20}, the powers of the other tests considered here decrease for $\omega > 3$. Some powers are given in Table 1. Figures 2 and 3 give a plots of the power functions based on a finer grid of ω values.

The simulations permit several conclusions.

- The tests based on V_1^2 and V_3^2 have higher powers when $f_Y(y;\omega)$ may be thought of as crudely a polynomial of odd degree. Their powers are greatest when $f_Y(y;\omega)$ may be thought of as crudely cubic.

- The tests based on V_2^2 and V_4^2 have higher powers when $f_Y(y;\omega)$ may be thought of as crudely a polynomial of even degree. Their powers are greatest when $f_Y(y;\omega)$ may be thought of as crudely quartic.

- The test based on S_4 often has power greater than the best of its component tests.

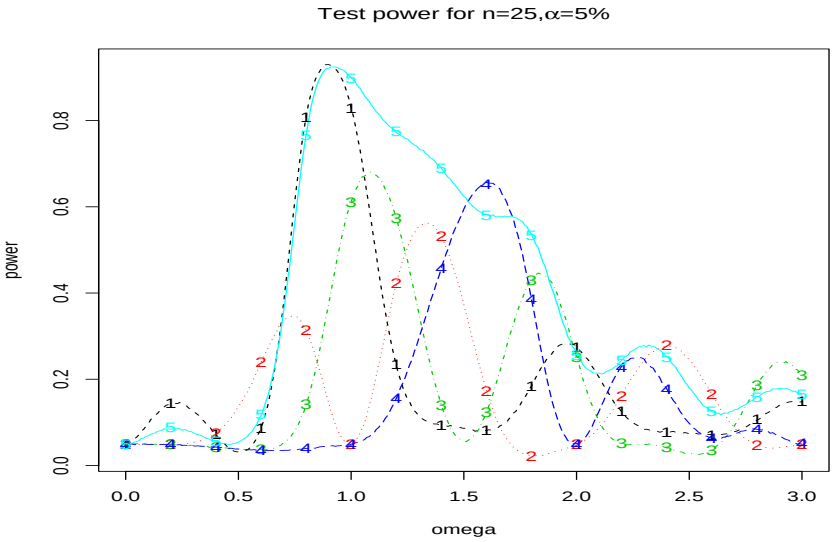
Table 1. Powers of various smooth and X^2 tests and the Anderson-Darling test for various periodic alternatives.

ω	V_1^2	V_2^2	V_3^2	V_4^2	S_4	AD	X^2_{P2}	X^2_{P5}	X^2_{P10}	X^2_{P20}
0.0	0.049	0.051	0.049	0.050	0.050	0.050	0.043	0.047	0.053	0.057
0.2	0.143	0.048	0.049	0.049	0.085	0.137	0.111	0.083	0.080	0.076
0.4	0.070	0.077	0.042	0.044	0.055	0.067	0.068	0.071	0.072	0.072
0.6	0.086	0.239	0.035	0.035	0.117	0.104	0.100	0.151	0.130	0.114
0.8	0.807	0.315	0.141	0.039	0.767	0.907	0.751	0.736	0.564	0.399
1.0	0.828	0.049	0.612	0.049	0.900	0.886	0.930	0.811	0.702	0.518
1.2	0.232	0.421	0.574	0.155	0.772	0.458	0.413	0.634	0.541	0.393
1.4	0.094	0.528	0.141	0.460	0.689	0.268	0.059	0.498	0.438	0.323
1.6	0.079	0.173	0.123	0.654	0.578	0.145	0.047	0.356	0.394	0.302
1.8	0.186	0.023	0.430	0.386	0.536	0.240	0.045	0.564	0.594	0.451
2.0	0.276	0.049	0.250	0.050	0.258	0.333	0.044	0.546	0.632	0.499
2.2	0.124	0.164	0.051	0.230	0.244	0.199	0.045	0.352	0.538	0.434
2.4	0.078	0.277	0.042	0.177	0.252	0.173	0.046	0.588	0.474	0.385
2.6	0.071	0.165	0.037	0.065	0.126	0.117	0.048	0.500	0.415	0.359
2.8	0.105	0.047	0.187	0.084	0.158	0.135	0.113	0.220	0.526	0.456
3.0	0.149	0.049	0.208	0.050	0.164	0.176	0.171	0.236	0.525	0.480

- The test based on X^2_{P2} divides the domain into two equal parts. It achieves its greatest power when $\omega = 1$, when the discrepancy between the two parts is greatest.
- For the X^2_{Pm} tests too few classes means the test cannot detect more complicated alternatives, while too many classes 'dilutes' the test, in that its ability to detect more complicated alternatives diminishes its ability to detect less complicated alternatives. Of these X^2_{Pm} tests, those based on 5 and 10 classes both give good results.
- The Anderson-Darling test performs reasonably well for most values of ω but usually does not give the most powerful test.

