ECONOMETRIC REVIEWS
Vol. 23, No. 4, pp. 341–370, 2004

The Sensitivity of Chi-Squared Goodness-of-Fit Tests to the Partitioning of Data

Gianna Boero,[1,2] Jeremy Smith,[1] and Kenneth F. Wallis[1,*]

[1]Department of Economics, University of Warwick, Coventry, UK
[2]DRES, University of Cagliari, Italy

ABSTRACT

The power of Pearson's overall goodness-of-fit test and the components-of-chi-squared or "Pearson analog" tests of Anderson [Anderson, G. (1994). Simple tests of distributional form. *J. Econometrics* 62:265–276] to detect rejections due to shifts in location, scale, skewness and kurtosis is studied, as the number and position of the partition points is varied. Simulations are conducted for small and moderate sample sizes. It is found that smaller numbers of classes than are used in practice may be appropriate, and that the choice of non-equiprobable classes can result in substantial gains in power.

Key Words: Pearson's goodness-of-fit test; Component tests; Monte Carlo; Number of classes; Partitions.

JEL Classification: C12; C14.

*Correspondence: Kenneth F. Wallis, Department of Economics, University of Warwick, Coventry CV4 7AL, UK; E-mail: k.f.wallis@warwick.ac.uk.

DOI: 10.1081/ETC-200040782
0747-4938 (Print); 1532-4168 (Online)

1. INTRODUCTION

The assessment of goodness of fit or the degree of correspondence between observed outcomes and expected outcomes based upon a postulated distribution is a central problem in classical statistics. The two classical nonparametric approaches to testing goodness of fit, as surveyed by Stuart et al. (1999, Ch. 25), for example, are Pearson's chi-squared (X^2) test, which involves grouping data into classes and comparing observed outcomes to those hypothesised under some null distribution, and the Kolmogorov–Smirnov (K–S) test, which involves comparing the empirical cumulative distribution function (cdf) with a cdf obtained under some null hypothesis.

Despite the widespread use of the X^2 test, little reliable guidance on the design of the test, namely how many classes, and of what relative magnitudes, seems to be available. Typical practitioner choices are round numbers such as 10 or 20 classes, invariably with equal probability under the null hypothesis. To use classes with equal, or nearly equal probabilities is the first of the practical recommendations Stuart, Ord and Arnold draw from their literature review, although they note that if lack of fit in the tails is of particular interest, this may lead to a considerable loss of power. In these circumstances, rather than redesigning the test, they suggest using the K–S test. Their second recommendation is to determine the number of classes, for sample sizes in excess of 200, by a formula akin to one originally developed by Mann and Wald (1942), although this gives many more classes than are commonly used.

In this article, we conduct a systematic analysis of the power of the X^2 test in relation to these design questions – the number of classes, and the location of the partition points. We first reappraise some of the literature reviewed by Stuart et al. (1999), and then present a comprehensive set of simulation experiments. We include in the X^2 family of tests the components-of-chi-squared or ''Pearson analog'' tests of Anderson (1994). The components focus on particular moments or characteristics of the distribution of interest, and are potentially more informative about the nature of departures from the null hypothesis than the single ''portmanteau'' goodness-of-fit test. The power of the X^2 test and its components is examined with respect to departures in location, scale, skewness and kurtosis from a null distribution, $N(0,1)$. The power of these tests is also compared to that of the K–S test and standard moment-based tests. No applications are discussed in this article; whereas the comparison of income distributions is a leading example in economics, our interest lies in the evaluation of density forecasts (Boero and Marrocu, 2004; Noceti et al. 2003; Wallis, 2003).

The article is organized as follows. Section 2 contains a brief discussion of the X^2 test and some relevant literature on the choice of the number and location of the partition points. We also describe Anderson's (1994) decomposition of the X^2 test, as well as the other tests used for comparative purposes. Section 3 is concerned with experimental design, and presents the various distributions used to generate artificial data under the alternative hypotheses against which the $N(0,1)$ null hypothesis is tested. Section 4 reports the simulation results on the ability of the X^2 test and its components to detect departures from the null hypothesis using both equiprobable and non-equiprobable partitions. Section 5 summarises the main results and offers some concluding remarks. It is seen that smaller numbers of classes than are used

in practice may be appropriate, and that the choice of non-equiprobable classes can result in substantial gains in power.

2. GOODNESS-OF-FIT TESTS

2.1. The Chi-Squared Goodness-of-Fit (X^2) Test

Pearson's classical goodness-of-fit test proceeds by dividing the range of the variable into k mutually exclusive classes and comparing the expected frequencies of outcomes falling in these classes given by the hypothesised distribution with the observed class frequencies. With class probabilities $p_i > 0,\ i = 1, \ldots, k$, $\sum_{i=1}^{k} p_i = 1$ and observed class frequencies $n_i,\ i = 1, \ldots, k,\ \sum_{i=1}^{k} n_i = N$, the test statistic is

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i}.$$

This has a limiting χ^2 distribution with $k - 1$ degrees of freedom if the hypothesised distribution is correct.

The existing literature on the power of the X^2 test with different numbers of classes ranges from early studies by Mann and Wald (1942) and Gumbel (1943) to more recent work by Kallenberg et al. (1985) and Koehler and Gan (1990). Most of these studies are based on asymptotic results and assume equiprobable classes, with $p_i = 1/k,\ i = 1, \ldots, k$. Mann and Wald (1942) suggested the use of equiprobable classes to resolve the question of how class boundaries should be determined, and developed a formula for the optimal choice of the number of classes, which depends on the sample size, N, and the level of significance. At the 5% significance level the formula is $k = 3.765(N-1)^{0.4}$, and Table 1 reports the resulting values of k for selected values of N. Although this rests on the asymptotic power function, Mann and Wald suggest that the results hold approximately for sample sizes as low as 200, and may be true for considerably smaller samples.

This procedure removes the subjective element from the choice of the number and width of the classes, moreover equiprobable classes are easy to use and lead to unbiased tests (Cohen and Sackrowitz, 1975). However, various numerical studies

Table 1. Mann and Wald's optimum k.

N	k
25	13
50	18
75	21
100	24
150	28
250	34
350	39

argue that the value of k proposed by Mann and Wald is too large, resulting in loss of power in many situations; for example Williams (1950) suggests that it may be halved for practical purposes, without relevant loss of power. See also Dahiya and Gurland (1973), who suggest values of k between 3 and 12 for several different alternatives in testing for normality, for sample sizes of $N = 50$ and 100.

Other studies have suggested that the best choice of k depends on the nature of the alternative hypothesis under consideration as well as the sample size. In a comparison of the power of the X^2 and likelihood ratio goodness-of-fit tests, Kallenberg et al. (1985) suggest that, particularly for heavy tailed alternatives, the X^2 test with equiprobable classes has the best power when k is relatively large ($k = 15$ and 20 when $N = 50$ and 100, respectively). These values of k are only slightly smaller than those given by the Mann and Wald formula, and are also suggested by Koehler and Gan (1990) as a good overall choice.

On the other hand, Kallenberg et al. (1985) argue that as the variance of the X^2 test increases with k, and this has a negative effect on the power, non-equiprobable partitions with moderate k are better than equiprobable partitions with large k. For example, partitions with some smaller classes in the tails and larger classes in the middle may lead to an important gain of power for alternatives with heavy tails, while for thin-tailed alternatives, unbalanced partitions often cause a loss of power.

Most of these results are based on asymptotic theory, and only a limited number of cases have been examined to validate the asymptotic theory. For our present purposes the Monte Carlo study of Kallenberg et al. (1985) is an interesting precursor, since they observe (p. 966) that ''the good power properties of unbalanced partitions in most examples is striking; a gain of power of 0.3 or 0.4 compared with balanced partitions is not uncommon''. This is surprisingly overlooked by Stuart et al. (1999), who instead state (p. 408) that Koehler and Gan (1990) ''show that unequal probability partitions may increase power for specific alternatives''. Unfortunately Koehler and Gan's Monte Carlo study uses only equiprobable partitions, and their remark about the non-equiprobable case is an assertion in their next-to-last sentence about procedures that are beyond the scope of their article. This deficiency is remedied below.

2.2. Components-of-Chi-Squared or "Pearson Analog" Tests

Anderson (1994) presents a rearrangement of the X^2 statistic into $k-1$ components, each independently distributed as $\chi^2(1)$, as follows:

$$X^2 = \sum_{j=1}^{k-1} \nu_j^2 / N\sigma_j^2$$

where $\nu_j = i_j'(x - \mu)$ and x is a $k \times 1$ vector of observed frequencies with mean vector μ under the null hypothesis. The variance is $\sigma_j^2 = 1 - (i_j' p)^2$ where $p = (p_1, \ldots, p_k)'$ is a vector of class probabilities. With k equal to a power of 2, the $k \times 1$ vectors i_j, $j = 1,\ldots, k-1$ have elements equal to 1 or –1 and are mutually orthogonal.

The rearrangement provides additional information about departures from the hypothesised distribution in respect of specific features of the empirical distribution

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

such as its location, scale, skewness and kurtosis. For example, with $k = 8$, we emphasise this interpretation by writing the first four i_j vectors with alphabetic subscripts, namely:

$$
\begin{aligned}
i'_m &= [1 \quad 1 \quad 1 \quad 1 \quad -1 \quad -1 \quad -1 \quad -1] \\
i'_{sc} &= [1 \quad 1 \quad -1 \quad -1 \quad -1 \quad -1 \quad 1 \quad 1] \\
i'_{sk} &= [1 \quad 1 \quad -1 \quad -1 \quad 1 \quad 1 \quad -1 \quad -1] \\
i'_k &= [1 \quad -1 \quad -1 \quad 1 \quad 1 \quad -1 \quad -1 \quad 1].
\end{aligned}
$$

Thus, with equiprobable classes, the first Pearson component test (PCM), using i_m, focuses on location shifts relative to the median. The second component test (PCSc), using i_{sc}, focuses on scale shifts to the inter-quartile range. The third component test (PCSk), using i_{sk}, detects asymmetries, represented as shifts between the first and third quarters and the second and fourth quarters of the distribution. The fourth component test (PCK), using i_k, detects kurtosis, with shifts towards the extremes and the centre of the distribution. The three remaining components have no obvious interpretation.

Boero et al. (2004) present a formal derivation of the component tests and show that the above decomposition into a sum of independent components only holds in the case of equiprobable classes, hence the procedure is less generally applicable than was originally claimed. Nevertheless with non-equiprobable classes each component test still has a $\chi^2(1)$ distribution, and they are retained in our experiments, which focus on specific departures with respect to location, scale, skewness and kurtosis one at a time.

Whereas for the X^2 test the choice of k is important, the individual component tests do not depend on k, once k is large enough to define them. When $k = 4$ only three components are defined, and these three component tests are unchanged when k is increased to eight, assuming that the eight classes are obtained by dividing each of the original four classes into two, without moving the partition points.

We define partition points implicitly, as the appropriate percentage points of the relevant cdf, F, the partition points being the corresponding x-coordinates. Denote the value of the cdf at the upper boundary of the jth class as F_j. The first and last classes are open-ended, thus with $F_0 = 0$ and $F_k = 1$ the class probabilities satisfy $p_j = F_j - F_{j-1}$, $j = 1, \ldots, k$, and a partition configuration is reported as the set $\{F_1, \ldots, F_{k-1}\}$.

The power of the individual component tests (and hence of X^2) depends on the location of F_j. For example, for unimodal distributions, if two distributions differ only in their median then with $k = 8$ classes, the power of the median component test is sensitive to F_4, which corresponds to the sign change in the vector i_m. If the distributions differ in the scale parameter (inter-quartile range), the power of the scale component test depends upon F_2 and F_6, which correspond to the two sign changes in the vector i_{sc}. If the distributions differ in skewness, the power of the skewness component test is reliant on F_2, F_4 and F_6, which correspond to the three sign changes in the vector i_{sk}. Finally, if the distributions differ in kurtosis, the power of the kurtosis component test is reliant on F_1, F_3, F_5 and F_7, which correspond to

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

the four sign changes in the vector i_k. See Anderson (1994) for further discussion of these partition points.

One major criticism levelled at the Pearson X^2 test is that the test is potentially inconsistent, since the power of the test does not approach unity as $N \to \infty$, see, for example, Barrett and Donald (2003) and Linton et al. (2002). The problem arises from the relation between the choice of classes and the nature of the specific alternative. For example, if the null and alternative distributions are both symmetric around the same mean and we choose two equiprobable classes, then the power of the X^2 test is zero irrespective of N. In general, the power of the test depends on the closeness of the intersection points of the two density functions to the class boundaries, and the inconsistency problem can always be overcome by choosing the number of classes to be greater than the number of anticipated intersection points. Or, even in the example above, by choosing non-equiprobable classes. The inconsistency of the X^2 test is not evident in our experiments as we always consider values of k greater than the number of intersection points.

2.3. Kolmogorov–Smirnov Test

The other nonparametric test used in the study is the K–S statistic

$$D = \max|A_i - Z_i|, \quad 1 < i < N,$$

where Z is the theoretical cdf under the null hypothesis, and A is the empirical cdf, with critical values taken from Miller (1956).

2.4. Moment-Based Tests

Grouping the data in general loses information, and tests based on grouped data can be expected to be less powerful, though possibly more robust, than tests based directly on the sample observations. For comparative purposes, and as possible benchmarks, we also include in our experiments the moment-based tests relevant to the specific departures from the null hypothesis of $N(0,1)$. For departures due to a non-zero mean, this is the standard normal test of a sample mean, calculated as the sample mean divided by its standard error $\sqrt{1/N}$. For departures due to a non-unit variance we test the sample sum of squares of mean deviations in the $\chi^2(N-1)$ distribution.

Tests of skewness and kurtosis are often combined, and sometimes called ''tests of normality''. Bowman and Shenton (1975) show that a combined statistic based on the skewness and kurtosis coefficients is asymptotically $\chi^2(2)$ under the null hypothesis. The test is commonly attributed to Jarque and Bera (1980), who show that it is a score or LM test and hence asymptotically efficient. We follow this practice, and refer to it as the J–B test below.

Apart from these simple examples, estimation of the parameters of the various distributions is not undertaken in any of our experiments, hence no modification of any test to allow for the effect of parameter estimation error is required.

3. EXPERIMENTAL DESIGN

The Monte Carlo experiments are designed to determine the power of the X^2 and its component tests to detect departures from the null distribution, $N(0,1)$, with respect to mean, variance, skewness or kurtosis. In this section, we outline the nature of these alternative distributions.

3.1. Experiment A: Non-zero Mean

$N(\delta/\sqrt{N}, 1)$, with δ varying from 0.1 to 2.0 through steps of 0.1.

3.2. Experiment B: Non-unit Variance

$N(0, \delta^2)$, with δ varying from 0.1 to 2.0 through steps of 0.1.

3.3. Experiment C: Skewed Distributions

$\boldsymbol{C}_1$: *Ramberg distribution* (see Ramberg et al., 1979), is a flexible form expressed in terms of its cumulative probabilities. The Ramberg quantile and density functions have the form:

$$R(p) = \lambda_1 + \left[p^{\lambda_3} - (1-p)^{\lambda_4}\right]/\lambda_2$$
$$f(x) = f[R(p)] = \lambda_2\left[\lambda_3 p^{\lambda_3 - 1} + \lambda_4 (1-p)^{\lambda_4 - 1}\right]$$

with $0 < p < 1$ being the cumulative probability, $R(p)$ the corresponding quantile, and $f[R(p)]$ the density corresponding to $R(p)$. Of the four parameters, λ_1 is the location parameter, λ_2 the scale parameter, and λ_3 and λ_4 are shape parameters. For the present purpose, we choose their values such that $E(X) = 0, V(X) = 1$, skewness $= \{0.00, 0.05, 0.10, \ldots, 0.90\}$ and kurtosis $= 3$. The median is then non-zero and it is an increasing function of the skewness. In order to concentrate on the effect of skewness alone, we shift the distribution by the empirically calculated median.

$\boldsymbol{C}_2$: *Two-piece normal distribution* (see Wallis, 1999), is used by the Bank of England and the Sveriges Riksbank in presenting their density forecasts of inflation. The probability density function is

$$f(x) = \begin{cases} [\sqrt{2\pi}(\sigma_1 + \sigma_2)/2)]^{-1} \exp[-(x-\mu)^2/2\sigma_1^2] & x \le \mu \\ [\sqrt{2\pi}(\sigma_1 + \sigma_2)/2)]^{-1} \exp[-(x-\mu)^2/2\sigma_2^2] & x \ge \mu \end{cases}.$$

The distribution is positively skewed if $\sigma_2^2 > \sigma_1^2$, and is leptokurtic if $\sigma_1 \neq \sigma_2$. As in the Ramberg distribution, the median is an increasing function of skewness and we again shift the distribution, to ensure a theoretical median of zero. In our simulations, we consider combinations of (σ_1, σ_2) that yield $V(X) = 1$ and skewness of $\{0.00, 0.05, 0.10, \ldots, 0.90\}$.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$\boldsymbol{C}_3$: *Anderson's skewed distribution* (see Anderson, 1994)

$$x = \begin{cases} (z/(1+d)) & z < 0 \\ z(1+d) & \text{otherwise} \end{cases}$$

where $z \sim N(0,1)$. Since skewness $\approx 2 \times d$, we set $d = \{0.00, 0.025, \ldots, 0.45\}$. The mean, variance and kurtosis of this distribution are all increasing functions of d, although the median is zero. The transformation is discontinuous at zero, hence the probability density function has a central singularity, unlike the two-piece normal distribution.

3.4. Experiment D: Distributions with Heavy Tails

$\boldsymbol{D}_1$: *Stable distribution* (see Chambers et al., 1976 for the code). General stable distributions allow for varying degrees of tail heaviness and varying degrees of skewness. They can be represented with the general notation $S(\alpha, \beta, \gamma, \delta)$, with four parameters: an index of stability (or characteristic exponent) $0 < \alpha \leq 2$, which measures the height of (or total probability in) the extreme tail areas of the distribution, a skewness parameter $-1 \leq \beta \leq 1$, a scale parameter $\gamma > 0$ and a location parameter $\delta \in \Re$. When $\alpha = 2$ and $\beta = 0$ the distribution is Gaussian with variance 2; when $\alpha = 1$ and $\beta = 0$ the distribution is Cauchy; when $\alpha = 0.5$ and $\beta = 1$, the distribution is Levy. When $0 < \alpha < 2$ the extreme tails of the stable distribution are higher than those of a normal distribution, and the total probability in the extreme tails is larger the smaller the value of α. In our simulations we use standardised symmetric stable distributions, by setting $\gamma = 1$, $\delta = 0$ and $\beta = 0$. One consequence of stable distributions is that, if $\alpha < 2$, moments of order α or higher do not exist. For $\alpha = 2$ we scale the distribution to have a unit variance.

Stable distributions have been proposed as a model for many types of variables, especially in physics, finance and economics (see, for example, Uchaikin and Zolotarev, 1999). In finance, for example, stock prices are often modelled as non-Gaussian stable distributions (see Fama, 1965; Mandelbrot, 1963; McCulloch, 1994; Rachev and Mittnik, 2000).

$\boldsymbol{D}_2$: *Anderson's kurtotic distribution* (see Anderson, 1994)

$$x = z(|z|^q)(1+t)$$

where $z \sim N(0,1)$ and t is a variance-shifting nuisance parameter. We take combinations of q and t that give $V(X) = 1$ and kurtosis in the range 2.0–7.0.

The Monte Carlo experiments are based on sample sizes of $N = 25$, 50, 75, 100, 150, 250, 350. With 5000 replications, and power reported as a percentage the estimation error is smaller than 1.5 percentage points with 95% confidence. All tests are undertaken at the 5% significance level. For equiprobable partitions, we take $k = 2, 4, \ldots, 40$. In all experiments, except in experiment $\boldsymbol{A}$, for non-equiprobable partitions we take $k = 4$, 8, 16, 32 and consider partitions which are symmetric around $F_4 = 0.5$, such that for $k = 8$, $\{F_1, F_2, F_3, 0.5, 1 - F_3, 1 - F_2, 1 - F_1\}$.

In experiment ***A*** we take $\{F_4/4, F_4/2, 3F_4/4, F_4, 1-F_3, 1-F_2, 1-F_1\}$, when $F_4 \leq 0.5$ and $\{(1-F_4)/4, (1-F_4)/2, 3(1-F_4)/4, F_4, 1-F_3, 1-F_2, 1-F_1\}$ when $F_4 > 0.5$.

4. THE POWER OF THE TESTS

4.1. Experiment *A*: Departure from Zero Mean

4.1.1. Equiprobable Classes

We first illustrate the relation between power and number of classes for the overall X^2 test in the equiprobable case. The results are reported in Fig. 1 for a wide range of alternative values of $(\delta/\sqrt{N})$, for $N=150$ and k ranging from 2 to 40. More extensive results for different sample sizes are summarised in Table 2 (first three columns), although as we consider standardised location departures we do not observe an increase in power as N increases.

From Fig. 1 it is evident that, for most alternatives, power is maximised for a value k in the interval two to four, with power falling steadily as a function of k for $k>4$. These findings are confirmed in Table 2 for all sample sizes.

The last two columns of Table 2 report the power of the K–S test and the standard normal (z-) test for a sample mean. Comparing the power of the X^2 test with the K–S test shows a superiority of the K–S test in all cases; however, both of these tests have markedly inferior power compared to the z-test.

4.1.2. Non-equiprobable Classes

As discussed above the power of the median component test (PCM) depends upon F_4. However, the preceding results demonstrate the superiority of $k=4$ over

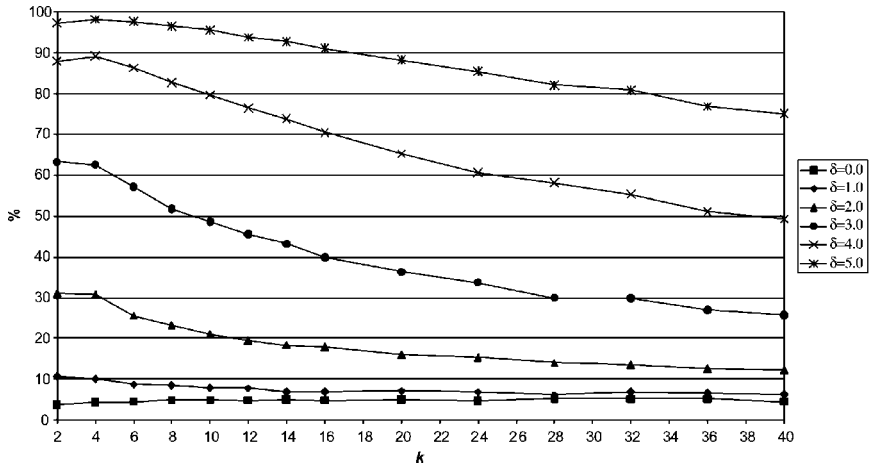

Figure 1. Power of X^2 test of $N(0,1)$ vs. $N(N^{-0.5}\delta, 1)$: Experiment $\boldsymbol{A}$, $N = 150$ (equiprobable).

Table 2. Power of the X^2 test of $N(0,1)$ vs. $N(\delta/\sqrt{N}, 1)$: Experiment ***A***.