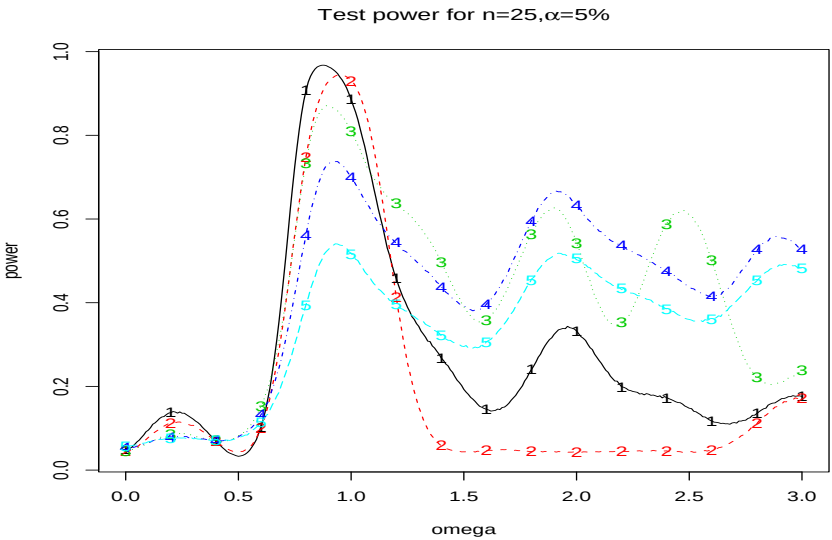
Quesenberry and Miller (1977) and Miller and Quesenberry (1979) both consider testing for uniformity, but neither considered the tests based on the components V_r, $r = 1, \ldots, 4$. Of several tests (including X^2_{Pm} with $m = 10$ and $m = 20$) they ultimately recommend the test based on S_4, in part because it "would have better power performance against further, perhaps more complicated, alternative classes"; see Miller and Quesenberry (1979, p.288). Several of our conclusions, such as the first two dot points immediately above, weren't addressed or could not be addressed in the Quesenberry and Miller studies. In addition, the conclusion that the low order tests considered cannot detect the more complicated alternatives, the key conclusion of the Quesenberry and Miller studies, isn't well addressed

Figure 2. Power curves for the tests based on V_1^2 (1=dashed), V_2^2 (2=dotted), V_3^2 (3=dotdash), V_4^2 (4=longdash), and S_4 (5=solid).



in their papers. If "more complicated" is taken to mean "$\omega > 2$", then the current study does address their key conclusion. Unlike the alternatives used here, the alternatives used in power studies by Quesenberry and Miller (1977), Miller and Quesenberry (1979) and most other authors, are indexed discretely rather than continuously. Also the range of alternatives accessed is not as broad as we are able to consider using our family.

Figure 3. Power curves for the Anderson-Darling test (1=solid) and tests based on X^2_{P2} (2=dashed), X^2_{P5} (3=dotted), X^2_{P10} (4=dotdash), and X^2_{P20} (5=longdash).



4. Discussion

An informed practicing statistician almost always has a *contextual* expectation of some of the basic characteristics of the data before it is seen. That is how good experiments are designed and tests of sensible hypotheses decided upon. More powerful Neyman smooth tests result from using fewer components: the ability to successfully use fewer components depending on the orthonormal system chosen. When you choose a particular orthonormal system (based on your contextual expectations about the data), you are choosing the alternatives you will best be able to detect, even though the data should not yet have been sighted. The results of our study demon-

strate the unsurprising fact that when using a polynomial orthonormal system, more complicated alternatives (here, higher frequency or larger ω) require many components in order to have reasonable power.

Ultimately, fewer components result in more powerful tests (by appropriately selecting a particular orthonormal system) and the statisticians contextual expectation's about the data are the basis on which this is achieved.

The statistic V_r optimally detects a particular order r polynomial alternative to uniformity. It is thus the basis of a very directional test, and could not be expected to detect more complex alternatives well. The statistic S_4 optimally detects polynomial alternatives to uniformity of degree up to four. It is thus the basis of a broadly focused test, being able to detect interesting and relatively complex alternatives. Attempting to detect even more complex alternatives results in less power for detecting alternatives up to degree four. This cost is often achieved with little gain, as a four dimensional parameter space is usually rich enough to detect most alternatives that arise in practice.

Generally, we recommend using a polynomial orthonormal system unless there is a reason not to (for example, if the context suggests periodic alternatives may be more suitable). The advantage of the polynomial orthonormal systems is the components may be interpreted as (roughly) detecting moment departures of the data from the null distribution. The interpretation for other systems is more problematic.

There may well be a loss if we have chosen the wrong orthonormal system. So here, if we had a contextual expectation of periodic alternatives it would be appropriate to use alternatives based on something like the orthonormal series $\{\sqrt{2}\sin(i\pi y)\}$. Using such periodic orthonormal functions would probably give good protection against such alternatives, (though no doubt these would produce poorer power against the alternatives the polynomial orthonormal system components have good power detecting). Our study assumes there are no such contextual expectations of periodic alternatives, so the alternatives of interest here are only weakly periodic ($0 < \omega < 3$).