	Equiprobable splits				
δ	$k=4$	$k=8$	$k=16$	z-test	K–S
			N=25		
0.0	4.5	4.7	4.4	5.0	4.9
1.0	9.1	8.4	7.1	26.0	14.1
1.5	16.3	14.2	10.9	44.2	25.0
2.0	26.9	22.6	17.7	63.9	39.6
2.5	43.2	36.3	27.9	80.4	57.6
3.0	60.1	53.2	41.6	91.2	73.9
3.5	74.2	67.4	56.6	96.8	85.2
4.0	85.5	82.1	72.3	99.1	93.2
4.5	93.7	90.7	82.6	99.8	97.1
5.0	97.2	96.0	91.8	100.0	99.1
			N=50		
0.0	5.6	4.9	4.8	5.0	5.4
1.0	10.4	7.5	6.5	26.0	13.6
1.5	18.7	13.6	9.9	44.2	24.9
2.0	30.8	22.8	16.7	63.9	39.7
2.5	46.4	35.5	25.3	80.4	57.2
3.0	63.7	51.9	39.3	91.2	73.6
3.5	77.8	67.9	55.0	96.8	85.8
4.0	88.6	82.2	70.4	99.1	93.5
4.5	95.2	90.9	82.2	99.8	97.3
5.0	97.9	95.7	90.9	100.0	99.0
			N=100		
0.0	5.0	4.6	5.4	5.0	5.3
1.0	10.2	8.0	7.5	26.0	13.4
1.5	18.4	13.2	11.0	44.2	24.4
2.0	30.9	23.4	18.4	63.9	39.6
2.5	45.4	35.5	27.2	80.4	56.2
3.0	63.5	52.3	41.1	91.2	73.5
3.5	77.9	67.8	56.2	96.8	86.4
4.0	88.7	82.3	71.1	99.1	93.6
4.5	95.4	90.5	82.5	99.8	97.6
5.0	98.3	96.5	91.4	100.0	99.2
			N=150		
0.0	4.7	5.1	5.4	5.0	4.7
1.0	10.8	9.0	7.4	26.0	13.2
1.5	17.9	13.6	10.6	44.2	24.7
2.0	31.7	23.5	17.6	63.9	40.3
2.5	46.0	36.6	27.1	80.4	57.2
3.0	63.3	53.3	40.0	91.2	73.9
3.5	77.9	69.3	55.6	96.8	86.4
4.0	89.5	83.5	71.3	99.1	94.5
4.5	95.2	91.1	82.7	99.8	97.5
5.0	98.2	96.5	91.4	100.0	99.2

(continued)

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 2. Continued.

	Equiprobable splits				
δ	$k=4$	$k=8$	$k=16$	z-test	K–S
			N=250		
0.0	4.5	5.0	4.9	5.0	4.6
1.0	10.1	8.7	7.1	26.0	13.4
1.5	18.2	14.0	11.1	44.2	24.1
2.0	30.9	23.2	17.9	63.9	39.2
2.5	46.6	36.9	25.9	80.4	57.4
3.0	62.5	51.8	39.8	91.2	72.7
3.5	77.7	67.8	54.4	96.8	85.9
4.0	89.1	82.7	70.5	99.1	94.2
4.5	95.0	90.7	82.4	99.8	97.5
5.0	98.2	96.4	91.0	100.0	99.3

$k=8$ classes, hence in experiment ***A*** we take $k=4$ and set the partitions such that $\{F_1,\ F_2,\ F_3\}=\{F_2/2,\ F_2,\ 1-F_2/2\}$, and take values of F_2 in the range 0.2–0.8 through steps of 0.05, for $F_2 \geq 0.5$ we take $\{(1-F_2)/2,\ F_2,\ 1-(1-F_2)/2\}$. For $k=4$, Fig. 2a and b plot the power of the X^2 and PCM tests, respectively, against values of F_2 for $N=150$ and $\delta \geq 0$.

Figure 2a shows that the power of the X^2 test is relatively insensitive to the location of the partition point F_2 for $F_2 \geq 0.4$, although there is some evidence that

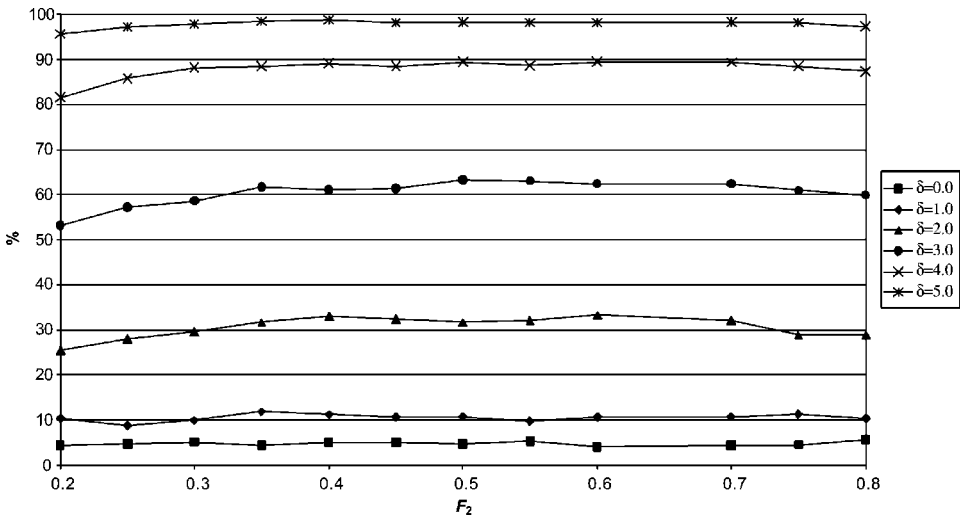

Figure 2a. Power of X^2 test of $N(0,1)$ vs. $N(N^{-0.5}\delta,1)$: Experiment ***A***, $N=150$ (non-equiprobable).

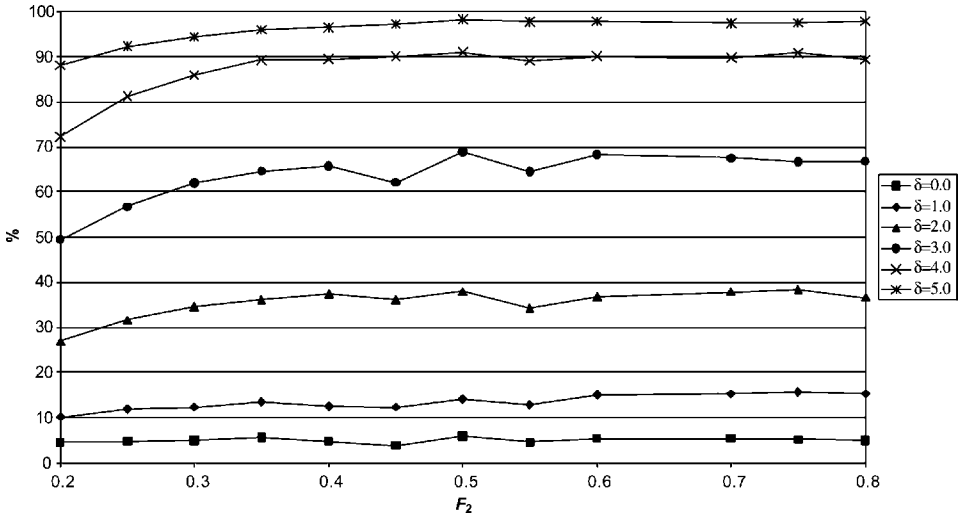

Figure 2b. Power of PCM test of $N(0,1)$ vs. $N(N^{-0.5}\delta,1)$: Experiment $\boldsymbol{A}$, $N=150$ (non-equiprobable).

power peaks close to the equiprobable partitions, $F_2 = 0.5$. For example, when $\delta = 2$, the power of the X^2 test is 32% for $F_2 = 0.5$, compared with 28% for $F_2 = 0.25$. Figure 2b shows a similar picture for the PCM test, with power largely insensitive to the location of F_2, providing $F_2 \geq 0.4$. Results not reported, show that non-equiprobable splits which are symmetric around $F_2 = 0.5$, of the form $\{F_1, 0.5, (1-F_1)\}$ for F_1 in the range 0.1–0.4, also show no significant power gains for the

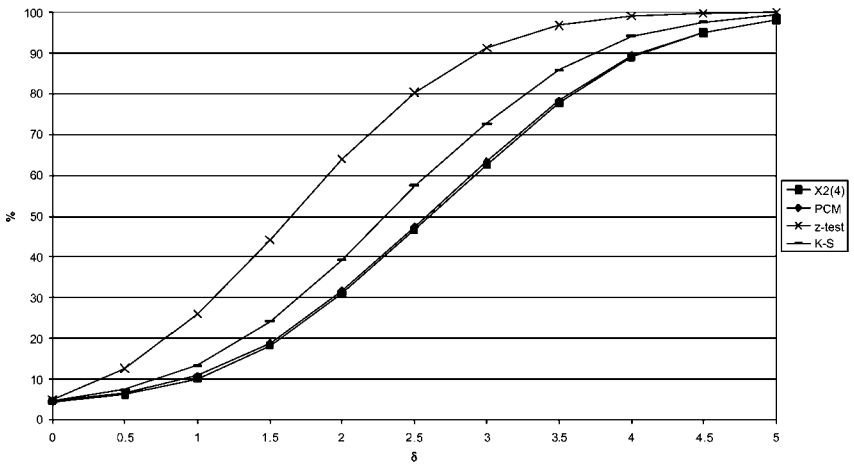

Figure 3. Power of the tests of $N(0,1)$ vs. $N(N^{-0.5}\delta,1)$; Experiment $\boldsymbol{A}$, $N=150$.

X^2 test over the equiprobable case $F_1 = 0.25$. Moreover, as expected, in this case the power of the PCM test was invariant to the location of F_1.

Finally, in Fig. 3 we summarise the power results of the X^2 test obtained for $\delta \geq 0$ and $N = 150$, using $k = 4$ equiprobable intervals (denoted X2(4)), we also report power of the PCM component test with equiprobable partitions ($F_2 = 0.25$), the K–S test and the simple z-test for the population mean. As we can see, the best performance is given by the z-test. The power of the X^2 test, with equiprobable partitions, is very similar to that of the PCM test, but both of these tests are inferior to the K–S test.

4.2. Experiment *B*: Departure from Unit Variance

4.2.1. Equiprobable Classes

We first illustrate the relation between power and number of classes, in the equiprobable case for the X^2 test. The results are reported in Fig. 4a for a wide range of alternatives with excess variance, and in Fig. 4b for alternatives with variance smaller than one, for $N = 150$ and k ranging from 2 to 40. More extensive results for different sample sizes are summarised in Table 3 (first three columns).

From Fig. 4a and b, it is evident that, for most alternatives, power is maximised for a value k in the interval 4 to 10. Values of k greater than 10 do not lead to further increases in power, rather, the performance of the test is more or less unchanged in the presence of excess variance, while there seems to be considerable loss in power when variance is reduced for values of k greater than 10.

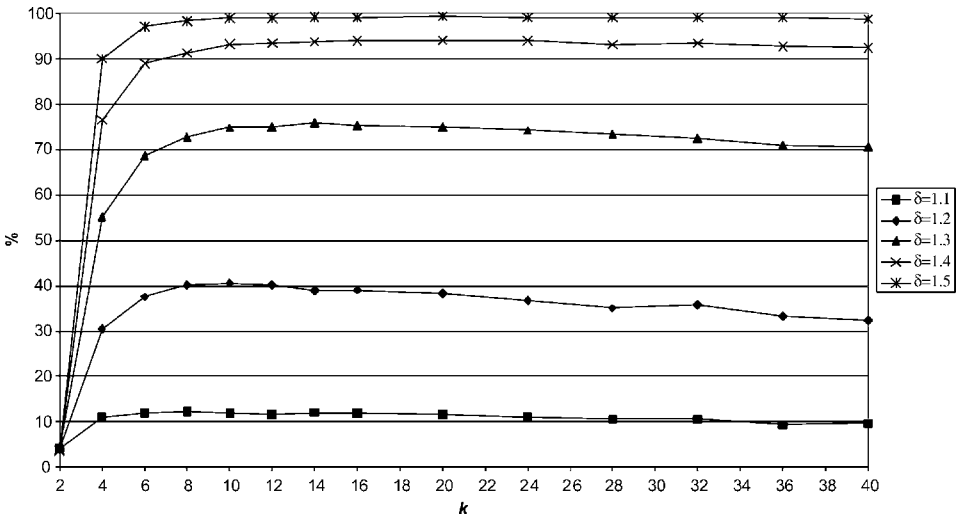

Figure 4a. Power of X^2 test of $N(0, 1)$ vs. $N(0, \delta)$: Experiment ***B***, $N = 150, \delta > 1$ (equiprobable).

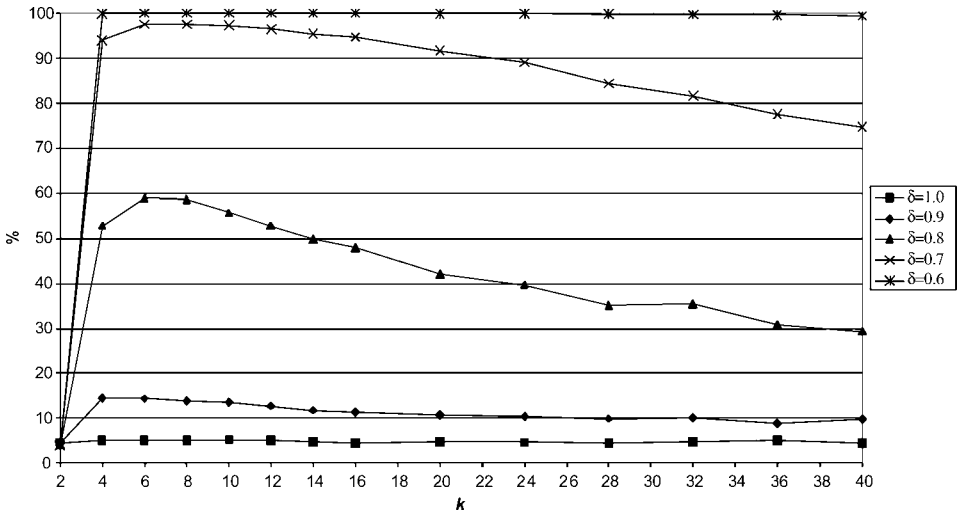

Figure 4b. Power of X^2 test of $N(0,1)$ vs. $N(0,\delta)$: Experiment ***B***, $N = 150, \delta > 1$ (equiprobable).

Table 3 reports results for different sample sizes and indicates that the power is maximised for all sample sizes at k around 4 and 8 for alternatives with $\delta < 1$, and at k around 8 and 10 for alternatives with $\delta > 1$. Moreover, the power of the X^2 test to detect departures from a unit variance is asymmetric around $\delta = 1$, showing more power to reject the null distribution for $\delta < 1$ compared with $\delta > 1$. In sum, for all alternatives and sample sizes considered in this section, the optimal value of k in the case of equiprobable classes appears to be much smaller than the values suggested by the Mann and Wald formula in Table 1.

The last two columns of Table 3 report the power of the K–S test and the moment-based test for the variance of a population. Comparing the power of the X^2 test with the K–S test shows a clear superiority of the X^2 test in all cases; however, both tests have power inferior to that of the moment-based test.

4.2.2. Non-equiprobable Classes

The power of the scale component test (PCSc) (and hence the X^2 test) depends upon F_2 and F_6, as noted above. In experiment ***B*** we set the partitions such that $\{F_1, F_2, \ldots, F_7\} = \{F_2/2,\ F_2,\ (0.5+F_2)/2,\ 0.5,\ 1-(0.5+F_2)/2,\ 1-F_2,\ 1-F_2/2\}$, and take values of F_2 in the range $0.15-0.3$ through steps of 0.025, where $F_2 = 0.25$ corresponds to the equiprobable case. For $k = 8$, Fig. 5a and b plot the power of the X^2 and PCSc tests, respectively, against values of F_2 for $N = 150$ and $k = 8$ for $\delta \neq 1$.

Figure 5b shows that the power of the PCSc test unambiguously falls as F_2 increases for all δ. By comparison, Figure 5a shows that the power of the X^2 test falls as F_2 increases only for $\delta > 1$, and is largely insensitive to F_2 for $\delta < 1$. For example,

Table 3. Power of the X^2 test of $N(0,1)$ vs. $N(0,\delta)$: Experiment ***B***.

	Equiprobable splits			Non-equiprobable splits				
δ	$k=4$	$k=8$	$k=16$	$k=4$	$k=8$	$k=16$	Chi-sq.	K–S
				N=25				
0.6	47.9	42.7	30.1	44.6	26.7	13.2	96.9	13.6
0.7	22.2	21.2	16.1	20.3	13.0	8.1	75.1	7.6
0.8	10.1	9.8	8.6	9.0	6.7	4.8	39.9	5.1
0.9	5.4	6.5	5.7	4.8	6.1	4.3	15.5	4.9
1.0	4.5	4.7	4.4	4.2	4.8	4.5	5.0	4.9
1.1	4.9	6.1	5.8	7.8	9.8	9.2	18.2	6.5
1.2	6.9	8.5	9.4	13.8	14.6	14.8	39.0	8.3
1.3	9.9	14.7	15.3	20.7	24.2	24.2	60.6	11.7
1.4	14.8	22.0	23.0	32.1	36.9	36.2	77.4	15.6
1.5	18.5	30.3	32.8	43.4	47.6	50.0	88.1	20.5
				N=50				
0.6	84.6	84.8	69.0	91.0	79.6	53.2	100.0	42.8
0.7	47.6	44.6	32.5	55.9	39.1	21.4	97.0	16.2
0.8	19.7	17.6	14.3	21.1	15.4	8.4	67.8	8.2
0.9	8.3	7.4	6.5	8.5	6.8	5.0	24.6	5.5
1.0	5.6	4.9	4.8	4.8	5.8	5.0	5.0	5.4
1.1	6.9	6.9	7.0	11.2	10.6	10.8	26.3	6.9
1.2	12.0	13.8	13.7	19.8	21.9	21.0	59.3	10.8
1.3	19.9	26.8	26.7	34.5	38.0	37.2	83.9	16.7
1.4	32.4	42.6	45.7	51.6	59.6	59.1	95.1	26.1
1.5	41.9	56.8	61.3	69.5	78.6	79.7	98.8	35.4
				N=100				
0.6	99.3	99.9	99.2	100.0	99.8	98.6	100.0	90.8
0.7	81.4	86.7	74.7	91.3	87.4	66.9	100.0	45.7
0.8	37.1	39.1	30.4	47.4	37.5	23.6	92.9	14.7
0.9	11.1	11.6	9.4	11.3	10.1	7.2	40.8	6.2
1.0	5.0	4.6	5.4	5.8	5.1	4.8	5.0	5.3
1.1	8.8	9.5	9.3	15.4	13.6	15.5	40.2	8.0
1.2	20.5	27.1	26.8	37.3	38.9	36.4	83.0	16.1
1.3	39.2	53.1	55.2	61.8	67.6	66.2	97.7	30.8
1.4	57.7	76.8	80.1	81.6	88.6	88.6	99.8	47.8
1.5	73.8	90.0	92.7	95.3	97.9	98.2	100.0	67.9
				N=150				
0.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.6
0.7	94.0	97.8	94.7	99.0	98.6	94.3	100.0	72.4
0.8	52.8	58.8	48.1	66.9	62.3	41.5	98.7	23.4
0.9	14.5	13.9	11.4	16.5	11.3	9.1	54.6	7.5
1.0	5.1	5.1	5.4	4.8	5.8	4.5	5.0	4.7
1.1	11.2	14.4	12.1	20.3	18.8	16.9	51.9	8.5
1.2	30.4	39.1	39.1	49.2	54.5	52.5	93.4	21.1
1.3	55.3	72.7	75.3	80.4	85.7	85.4	99.7	44.8
1.4	76.5	91.3	93.9	94.4	97.0	97.6	100.0	68.2
1.5	89.9	98.3	99.1	99.0	100.0	100.0	100.0	87.7

(continued)

Table 3. Continued.

	Equiprobable splits			Non-equiprobable splits				
δ	$k=4$	$k=8$	$k=16$	$k=4$	$k=8$	$k=16$	Chi-sq.	K–S
				N=250				
0.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
0.7	99.8	100.0	100.0	100.0	100.0	100.0	100.0	97.6
0.8	77.1	84.8	77.7	90.8	89.7	76.8	100.0	45.7
0.9	20.7	21.9	17.6	27.1	22.5	15.5	74.8	10.2
1.0	4.5	5.0	4.9	4.3	5.9	5.7	5.0	4.6
1.1	15.8	18.7	17.8	24.9	27.0	24.5	69.7	11.1
1.2	48.4	62.8	63.9	71.1	75.1	73.9	99.1	34.3
1.3	80.2	93.6	94.9	95.6	97.7	97.7	100.0	70.6
1.4	95.5	99.4	99.8	99.5	99.7	99.9	100.0	93.6
1.5	99.1	100.0	100.0	99.9	100.0	100.0	100.0	99.2

for the X^2 test when $\delta = 1.2$, power increases to 55% for $F_2 = 0.15$, compared with 39% for the equiprobable $F_2 = 0.25$, whereas when $\delta = 0.8$ power is roughly 60% irrespective of the value of F_2. The explanation for the insensitivity of the X^2 test to F_2 for $\delta < 1$ lies in the performance of the other component tests for $\delta < 1$. As expected we find that both the median (PCM) and skewness (PCSk) component tests have power around nominal size for all N and for all values of δ, irrespective of F_2 (these results are omitted from the figures). However, the kurtosis (PCK) component test has some power to detect $\delta \neq 1$, and for $\delta < 1$ this power increases as F_2 increases, which tends to offset the falling power of the PCSc component in the X^2 test, although exact

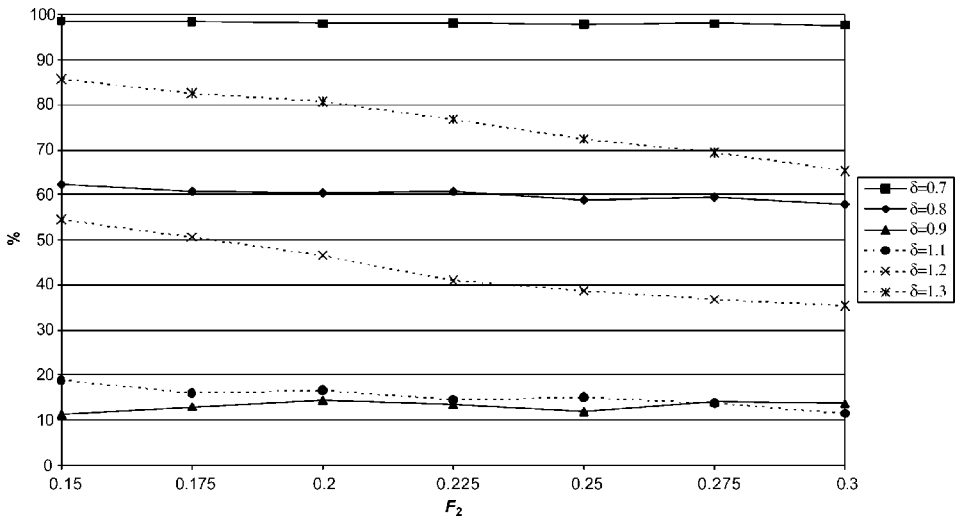

Figure 5a. Power of X^2 test of $N(0,1)$ vs. $N(0,\delta)$: Experiment $\boldsymbol{B}, N = 150$ (non-equiprobable).