In the size and power study here, it seems that the tests based on V_r^2 with $r = 1$ outperforms the tests based on larger r for smaller values of ω, although this is not uniformly true. Tests based on V_r^2 with larger r are more powerful for larger values of ω, but again, this is not uniformly true. The test based on S_4 is sometimes more powerful than all the V_r^2 tests, and is always a good compromise.

The Pearson tests are tests for discrete alternatives. The test based on X_{Pm}^2 may be thought of as optimally detecting 'order' $m-1$ alternatives; see Rayner and Best (1989, Chapter 5). Order in this sense reflects the complexity of the alternative. From the simulations it is clear that the Pearson

test with $m = 2$ classes is unable to detect the more complex alternatives, while that with $m = 20$ protects against quite complex alternatives − that we don't have here − at the cost of a loss of power for the less complex alternatives. The X^2_{Pm} tests with $m = 5$ and $m = 10$ outperform S_4 for $\omega \geq 2$. Presumably the X^2_{Pm} test with $m = 5$ is able to detect alternatives of similar complexity to the S_4 test, and so will sometimes do better and sometimes worse. The X^2_{Pm} test with $m = 10$ is able to protect against more complex alternatives than the S_4 test, but is clearly inferior when the alternatives are less complex: for $\omega < 2$.

We can predict the outcome of assessing the Best and Rayner (1985) idea of looking at the residual from X^2_{P20}. The residual will always include higher order components. If the alternative is not complex (say $0 < \omega < 1$) the tests based on these components will have little power, as will a test based on a residual involving these components. If the alternative is complex we probably should be using a different orthonormal family. For example, the more apparently periodic alternatives that occur for $\omega > 3$ would imply the use of something like the periodic orthonormal series given above.

If we are looking at residuals from X^2_{p5} and X^2_{P10}, what may be of interest is to combine later order components. The chi-squared components will be similar to the smooth test components, and corresponding to residuals of the the chi-squared tests are tests based on sums of V^2_r such as $V^2_{r+1} + \ldots + V^2_s$. Again we can predict what will happen. If S_r is powerful we would expect a residual like $V^2_{r+1} + \ldots + V^2_s$ not to be, and conversely. A good question here is what smooth test residual should we use: $V^2_3 + V^2_4 + V^2_5$ or perhaps $V^2_5 + \ldots + V^2_{10}$? Consideration could also be given to sums of squares of odd and sums of squares of even components.

To some extent the chi-squared components duplicate the smooth test components, and since we are testing for a continuous null, the smooth test is more appropriate. We *are* still advocating looking at the components.

The key point from this study are that if we seek broadly focused tests rather than strongly directional or weakly omnibus tests, then the tests based on S_r with r about 4, or on X^2_{Pm} with m in the range 5 to 10 will perform well. With the S_r tests the orthonormal system should be chosen so that relatively few components are required to detect important alternatives. Given that the Pearson tests can perform well, it would be useful to look again at the class formation options. Can the equiprobable class construction used here be improved upon? We will consider this question in a subsequent paper.

Acknowledgments

We would like to thank Dr John Best for his helpful comments and criticism.

References

1. D. J. Best and J. C. W. Rayner. Uniformity testing when alternatives have low order. *Sankhya A*, 47(1):25–35, 1985.
2. A. M. Carolan and J. C. W. Rayner. Interpreting the components of a smooth goodness of fit test for normality. *Submitted*, 2000.
3. A. Cohen and H. B. Sackrowitz. Unbiasedness of the chi-squared, likelihood ratio, and other goodness of fit tests for the equal cell case. *Ann. Statist.*, 3:959–964, 1975.
4. R. B. D'Agostino and M. A. Stephens. *Goodness-of-fit techniques.* New York: Marcel Dekker, Inc, 1986.
5. W. C. M. Kallenberg, J. Oosterhoff and B. F. Schriever. The number of classes in chi-squared goodness-of-fit tests. *J. Amer. Statist. Ass.*, 80:959–968, 1985.
6. K. Lange. *Numerical Analysis for Statisticians.* New York: Springer-Verlag, 1999.
7. F. L. Miller and C. P. Quesenberry. Power studies of some tests for uniformity II. *Commun. Statist.-Simul. Comp.*, 8:271–290, 1979.
8. J. Neyman. 'Smooth' test for goodness of fit. *Skand. Aktuarietidskr.*, 20:150–199, 1937.
9. C. P. Quesenberry and F. L. Miller. Power studies of some tests for uniformity. *J. Statist. Comp. Simul.*, 5:169–191, 1977.
10. J. C. W. Rayner and D. J. Best. *Smooth tests of goodness of fit.* New York: Oxford University Press, 1989.
11. J. C. W. Rayner and D. J. Best. Goodness of fit: methods and models. To appear in the *International Encyclopaedia of Social and Behavioral Sciences*, 2000.
12. J. C. W. Rayner, D. J. Best, and K. G. Dodds. The construction of the simple X^2 and Neyman smooth goodness of fit tests. *Statistica Neerlandica*, 39:35–50, 1985.
13. J. C. W. Rayner, D. J. Best, and K. L. Mathews. Interpreting the skewness coefficient. *Commun. Statist.-Theor. Meth.*, 24(3):593–600, 1995.