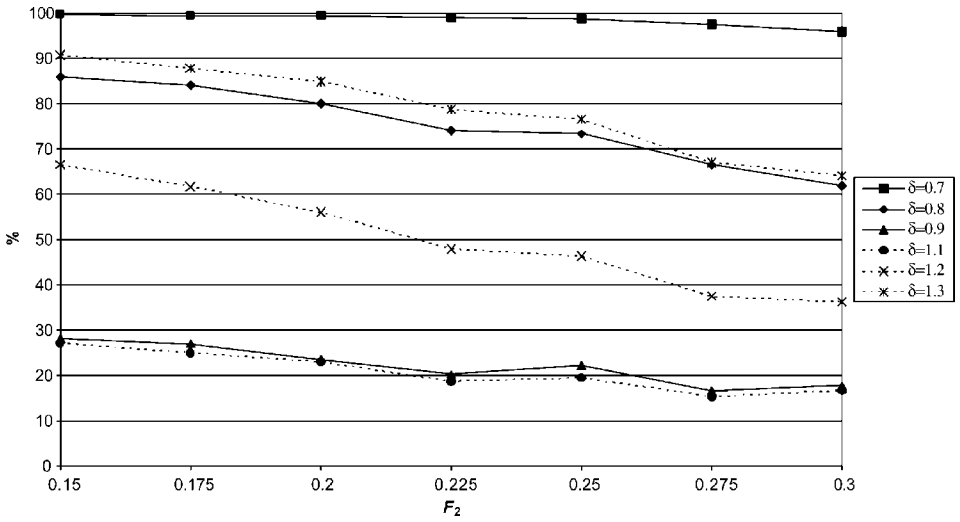

Figure 5b. Power of the PCSc test $N(0,1)$ vs. $N(0,\delta)$: Experiment $\boldsymbol{B}, N = 150$ (non-equiprobable).

additivity of the components does not hold in this case. For $\delta > 1$ there is some increase in power for the PCK test as F_2 increases, but this increase is small compared to that observed for $\delta < 1$.

As anticipated above, Figure 5b shows that for both $\delta > 1$ and $\delta < 1$ there are clear gains in power for the PCSc test when the test is computed using non-equiprobable classes. For example, when $\delta = 1.2$, power increases to 67% for $F_2 = 0.15$, compared with 46% for the equiprobable $F_2 = 0.25$, and to 86% from 73% when $\delta = 0.8$.

Using $F_2 = 0.15$, Table 3 (columns 4–6) reports the power of the X^2 test for different sample sizes. We note that, for all sample sizes, the power of the X^2 test using non-equiprobable partitions is significantly higher than for equiprobable partitions for $\delta > 1$, while there is little difference for $\delta < 1$. The results indicate that with non-equiprobable partitions the power of the X^2 test remains largely insensitive to increasing k from 4 to 8 or 16 for $\delta > 1$.

Finally, in Fig. 6 we summarise the power results of the X^2 test obtained for $\delta \neq 1$ and $N = 150$, using $k = 10$ equiprobable intervals (denoted X2(10)) and $k = 8$ non-equiprobable intervals (denoted X2(8-ne) with $F_2 = 0.15$). We also report the power of the PCSc component test with non-equiprobable partitions ($F_2 = 0.15$), the K–S test and the simple test for the population variance (chi-squared). As we can see, the best performance is given by the latter. The power of the X^2 with non-equiprobable partitions is very similar to that of the PCSc test and both tests have higher power than the K–S test.

These results complement various suggestions from previous findings, summarised in Koehler and Gan (1990), that the best choice of k may depend on the alternative under consideration, as well as the sample size N, and provide a further contrast to the results of Mann and Wald (1942). Moreover, our results clearly show

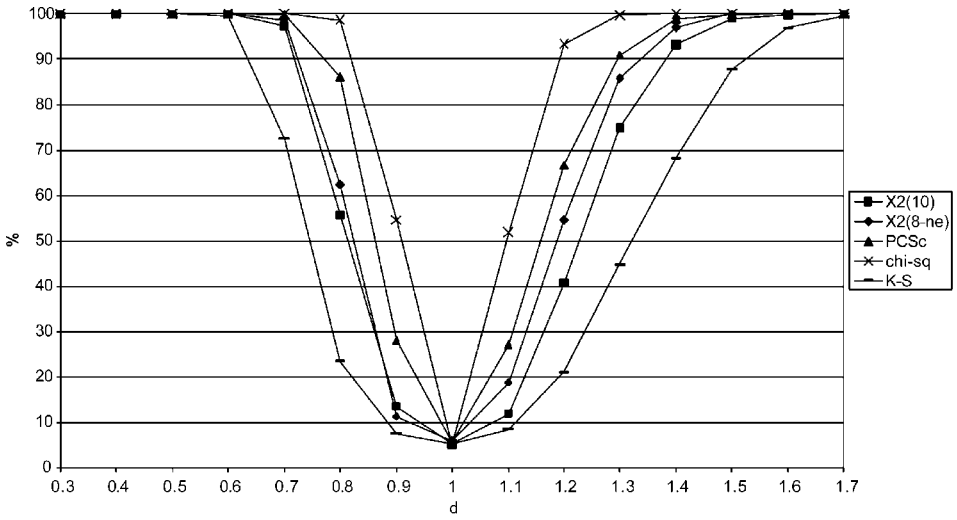

Figure 6. Power of the tests of $N(0,1)$ vs. $N(0,\delta)$: Experiment $\boldsymbol{B}, N = 150$.

that unbalanced partitions with some small classes in the tails lead to important gains in power when $\delta > 1$.

4.3. Experiment *C*: Departures from Symmetry

4.3.1. Equiprobable Classes

We now look at departures from the null hypothesis of normality due to skewness. The relation between power and number of classes in the equiprobable case is illustrated in Figs. 7–9. Figure 7 reports the results for the Ramberg distribution ($\boldsymbol{C}_1$), with skewness ranging from 0.2 to 0.8, for $N = 150$ and k ranging from 2 to 40.

First of all, as expected, we observe noticeable gains in the power of the X^2 test for increasing degrees of skewness. Moreover, the power of the test is sensitive to the choice of k. In particular, there are significant gains in power when the number of classes increases from $k = 4$ to $k = 8$, 10 and 12, with power increasing at a faster rate the higher the degree of skewness. For example, from Fig. 7 we observe that, for $N = 150$, power increases from approximately 14% ($k = 4$) to 56% ($k = 10$) when skewness is 0.6, and from 28% ($k = 4$) to 91% ($k = 6$) and to 100% ($k = 8$) when skewness is 0.8.

Table 4 reports results for different sample sizes and these are qualitatively similar to those discussed above. The results show that power of the X^2 test is not very sensitive to changes in k in the range 24–40. Again the optimal values of k suggested by our experiments under skewed alternatives are significantly smaller than the values suggested by Mann and Wald.

Qualitatively similar results are obtained for the two-piece normal distribution ($\boldsymbol{C}_2$), see Fig. 8, which reports the power of the X^2 test for skewness = 0.6 and

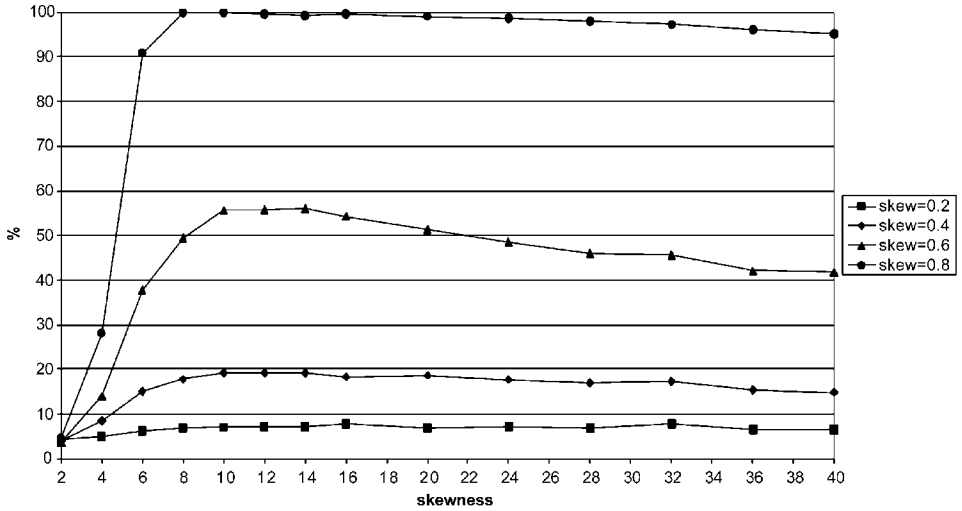

Figure 7. Power of X^2 test against skewness ($\boldsymbol{C}_1$: Ramberg distribution) $N = 150$ (equiprobable).

different sample sizes. Figure 9 reports the power of the X^2 test for Anderson's skewed distribution ($\boldsymbol{C}_3$), for $N = 150$ and different degrees of skewness. In this case we observe a less prominent impact on power by increasing k from 4 to 8 or 10.

4.3.2. Non-equiprobable Classes

As discussed above for $k = 8$, the power of the skewness component test (PCSk) depends upon the three points, F_2, F_4 and F_6. Given our symmetry assumption on the

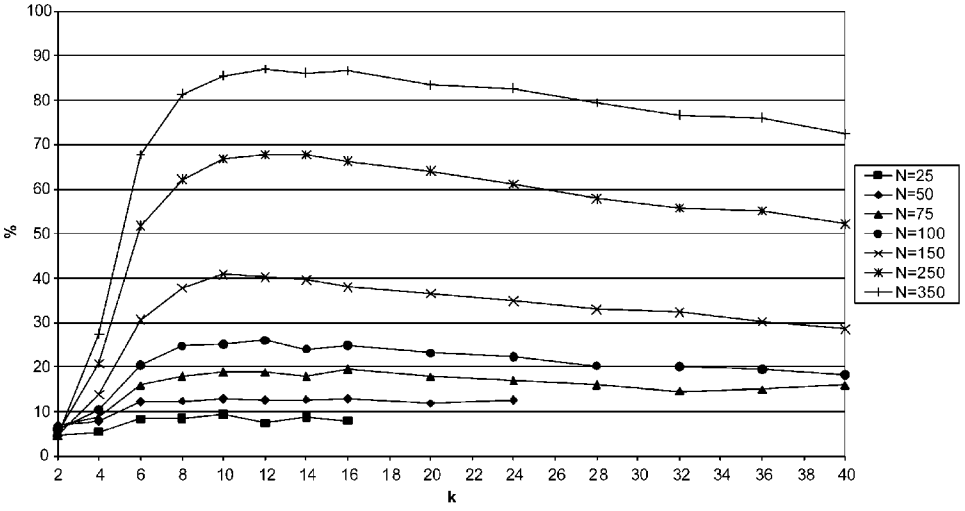

Figure 8. Power of X^2 test against skewness ($\boldsymbol{C}_2$: two piece). Skew = 0.6, (equiprobable).

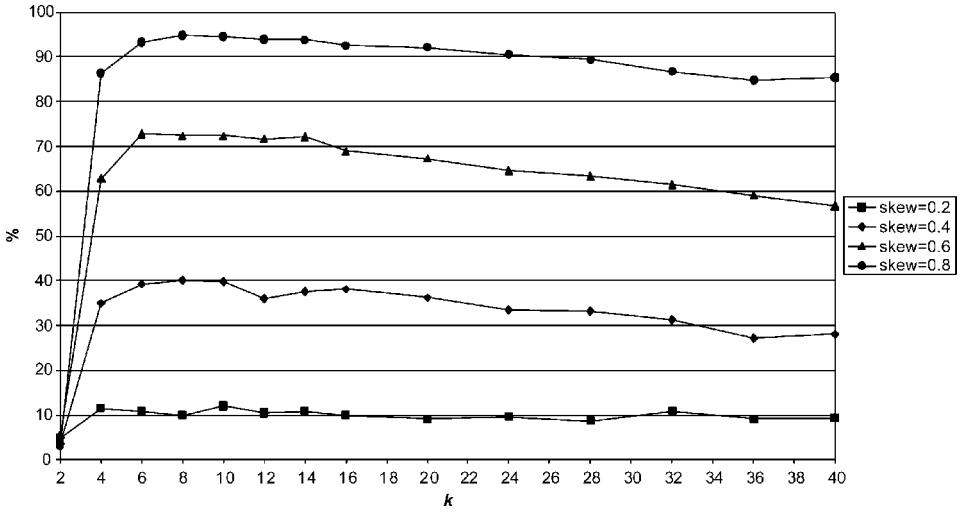

Figure 9. Power of X^2 test against skewness ($\boldsymbol{C}_3$: Anderson) $N = 150$ (equiprobable).

partitions, we set $\{F_1,\ F_2, \ldots, F_7\} = \{F_2/2,\ F_2,\ (0.5+F_2)/2,\ 0.5,\ 1-(0.5+F_2)/2,\ 1-F_2,\ 1-F_2/2\}$, and conduct experiments for values of F_2 in the range 0.15–0.3 in steps of 0.025, where $F_2 = 0.25$ again corresponds to the equiprobable case. In Fig. 10a and b we plot the power of the X^2 and PCSk tests, respectively, against values of F_2 for $N = 150$ and $k = 8$, for $\boldsymbol{C}_1$ (Ramberg) with skewness ranging from 0.2 to 0.8.

From Fig. 10a and b it can be seen that the power of both the X^2 and the PCSk increases as F_2 becomes smaller. At $F_2 = 0.15$ ($N = 150$ and skewness $= 0.6$) the power for PCSk is 56% (67%) compared to 25% (51%) when using the equiprobable $F_2 = 0.25$. In general, the use of $F_2 = 0.15$ (instead of $F_2 = 0.25$) nearly doubles the power of the PCSk test.

The results for $\boldsymbol{C}_2$ (two-piece normal) and $\boldsymbol{C}_3$ (Anderson's skewed distribution) are qualitatively similar to those reported in Fig. 10a and b, although for $\boldsymbol{C}_3$ the gains to using non-equiprobable partitions are smaller than observed for $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$. For all experiments we find that the PCM, PCSc and PCK component tests exhibit power approximately equal to nominal size for all values of skewness, N and F_2 considered.

Figure 11a and b plots the power of the X^2 test for $\boldsymbol{C}_1$ and $\boldsymbol{C}_3$ with $N = 150$, for $k = 8$ equiprobable partitions (X2(8)), $k = 8$ non-equiprobable partitions ($F_2 = 0.15$) (X2(8-ne)), and the PCSk test using non-equiprobable partitions ($F_2 = 0.15$). In the same figures we also report the results from the K–S and J–B tests. Results for different sample sizes are summarised in Table 4 for $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$.

From Fig. 11a and b and Table 4, we observe for all tests a significant increase in power for increasing degrees of skewness. Figure 11a and Table 4 show that, overall, the performance of the X^2 test is maximised with the use of $k = 8$ non-equiprobable partitions. These results apply to all sample sizes, as shown in Table 4, and both alternatives $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$. Figure 11a and Table 4 also report the performance of the K–S and J–B tests. The results indicate that the best performance for $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ is achieved by the J-B test and the worse performance by the K–S test. For example,

Table 4. Power of the X^2 test against skewness.

	Equiprobable splits			Non-equiprobable splits				
Skew	$k=4$	$k=8$	$k=16$	$k=4$	$k=8$	$k=16$	K–S	J–B
Experiment C_1: Ramberg distribution								
				N=50				
0.2	5.1	5.1	5.4	6.1	5.7	6.1	5.5	4.3
0.4	6.6	7.8	8.1	8.7	9.3	8.7	7.7	8.2
0.6	7.5	15.3	16.2	16.8	20.2	18.1	11.7	16.6
0.8	12.3	61.6	48.2	51.6	52.3	47.0	19.8	36.2
				N=100				
0.2	5.5	6.3	7.2	6.6	6.9	7.0	6.2	6.6
0.4	7.3	13.1	13.8	14.1	16.1	14.6	10.4	18.7
0.6	10.8	32.6	34.5	32.8	43.2	36.9	19.6	50.7
0.8	19.0	97.5	92.1	88.3	95.1	92.0	36.3	93.5
				N=150				
0.2	5.1	6.8	7.7	7.7	7.9	7.7	6.9	8.9
0.4	8.4	17.8	18.4	19.8	24.1	21.1	12.0	32.1
0.6	14.0	49.5	54.3	49.7	67.5	58.6	26.5	80.3
0.8	28.1	99.9	99.6	98.7	100.0	99.9	49.9	100.0
				N= 250				
0.2	6.1	8.7	9.2	9.0	10.6	9.8	8.0	15.0
0.4	12.6	29.0	31.6	31.9	41.0	37.2	16.9	60.7
0.6	22.6	79.1	85.1	75.8	93.8	89.3	40.0	98.7
0.8	45.5	100.0	100.0	100.0	100.0	100.0	80.0	100.0
Experiment C_2: Two-piece normal								
				N=50				
0.2	5.1	5.2	5.0	5.0	5.2	5.5	5.1	4.7
0.4	7.0	7.7	7.1	8.6	8.7	8.2	7.7	9.6
0.6	7.7	12.4	12.8	14.3	16.0	13.8	10.8	19.4
0.8	8.6	25.1	21.6	25.9	28.2	23.5	13.8	34.4
				N=100				
0.2	5.6	6.3	6.1	6.1	6.7	6.2	5.8	7.9
0.4	7.6	12.4	11.6	12.1	13.3	11.9	10.1	20.7
0.6	10.4	24.8	24.9	26.8	31.7	26.3	15.4	46.7
0.8	13.6	56.0	51.5	54.6	65.3	53.4	24.5	77.9
				N=150				
0.2	6.0	7.1	6.9	7.2	8.1	7.3	6.0	10.6
0.4	8.7	15.3	15.6	16.7	19.5	16.7	11.5	32.8
0.6	13.9	37.8	38.0	40.2	50.7	42.2	21.1	71.8
0.8	17.6	79.0	78.1	76.2	90.1	80.1	33.5	96.5
				N=250				
0.2	7.2	8.3	7.8	8.9	10.4	10.1	7.4	16.3
0.4	12.2	24.0	25.4	26.2	32.8	29.4	16.0	58.6
0.6	20.8	62.2	66.3	63.4	80.3	71.9	32.6	95.7
0.8	28.2	97.2	98.3	95.1	99.9	98.8	50.4	100.0

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

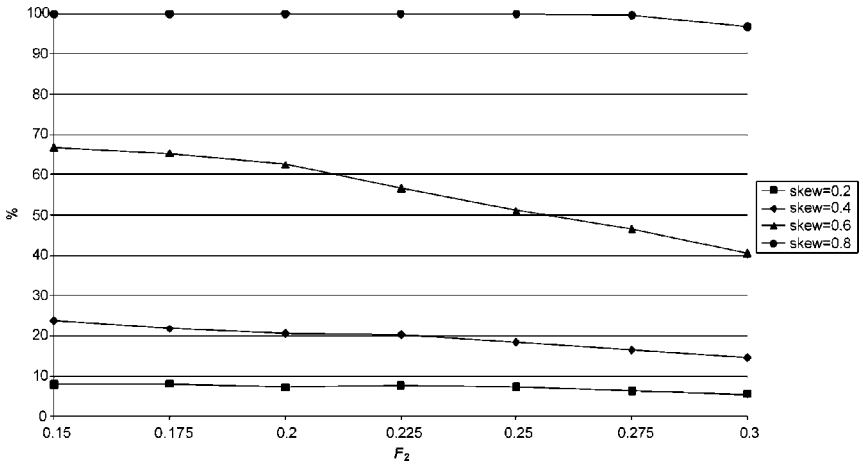

Figure 10a. Power of X^2 test against skewness ($\boldsymbol{C}_1$: Ramberg), $N = 150$ (non-equiprobable).

for alternative $\boldsymbol{C}_1$, and skewness 0.6 ($N = 150$), the power of the J–B test is about 80%, while it is only 27% for the K–S test. It is also clear that, with $k = 8$ non-equiprobable classes, the power of the X^2 test is only marginally below that exhibited by the J–B test (the largest differential between the power of the two tests is 13 percentage points for skewness of 0.6).

A different ranking of the tests is obtained from the experiments conducted under $\boldsymbol{C}_3$. Figure 11b shows that, in this case, both the equiprobable

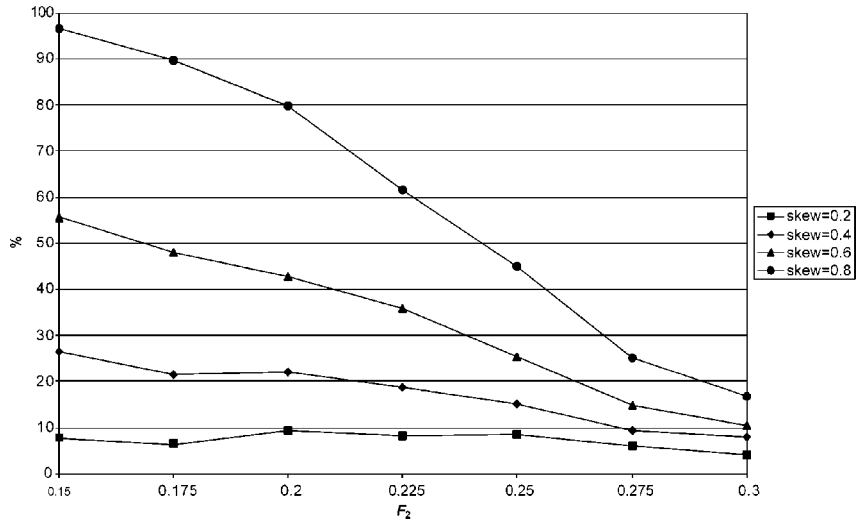

Figure 10b. Power of the PCSk test against skewness ($\boldsymbol{C}_1$: Ramberg), $N = 150$ (non-equiprobable).

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

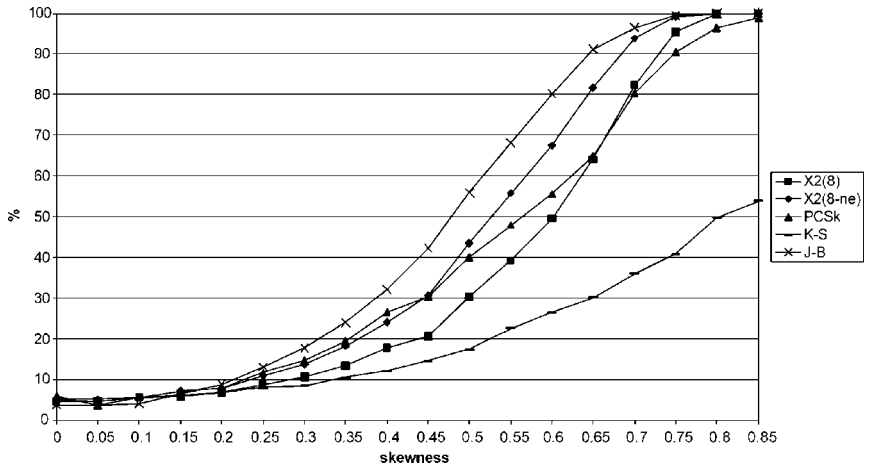

Figure 11a. Power against skewness ($\boldsymbol{C}_1$: Ramberg), $N = 150$.

and non-equiprobable X^2 tests clearly dominate the J–B test as does the PCSk test.

4.4. Experiment *D*: Departures Due to Kurtosis

4.4.1. Equiprobable Classes

The relation between power and number of classes in the presence of kurtosis when the test is computed using equiprobable partitions is illustrated in

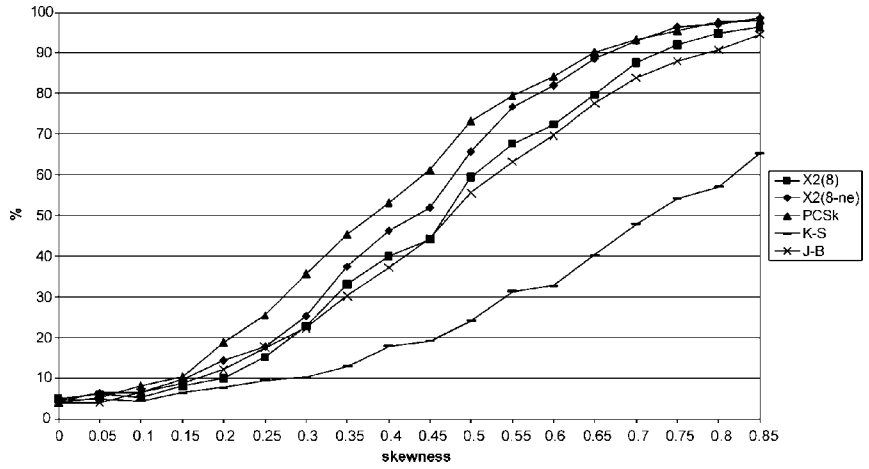

Figure 11b. Power against skewness ($\boldsymbol{C}_3$: Anderson), $N = 150$.

Figs. 12 and 13 for $\boldsymbol{D}_1$ (Stable distribution) and $\boldsymbol{D}_2$ (Anderson's kurtotic distribution), respectively, for k varying from 2 to 40 and $N=150$. As we can see from Fig. 12, the power of the X^2 test to detect departures from $N(0,1)$ is very high, approaching 100% when the stability parameter α describing the alternative distribution is less than 2. Similar findings are obtained in the simulations with Anderson's distribution, reported in Fig. 13, where we see that the power of the X^2 test approaches 100% for kurtosis of about 6.4. For both alternatives, we notice a continued improvement in power for k in the range 12–16, after which the test does not, in general, seem to be sensitive to further increases in k, although when kurtosis is less than 3 there is a deterioration in power. A major difference between the two distributions is that whereas for $\boldsymbol{D}_1$ the X^2 test has power at $k=4$, possibly due to the infinite variance of the Stable distribution, for $\boldsymbol{D}_2$ power is approximately equal to nominal size at $k=4$, as might be expected for a purely kurtotic distribution.

4.4.2. Non-equiprobable Classes

As discussed above, for $k=8$ the power of the kurtosis component test (PCK) depends upon the four points, F_1, F_3, F_5 and F_7. Given our symmetry assumption the partitions are $\{F_1, F_2, \ldots, F_7\} = \{F_1, (F_1+F_3)/2, F_3, 0.5, 1-F_3, 1-(F_1+F_3)/2, 1-F_1\}$, and we set values of F_1 and F_3 in the range 0.05 to 0.2 and 0.25 to 0.45, respectively, where $F_1=0.125$ and $F_3=0.375$ corresponds to the equiprobable case. In Fig. 14a and b we plot the power of the X^2 and PCK tests, respectively, against values of F_1 and F_3 for $N=100$ and $k=8$ classes, for $\boldsymbol{D}_1$ (stable distribution) with $\alpha = 1.975$.

Figure 14a shows that the power of the X^2 test is maximised at $F_1=0.05$ and $F_3=0.45$, whereas the power for the PCK component test (Fig. 14b) is maximised

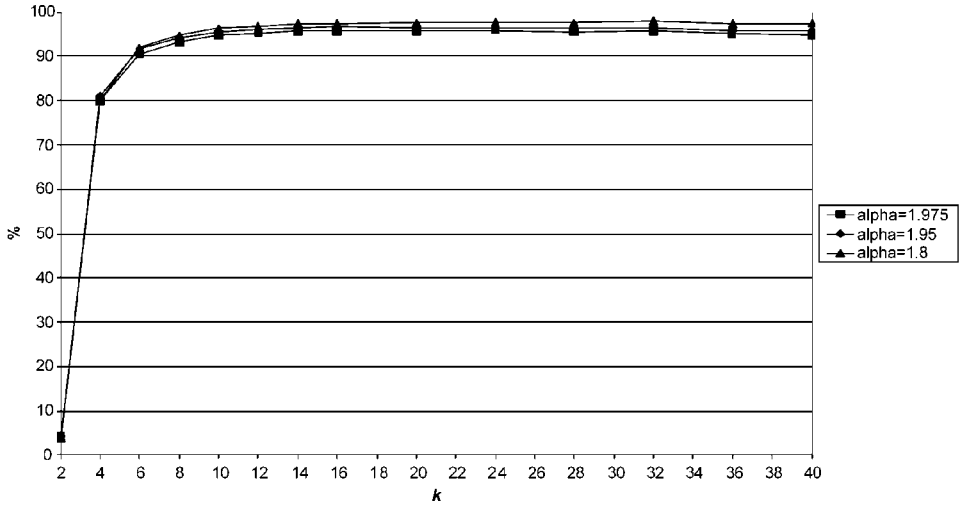

Figure 12. Power of the X^2 test against kurtosis ($\boldsymbol{D}_1$: Stable distribution), $N = 150$.

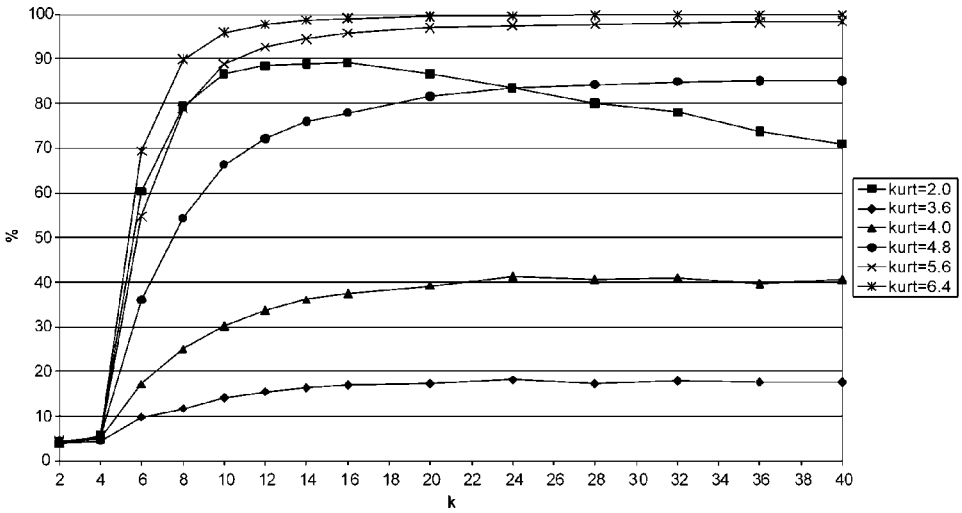

Figure 13. Power of the X^2 test against kurtosis ($\boldsymbol{D}_2$: Anderson), $N = 150$ (equiprobable).

at $F_1 = 0.1$ and $F_3 = 0.45$. The difference between the two tests are accounted for by the fact that the PCSc component test has considerable power (as the variance is infinite) and this is maximised at $F_1 = 0.025$ and $F_3 = 0.3$. Both the PCM and PCSk component tests have power equal to nominal size for almost all values of kurtosis (results omitted from the figures).

In the simulations with $\boldsymbol{D}_2$ (Anderson's kurtotic distribution), we find that the power of both the X^2 test and the PCK component test is maximised at $F_1 = 0.05$ and $F_3 = 0.45$. In this case the PCSc test has little power, as the sample variance for this experiment is approximately unity. The respective power of the PCSc tests for experiments $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ helps to explain the performance of the X^2 test using $k = 4$ equiprobable classes, discussed above.

Figures 15 and 16 explore further the performance of the tests. Figure 15 (16) plots the power of the X^2 test for $\boldsymbol{D}_1$ ($\boldsymbol{D}_2$) with $N = 100$ (150), for $k = 16$ equiprobable classes (X2(16)), $k = 8$ non-equiprobable classes ($F_1 = 0.05$ and $F_2 = 0.45$) (X2(8-ne)), and the PCK test using non-equiprobable classes. In the same figures, we also report the results from the K–S and J–B tests. Figure 15 shows that both the equiprobable and non-equiprobable X^2 tests and the non-equiprobable PCK test dominate the other tests in terms of power. In particular, we see that the maximum power exhibited by the K–S test is 60%, while the power of the J–B test ranges between 15% and 60% for the stability parameter α between 2 and 1.85, and becomes comparable to that of the X^2 tests for smaller values of α. Figure 15 illustrates some gains in power for the X^2 test by using non-equiprobable partitions. The power of the equiprobable X^2 test is around 85% compared with 95% for the non-equiprobable X^2 test.

From Figure 16, it is clear that greater gains are achieved by the X^2 test for this distribution (compared to $\boldsymbol{D}_1$) with the use of non-equiprobable classes relative to equiprobable classes. For example, for kurtosis of 4.8 equiprobable classes delivers

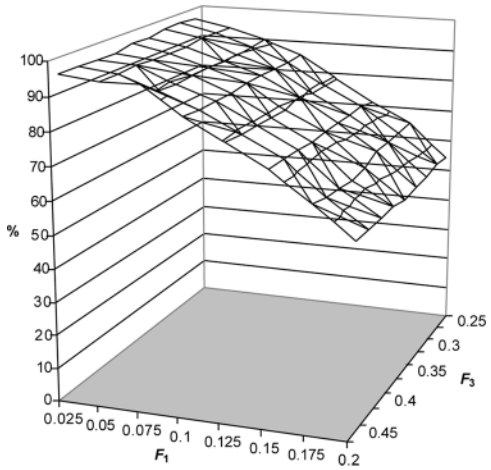

Figure 14a. Power of the X^2 test against kurtosis ($\boldsymbol{D}_1$: Stable, $\alpha = 1.975$), $N = 100$.

power of 78%, while power reaches 93% with non-equiprobable classes. We see that both the non-equiprobable X^2 and PCK tests are superior to the K–S statistic and the J–B test for all values of kurtosis. This is also evident from Table 5, which presents results for different sample sizes.

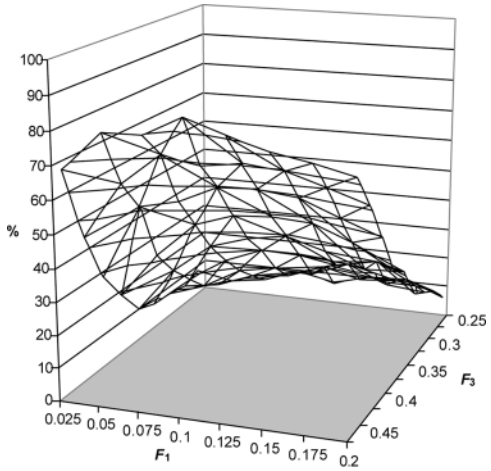

Figure 14b. Power of the PCK component test against kurtosis ($\boldsymbol{D}_1$: Stable $\alpha = 1.975$), $N = 100$.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

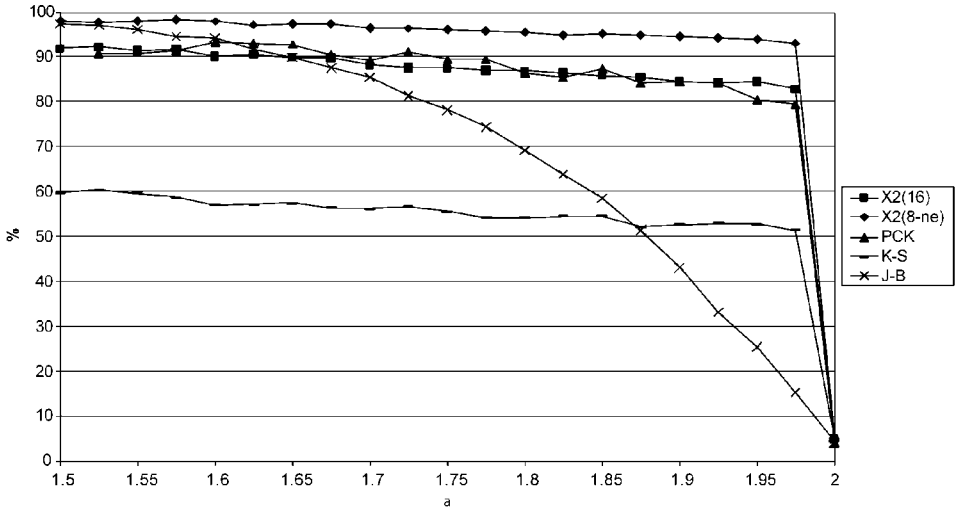

Figure 15. Power against kurtosis ($\boldsymbol{D}_1$: Stable), $N = 100$.

5. CONCLUSION

This article presents the results of a Monte Carlo study of the power of the X^2 test, considering small and moderate sample sizes, various values of k, and both equiprobable and non-equiprobable partitions. The simulations are designed to determine the ability of the tests to detect departures from a standard normal distribution, in the form of changes in mean, variance, skewness and kurtosis. The

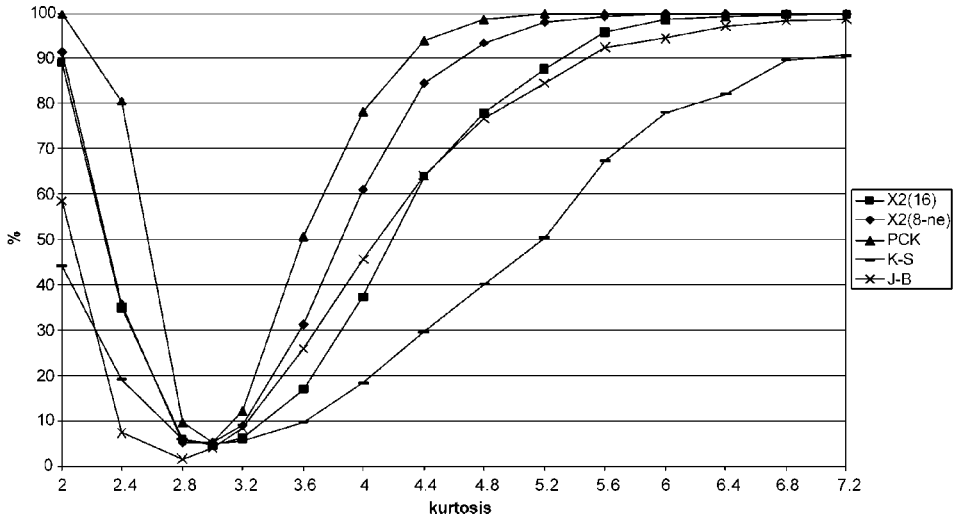

Figure 16. Power against kurtosis ($\boldsymbol{D}_2$: Anderson) $N = 150$.

Table 5. Power of the X^2 test against kurtosis.

	Equiprobable splits			Non-equiprobable splits				
Kurtosis	$k=8$	$k=16$	$k=32$	$k=8$	$k=16$	$k=>32$	K–S	J–B
Experiment $\boldsymbol{D}_2$: Anderson's kurtotic distribution								
				N=25				
2.0	15.6	15.8	13.8	4.7	3.3	1.6	12.6	0.1
2.8	5.2	5.3	6.2	4.1	4.7	5.1	5.3	1.6
3.2	4.7	5.9	6.2	7.8	8.7	11.0	5.3	3.2
4.0	7.2	8.4	10.3	19.3	23.6	28.6	5.7	10.9
4.8	11.2	15.1	18.6	35.4	42.3	48.1	8.1	19.3
5.6	14.9	23.0	30.1	49.2	59.6	66.6	11.4	26.6
6.4	17.9	29.5	38.5	61.0	70.4	76.5	13.8	34.2
7.2	20.3	34.0	44.9	65.6	76.1	82.3	16.5	37.2
				N=50				
2.0	28.9	28.4	20.4	16.7	11.1	4.4	18.9	0.0
2.8	4.9	4.7	4.9	3.7	4.0	3.9	5.3	1.7
3.2	4.8	5.6	4.6	7.8	8.7	9.5	4.7	5.0
4.0	8.7	12.2	12.4	28.0	32.1	36.1	7.4	20.1
4.8	18.1	27.5	31.4	53.9	60.8	65.5	14.5	36.2
5.6	28.3	44.9	53.1	74.5	82.6	86.3	23.0	49.7
6.4	35.8	58.7	69.0	85.5	91.6	94.3	30.2	61.6
7.2	42.0	66.9	77.3	89.8	95.2	96.8	38.2	67.9
				N=100				
2.0	59.9	64.8	49.4	62.8	44.6	18.9	30.4	14.2
2.8	5.9	5.7	5.0	4.4	4.5	4.5	5.8	1.8
3.2	5.4	5.5	5.7	8.8	9.6	10.6	5.3	7.3
4.0	17.5	24.7	26.7	45.3	51.3	54.3	12.9	34.2
4.8	37.2	56.7	63.7	80.4	86.5	88.7	26.6	62.0
5.6	57.6	82.4	89.1	96.4	98.1	98.6	47.2	78.9
6.4	71.4	92.4	96.6	98.9	99.7	99.9	61.7	88.1
7.2	80.6	96.2	98.6	99.5	99.9	99.9	72.5	93.2
				N=150				
2.0	79.4	89.0	78.0	91.4	81.4	48.0	44.1	58.6
2.8	5.6	6.0	6.5	5.2	4.3	3.9	6.0	1.6
3.2	5.1	6.1	6.8	8.9	10.0	10.3	5.6	8.4
4.0	25.1	37.4	41.0	61.1	67.1	68.4	18.3	45.7
4.8	54.4	78.0	84.9	93.4	96.5	97.0	40.1	76.8
5.6	78.8	95.8	98.1	99.3	99.8	100.0	67.3	92.5
6.4	89.9	99.1	99.9	99.9	100.0	100.0	82.2	97.2
7.2	94.8	99.8	100.0	100.0	100.0	100.0	90.7	98.7
Experiment $\boldsymbol{D}_1$: Stable distribution								
				$\alpha = 1.975$				
25	25.3	27.9	23.4	42.7	46.0	48.1	16.5	5.3
50	47.2	48.3	46.3	69.1	69.6	68.8	26.2	9.7
75	66.9	70.0	66.7	84.6	85.7	84.0	38.2	12.6
100	81.0	83.1	80.4	93.1	93.8	92.2	51.4	15.3
150	94.8	95.8	95.2	98.9	99.1	98.8	73.0	19.9
250	99.8	99.9	99.9	100.0	100.0	100.0	94.9	27.9
350	100.0	100.0	100.0	100.0	100.0	100.0	99.6	34.3

relative performance of the X^2 test is compared to that of the K–S statistic and standard moment-based tests.

In summary, three main results seem to apply in general to small and moderate sample sizes, and stand against common practical recommendations. First, our simulations indicate that with equiprobable classes the optimal number of classes is smaller than recommended in previous studies, and smaller than the number of classes typically used by practitioners. We find that k in the range 8–12 maximises the power of the X^2 test against shifts in variance, skewness and kurtosis, and $k = 4$ for shifts in mean.

Second, we find convincing evidence that the choice of non-equiprobable classes can increase the power of the X^2 test significantly, for departures from the $N(0,1)$ null due to shifts in variance, skewness and kurtosis. In particular, we find that a choice of $k = 8$ with partitions which concentrate on the tail behaviour of the distribution in general yields substantial power gains. For shifts in both scale and skewness this entails setting the partition points at $\{0.075, 0.15, 0.325, 0.5, 0.675, 0.85, 0.925\}$ compared to the equiprobable set $\{0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875\}$, whereas for shifts in kurtosis we use $\{0.05, 0.25, 0.45, 0.5, 0.55, 0.75, 0.95\}$. These non-equiprobable partitions improve the performance of the X^2 test over the K–S test, and reduce its disadvantage with respect to the moment-based tests considered in this article.

Thirdly, a choice of $k = 8$ enables practitioners to decompose the X^2 test into its component tests for location (PCM), scale (PCSc), skewness (PCSk) and kurtosis (PCK). We have shown that these component tests have power to detect the relevant departures from the null of $N(0,1)$ due to shifts in location, scale, skewness and kurtosis and their power can be substantially improved by the use of non-equiprobable partitions, of a similar nature to that described above.

ACKNOWLEDGMENTS

The helpful comments of Gordon Anderson, Peter Burridge, and two anonymous referees and an associate editor of this journal are gratefully acknowledged.

REFERENCES

Anderson, G. (1994). Simple tests of distributional form. *J. Econometrics* 62: 265–276.

Barrett, G. F., Donald, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica* 71:71–104.

Boero, G., Marrocu, E. (2004). The performance of SETAR models by regime: A conditional evaluation of interval and density forecasts. *Int. J. Forecast.* 20:305–320.

Boero, G., Smith, J. P., Wallis, K. F. (2004). Decompositions of Pearson's chi-squared test. *J. Econometrics* 123:189–193.

Bowman, K. O., Shenton, L. R. (1975). Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika* 62:243–250.

Chambers, J. M., Mallows, C. L., Stuck, B. W. (1976). A method for simulating stable random variables. *J. Amer. Statist. Assoc.* 71:340–344.

Cohen, A., Sackrowitz, H. B. (1975). Unbiasedness of the chi-square, likelihood ratio, and other goodness-of-fit tests for the equal cell case. *Ann. Statist.* 3:959–964.

Dahiya, R. C., Gurland, J. (1973). How many classes in the Pearson chi-square test? *J. Amer. Statist. Assoc.* 68:707–712.

Fama, E. F. (1965). The behaviour of stock market prices. *J. Business* 38:34–105.

Gumbel, E. J. (1943). On the reliability of the classical chi-square test. *Ann. Math. Statist.* 14:253–263.

Jarque, C. M., Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Econ. Lett.* 6:255–259.

Kallenberg, W. C. M., Oosterhoff, J., Schriever, B. F. (1985). The number of classes in chi-squared goodness-of-fit tests. *J. Amer. Statist. Assoc.* 80:959–968.

Koehler, K. J., Gan, F. F. (1990). Chi-squared goodness-of-fit tests: Cell selection and power. *Commun. Statist. B* 19:1265–1278.

Linton, O. B., Massoumi, E., Whang, Y.-J. (2002). Consistent testing for stochastic dominance: A subsampling approach. Cowles Foundation Discussion Paper No. 1356, Yale University.

Mandelbrot, B. (1963). New methods in statistical economics. *J. Political Econ.* 71:421–440.

Mann, H. B., Wald, A. (1942). On the choice of the number of intervals in the application of the chi-square test. *Ann. Math. Statist.* 13:306–317.

McCulloch, J. H. (1994). Financial applications of stable distributions. In: Maddala, G. S., Rao, C .R., eds. *Statistical Methods in Finance.* Elsevier, pp. 393–425.

Miller, L. H. (1956). Table of percentage points of Kolmogorov statistics. *J. Amer. Statist. Assoc.* 51:111–121.

Noceti, P., Smith, J., Hodges, S. (2003). An evaluation of tests of distributional forecasts. *J. Forecast.* 22:447–455.

Rachev, S. T., Mittnik, S. (2000). *Stable Paretian Models in Finance*. Wiley.

Ramberg, J. S., Dudewicz, E. J., Tadikamalla, P. R., Mykytka, E. (1979). A probability distribution and its uses in fitting data. *Technometrics* 21:201–214.

Stuart, A., Ord, J. K., Arnold, S. (1999). *Kendall's Advanced Theory of Statistics.* Vol. 2A. 6th ed., London: Edward Arnold.

Uchaikin, V. V., Zolotarev, V. M. (1999). *Chance and Stability: Stable Distributions and Their Applications*. Utrecht: VSP BV.

Wallis, K. F. (1999). Asymmetric density forecasts of inflation and the Bank of England's fan chart. *Nat. Inst. Econ. Rev.* 167:106–112.

Wallis, K. F. (2003). Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts. *Int. J. Forecast.* 19:165–175.

Williams, C.A. Jr. (1950). On the choice of the number and width of classes for the chi-square test of goodness-of-fit. *J. Amer. Statist. Assoc.* 45:77–86.